Thesis for the degree of Doctor of Philosophy

Diagnostic Review with Digital Pathology

Design of digital tools for routine diagnostic use

Jesper Molin

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2016

Diagnostic Review with Digital Pathology
Design of digital tools for routine diagnostic use
Jesper Molin
ISBN 978-91-7597-497-2

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4178
ISSN 0346-718X

Department of Computer Science and Engineering
Chalmers University of Technology
SE–412 96 Göteborg, Sweden
Telephone + 46 (0) 31 – 772 1000

Cover: A Scale-Stain visualization of a breast cancer tumor. The method used to generate the image is presented in Paper 7.

Typeset using LaTeX, template by https://github.com/alexvoronov/chalmers-thesis.

Printed by Chalmers Reproservice
Göteborg, Sweden 2016

Diagnostic Review with Digital Pathology

Design of digital tools for routine diagnostic use

Jesper Molin

Department of Computer Science and Engineering
Chalmers University of Technology

Abstract

Digital pathology is a novel technology currently being implemented world wide. This thesis summarizes four years of HCI and visualization research and provides an overall understanding of designing workstation software for pathologists. A human-centered design approach has been used to create a number of design interventions.

The thesis covers three main areas of inquiry: Understanding pathologists' problem solving processes during diagnostic review, how to build different digital tools to support those processes, and how to incorporate digital image analysis algorithms when building these tools.

The thesis consist of a kappa that provides background and context, to the remaining appended papers. The papers describe studies covering, pathologists' navigation strategies in gigapixel sized images, the usability of different input devices and structured reporting interfaces, how principles from volume rendering can be used for multi-scale images, and how make to use of machine learning algorithms to support pathologists' diagnostic processes.

Together, these design projects show how digital pathology images can be used to create tools to make pathologists more productive. This will make it possible for pathology laboratories to replace their diagonstic workflow using glass slides, with a workflow based on digital images.

Keywords: Digital pathology, Human-computer interaction, Visualization.

Acknowledgments

Here is a modest attempt of remembering everyone who has made this thesis possible.

To my great supervisors Claes Lundström and Morten Fjeld and my co-authors: Anna Bodén, Darren Treanor, Sten Thorstenson, Paweł W. Woźniak, Claudia Mello-Thoms, Ida Cervin, Hanna Källén, Anders Heyden, Kalle Åström, Kavitha Shaga Devan, and Karin Wårdell.

To everybody working with developing digital pathology workstations at Sectra: Olle Westman, Tobias Dahlberg, Andreas Vendel, Jens Clarholm, Morgan Filipsson, Martin Svensson, Erik Sjöblom, Kristian Köpsén, Jocke Pennerup, Mimmi Vrang, Elin Kindberg, and Simon Häger.

To colleages at Kuggen for interesting conversations: Alexandru Dancu, Tomasz Kosiński, Zlatko Franjcic, Khanh-Duy Le, and Kristina Knaving.

This thesis would not have been possible without the support of many pathology departments that have participated in experiments and interviews. Thus, special thanks to the pathology departments at Linköping University Hospital, Gävle Hospital, Kalmar Hospital, Karlstad Central Hospital, Helsingborg Hospital, Stockholm South General Hospital, Karolinska University Hospital, Sahlgrenska University Hospital, Rigshospitalet Copenhagen, Salford Royal NHS Foundation Trust, and Leeds Teaching Hospitals NHS Trust.

Jesper Molin
Göteborg, October 2016

List of Publications

This thesis is based on the following appended papers:

Paper 1. Sten Thorstenson, Jesper Molin, Claes Lundström. *Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: Digital pathology experiences 2006-2013.* Journal of Pathology Informatics 5(14), 2014.

Paper 2. Jesper Molin, Morten Fjeld, Claudia Mello-Thoms, Claes Lundström. *Slide navigation patterns among pathologists with long experience of digital review.* Histopathology 67(2), p185-192, 2015.

Paper 3. Jesper Molin, Claes Lundström, Morten Fjeld. *A comparative study of input devices for digital slide navigation.* Journal of Pathology Informatics 6(7), 2015.

Paper 4. Ida Cervin, Jesper Molin, Claes Lundström. *Improving the creation and reporting of structured findings during digital pathology review.* Journal of Pathology Informatics (2016) 7(32).

Paper 5. Hanna Källén, Jesper Molin, Anders Heyden, Claes Lundström, Kalle Åström. *Towards grading Gleason score using generically trained deep convolutional neural networks.* 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), 1163-1167.

Paper 6. Jesper Molin, Kavitha Shaga Devan, Karin Wårdell, Claes Lundström. *Feature-enhancing zoom to facilitate Ki-67 hot spot selection.* Proceedings of SPIE Medical Imaging (2014): Vol. 9041 90410W-1.

Paper 7. Jesper Molin, Anna Bodén, Darren Treanor, Morten Fjeld, Claes Lundström. *Scale Stain: Multi-Resolution Feature Enhancement in Pathology Visualization.* Submitted to IEEE Transactions of Visualization and Computer Graphics.

Paper 8. Jesper Molin, Paweł W. Woźniak, Claes Lundström, Darren Treanor, Morten Fjeld. *Understanding Design for Automated Image Analysis in Digital Pathology.* To be presented at 9th Nordic Conference on Human-Computer Interaction Proceedings (2016).

Author contribution

Main supervisor: Morten Fjeld

Co-supervisor: Claes Lundström

Examiner: Devdatt Dubhashi

Paper 1. The author contributed with most of the literature review, questionnaire design, statistical analysis and the design and preparation of the manuscript, whereas most ideas and the work described in the paper was performed by S. Thorstenson.

Paper 2. The author was the main contributor to the design and execution of the research study, the analysis of the data, and the preparation of the manuscript.

Paper 3. The author was the main contributor to the design and execution of the research study, the analysis of the data, and the preparation of the manuscript.

Paper 4. The author supervised the master thesis that the paper is based on. The author also contributed to the re-analysis of the data, the additional literature review and the discussion that were added to the journal paper.

Paper 5. The author contributed to the development of the system design, the literature review and the preparation of the manuscript. H. Källén was the main contributor and performed all the experiments and the analysis of the results.

Paper 6. The author was the main contributor to the design and execution of the user study, whereas the studied system was equally developed and designed by the author and K.S Devan, as well as the analysis of the data and preparation of the manuscript.

Paper 7. The author was the main contributor to the design and execution of the research study, the analysis of the data, and the preparation of the manuscript. The studied system was designed and implemented by the author.

Paper 8. The author was the main designer of the studied systems. The analysis of the data and the preparation of the manuscript was equally performed by the author and P. W. Woźniak.

Contents

Part I

Kappa

Chapter 1

Introduction

The diagnostic review of cancer could be greatly improved with a new set of tools. Pathologists have been reviewing tumor tissue samples under the light of the microscope for over a century, but that is changing. Digital scanners are able to create gigapixel-sized digital images of the tissue in a few minutes, computer software can present the digital images on a computer display in real time, and algorithms can automatically detect and count the number of cells in the tissue. The technology already exists, but has not yet been transformed into effective tools that are widely used. The aim of this thesis is to narrow that gap by contributing knowledge needed to design useful digital diagnostic tools.

Around 60,000 new cancer patients are diagnosed in Sweden every year. This means that every third person can expect to be diagnosed with cancer in their lifetime. However, two out three cancer patients survive thanks to improving treatments [54].

One of the most successful strategies has been to divide tumors into different subtypes and tailor the treatment for each specific tumor. Currently, cancer tumors are divided into around 200 subtypes. This strategy has put a tremendous burden on pathologists, the specialists that perform the subtyping, with average turn-around times to review a case being around four weeks in some Swedish regions. In other words, every other patient has to wait over a month to get a diagnosis, which delays the treatment and causes a great deal of unnecessary worry.

The pathology lab and the pathologists play a crucial role in cancer care. Based on analysis of tissue specimens and other cell samples, they characterize cancer disease and deliver diagnostic reports. Such a report is a primary source of information to guide the oncologist's decision on treatment options for the patient, such as chemo-, radiation- and/or immunotargeted therapies.

A recent 2012 report commissioned by the Swedish government [105], concluded that the resources provided to Swedish pathology departments are insufficient to ensure an adequate number of pathologists, both for clinical work and research. In fact, neighboring countries have the double amount of pathologists per capita compared to Sweden. At the same time, the annual amount of cancer diagnoses per capita has increased with 1-2%. This steady increase is mainly ascribed to an ageing population [54].

It is in this context that digital pathology comes in, and it has recently gained traction as a serious technology for use within routine diagnostics. However, the first commercially

available digital scanner was launched already around 15 years ago. Today's scanners can be loaded with hundreds of glass slides that are scanned at a pace of a few minutes per slide. The scanner creates images in the gigapixel range, at least two orders of magnitude larger than a modern digital camera. The compressed size of a typical image scanned in 400 times magnification is around 1 gigabyte. This means that the yearly digital storage needs of a medium-sized pathology laboratory is counted in hundreds of terabytes. This is a large amount of data, but it becomes increasingly manageable as storage prices continue to decrease.

Thus, the basic technology needed to perform diagnostic review digitally already exists. It is however still unclear exactly what the technology could be most useful for. This thesis project was started in order to shed light on this problem and to use a systematic approach to design innovations using digital pathology images.

1.1 Thesis overview

The systematic approach used is human-centered design. A human-centered design approach starts with the user in focus and builds systems tailored to typical tasks and environments. It usually involves a number of prototypes before the final systems are built, which are evaluated in order to improve the next iteration of the prototype.

The thesis describes this approach applied to the digital pathology field. The presented work is multi-disciplinary research based on design processes and evaluations from established HCI research methodology.

The thesis consists of an extended summary and background, a so called *Kappa*, and of eight appended papers. The *Kappa* consists of an introduction to the diagnostic work of pathologists and to what digital pathology is. This is followed by an introduction to the three main areas of research that are relevant to this thesis. First, digital pathology is described from the aspect of visual search and exploration. This includes sensemaking theory, perception research on visual search and multi-scale visualization systems that have been studied within HCI. The second area of research is within computer vision and image analysis. The description covers existing algorithms and applications. The aim with this section is to get a sense of how well existing image analysis algorithms can perform in a pathology context. The third area compares how HCI and human-centered design research differs from common research methodology used within the medical domain. The aim here is not to give a comprehensive overview of all research methods, but to highlight differences. This methodological background is needed in order to understand why the prototypes presented are quite different from those developed under other paradigms.

The results chapter then presents a review of the appended papers under three different themes that match the different theoretical backgrounds. The first theme covers results concerning the diagnostic problem solving in multi-scale images. The findings in our papers are presented and related to relevant literature in the *Visual search and exploration* chapter. The second theme concerns how to design systems that measure quantitative properties of the images using machine learning and visualization. The results cover a human-centered approach to automation in digital pathology. The third

theme compares the design process used in this thesis with a similar process found in the literature.

The purpose of the Kappa is to give an overview of the literature available within digital pathology and to provide an overview of the results reported in the separate papers. The papers upon which this thesis is built, are attached as an appendix in the end.

1.2 Digital Pathology

The pathology discipline has its basis in the over 150 year old tradition of microscopic diagnostic review of tissue samples. Tissue samples are collected by surgeons or other clinicians and sent to the pathology lab for analysis. In Sweden, typically, the largest hospital in each county is equipped with a pathology lab that receives tissue specimens or a bodily fluid from other departments or hospitals. The pathology lab processes the material, creates glass slides with pieces of the specimens, and finally a pathologist reviews the content on the glass slide and dictates a report that is sent back to the referring physician. As this thesis focuses on histopathology, the microscopic study of diseased tissue, the process on of how to analyze bodily fluids (cytology) has been omitted in the following description.

The typical tissue specimens arrive at the pathology lab in a plastic box filled with formalin. A pathologist or lab technician performs a grossing examination of the specimen. This examination consists of measuring, sketching, and sometimes taking photos of the specimen. Small pieces of tissue are cut out for further analysis, see Figure 1.1(a).

These pieces are further processed in a sequence of chemical solutions by an automatic processing machine (b) and then embedded into paraffin blocks (c). From these paraffin blocks, small sections of tissue are sliced using a microtome. The microtome is able to create micrometer thin sections, which are placed on glass slides. At this stage the resulting sections are almost transparent. In order to reveal the tissue structures, the glass slides are stained with different colors that attach to different types of structures. For each specimen, this process can result in a large set of glass slides (d). Finally, the pathologist reviews the content of the glass slides with a microscope and dictates a report. This work is typically organized as a sequential production line in multiple steps, consisting of different kinds of processing and selection steps as in Figure 1.2.

If the patient has given written consent, the glass slides and the embedded blocks are stored for future research and follow-up purposes. The initial processing and the staining have become automated and are nowadays performed by a set of machines. However, the workflow is still driven and supervised by lab technicians. In a digital workflow, the produced glass slides are scanned in a digital scanner, creating digital slides for review with digital workstations.

While the use of digital images is pervasive in other medical disciplines, pathology is still, with a few exceptions, performed by reviewing glass slides with microscopes. For a long time, the main reason has been technical limitations related to scanning technology and storage capacity. The image size of a scanned slide has been considered enormous. In the maximum scanning magnification, the images are sampled at a pixel density of 0.25 microns per pixel. This means that a standard sized glass slide, 75x25

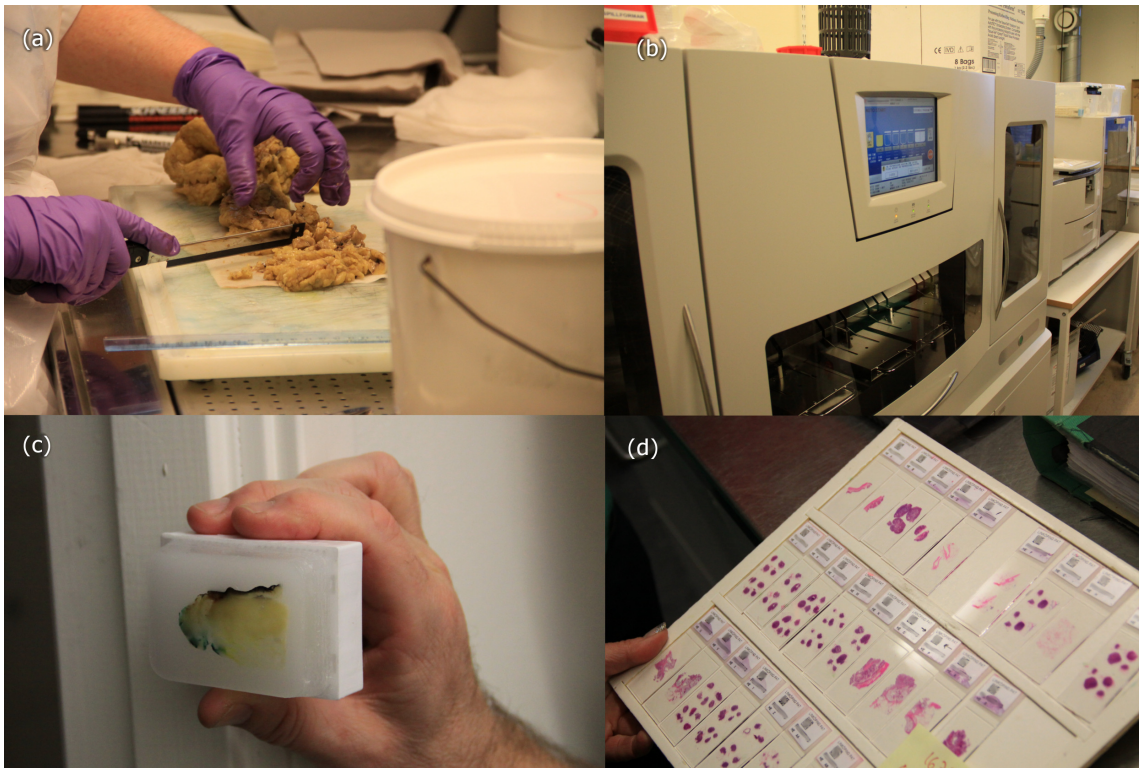


Figure 1.1: The process of producing glass slides. (a) Small pieces of tissue are selected for processing. (b) The machine processing the pieces of tissue. (c) A piece of tissue embedded in a paraffin block. (d) A tray of glass slides ready for review.

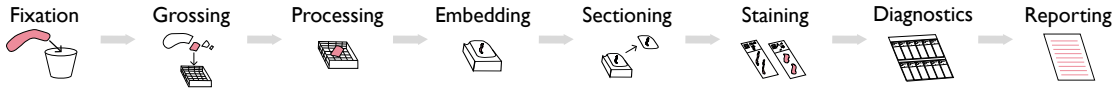


Figure 1.2: The histopathology lab workflow.

mm, requires a digital image of 30 gigapixels to be fully reproduced. This size is three orders of magnitude larger than a current high-end digital camera. However, since most glass slides contain large areas without pieces of tissue on them, and since lossy image compression algorithms can be used, it is possible to reach an average file size of about 1 GB per scanned slide in full resolution. While this file size is still large, it is manageable considering today's continuously falling prices on storage.

Paper 1 presents the current state-of-the-art in digital pathology in depth. It is worth mentioning that only a few years back it became possible to produce digital images of all the slides in a pathology lab. A recent review presented a timeline of digital images within pathology: The first digital image of a pathology slide was taken in 1968, the first digital remote viewing system (telepathology) was demonstrated in 1986, the first frozen section service using telepathology came in 1987, the first commercial system for telepathology was available in 1994, and the first commercial whole slide scanner was launched year 2000 [75].

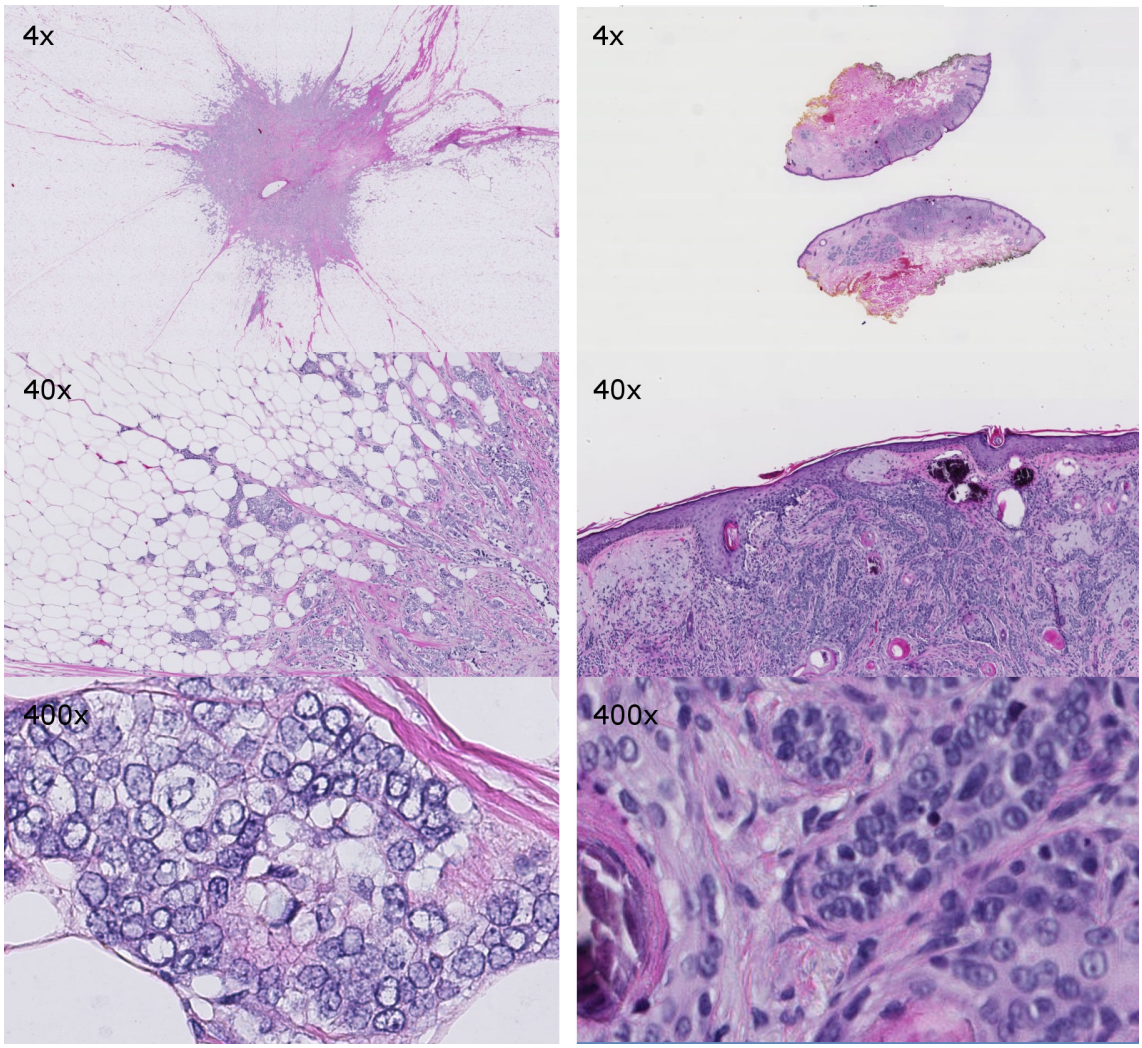


Figure 1.3: Examples of histology images of breast cancer (left) and skin cancer (right).

Before the age of whole slide scanners, it had only been possible to take snapshots of small regions. Whole slide scanners, the dominating technology today, brought a capability to create full resolution images of glass slides, generally called whole slide images (WSI). WSI files contain thousands of small image tiles, in multiple magnifications, thereby supporting fast retrieval and display in any field of view. Multiple magnification levels enable quick panning and zooming, very similar to the navigation experience of using a microscope. The software accompanying the scanners has therefore been referred to as virtual microscopes. To show the general appearance of WSIs, samples of a breast cancer and a skin cancer are provided in Figure 1.3.

Pathologists review these images and create a report that is sent back to the clinician ordering the study. The pathologist performs this analysis by detecting visual features that are associated with different conditions. A number of common visual features can be seen in Figure 1.3. The breast cancer to the left is, for example, clearly invasive since the dark blue malignant cells grow into to the surrounding fatty tissue (white bubbles). This visual feature is associated with a higher risk of metastasis. Both tumors

are hyperchromatic, meaning that the nuclei belonging to the cancerous cells are more dark blue than normal. In the skin cancer to the right, the blue islands of cancer cells can be seen even in the lowest magnification.

The pathology report always contains a diagnosis and a short description of the microscopic and macroscopic findings. Many types of cancer reports also include a histological grade, predictive values, and surgical margins. A histological grade, often based on the appearance of the cancer cells and their growth pattern, estimates overall prognosis for a patient's survival. Predictive values are used to guide whether a specific treatment option is suitable for a specific tumor and is often based on special stainings. A surgical margin is a measure of the shortest distance between a tumor cell and a border marked with ink, visible both macroscopically and microscopically.

Another important function is to participate in multi-disciplinary conferences, to discuss findings for complicated cases with Radiologists, Oncologists, Surgeons and other medical specialties involved in cancer care.

In Sweden, licensed medical doctors become pathologists by going through a five year long trainee program at a pathology department. The program is aimed at learning all the skills necessary to be able to perform macroscopic examination, microscopic diagnostics, and autopsies. Current trainee program guidelines from the professional organization in Sweden state, among many other things, that at least 6,700 cases should have been reviewed under supervision, seven courses in morphological diagnostics attended and a scientific project of 10 weeks performed [57]. Reference books and medical literature are used extensively as study material, but the skill is mostly learned directly from experienced pathologists. For a beginner's introduction to the diagnostic review skill, the author recommends Molavi [69].

Digital pathology is the domain that the work in the thesis has been performed in, and especially the diagnostic work with microscopic images. The next section will introduce three theoretical areas and how they relate to this domain.

Chapter 2

Background and theory

Three major areas form the theoretical basis of the results in this thesis. The first area is visual search and exploration, which are common activities in many domains, and established research can help us understand how this works within digital pathology. Second, since computer vision and image analysis will be used throughout the thesis, this area will be introduced from a human-centered perspective. A number of application areas and their state-of-the-art accuracy will be presented. And third, human-centered design practice will be compared to medical research practice to give background to the methods used in this thesis.

2.1 Visual search and exploration

The focus of this thesis is the pathologists' work at the digital workstation, which naturally is very similar to the work performed when reviewing at the microscope. When using the microscope, a pathologist reviews cases by visually inspecting glass slides. The microscope allows them to search for diagnostic features. The pathologist can pan around within a glass slide by moving it using the stage knob or directly with a finger if the slide holder is removed. The microscope is equipped with multiple objective lenses that are mounted on a rotating wheel. The magnification of the mounted objectives typically range from 1x to 40x, and, adding the static eyepiece magnification of 10x, the pathologist can quickly adjust the total magnification between 10x and 400x. It is also common to have a microscope with a secondary eye-piece in order to facilitate collaboration and tutoring. A picture of such a microscope is given in Figure 2.1.

In a workplace study [88, 85], this review process of panning and zooming with the lenses was analyzed based on recorded videos of working pathologists. They highlighted the pathologist's need to organize and collect information from diverse sources, to collaborate with colleagues, and to ensure patient safety by double checking the details on the request and the final report.

While certain parts of the pathologists' work and artefacts are domain specific, many generic traits are shared with other professions. Consider for example the stunning similarity between the histology image of an invasive breast cancer and satellite imagery of Stockholm in Figure 2.2. The resolution has been matched so that the size of a typical

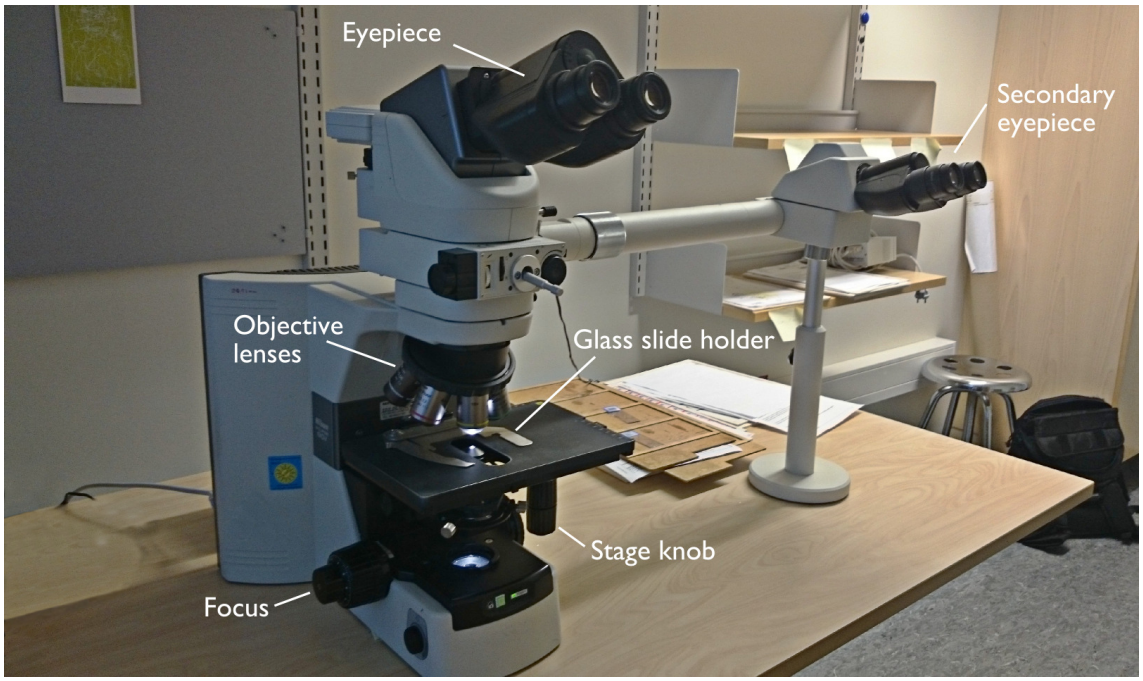


Figure 2.1: The pathologist's microscope describing the parts used when reviewing cases. Photo by Anna Bodén and used with permission.

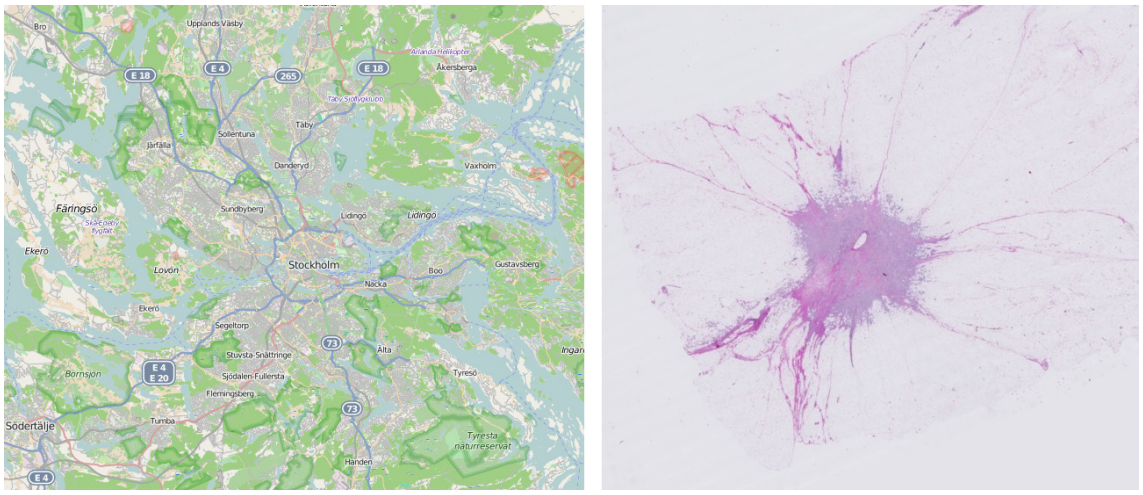


Figure 2.2: Interactive map of Stockholm compared to an image of a breast cancer tumor, equally zoomed out from the maximum zoom level. The map to the left was taken from openstreetmap.org, © OpenStreetMap contributors, and is used under the CC BY-SA-license (creativecommons.org) and the ODBL (opendatacommons.org).

building is the same as the size of a cell. Considering that both are biological phenomena, growing within the constraints of its environment, the similarity should not be surprising.

While the imagery is very similar, the typical user of satellite images use them as maps, in order to find locations or visualize directions in their daily life. In contrast, the pathologist tries to analyze the images in order to derive a diagnosis and to guide treatment strategies. More similar users are found within oceanography or geology [82], where satellite images are used to investigate, analyze, and to produce reports. Another

similar user is the intelligence analyst [81], who investigates, analyzes and produces reports as well, but mostly based on textual data.

Considering these similarities, it becomes rational to connect the work of the pathologists to more general theories or disciplines of research. This thesis embraces the idea that theories are excellent tools for connecting empirical observations between multiple disciplines.

In the coming sections, related work will be presented within three fields of research, which are the most relevant to the findings in this thesis: Sense-making, Visual search and Multi-scale user interfaces.

2.1.1 Sensemaking

Sensemaking theory has its foundations in studies of how humans process information and solve problems. A foundational approach to problem solving was introduced by Newell, Simon [71] that defines the general problem space as consisting of:

1. An initial state
2. A goal state that is to be achieved.
3. Operators that transform the problem from the initial state to the goal state in a sequence of steps.
4. Constraints on the application of operators that must be satisfied.

In an initial attempt to model the problem of diagnosing a set of glass slides this typically means that:

Initial state: There is a set of glass slides with possibly diseased tissue sections from a human being.

Goal state: The goal is to characterize the disease, if any, and communicate this in the form of a written report.

Operators: Operators are different actions, such as panning, zooming, and other procedures that are used to diagnose the disease.

Constraints: External constraints of the imaging technique like the staining or the microscope or internal constraints like lack of knowledge, limited human short-term memory etc.

These categories are interdependent, so that if the external constraints change, e.g. if microscope is replaced by a computerized workstation, action operators like physically organizing the glass slides or changing focus might be superfluous, whereas other action operators like reviewing two slides side-by-side or in stack mode, might be in demand.

To study a specific problem space, different recording devices can be used. Randell et al. [88] used video recording of pathologists working at their desks to study which tools they use at a macro level, excluding the detailed glass slide review. Eye-trackers can be used to study operators' eye-movements [107], and with a digital workstation it is convenient to record the navigation made within the digital slides [109]. To study how pathologists think is more complicated since current technology does not offer automatic elicitation of human reasoning. Ericsson and Simon [23] describes an alternative method to systematically study human thinking called protocol analysis. The method consists of asking participants to think out loud while performing a task, recording their verbal

Table 2.1: Protocol analysis codes identified in a study by Crowley and Naus [17] of expert pathologists reviewing cases on a microscope.

Category	Most common operators	Occurrence [%]
Data examination:		
- Identification	Identify-structure, Locate-lesion, Identify-histopathologic-finding	29
- Comparison	Compare-cue-to-normal, Compare-findings-from-multiple-locations	3
- History	Read, Identify-historical-cue	4
Data exploration and explanation	Associate-location, Note-absent-finding, Evaluate-salience	7
Data interpretation	Statement-of-hypothesis, Confirmation-with-present-finding, Set-goal-identify-or-examine	31
Control processes	Evaluate-own-ability-or-knowledge, Assess-difficulty	1
Operational processes	Operator-verbalization	5
Goal-setting	(part of other codes)	7
Unique hypotheses	(part of other codes)	12

statements and encoding the statements iteratively to make sense of the statements. This method was used by Crowley and Naus [17], to map out pathologists' verbal reasoning when reviewing cases with a microscope. They identified numerous thinking operators that were classified into different categories and measured the difference in the use of them between novice, intermediate and expert pathologists. To exemplify the different ways of thinking that a pathologist might use, Table 2.1 provides a sample of the thinking operators used by the expert group participating in that study.

This type of modelling of operators has also been performed for other medical disciplines. Within medical imaging diagnostics, there are for example a study of the verbal reasoning of cardiologists [41], and a study of the tasks of radiologists interacting with a CT workstation [7].

The main principle is to divide a problem solving task into smaller concepts. This type of modelling can be performed in many different ways. Wei and Salvendy [114] divide cognitive task analysis models into four families: (1) observations and interviews, (2) process tracking, (3) conceptual techniques and (4) formal models. In Table 2, a few typical methods from each category have been listed.

Crowley and Naus [17] point out that histopathological review is a problem solving process that is especially suitable to model with the information processing approach. This is because the microscope invites the pathologists to perform the review by serially searching the slides in a single field of view. When (or if) the pathologists start to perform this review with digital workstations, this assumption might no longer be true. Digital workstations might invite the user to perform visual comparisons, jumping between distant locations or manipulate the display in different manners that make the review process become more parallel, and therefore more complicated to model. A

Table 2.2: Different knowledge elicitation methods that can be used perform a cognitive task analysis.

Observation and interviews	Process tracking	Conceptual techniques	Formal models
Observation	Cognitive walkthrough	Error analysis	ACT(-R)
Unstructured interview	Verbal reports	Conceptual graph analysis	GOMS
Structured interview	Protocol analysis	Diagraming	ARK
			Questionnaires

more appropriate approach would therefore be to use sensemaking theory, which is more targeted towards describing the process of analytical thinking using interactive visualizations on computer workstations.

Pirolli and Card [81] describe sensemaking as a task that consists of "information gathering, re-representation of the information in a schema that aids analysis, the development of insight through the manipulation of this representation, and the creation of some knowledge product or direct action based on the insight." For a pathologist this process could be represented by the following process:

Histology slides (information) →Findings (Schema) →Conclusions and hypothesis (Insight) →The pathology report (Product).

Pirolli and Card [81] use this model to describe the typical work of an intelligence analyst. They model the analytical process in two main loops, the foraging loop, where the analyst searches for information, and the sense-making loop, where the analyst makes sense of the information found in order to produce a report.

Arguably, this process can be adapted to the histopathology diagnostic process in a straightforward manner, as shown in Figure 2.3. However, similarly to the case of the intelligence analyst, it should be remarked that:

1. Most of the pathologist's work consists of extracting information and repackaging it as a report without much analysis, i.e., easy cases where the diagnosis becomes evident as soon the specimen is put under the light of the microscope.

2. Higher-level thinking such as hypothesizing and schematizing can be used to reject findings in early stages. Treanor et al. [109] argue that this process helps the pathologist to avoid cluttering the mind with less important findings.

Studies comparing novice and expert pathologists, have found some evidence to support the view that novice pathologists are able to detect large part of the findings in the foraging loop, but are unable to create sufficient schema to produce an accurate diagnosis [17, 66].

Since the microscope does not offer convenient means of storing image information throughout the analytical process (that is: foraging and sense-making), pathologists memorize most of the found evidence, schema, and generated hypotheses. One way to overcome this limitation is to verbalize the findings and to keep scribbles on the glass

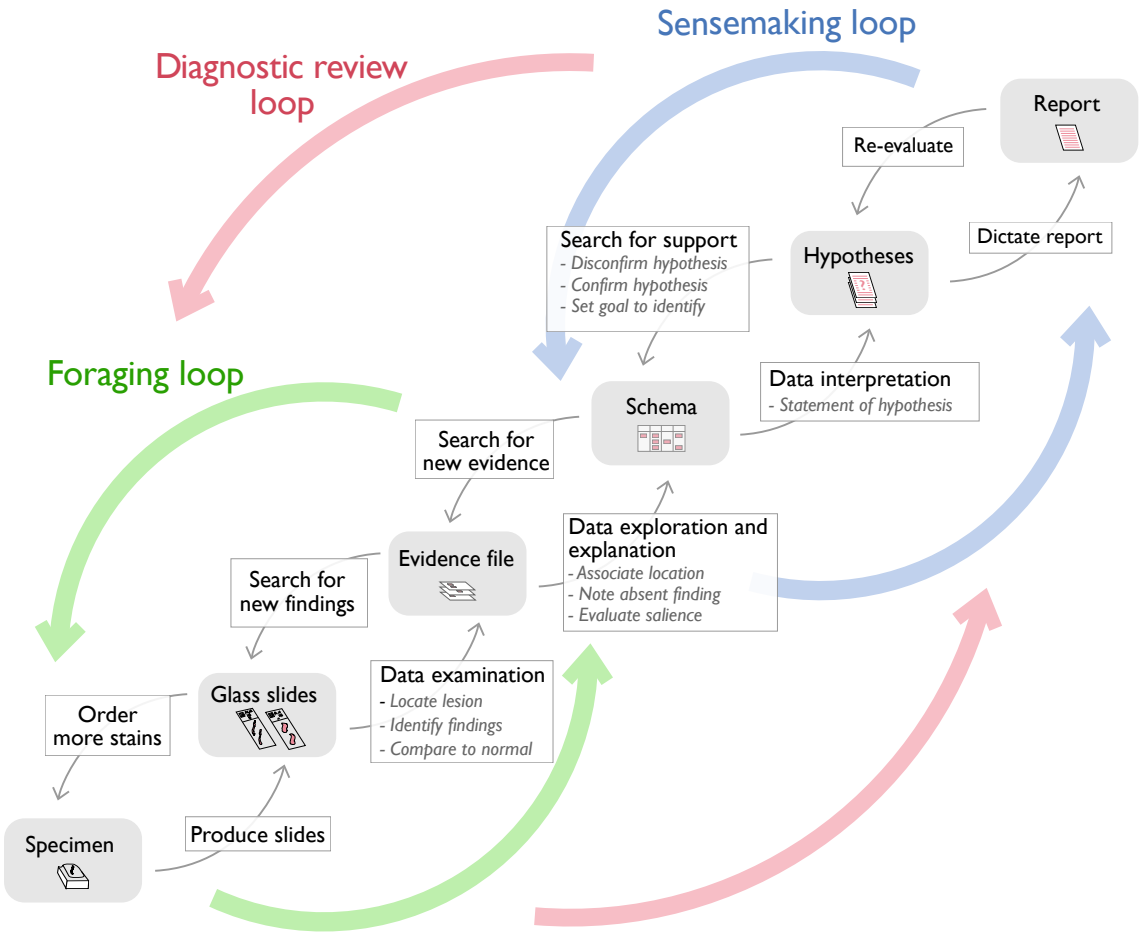


Figure 2.3: Sense-making for pathologists, in a diagram style adopted from Pirolli, Card [81]. Arrows pointing upward represent forward or bottom-up reasoning, whereas arrows pointing downward represent backward or top-down reasoning.

slides [88]. Considering the limited capacity of the visual working memory [64, 116], this represents an important opportunity when designing a digital workstation for pathology. Other opportunities for digital pathology can be found in Shneiderman's taxonomy [95], which lists operations common when exploring data within information visualization:

1. *Overview:* Gain an overview of the entire collection.
2. *Zoom:* Zoom in on items of interest.
3. *Filter:* Filter out uninteresting items.
4. *Details-on-demand*: Select an item or group and get details when needed.
5. *Relate:* View relationships among items.
6. *History:* Keep a history of actions to support undo, replay, and progressive refinement.
7. *Extract:* Allow extraction of sub-collections and of the query parameters.

Current digital pathology workstations usually provide means to support operation (1), (2), and (4), whereas the ability to filter, relate, keep a history of, and extract data when exploring pathology images represent opportunities for research.

2.1.2 Visual search

In histological review, the information foraging loop is driven by visual search processes in the form of navigation and vision. There, more specifically, navigation consists of virtual navigation, i.e. panning, zooming, and rotation of the image slide, as well as physical navigation in the form of eye and head movements that move relative to the screen. If the screen is large, physical navigation also include movements of the full body [5].

The vision system can be divided into foveal and peripheral vision. The fovea is the small part of the retina where it is possible see in the eye's full resolution, whereas the remainder of the retina corresponds to the peripheral vision. Ware [113] provides the following rule of thumb, the size of the fovea is the size of the nail of the thumb when held at arm's length distance from the eye. Ware [113] further lists the three types of eye-movements in humans: saccadic, smooth-pursuit, and convergent movements. Saccadic eye-movements are rapid moves between fixations. The time it takes to move between fixations varies between 20 and 180 ms, whereas the fixations last between 200 and 400 ms. Smooth-pursuit movements enable the eye to lock onto a target, and to follow it when it moves within the visual field, or if the target is fixed and the body is moving. Convergent movements deal with objects in visual field moving towards or away from us.

Performing visual search is a complex task that involves at least all of the mechanisms described above. There exist multiple empirically based models that describe this visual search process. These theories share a view that the visual system first has a parallel processing that triggers saccadic movements acting as starting points for more accurate processing. The parallel processing step uses peripheral vision to detect either high local contrast from the surrounding or using top-down processes to detect specific items that are different from the surrounding independent of the local situation [113].

A number of studies have made use of eye-tracking to study pathologists' eye movement during slide review. Braxton et al. [13] tested the correlation between mouse pointer movement and gaze. Brunyé et al. [14] used gaze as an index of expertise. Tiersma et al. [107] characterized two types of eye-movement: a scanning type and a selective type. Krupinski [60, 61] attributed these characteristics to a difference in expertise since experts were more selective and spent more time on relevant features instead of scanning quickly through everything.

Another interesting effect was reported by Bombari et al. [11], who showed that top-down visual processes might cause confirmation bias. In their experiment, each participating pathologist was asked to grade the malignancy of individual nuclei while first being shown an image of the architecture of the tumor. A high-grade architecture caused the participant to unconsciously search for nuclei that looked more high-grade and a low-grade architecture caused them search for low-grade nuclei.

Most of the above research addressed visual search patterns in regards to static images. However, due to the sheer size of pathology slides, case reviewing requires virtual navigation. A number of studies have been performed on virtual navigation in pathology: comparing navigation between novice and expert pathologists [66, 76, 109], studying students taking an examination [112], and correlating gaze coordinates with digital slide coordinates [14]. These studies mainly used recordings of navigation behavior, represented as heat maps, to compare the difference in explored regions between different study groups. Heat maps display what the pathologists have looked at, but does not display the dynamics of navigation, ignoring sequences of movements and fixations. In a single study [67], programmatic rules of the dynamic properties of the navigation were used to automatically locate regions of interest in pathology slides, but the dynamic behavior was not studied in itself. The dynamic behavior is important to understand in order to design the user interface, especially with respect to the input device used to navigate the slides.

A number of recent studies by Jaarsma et al. [52, 51, 50] combined eye-tracking recording and navigational tracking to the behavior of pathologists with different levels of expertise. They were not able to reproduce eye-tracking findings recorded on static images, and recommended focusing problem-solving studies on navigational tracks [51]. They further highlighted a number of general differences. Experts were found to spend more time in low magnification than novices. Experts also tended to express their problem solving in comparative terms (normal, abnormal, atypical), whereas intermediates described specific pathologies (invasion, infiltration, lymphocytes), and novices described what they saw using descriptive words about color and shape (purple, pink, round). The design of digital workstation tools should take these findings into consideration and support pathologists at all levels of expertise. A good design should also aid pathologists to become experts, and should give incentive to the development of expert traits. As was remarked earlier, expert traits might change as the design of the diagnostic review tool changes. In other words, what characterizes an expert today might not be same with a new set of tools.

Another visual search phenomenon worth highlighting is that of inattentional blindness. In a well-known study, a group of participants were asked to count the number of ball passes between a group basketball players in a recorded video. However, the participants were not told the video also showed a person dressed in a gorilla costume traversing the scene. Since the participants were focused on counting, around half of them failed to report seeing the gorilla [70]. This study was recently reproduced in a medical setting. A picture of a gorilla was inserted into a stack of CT images of a lung. Trained radiologists were then asked to search through the images for lung nodules. The detection rate of the gorilla was very low—20 out of 24 radiologists failed to detect it even though the gorilla was visible for an average of 5.8 seconds. Furthermore, 12 of the radiologists directly looked at the gorilla image for an average of 0.5 seconds but still failed to detect it [21]. As a consequence it is thus not safe to assume that all visible areas on a display are being looked at or that visual features being looked at have been attended to.

Efficient and accurate histological reviews depend on the images displayed. Knowledge about the human visual system should thus be incorporated into workstation design.

Large high-resolutions screens have been shown to reduce review time when using digital workstations, probably due to less time spent performing virtual navigation [84]. This reduction has been shown to make digital case review as fast as review with the microscope [86, 108]. However, it is not evident that total review time will decrease, even if the work at digital workstations becomes more efficient at visual search than with a microscope. Randell et al. [87] recorded use patterns indicating that the increased comfort of digitally navigating between slides, may cause pathologists to more often revisit previously reviewed slides. This is problematic considering the current lack of pathologists, since increased efficiency in navigation is being used to increase their overall accuracy rather than having them report on more cases.

To design workstations that support visual search in the way pathologists need, their diagnostic review process needs to be studied in further detail, especially the dynamic behavior of virtual navigation. The design of search within large data spaces is not a new topic. A large body of research already exists within the field of human-computer interaction (HCI). In the following section, related literature on multi-scale (or zoomable) user interfaces will be presented.

2.1.3 Multi-scale visualization

Virtual navigation has for long been needed for information spaces larger than can be completely shown on a display. Beard and Walker [7] divide the information spaces into those that are naturally meaningful to map two-dimensionally such as large images, maps, or medical images, and those without this natural mapping such as different kinds of graph data. Since pathology images are natural to map two-dimensionally we will here focus on that part of the HCI research.

In order to display images too large for the display, you either have to truncate it by only showing a part of the image, or to shrink it by showing a smaller version of the image, or to distort it by showing certain parts in higher magnification, using for example fish-eye views [26]. A convenient way to express different display modes for multi-scale data spaces is by using scale-space *(or space-scale*[1]*)* diagrams [27]. In scale-space diagram, a fixed-size viewport is used to represent the device's field of the view in a three-dimensional space, where two dimensions represent the extent, and the remaining dimension represents the zoom level. An example of a scale-space diagram is depicted in Figure 2.4, where panning and zooming with a microscope is shown.

To select what is displayed the user can then use different input devices or soft interfaces. Early GIS systems made use of SQL like query languages for navigation [25] to specific addresses within a large database of maps. A similar interface is still used extensively to search for locations in online maps, however the SQL queries have been replaced by keyword and natural language search queries. Early display manipulation techniques include using scroll bars or cursor keys for panning (or roaming) within the data space. Scroll wheels on mice were not popularized until mid-90s, so zooming had to depend on other means.

[1] In Furnas et al. [27], the term space-scale is used, but in this thesis the word order has been reversed to avoid confusing the technical reader.

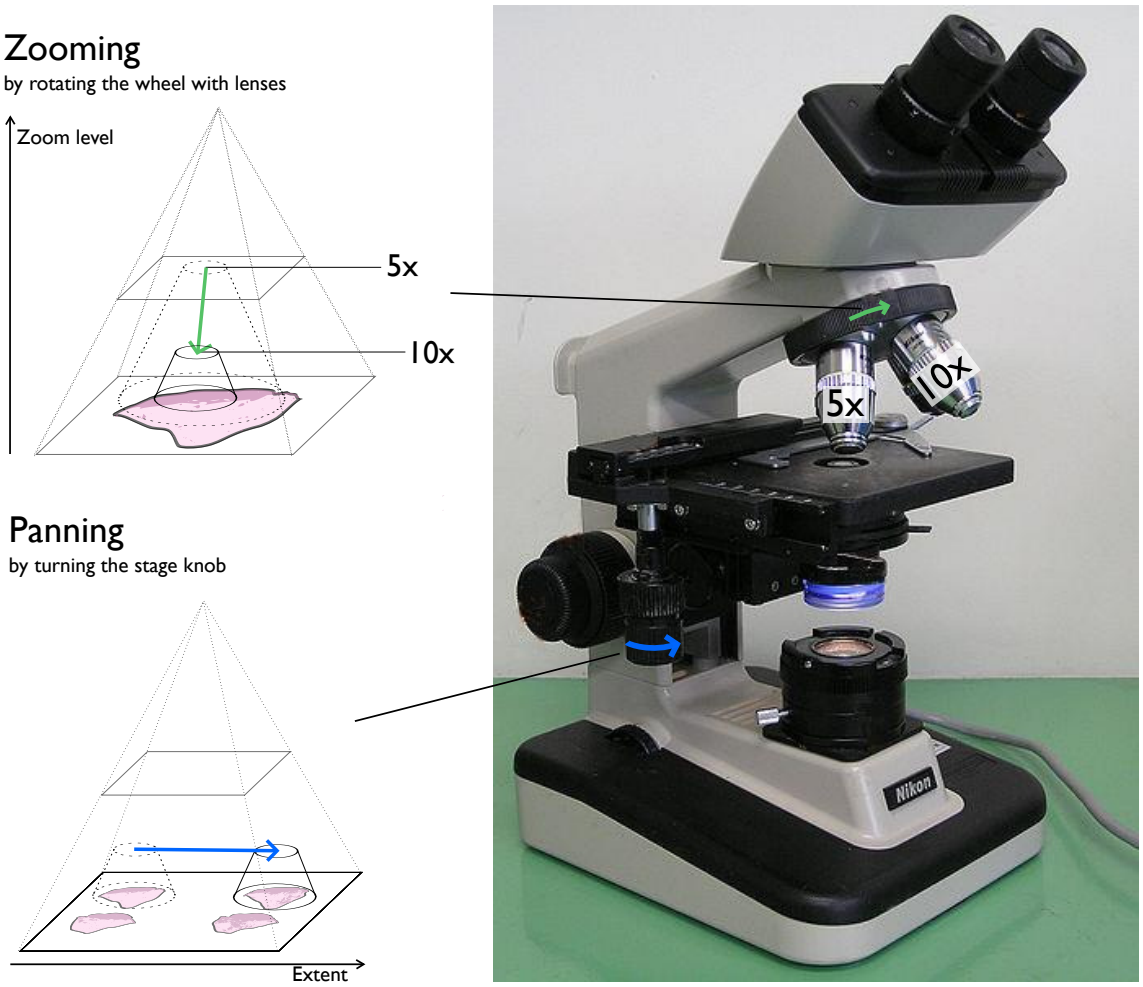


Figure 2.4: Depicting zooming and panning actions on a microscope using scale-space diagrams. The microscope's circular field of view remains the same size, but when zooming in by switching the lens it projects to a smaller area of the tissue. Photograph of microscope is used under CC-BY-SA license.

HCI research within this field has typically compared alternative interface design solutions for different specific applications, in different laboratory tasks, or both. The studies have, e.g., investigated specific interfaces for visual comparison [56, 82, 63], interactions that are relevant when using different kinds of tools [38, 96], displaying the location of off-screen items [6], changing representation of items depending on the zoom level [78], specific input devices for navigation [31, 101, 103, 100, 102], and how to smoothly perform a run-over from point A to B in scale space [115]. Experimental navigation tasks are of particular interest in this thesis. These laboratory tasks can be divided into pure navigation tasks in which a user knows where to navigate, and into search and exploration tasks.

Pure navigation tasks can, for example, include techniques for quick navigation in large documents [46] or navigation to targets within a map [43]. Pure navigation tasks were formalized into a multi-scale version of Fitt's law (a law to predict pointing performance [24]) by work of Bourgeois [12] and Guiard [36]. The typical pathologist does not, however, know where to go since the diagnosis is generally performed on a new

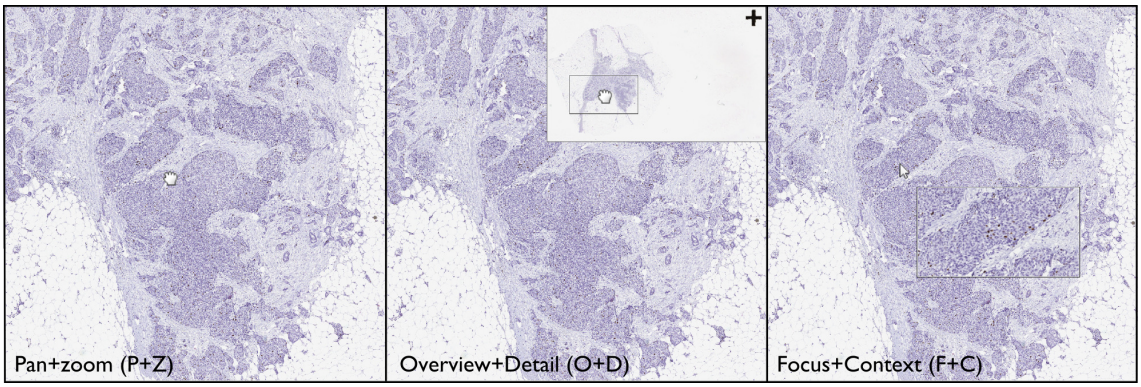


Figure 2.5: Three common zoomable interfaces. Pan+Zoom is a basic interface where you pan by dragging and zoom using the scroll wheel of the mouse. Overview+Detail enable navigation by clicking or dragging in an overview where the full information space is visible. Focus+Context is when it possible to magnify the area around the mouse pointer, this may for example be implemented as a distorted fish-eye view, or as in the figure, with a small window following the mouse pointer.

piece of tissue every time. Therefore, studies within multi-scale search and exploration are more relevant.

Search tasks have been performed by asking participants to find a specific target within an information space using different navigation techniques either in laboratory tasks or for different applications. The predominant interaction techniques consist of Pan+Zoom (P+D), Overview+Detail (O+D), and Focus+Context (F+C); all shown in Figure 2.5.

Pietriga et al. [79] proposed a laboratory search task in order to perform application independent evaluation of navigation techniques, and compared the three techniques presented above. In their study, O+D performed best in terms of task completion time. In a related laboratory study, Hemminger [42] compared two keyboard based methods, scrollbars, P+Z, and a fisheye F+C for a systematic search task common in medical imaging. They found that P+Z outperformed the others in user satisfaction, whereas F+C was the least preferred technique.

Hornbæk and Frøkær [45, 44] studied navigation techniques and patterns for reading large electronic versions of academic papers. The same three techniques were evaluated in two different tasks: a reading task followed by writing an essay, and a question answering task, searching for answers in the text. Users were most satisfied with the using the O+D interface. Users read faster with the F+C interface, but this was at the cost of a lower degree of text retention. When analyzing the reading patterns a few key navigation tasks were identified. For the essay task, most participants started with an initial orientation by quickly flipping through the full paper followed by a linear reading of the full paper from start to end, and finished by reviewing the paper. In the question answering task, the navigation instead consisted of searching for targets, reading the found targets and, in the overview condition, jumping directly to parts already visited. In another study of navigation in maps [43], similar results were found, the participants were slower but reported higher levels of satisfaction when navigating with an overview. Hornbæk and Frøkær [44] proposed that this is because an overview invites users to perform additional, perhaps unnecessary, explorations and that the overview helps users to remember what

areas that have already been visited. These behaviors increase the quality of the task performance but could potentially increase task completions times. This finding is similar to the finding within pathology presented in the previous section that the revisitation rate of already reviewed slides increased when the digital interface was used, which provided more convenient means to switch between slides [87].

A recent development in multi-scale navigation are techniques that take advantage of information in the underlying image, so called topology-aware navigation. This can for example be a fisheye that adapts to the size and shape of the object under the lens [80] or a gravity model that slightly steers the navigation onto targets [55].

A limitation of the laboratory search tasks above is that the goal has been to find a specific target, which is, at least, qualitatively different from a search task where a target is not present. Another issue that has received little attention is what effects that these techniques might have when being used extensively, and for long durations of time.

Overall, the HCI field contains a number of studies that like pathology deal with large amount of data in multi-scale information spaces. The theoretical frameworks that have been developed for these other domains can be reused within digital pathology, and can be used to feedback contributions and insights from digital pathology research. One major difference is the assumption in existing HCI research that there exists large amounts of conceptualized information that can be visualized and filtered. With a raw digital pathology image there is only access to the pixels. This means that if only the pixels could be turned into objects, it would possible to reuse a great deal of the existing literature. This mapping is usually performed by making use of a computer vision system. The get an understanding what current systems are capable of, the next section will go through what these could potentially be capable of doing within the digital pathology domain.

2.2 Computer vision and image analysis

Computer vision is a field of computer science aiming at generating descriptions or models of the world using one or more images of it [106]. Digital Image Processing on the other hand refers to processing digital images by means of a digital computer [32]. Comparing these two approaches, computer vision is more related to artificial intelligence, with the aim to trying to reproduce human perception in the machine. A third term, Digital Image Analysis lies somewhere in-between the two. Gonzalez, Woods [32] provide a useful separation of digital image processing into three levels: Low-level processing only uses primitive operators and both the input and output of the processing is an image. This includes simple forms of contrast and color enhancements. Mid-level processing takes an image as input and detect intermediate *features* or segments, such as edges, textures, contours, or coherent areas. High-level processing recognizes objects and relates more to computer vision and digital image analysis.

The different levels of processing can be performed in many ways, but machine learning is now perhaps the most common technique to implement the mapping of the result, especially from the mid-level output to a high-level output. In machine learning [10], the idea is that instead of trying to figure out a deterministic function or algorithm to solve a certain problem, you provide a set of examples, here in the form of image features,

together with a set of true values, training data. The machine learning then figures out how the algorithm or function should look like. Two types of machine learning are particularly relevant to digital pathology: classification and regression. In classification, you take a set of image features together with a set of classes and you automatically construct the function. For example, to build a breast cancer detector system, you need images from two classes, both cancerous tissue and non-cancerous tissue, to feed into the system. In regression, you take set of image features together with a set of continuous values and you automatically construct the function. For example, a survival risk score estimator system, then you need a set of images, where each has a risk score attached to it. The differences between the two concern what training mechanisms that are most suitable to use and how the training data is structured.

When automatic systems are built using these systems, the output properties are what matters. Exactly how that output is derived within the system is secondary. This means for example that accuracy and speed are important properties for the design and end-user, but whether it uses a neural network or a random forest to derive that output can be left to the engineer to figure out.

A common distinction among image analysis practitioners, is between unsupervised features and handcrafted features. Unsupervised features mean that the image analysis practitioner do not explicitly choose the mid-level features, and this is one of the main advantages with so called deep learning approaches that have recently been developed [59]. This can be achieved by training on a large dataset without a ground truth [2] or by training on a similar dataset where a large number of labels are available [3].

Handcrafted features are mid-level features that the image analysis practitioner actively chooses, for example, edge structures or color hue. Madabhushi et al. [65] argue for a third distinction of features: domain inspired features that are a subclass of handcrafted features specific to the application domain. This includes for digital pathology, for example, that the entropy of gland angularity is a good feature for the analysis of prostate cancer [65]. However, this distinction does not say anything about whether the features are interpretable by pathologists. The view in this thesis is instead, that these distinctions are useful for image analysis research, but are less valuable for the end-user to understand. In the following section, this user-centric view will be used to summarize what existing image analysis algorithms can do within digital pathology.

2.2.1 Image analysis in digital pathology

A number of reviews describe the types of algorithms that are available today. Gurcan et al. [37] reviewed image analysis methods aiming at creating Computer Aided Diagnostic tools, describing the whole chain from useful pre-processing methods such as color normalization and image registration, to feature extraction and object detection methods. Di Cataldo et al. [19] reviewed common steps involved in scoring of IHC slides as well as algorithm used to score these slides. Riber-Hansen et al. [89] reviewed the use of image analysis in a clinical setting whereas Hamilton et al. [39] reviewed use of image analysis for biomarker research, and Veta et al. [111] focused on image analysis for breast cancer pathology. Irshad et al. [47] provided a deep technical account of methods used for detecting and classifying nuclei, and a recent smaller review by Madabhushi, Lee

[65] discussed the use of *Big data* and Deep learning within digital pathology. Based on these reviews, we will now discuss the state-of-the-art for different digital pathology applications, the algorithms' performance and their shortcomings from a human-centered perspective. To compare the different algorithms, the F1-score will be used when available for the cited sources.

Nuclei detection, counting and scoring Immunohistochemistry staining attaches to the nucleus, the cytoplasm, the cell membrane or combinations of these. Nuclei detectors and cell membrane scoring are the most commonly used algorithms. Nuclei detectors of IHC-stained images perform at an accuracy level around 80-90%. Qi et al. [83] compared different segmentation scores and reached an F1-score (recalculated from the reported precision and recall values) of 0.83 with their own algorithm, and others' F1-scores ranging from 0.60-0.81. For a task of counting 200 nuclei, this means that for the top score 0.83, the end-user would have to remove 34 nuclei that have been falsely counted and add 34 nuclei that have been missed to reach a perfect score, if we assume that the algorithm is balanced. A related task on detecting nuclei on H&E images reaches detection rates slightly higher than that, when using more sophisticated methods. Sirinukunwattana et al. [98] used Deep Learning methods to reach an F1-score for epithelial nuclei of 0.88 (extracted from their chart in Fig 7). The accuracy depends heavily on the evaluation methods used and the dataset that was used for evaluation, but these numbers give an approximate performance of the state-of-the art performance of nuclei detectors.

However, to report for example the Ki-67 label index, it is also possible to estimate the percentage value directly without using nuclei detection as an intermediate state. Tuominen et al. [49] detected pixels belonging to epithelial nuclei, without segmenting them, and used the ratio of positive and negative pixels to calculate the Ki-67 positivity. This decreases the sensitivity of the algorithm to the difficult task of separating fused nuclei, but also make the result more abstract since it is not directly connected to manual cell counting. This could potentially make it harder to detect systematic errors. Gudlaugson et al. [35], use a well calibrated nuclei area ratio based system, and reach good concordance with a manual ground truth based on systematic sampling (stereology), and better prognostic power than a manual approach. Specific errors of nuclei area systems have not been studied, but for example Tuominen et al. [49] removed two "outliers" from their study due to weak staining and due to bad camera contrast. On the other hand, the high prognostic power of these systems has been reproduced [99].

Hotspots and tumor hetereogeneity Tumors can often be heterogeneous. This means that if the nuclei is counted in only a small area, the selection of the area to measure can induce quite substantial error [9]. Once the nuclei have been detected it is possible to model the heterogeneity and make hotspot selections as is used by the software in [99]. It is also possible to make hotspot selection based on more direct measurements, such as the staining component, which is much faster than detecting every individual nuclei [72]. Laurinavicius et al. [62] used calibrated nuclei detectors to model the whole tumor to investigate different heterogeneity measurements, and established that bimodal positivity was an independent predictor of overall patient survival. Bimodal positivity

refers to that there are two dominant populations of cells that have significantly different positivity.

Mitosis detection Mitosis detection algorithms go beyond cell detection algorithms in that when cells are detected, they also need to be classified as being a mitotic figure. Mitosis detection are used in many types of cancer as a diagnostic feature, and is also part of the Nottingham Histological Grade that is used to grade breast cancer tumors. In the MITOS-challenge [90], the best performing algorithm reached an F1-score of the detection rate of 0.78. The state of the art is arguably better represented by the AMIDA13 challenge, which used a more realistic setting where 14 different research groups submitted algorithms for evaluation on a common dataset [110]. The best performing algorithm (from the same team as MITOS) in the evaluation reached a F1-score of 0.61, using a deep learning approach. In the latest challenge, TUPAC16, the best performing algorithm reached a F1-score of 0.65[2]. It is natural that these results are lower than pure nuclei detection considering that this is both a detection and a classification task.

Prostate cancer grading Prostate cancer grading is usually performed using the Gleason system [18], which is based on classifying glandular growth patterns disregarding the appearance of the cells making up the glands. Note that the latest ISUP consensus meeting have modified the system and recommend using the term Grade Groups 1-5 [22]. The glandular patterns that should be detected have not changed, so existing image analysis research based on the old classification scheme are still valid. A number of algorithms for this purpose have been developed. For example, Doyle et al. [20] first detected glandular nuclei using the same approach as is used within IHC counting, to derive architectural features, and combined these with texture features. The algorithm was then evaluated on hand-selected image patches with a single homogeneous growth pattern at the Gleason level. This resulted in a classification system with accuracies of 0.77 for Gleason 3, 0.76 for Gleason 4, and 0.95 for Gleason 5 with the most balanced of the evaluated systems. Both Gorelick et al. [33] and Jacobs et al. [53] used super-pixels and derived features to use in a classifier. The former study reached an accuracy of 0.72 in their most difficult experiment (leave-one-patient out), and the latter study reached a Gleason 3 accuracy of 0.85 and a Gleason 4 accuracy of 0.77. Again it is hard to compare different studies, since they use different datasets and evaluation metrics. If these accuracies were directly translated to a visualization where the predicted class would correlate to the underlying image the user would consider around 10-30% of the area to be wrongly classified.

Metastasis detection in lymph nodes Perhaps the most promising algorithmic performance to date is given by the CAMELYON16 challenge [15], that compared algorithms for detection of metastases in lymph nodes. The best performing algorithm had a AUC-value of 0.92 at the submission deadline. If submissions after deadline are included, the best performing algorithm had a AUC of 0.97. Note that AUC values do not translate directly to F1-score, which has been used so far.

[2] **Personal correspondence**.

Tissue characterization and Prognostication The previous paragraphs report a number of detection level accuracies. This represents the accuracy of when pixels are turned into objects such as cells and glands. These results are not very encouraging, and a pathologist would likely outperform these algorithms if given an appropriate protocol definitions and the necessary amount of time to perform the detection tasks. However, this does not mean that these algorithms cannot be useful. An algorithm can be used to process a much larger amount of information than a pathologist has the time to go through, and this information can then be averaged together and statistically outperform the pathologist on high-level tasks. A number of studies have evaluated such approaches.

Beck et al. [8] built a prognostic model directly from mid-level image features and skipped the normal middle steps of mitoses, glandular structures and cell morphology, that added prognostic information beyond the manual grading system. Laurinavicius et al. [62] showed that bimodality features built from non-perfect nuclei detectors also added prognostic information. Stålhammar et al. [99] also used nuclei detectors to create a proliferation score that was at least on par with the manual method. These studies show that automation can be helpful in creating diagnostic systems in different modes of operation. It is, however, still unclear whether these modes of operation are the best from a usability perspective, since non of them are fully automatic. The first mentioned study [8] used TMAs to perform the analysis, meaning that a pathologist had already chosen the representative area for analysis. The second study [62] used a manual time-consuming calibration procedure to derive its values, and the third study [99] added the manual work of creating consequent cytokeratin stain, which was co-registered with the Ki-67 staining in order to easily distinguish tumor nuclei. This thesis tries to bridge this gap by building systems based on semi-automatic tools with good usability and efficiency for the end user.

2.3 Human-centered design in medicine

The past sections described different areas of research providing knowledge and building blocks necessary to design a digital workstation for pathologists. Design interventions within medicine need to be understood both from a medical perspective but also from a design perspective. These two aspects are not always easy to combine, since they have arised from different traditions. In this final section in the background, these two perspectives and how they might be combined will be described.

2.3.1 Building evidence in medicine

Evidence-based medicine is the current predominant view within western medicine. It means that the practitioner should strive towards including as much external clinical evidence as possible, combined with individual expertise to make clinical decisions with the patient in focus. Evidence is used to decide whether a novel practice is better than the current best practice. Evidence is commonly seen as, but not limited to, randomized trials, meta-analyses and other types of controlled experiments. Even though the term was first formally defined in 1996 by Sacket et al. [93] as “the conscientious and judicious use of current best evidence from clinical care research in the management of individual

patients", Claridge et Fabian [16] notes that evidence-based medicine has been developed through centuries of medical practice. They further list how evidence should be valued. One or many large randomized trials are valued the most, whereas descriptive studies, clinical experience, and outcomes from expert committees should be less valued. Evidence-based medicine has also been criticized. Greenhalgh et al. [34] highlight problems this system: that industry distorts evidence and increasingly sets the research agenda, that the vast amount of evidence is difficult to overview, that statistical significance takes precedence over actual significance, and that algorithmic rules are emphasized over good judgement.

The evidenced-based medicine view has a large impact on what scientific studies are seen as important and how different outcomes are judged. For example, one of the most common types of study performed within the digital pathology domain is a comparative validation of the review with a microscope and digital workstation. A recent meta-review found 38 such comparative studies, covering over 5000 case reviews, but concluded that more research was needed since the studies were too heterogeneous to perform a meta-analysis [30].

Statistical testing of quantitative measurements are favored. Within digital pathology, and pathology in general, a large number of studies are measured either by its predictive value, prognostic value, or by its concordance. Predictive value translates to whether the diagnostic categorization is able to predict if the patient responds to a particular drug, treatment or similar. Prognostic value is from a statistical point of the view the same type of measurement, but the term is reserved for predictions about outcome in untreated individuals [48]. Concordance refers to the degree to which different pathologists make the same diagnostic decision when presented with the same case. These concepts relate to validity and reliability. If the predictive or prognostic value is high, it means that the diagnostic method has a high validity and is able to inform actual clinical decisions at a population level. High concordance means high agreement between multiple pathologists doing the same diagnostic task. This increases the reliability for the individual patient, i.e., the probability of an erroneous diagnosis for an individual patient is low. A good diagnostic test should both have high predictive value and concordance, even though the measures can be interdependent.

When evaluating digital image analysis algorithms, the methods are very similar to when evaluating diagnostic procedures. However, concordance cannot meaningfully be measured, since most algorithms produces the same result every time. Instead the variability of algorithms is estimated by assembling ground truth datasets. A ground truth dataset consists of a number of images that have been manually annotated by one or more pathologists. The type of image and the type of annotation differ from case to case. For different prostate cancer detection algorithm the annotations types differ: one study meticulously contoured all areas by a physician into nine different classes [33], whereas another study focused on contouring areas of homogenous tissue of six classes [20]. These annotations are then considered being the ground truth, and an automatic algorithm is evaluated against how well it can reproduce that ground truth. This type of experiment is not able to say anything about whether an algorithm could possibly outperform the pathologist, since the ground truth dataset is created by the pathologists. To do that, studies comparing the predictive value of the automatic method against the

manual method need to be performed.

These statistical metrics form the basis of evaluation in the digital pathology domain, and papers are judged on how well they perform on these metrics, and how sound the statistical method of the particular study is. And the idea is that once a diagnostic method or procedure is better than current best practice, that practice should be updated with the new findings. This forms a sort of community level design process were poor designs are slowly replaced with new better ones.

2.3.2 HCI and design research

HCI and design research concern the creation of new artefacts and how well they work in the larger context. Simon [97] sees design as *courses of action aimed at changing existing situations into preferred ones.* This is based on an idea that design and other artificial sciences can be considered being science if they are seen as reproducible processes. This means that a certain design process can be seen as useful if it again and again produces successful outcomes, even though the actual outcomes are different. The outcomes instead depend on the context within which the design process is carried out. This view fits well into the problem of creating new tools for pathologists. Another view on design is that design practice and science should be kept separate, since science tries to be true at all times [104]. With this strict view on science, medicine cannot be seen as a science either, since evidence-based medicine is based on continuously updating the current best practice. For this thesis, this distinction is less important and instead the view is that design can inform medical practice and vice versa. But when it comes to the research practice within HCI and design compared to medicine, there are a number of larger differences, which the next sections will try to highlight.

2.3.3 The role of iteration

Design improvements are the product of an iterative process. The development of diagnostic methods can also be seen as iterative; a new method can build on an old method. The main difference between design and medicine lies in the pacing of the iterations. To put it bluntly, from a design perspective it seems outrageous that the medical community has spent the time to perform 38 studies [30], comparing basically the same digital workstation with the microscope, when the same amount of time could have been spent developing a better one! In design, the number of iterations are more important, and only small studies are performed that are aimed at gaining as much information in order to inform the implementation the next iteration. A now established rule of thumb in design literature states that five users is enough for evaluations. This rule has its basis in a cost-benefit analysis investigating how many new usability issues are found in a system when more users are added to a study [73]. When performing usability studies to support an underlying design process it can, for example, be better to perform five studies with five participants, than one study with 25 participants.

Digital pathology is considered to be the domain of inquiry for this thesis. When designing for a novel domain, it is common in the HCI literature to first perform a grounding study, followed by an iterative design process and evaluation of a final prototype.

For example, Santosa, Wigdor [94] first performed a field study to explore how people used multiple devices such as computers, tablets and smartphones together. This was then followed up by making an iterative design intervention, to make a software that leveraged the ground work, and making a multi-device problem solving software in a second study [40]. Similarly, Knaving, Wozniak et al. [58] first presented a grounding study in the form a ethnographic study of amateur runners and their motivations. This was then followed up with a design intervention in the form a portable device that could be worn during races [117]. Within digital pathology, a similar design approach was used by Randell et al. who tried to understand pathologists working with microscopes [85, 88], and then used that knowledge to design and evaluate a digital workstation (e.g [87]).

2.3.4 The context of studies

Medical studies of diagnostic techniques are quite good at creating a realistic evaluation setting for their controlled studies. The most common approach is to validate novel methods for large cohorts of patients, where the diagnostic task is incorporated into everyday work. A common drawback from the HCI perspective is that while the number of cases is large, the number of pathologists participating in the studies is usually low. However, concordance studies usually contain a large number of participants. Because these are performed as controlled studies, where the response is the most important, the process of how the response is generated is not recorded. This means that when a medical study finds a problematic situation or design, it is difficult to understand the underlying reason. In an HCI study, dependent variables are almost always recorded along with some sort of qualitative data to aid understanding of the process by which the response is generated. In the controlled studies performed within the scope of this thesis, both quantitative and qualitative recordings have been performed, which help us understand how diagnostic decisions are made.

This summarized the overall design approach used throughout the different studies. Note that the words HCI and design has been used interchangeably and that there is a lot more nuance to these concepts if you look more deeply into to the design literature. A more detailed record of the methodology that has been used is provided within each paper.

Chapter 3

Paper summary

Based on the presented background, we will now shortly summarize the eight included papers, P1-P8, in this thesis. The topics of the papers were chosen iteratively throughout the PhD project. Roughly, two or three projects were developed simultaneously and when one paper was finished, another project was initiated based on the previous projects. This approach created a somewhat coherent story between the different papers. The papers are presented in a rough chronological order, except P6, that was one of the first projects but is easier to understand if it is read later.

In total, five evaluations of high fidelity prototypes are presented, these have five (in P2), six (in P3), two (in P4), four (in P6), and eight (in P7) pathologists participating in the study. This is sufficient to provide an acceptable understanding of the use of the prototype, but the statistical tests in the studies are quite low-powered and are only able to detect large differences in the quantitative measurements. The aim was to gain a rich understanding of the digital pathology domain, many smaller studies were therefore prioritized over one or two larger studies.

The type of studies differs between the papers. P1 and P2 can be seen as grounding studies, and P3-P7 represent the development and evaluation of different design interventions. The last paper (P8), is instead based on design reflection, where we try to look back at the earlier work and derive generalizable design advice.

This work builds upon the work by Randell et al. who grounded their investigations by studying the work with microscopes. In contrast, our work is based on user studies of the first generation of digital workstations instead of with microscopes. This made it possible to perform more iterations *further into the future* within the limited time of the thesis work. This also limits our design space so that, for example, telepathology with video cameras mounted on microscopes are not even considered [77].

P1: Implementation of large-scale routine diagnostics using whole slide imaging in Sweden. This paper presents hands-on experiences from implementing whole-slide imaging within a medium-size pathology laboratory. The paper describes how digital pathology can be practically implemented in order to perform diagnostic review of routine cases. This means that the laboratory processes need to be adapted because the scanners are more sensitive to artefacts than when pathologists review slides with the microscope. The main benefits of performing the review with computer workstations

have been the improved ergonomic situation (especially with respect to neck problems), the improved overview of a glass slide, and the possibility to review cases from a distance. In a questionnaire sent out to the users of the system, a number of drawbacks with the first generation of the digital solutions were identified. Some of the drawbacks were later addressed within projects of this thesis:

1. The speed of digital review was not faster than with the microscope.
2. Digital annotations were not used as much as expected.
3. The digital reporting of a case was not more convenient.
4. The automatic counting algorithms were not used as much as expected.
5. The navigation of the digital slides could be improved.

P2: Slide navigation patterns among pathologists with long experience of digital review In order to improve the navigation of digital slides, it was important to understand what normal navigational patterns were and how they could vary. This could potentially guide a number of design interventions. Since the aim was to support the design process, the study focused on dynamic properties of the movement. This meant that heatmaps and similar that had already been studied were left out. Instead the navigation tracks of five pathologists reviewing a set of representative cases were recorded together with their think-aloud statements. Using a thematic analysis of the movements, six different types of movements were identified and analyzed. A main result that is relevant for the following studies was that more than half of the navigation time was spent on tedious panning back and forth within the digital slide, a pattern we named *Cover panning*. This exploration pattern was used to determine whether a diagnostic finding was absent in a slide. This is distinctly different from what we called *Sporadic panning*, which is also pure panning but with a more random direction of search. This type of movement was associated with an opportunistic search, meaning that the pathologist believed that a particular finding should be present in a slide and tried to find it as soon as possible. Another important navigation strategy that was used was to point at findings, and then quickly zoom in and then zoom out, which was named *Dip-zooming*. These highlighted navigation strategies can support the design of the digital workstation in many ways. In P3, these insights were used to design navigation with a number of input devices. In P7, visualization techniques was used to reduce the need for cover panning, replacing it with Dip-zooming.

P3: A comparative study of input devices for digital slide navigation. The purpose of this study was to evaluate different input devices for navigation in diagnostic review. An initial limitation was set to input devices already available on the market to increase the likelihood of creating something that could actually be used. The design process aimed at adapting each of the selected input devices to support the exploration patterns described in the previous paper. A standard computer mouse, a portable touchpad device, and a 6-DOF controller (a puck that senses when it is pushed and twisted in different directions) were used as a basis for different software implementations. Each implementation went through an iterative design process to ensure good performance before they were evaluated. The evaluation estimated the difference in physical and cognitive load between the devices using the established NASA-TLX method. The study

also captured the six pathologists' user experience performing diagnostic review with the devices. Both the computer mouse and the 6DOF-controller performed well, whereas the touchpad performed poorly. When asked directly after the study, five out of six pathologists preferred the 6DOF-controller over the computer mouse. Informal observations when visiting pathology departments after the publication of the study seems to support the finding that using a touchpad for digital pathology is generally a bad idea, and it is hardly used by anybody. A number of pathologists still like the 6DOF-controller and use it in their daily work, but many prefer using the computer mouse. A few users have acquired a 6DOF-controller and uses it for zooming and for easy access to hotkeys, whereas panning is performed with the computer mouse. A future follow-up study could look further into this bimanual control, but could also evaluate the NASA-TLX method as a predictor of future performance.

P4: Improving the creation and reporting of structured findings during digital pathology review. The idea with this paper was to combine structured reporting with a shoe-boxing interface adapted for diagnostic review. Historically, the main purpose of structured reports has been to categorize measurements in order to allow for automatic data extraction, in contrast to the free-text reports that are predominant at most hospitals. Shoe-boxing interfaces were described in the sensemaking section as an interface that facilitate the temporary storage of findings, such as diagnostic key images or different tumor measurements. This paper reports on a design-based research study with the aim to develop an understanding of the design space for structured reporting within digital pathology. The paper reports findings from the design process and a small user study with the final prototype. This paper touches upon two of the user needs reported in the paper P1: that digital annotations were not used as much as expected and that digital reporting was not more convenient than reporting at the microscope. To simplify the study, we decided to make an optimized interface to only review and report on a particular kind of diagnoses: partial mastectomies. The final evaluation resulted in very optimistic users, but the study does not tell us much about whether this approach could be generalized to other pathology domains. This shows a very interesting future avenue of research since this way of reporting could be relevant also for domains other than pathology within medical imaging.

P5: Towards grading Gleason score using generically trained deep convolutional neural networks. This study was a first of two studies investigating how image analysis can be developed and used within digital pathology. This describes the algorithm development of a larger project, where an interactive prototype using the algorithms also were developed. The algorithm uses a deep learning network that had been trained on photographic images. The top of the network was then discarded and replaced with a random forest machine learning component that uses the deep learning features to re-train on prostate cancer growth patterns. The paper describes a purely technical evaluation of the approach. The technical validation used cross-validation of the algorithm's predictions on a ground-truth dataset, which consisted of image patches of prostate cancer with a homogeneous grade for all glands. The most interesting result from this thesis's point of view, is that when the algorithm is evaluated in full resolution, the patch

accuracy was 81% compared to ground truth. When the overall score was calculated, combining the classification result from multiple patches, the overall accuracy increased to 89%. This highlights that there is a trade-off between resolution and overall accuracy.

P6: Feature-enhancing zoom to facilitate Ki-67 hot spot selection. This paper was the first of two to investigate an idea to use image filters to modulate the visibility of different image features in low magnification in a pathology slide. The purpose with this approach was to visualize the pathology slides in a way that different diagnostic tasks become easier to perform. In this first paper, the focus was on selecting a Ki-67 hotspot. This is a common task, at least in Sweden, where breast cancer pathologists are supposed to select an area consisting of 200 cells, where the percentage of positive cells is as high as possible. This is a challenge, since the brown positive nuclei are not visible in low magnification, which makes it impossible to see the whole slide at the same time as the nuclei is visible. The paper presents a feasibility study of an attempt at making the nuclei visible in an interactive prototype. The presented method included color deconvolution to separate out the positive staining component, and Gaussian kernel filtering combined with a linear contrast enhancement to make the nuclei larger. The method was applied on the full size of the image in the highest magnification, which was then subsampled. The subsampled separated stain channel was then combined so that the staining component was visible in low magnification whereas the original image was maintained in high magnification. This created an effect of *zooming in into the original image*. The visualization system was evaluated in a user study with four pathologists. The study compared the accuracy of detecting hotspots with and without the system, by letting the users select between pairs of possible hotspots in low magnification. The study also included a short debriefing interview after the trials with each participant. A number of important findings were reported that were used to support the next design iteration in the subsequent paper.

1. The design pattern to zoom into the original image was well received.
2. The transition between low and high magnification was not smooth enough.
3. It was difficult to see what was tumor cells, since the background intensity had been decreased too much.

P7: Scale Stain: Multi-Resolution Feature Enhancement in Pathology Visualization. This paper used the understanding gained in the first study (P6) to create a new and improved version of the visualization. The enlargement mechanism in the previous version was replaced with max-value subsampling, which performed the enlargement and subsampling in the same processing step. This created a smooth transition between low and high magnification. The tumor cell visibility problem was solved by making the method interactive: it is possible to modify the sensitivity of the visibility enhancement and how much of the background staining that should be visible. The processing method also became computationally parallelizable, which was necessary in order to make the method suitable for routine deployment. The final system was evaluated in a usability study with eight pathologists. The study measured efficiency and accuracy of two diagnostic tasks (hotspot selection and bacteria detection), recorded navigation using the same method as in P2, and included a short debriefing interview. The use of the tool

for two diagnostic task resulted in a 15% efficiency increase at maintained accuracy. As predicted, the behavior of the participants changed completely when the visualization tool was used. Without the tool, the participants in the study used Cover-panning in medium magnification in order to conclude absence of findings. When the targets was visible in low magnification with the tool, a number of Dip-zooms were instead used to check the most suspected regions of interest.

P8: Understanding Design for Automated Image Analysis in Digital Pathology. The last paper in the thesis, in contrast the other papers, was based on design reflection. The aim with this study was to look back at a number of design activities and try to extract generalizable results. To make this study possible, design artefacts that were created throughout the PhD project was saved and organized in order to make a retrospective analysis possible in line with Obrenovic [74]. The analysis generalizes on experience from three image analysis projects. The first project included corresponds to P6 and P7, which was especially well documented. The prostate cancer project presented in P5 included a few user studies that were performed within the PhD projects but have not been published separately. I was also involved in a commercial project at Sectra aiming at creating an automatic cell counting tool, which included substantial user study data that have not been published either. These three projects make up a large basis of data. The data comes from different projects, but are still from within a confined scope.

The paper presents a retrospective analysis of the different design artefacts that were created during these three projects in order to be able to distill actionable design considerations that another designer could reuse in a similar scenario. Four main considerations were derived:

The first consideration, *Design to improve verification and correction*, is a consequence of the fact that when automation is implemented in practice, the manual task is transformed into a verification and correction task. This means that in order for automation to be beneficial, the time it takes to verify and possibly to correct the result will determine the efficiency of the whole system. There are many ways to visualize a result, and the interaction design should aim at exploring different visualization, calibration, and exploration interfaces to achieve good efficiency. The need for verification is independent of the accuracy of the algorithm, if the algorithm is below the non-perfect level.

The second consideration, *Design for algorithm transparency*, communicates that it is still important to consider the errors that naturally occur. This means that the designer should make sure that errors that occur are transparent to the user. A major problem with creating such designs is to be able to communicate the algorithmic results without occluding important information in the original image. This becomes even more challenging, if the algorithm outputs useful uncertainty information, which can increase the amount of occlusion even more.

The third consideration, *Support verification on different levels of detail*, talks about how to support the speed-accuracy trade-off in a verification scenario. Depending on how close a certain score or estimated percentage value is to a clinical decision border, the need for good accuracy varies. This means that the verification task that the pathologist performs needs to be possible to perform quickly in some cases, and more careful in other

cases. This can be implemented, by visualizing the result on different levels of detail, that the user can choose between in order to modulate the speed.

The last consideration, *Design for communication with clinicians*, echoes the conclusions in P1 and P5 that reporting findings is important for a digital workstation in general. But in the context of automation reporting can be even more important. Assume that the time it takes to automatically count a number of nuclei and verify and correct the result is the same time as it takes to manually count all the cells. If the task is performed in the first way, the value can automatically be inserted in the final report and be traceable back to the original count. If it is done manually, the value instead needs to be dictated and transcribed, which decreases the overall efficiency. This means that designing how the final values are reported can be one of the major time savers and not the task itself.

Chapter 4

Discussion

By reviewing the papers that are included in this thesis, a number of interesting discussions can be derived. The following sections will discuss reoccurring themes and relate them to the presented background.

4.1 Cognitive strategies in multi-scale images

A number of the presented papers study how to solve problems in large multi-scale images. This relates to a number of areas covered in the background section: theory of problem solving, how human visual search process work, and how existing design solutions for multi-scale images look like.

The results from P1 and P2 pointed towards that systematic search by panning were important tasks for pathologists. Current digital interfaces do not support this task well. A standard P+Z interface requires the user to restart the panning movement over and over again, which is not a sustainable task for pathologists to perform. An O+D interface increases the convenience slightly, since the user can hold the mouse button down in the overview in order to pan over large areas using the increased panning sensitivity. Users reported in P1, that this was the major panning strategy to get past limitations of the interface. In P3, a number of input devices were tailored to support this task well. A lock function was introduced for the computer mouse, so that clicking the mouse wheel increased the sensitivity to a similar level to that of the overview. The 6DOF-controller has a high sensitivity mode close to the end of the puck's movement range, and the touchpad has a double click and hold function that also increased the sensitivity. All these individual features were well received by the users, even if the touchpad solution overall performed poorly in the evaluation.

These design solutions could also work well within similar domains working within large two-dimensional spaces. Many applications today include similar lock functionality, where the mouse controls the scrolling using rate control instead of using position control. In future research, it would be interesting to investigate the difference in document navigation performance between the two types of lock modes. That is, to examine the apparent similarity between the navigation patterns that pathologists performed in Paper II and the reading patterns described in the study by Hornbaek, Frokaer [44]. The behaviors described in their paper (quickly flipping through a paper, systematically

searching for targets, and jumping to already visited part in the overview) all have analogs with pathology navigation.

In the same paper [44], it was suggested that an overview invites readers to perform additional, perhaps unnecessary, explorations. A recent study with pathologists came to the same conclusion, where the additional explorations were seen as increasing the quality of the review [87]. Increasing the quality is of course beneficial but it does not help pathologists to produce more reports with less resources. This phenomenon might therefore be problematic, considering the current lack of pathologists in Sweden. An import research question is therefore: How can interfaces be designed that increases the efficiency while the quality is maintained?

Designing to support these systematic search tasks is crucial to improve the pathologists' work since it is made up to a large extent of time spent examining tissue. This can be performed as in P3 by implementing support for the task manually using input devices, or as in P6 and P7 by using an alternative visualization. A third approach, which is more commonly mentioned in the literature, is to use different digital image analysis algorithms. In practical applications, as is argued in P8, it is a good design idea to stack the automatic approach on top of the manual and visualization approaches so that there is a fallback in place when the automation fails.

One of the most important findings from P2 was the role of cover panning in order to conclude the absence of a particular finding. Knowing that something is absent requires considerably more work than knowing that something is present. Therefore, for many tasks the pathologist needs to exhaustively search an area, making sure that everything has been covered. Because of the top-down processes in the visual system, it is only possible to search for one feature at the time. On the other hand, if something is present, the pathologist only needs to perform opportunistic search in the form of sporadic panning. Only if that something is not found, it is needed to perform an exhaustive search. This was reflected in P2, where *cover panning* represents systematic search, and *sporadic panning* represents opportunistic search.

Interestingly, this distinction of the movement were only based on studying the navigational patterns within the image, and not using any of the think-aloud data. But when the think-aloud data were correlated with navigation patterns, this distinction became clearer.

This type of problem-solving can also be compared with the haystack problem; finding a needle in a haystack is hard, but to conclude there is no needle at all is a lot harder. In reality it is difficult to guarantee that something very small is not present, so in the pathology report it is often written that an exhaustive search have been performed for X, but that there is no guarantee that X has not been missed. The visualization approach presented in P7 tries to remedy this problem by guaranteeing a close to 100% sensitivity for stained objects.

4.2 Image to object mapping and how to deal with its errors

In the background section it was established that there exists a large body of research dealing with information visualization. In order to make use of those techniques in digital pathology images, the image pixels need to be turned into objects. Computer vision and machine learning are techniques that can be used for this purpose. State-of-the art algorithms can turn pixels into cell nuclei with around 80-90% accuracy and detect malignant prostate glands at a similar level of accuracy. This level of performance is impressive, but it still means that many errors need to be corrected in order to reach a perfect mapping between the pixel-based image and the detected objects.

Our first application dealing with this problem was the application described in P6: a prototype that enhances the pixels belonging to positive nuclei without turning them into detected nuclei. In P6, a feasibility study with users was performed of a technically simple algorithm to perform this enhancement. In P7, this prototype was then further developed by improving the enhancement algorithm and by adding different interactive patterns that was learned during the feasibility study.

A comparison of the enhancement method presented in P6, P7 and a reference color deconvolution is given in Figure 4.1. The figure highlights the progressive design improvements. The method in P6 increased the comparative performance between areas of different positivity. However, it lacked the ability to maintain visibility for areas with low positivity. This was improved in P7, where the area with 9% positivity has a much higher visibility.

Something that is not visible in the figure, and difficult to reproduce on paper altogether, is the improved continuity between the low and high magnification representation. This was something that users complained about in P6, but did not bring up even when asked for in P7.

The enhancement methods improved the task that was performed, since they made the positive nuclei more visible. They make use of what normally seen as a preprocessing step for nuclei segmentation, but the methods do not take the step to make objects out of the pixels. Maintaining this conceptual connection between the original image and the output of the automatic system does however not require creating a smooth transition between the two. In a development project at Sectra, which is briefly mentioned in P8, two types of cell counting visualizations were highlighted, see Figure 4.2. With one of the types, it is very hard to know whether the detected cells are actual cells, and it is not possible to detect whether a nucleus is positive or negative, because the visualization has lost the conceptual connection with the original image, which means that the user has to think or act in order to verify the result. This increases the cost of comparing the class of the detected object and the class of the image.

As soon as the pixels are turned into detected nuclei, existing information visualization techniques can be used. The selection of techniques is however limited by two factors. First, the spatial correlation between the visualization and the background image needs to be maintained similarly to geographical information systems. Second, the conceptual connection to the background needs to be maintained, i.e., it cannot be occluded.

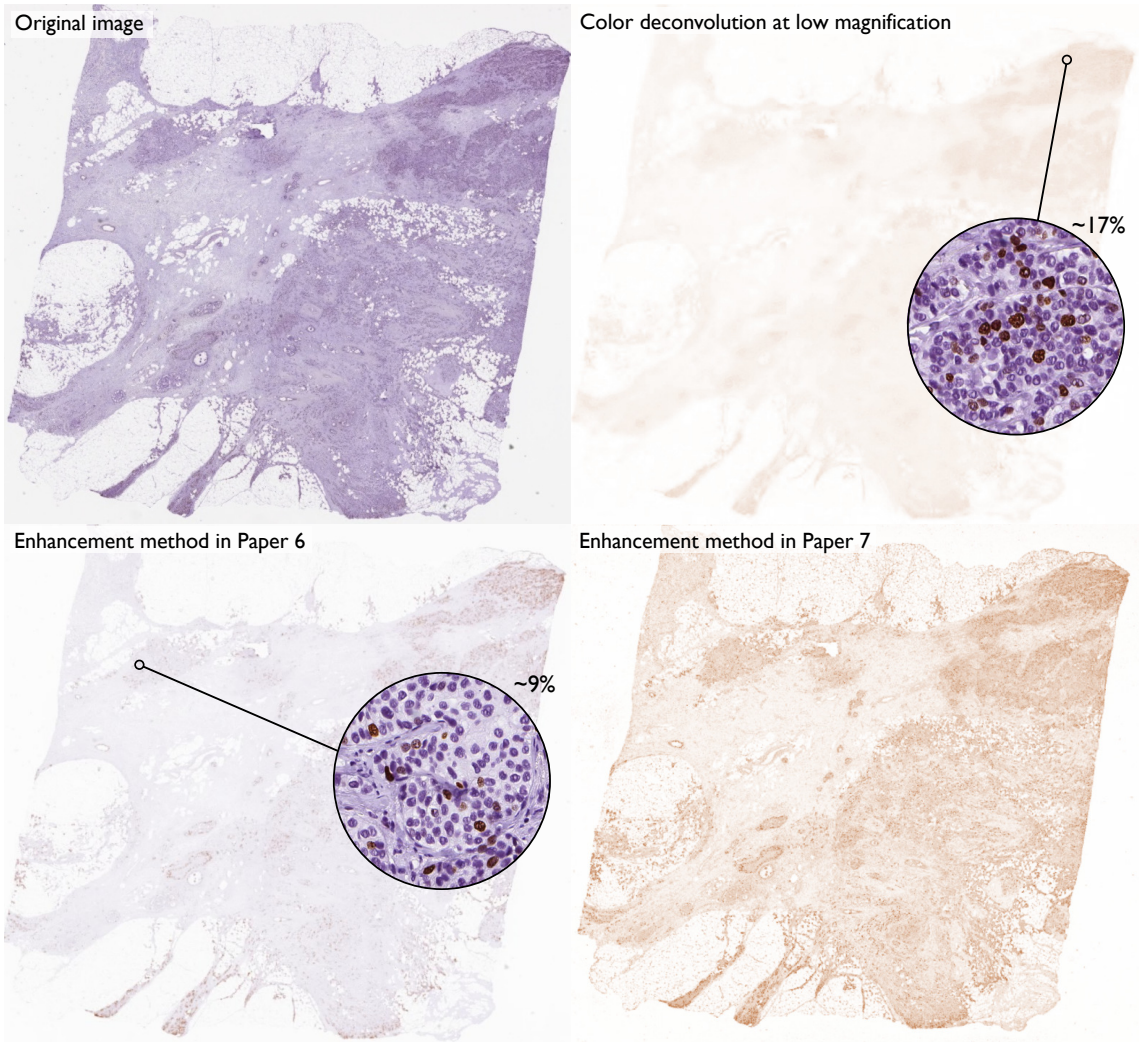
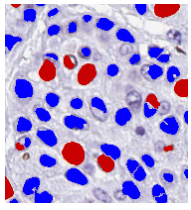
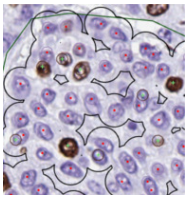


Figure 4.1: Comparison of two enhancement methods and standard color deconvolution [92]. Compare how the two highlighted areas are reproduced in low magnification.



(a) Result visualization that occludes cells



(b) Result visualization where it is possible to see the result

Figure 4.2: Comparing two types of visualizations to report automatic cell counts.

Lobo et al. [63] compared different interfaces to detect errors between spatially matched satellite images and maps. This is a very similar situation, since the satellite images have a photographic nature similar to the pathology images, and a map has a graphical nature in the same way as a counting overlay. That study recommends using translucent overlays to make these comparative tasks over showing the graphics and the image side-by-side, or different interactive blending and lens methods. Our design is most similar to the translucent overlay but blends the graphics with the image by local translation. This points towards that findings in digital pathology can be translated to GIS systems and back. A future study interesting for both domains would be to reproduce the Lobo study but focus on different ways of creating overlays, not only translucent ones.

It is possible to envision a number of scenarios where being able to detect errors in the pixel-to-object mapping is a crucial task. First, algorithms tend to fail unpredictably and non -linearly [4]. Figure 4.3 displays an algorithm that has an estimated F1-score of 0.80 in line with the state-of-the-art. It completely fails when there is a non-specific staining component that a pathologist would not even think about if counting manually.

An appropriate visualization method can ensure that a large number of errors are detected, which is especially important in the pathology setting described in this thesis. The detected error can then be used in many ways. The most common way is probably to alert the algorithm designer who can then make sure that the same error does not occur again. In the example in the figure, an edge filter solved the problem. The role of the visualization is here to make sure that these errors are discovered in the first place.

The visualization could also be used to improve a machine learning algorithm more directly. By letting the user record the errors and correct them, the corrections could be used to improve the algorithm themselves. This approach was recently used for example to label photographic images [68], and involves modeling the inherent noise that arise from user corrections. One of our collaborators are exploring this approach based on pathologists' corrections, with promising preliminary results, pointing towards an exciting future avenue of research.

Visualization can also preserve a richness of the data, which an automatic decision or descriptive statistic is not able to do. As reported in the background section, Laurinavicius et al. [62] used nuclei detection algorithms to detect and classify all the tumor nuclei in a large set of Ki-67 breast cancer slides. This was then used to look at different heterogeneity measurements and their connection with overall patient survivals. Their carefully calibrated nuclei detection algorithm made it possible to detect these patterns. By using the method developed in P7, this tumor heterogeneity can be visualized without detecting the nuclei in the first place, which results in a rich understanding of the positivity distribution. In Figure 4.4, two tumors with different bimodality using the novel visualization technique are displayed.

I would argue that there many similar opportunities to visualize pathology images in different ways, to perform exploratory research based on digital images. There are also many opportunities in designing different clinical tools that make use of visualization techniques. In P8, the aim is to present advice to perform design work in building such tools. There are probably many solutions to the design problems that are described. The point is that by using the developed design guidelines, it is possible to speed up the design

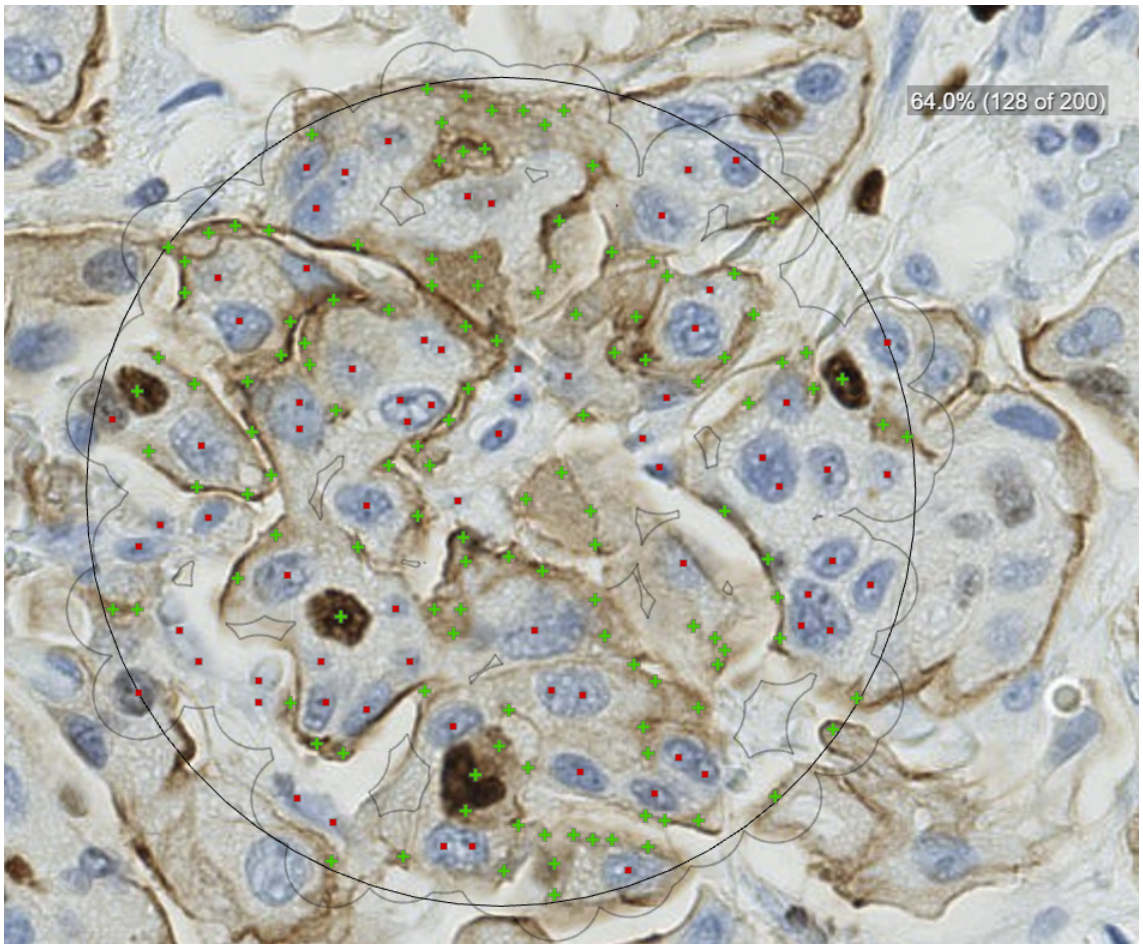


Figure 4.3: Automatic algorithms can fail unexpectedly when in front of an unknown situation. The image displays automatically counted positive (brown) nuclei by an algorithm with an estimated F1-score of 0.8. The non-specific staining component is counted as positive nuclei and the estimated percentage reaches 64% instead of a percentage below 10%.

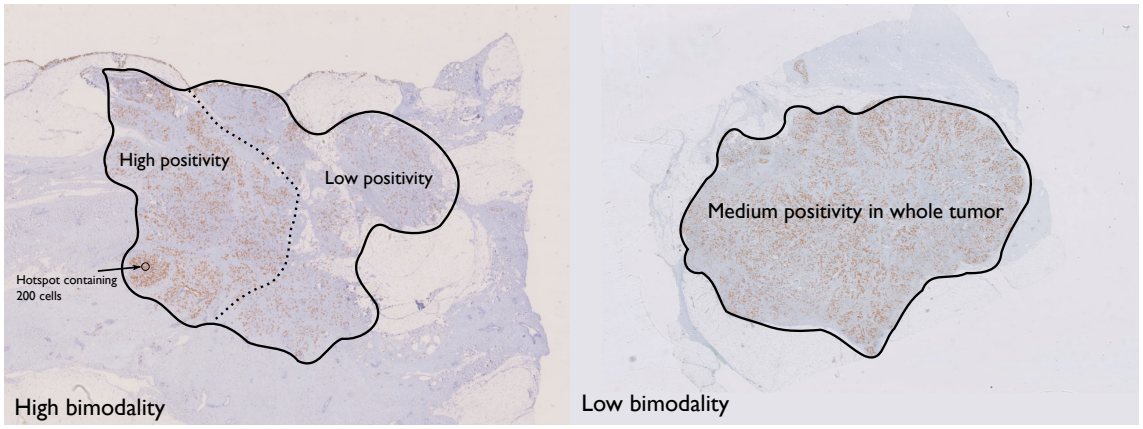


Figure 4.4: Example of different bimodalities as described by Laurinavicius et al. [62] of two breast tumors with similar overall Ki-67 positivity. The staining expression has been enhanced by the method described in P7.

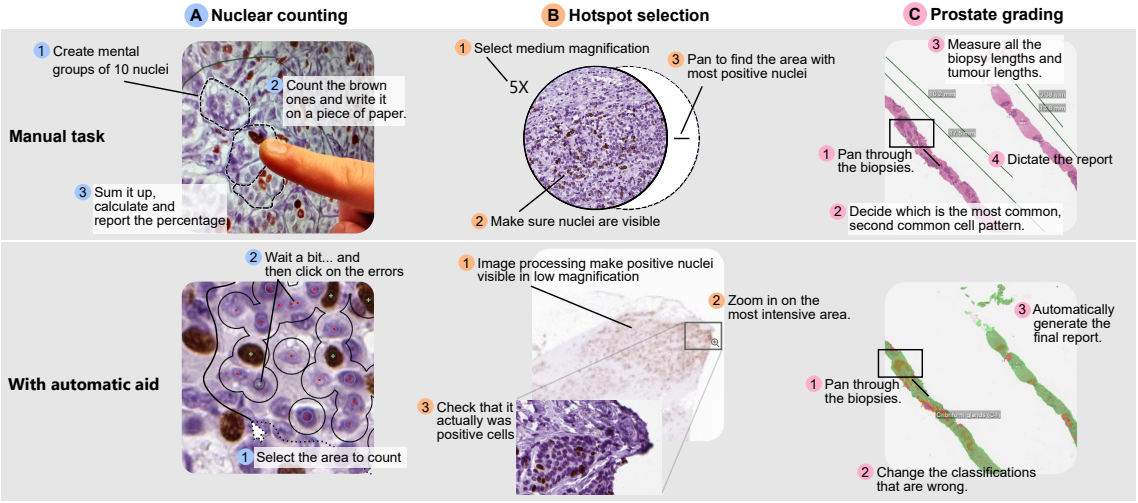


Figure 4.5: Figure from P8, comparing manual tasks (A-C) and how those tasks can be transformed when an automatic algorithm is used to build an interactive tool.

process by first trying out solutions that have worked before. During the development of a digital workstation, the reformulation tactic also apply to the tools that are developed because digital images are used instead of glass slides.

As a general rule, the visualization and the interface design for presenting and interacting with algorithmic output should be a designed experience where the algorithmic characteristics are combined with efficient visualizations and pathologists' capabilities. When this design work is performed, it is useful to think about how different diagnostic tasks can be reformulated. In P8, three examples are given how this could work: for cell counting, hotspot selection and the prostate cancer grading. The corresponding figure is provided in Figure 4.5.

Overall there are many possibilities to perform HCI or visualization research within digital pathology to make sure that the power of automatic algorithms can be turned into patient benefit by being implemented in clinical routine pathology.

4.3 Comparing design processes

A human-centered design process starts with the user in focus, but there is no guarantee that a process carried out at one place, can be reproduced within another context. This thesis together with its appended papers, form a design process that has been publicly documented, which has been carried out as a research process with industry connection in Linköping. Interestingly, a similar human-centered design process was carried out to design a workstation for digital pathology called the Leeds Virtual Microscope (LVM) [87], which largely predates this thesis. The leader of that process has co-authored P7 and P8, but the design processes have for the most part have been carried out independently. Now there exists an in-depth description of the design decisions taken in their process, published by Ruddle et al. [91]. This gives us a great opportunity to compare two separate design processes sharing the same purpose.

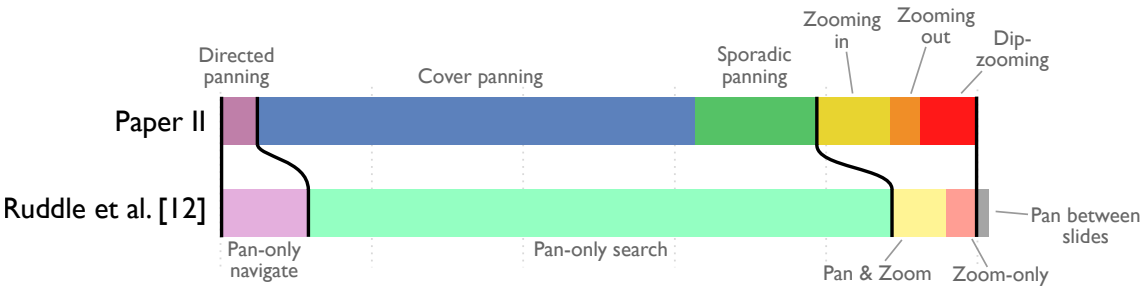


Figure 4.6: Comparing the time spent on different navigation strategies between the study presented in P2 and Ruddle et al. [91]. In order to match the percentages, a few categories have been merged.

In P2, the pathologists' movements across the digital images were conceptualized in categories called navlets. A similar conceptualization was published by Ruddle et al. [91], who defined meta-actions similar to the navlet concept. Comparing the definitions, the *Directed panning* navlet is almost exactly the same as the *Pan-only navigate* meta-action. *Cover panning* and *Sporadic panning* navlets can be combined into the *Pan-only search* meta-action. With respect to zooming however, all actions need to be merged into a single category in order for them to be comparable. A comparison of the reported percentages is given in Figure 4.6.

The statistics from the two different workstation with similar cases are largely reproduced between the two studies. The largest difference is that the amount of zooming was higher in P2 and that the amount of directed panning was higher in the Ruddle study. This can be due to differences in the user interfaces, that different cases were used, or that different pathologists participated in the studies. Overall, the navigation behavior is very similar considering the differences.

Slightly different evaluation methods were used to study the navigation. In P2 the navigation patterns were recorded together with think-aloud statements. This means that we know more about the participant's goal setting behavior. Our view was that the main reason why a participant revisited a certain area was to search for one thing at the time. This type of goal setting data is harder to record otherwise. Another insight made possible by the think-aloud statements were that we saw cover panning as something associated with concluding absence, whereas sporadic panning is a type of opportunistic search, when there is perceived high prevalence of the sought-after feature. However, the Ruddle et al. [91] study has higher validity since their method is less intrusive; the think-aloud method is known to slow the review down and could have an effect on the time spent on different activities [23]. This means that their recorded behavior is likely to be more accurate than the study presented in P2.

When comparing the design process of both workstations a number of shared design goals can be detected that have resulted in different design solutions. To highlight these differences, a number of different common challenges are listed next, along with their respective design solutions.

Cognitive lag A cognitive lag is introduced when pathologists have to jump between regions of interest by panning and zooming.

LVM Regions of interest can be identified and retained in a large thumbnail. This makes it possible for the user to click in the thumbnail to quickly jump between them and remember where they are.

Linköping Supports shoe-boxing of regions of interest. One button is designated to store a field of view and two other buttons are designated to cycle through stored views.

Systematic search The diagnostic review is mostly made up by systematically searching for findings with panning.

LVM A gaming mouse is used to make sure this task can be performed as fast as possible, and a large display is used so that the pathologist see as much information as in the microscope.

Linköping The visualization method presented in P6 and P7 reduces the need to perform for a number of tasks. For the remaining tasks the 6DOF controller is configured to optimize comfort and efficiency.

Panning sensitivity It is problematic to make small panning movements when the sensitivity is high due to the large magnification.

LVM The large thumbnail decreases the size ratio between the detail view and the overview. This makes it easier to control the panning movements, since the magnification differences are less than normally provided by digital pathology interfaces.

Linköping The mouse-lock functionality uses a multiplier of the panning which is less than the ratio between the main view and the overview. This makes it possible to steer the movements by looking in the overview but controlling the main view.

Rate and position control Rate control is good to support systematic search tasks, but fast position control is preferred when working fast.

LVM The game device that can be used in a portable setting perform well in an evaluation, and the mouse performed well in another controlled study.

Linköping Both the The 6-DOF controller and the position controlled mouse performed well in a comparative evaluation.

Microscope similarity in navigation

LVM The LVM is like a microscope in the way that zooming is performed in the middle of the display. This increases the cost of performing dip-zooming but is a more intuitive starting point for a pathologist used to working with a microscope.

Linköping To get an overview, the design forces the pathologist to zoom out before zooming in again since the large thumbnail is missing from the design. This is the same mechanisms used in the microscope, when you switch to low magnification lens to get an overview before you zoom in again.

Microscope dissimilarity in navigation

LVM The thumbnail system makes it really fast to jump between different ROIs, which also was a common behavior and increased compared to when the microscope was used in their evaluation study.

Linköping The mouse pointer based interface makes it easy to do frequent dip-zooms in the image, quickly switching between low and high magnification and making the connection between them. This design is highly beneficial in order to be able to support the visualization method described in P6 and P7.

The two design processes have chosen to focus on different aspects of the usage, and have made different trade-offs that have resulted in different design solutions. Both solutions show what a human-centric design process can accomplish and they both perform well when evaluated with users. In contrast, the predecessor system used in P1 appears to have been developed from a technical perspective, without consideration of the design goals that the Linköping and Leeds efforts targeted.

Chapter 5

Conclusions and suggestions

Each paper has its own conclusion, but to make the results more accessible, this chapter describes what different roles within the digital pathology community should take away from this thesis.

5.1 Implications for pathologists

Digital images have the potential to transform how diagnostic review in pathology is done. It is the novel technology paradigm that makes this possible, but the new solutions and tools also has to be purposefully designed.

It seems that a digital workstation can today be built to improve the ergonomics of the image review compared to with the microscope. This was the main reason behind the initial transformation as is described in P1. It is also possible to personalize the way that slides are navigated using different input devices as was described in P3. Both the specific implementation of the computer mouse and the 6-DOF controller got a task-load index below 50, which is considered a good score compared to similar studies. This makes it possible to switch between devices for different tools, but also throughout the day. This further improves the ergonomic situation.

The ergonomics can also be improved by removing tedious tasks. In P2 it was shown that over half of the time was spent on systematic search tasks that required large amounts of panning. In P7, we designed a prototype that removes the need for systematic panning, which not only increased the efficiency but also should be beneficial for the ergonomic situation.

Digital tools could also make you as a pathologist smarter. In this thesis a number of methods have been presented that could reduce the cognitive burden during the diagnostic process. In P4, a model that supports shoe-boxing of diagnostic findings was presented. This means that findings that need to be kept in mind when a microscope is used, can efficiently be stored as an annotation within the reporting structure, releasing cognitive resources for the diagnostic review. The enhancement method that was described in P6 and P7 can make positively stained nuclei visible in low magnification. This means that a complicated comparison task can be performed by direct visual comparisons instead of trying to remember the positivity ratio of different fields of view. In P8, a number of diagnostic aids were presented that show feasible ways to delegate tedious tasks to the

computer to increase the understanding of a specific case. If it only takes 20 seconds to count the number of nuclei compared to a couple of minutes, it will be possible to count five different areas and compare the different counts. These things and many others could contribute to a better efficiency and accuracy of the diagnostic review. Digital images still have some limitations when it comes to image quality, which needs to be addressed in order to detect very small image features. These special cases are fairly predictable, but it is important as a pathologist to be informed about these limitations by reading or performing validation studies. Once these minor issues have a technological solution, it is therefore likely that the overall accuracy and efficiency of digital review can outperform traditional review with a microscope.

This improved diagnostic power be used either to make the diagnostic review faster or to make it more accurate. Considering that the number of treatment options for cancer care is currently increasing, I think it is most likely in the short term that the increased diagnostic power will be used to improve the accuracy with a maintained number of pathologists diagnosing the same amount of cases per year.

5.2 Implications for image analysis practitioners

The development and evaluation of image analysis algorithms for pathologist use could be improved. Overall, better performing algorithms are of course beneficial, but the development of faster algorithms could be just as important, especially for interactive use. In interactive systems it is possible to divide the algorithm into a pre-processing part and a processing performed on the fly, which is possible with the algorithm presented in P6. This decreases the amount of time that pathologists have to wait.

To support the design of result visualization, an error analysis of the finalized algorithm would help tremendously. This could be incorporated into the normal quantitative evaluation that is usually performed. A novel algorithm is usually evaluated on unseen data and the errors are reported statistically using F1-score, AUC or similar. When this evaluation is performed, it could also be possible to perform a qualitative evaluation of the errors. In order to help the design of the result visualization, it would be beneficial to know: How often do different types of errors occur? When and why do they occur? Each error type could then be summarized in a small description and reported together with an example image. Even though this type of analysis is commonly carried out by practitioners, it is not common practice to report it. This is unfortunate since it could be tremendously helpful to support the development of systems in accordance to guidelines in P8. The qualitative error analysis of pathologists' diagnostic review by Gilbertsson et al. [29] could be used as an inspiration. A number of recent image analysis studies do provide a small qualitative error analysis, which could also be useful to look at. Ali, Madabhushi [1] and Gertych et al. [28] provide a number of example images comparing their presented method with a reference method, but it is only focused on when the reference method fails. A better approach from the interaction designer's perspective is shown by Veta et al. [110] who presents a number of example mitoses that none of the evaluated algorithm were able to detect, and by Xu et al. [118] who shows a visualization of different false positives and false negatives. These qualitative analyses are encouraged

and could be improved by some kind of systematic sampling of examples where one example for each error type is provided.

The performance measurement of an algorithm depends on its intended use. For example, if an algorithm is designed to exclude benign cases, the evaluation needs to focus on the performance of the algorithm when operating at a sensitivity close to 100%. If a system is built to detect nuclei and to show the result to the users for them to correct, the accuracy of the nuclei detection is the most important to evaluate. When designing a system for end-users it can therefore be a good idea to start with figuring out how the result will be communicated, and then focus the development of the algorithm to perform well in that context.

5.3 Implications for HCI

Pathologists are experts at problem solving in scale-space images, and as a profession their review practice has been developed through 150 years. This means that younger domains working with scale-space images probably can learn a great deal by studying how they work.

They have a vast vocabulary of describing different morphological phenomena that are common, which helps them remember what they see when reviewing cases. They are able to navigate extremely quickly and prefer a much higher panning and zooming sensitivity than the one used in, for example, Google Maps.

Most HCI research for scale-space images has focused on the usability of using GIS systems, and it is uncertain whether guidelines derived from studies in that single domain are generalizable to other scale-space images. Digital pathology could therefore be an especially interesting domain to perform HCI research in, in order to validate established principles. Ruddle et al. [91] quite clearly showed that guidelines concerning the size of the overview in a zoomable workspace do not hold up for digital pathology images. In P7, we increased the efficiency with only 15% for the hotspot selection task, which involved a quantitative visual comparison of multiple regions of interest. This results seems low considering the complexity of the task, compared to a visual comparison study with a laboratory task described by [82].

Digital pathology is therefore a great domain to test established ideas within HCI. The domain is also accessible to most HCI practitioners – most university hospitals have a pathology department, and soon they will be using digital images in their daily practice.

Bibliography

[1] Sahirzeeshan Ali and Anant Madabhushi. "An integrated region-, boundary-, shape-based active contour for multiple object overlap resolution in histological imagery". In: *IEEE Transactions on Medical Imaging* 31.7 (2012), pp. 1448–1460 (cit. on p. 46).

[2] I Arel, D C Rose, and T P Karnowski. "Deep Machine Learning - A New Frontier in Artificial Intelligence Research". In: *IEEE Computational Intelligence Magazine* 5.4 (2010), pp. 13–18. arXiv: 1209.5467 (cit. on p. 21).

[3] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, et al. "Factors of Transferability for a Generic ConvNet Representation". 2014. URL: http://arxiv.org/abs/1406.5774 (cit. on p. 21).

[4] Lisanne Bainbridge. "Ironies of Automation". In: *Automatica* 19.6 (1983), pp. 775–779 (cit. on p. 39).

[5] Robert Ball and Chris North. "The Effects of Peripheral Vision and Physical Navigation on Large Scale Visualization". In: *Proceedings of Graphics Interface 2008*. 2008, pp. 9–16 (cit. on p. 15).

[6] Patrick Baudisch and Ruth Rosenholtz. "Halo: a Technique for Visualizing Off-Screen Locations". In: *Proceedings of the 2003 ACM annual conference on Human Factors in Computing Systems - CHI '03*. 2003, pp. 481–488 (cit. on p. 18).

[7] David V Beard and John Q Walker. "Navigational techniques to improve the display of large two-dimensional spaces." In: *Behaviour and Information Technology* 9.6 (1990), pp. 451–466 (cit. on pp. 12, 17).

[8] Andrew H Beck, Ankur R Sangoi, Samuel Leung, et al. "Systematic analysis of breast cancer morphology uncovers stromal features associated with survival." In: *Science translational medicine* 3.108 (2011), 108ra113 (cit. on p. 24).

[9] Justinas Besusparis, Benoit Plancoulaine, Allan Rasmusson, et al. "Impact of tissue sampling on accuracy of Ki67 immunohistochemistry evaluation in breast cancer." In: *Diagnostic pathology* 11.1 (2016), p. 82 (cit. on p. 22).

[10] Christopher M Bishop. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006. ISBN: 9780387310732. arXiv: 0-387-31073-8 (cit. on p. 20).

[11] Dario Bombari, Braulio Mora, Stephan C. Schaefer, et al. "What was i thinking? Eye-tracking experiments underscore the bias that architecture exerts on nuclear grading in prostate cancer". In: *PLoS ONE* 7.5 (2012), e38023 (cit. on p. 15).

[12] Frédéric Bourgeois and Y Guiard. "Multiscale pointing: facilitating pan-zoom coordination". In: *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. 2002, pp. 758–759 (cit. on p. 18).

[13] Melissa O Braxton, TadT Brunyé, JoannG Elmore, et al. "Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images". In: *Journal of Pathology Informatics* 3.1 (2012), p. 43 (cit. on p. 15).

[14] Tad T. Brunyé, P. a. Carney, K. H. Allison, et al. "Eye movements as an index of pathologist visual expertise: A pilot study". In: *PloS one* 9.8 (2014), e103447 (cit. on pp. 15, 16).

[15] *CAMELYON16 - Results*. URL: https://grand-challenge.org/site/camelyon16/results/ (visited on 10/18/2016) (cit. on p. 23).

[16] Jeffrey A. Claridge and Timothy C. Fabian. "History and development of evidence-based medicine". In: *World Journal of Surgery* 29.5 (2005), pp. 547–553 (cit. on p. 25).

[17] Rebecca S Crowley, Gregory J Naus, Jimmie Stewart, et al. "Development of visual diagnostic expertise in pathology-an information-processing study". In: *Journal of the American Medical Informatics Association* 10.1 (2003), pp. 39–51 (cit. on pp. 12, 13).

[18] Brett Delahunt, Rose J. Miller, John R. Srigley, et al. "Gleason grading: Past, present and future". In: *Histopathology* 60.1 (2012), pp. 75–86 (cit. on p. 23).

[19] Santa Di Cataldo, Elisa Ficarra, and Enrico Macii. "Computer-aided techniques for chromogenic immunohistochemistry: Status and directions". In: *Computers in Biology and Medicine* 42.10 (2012), pp. 1012–1025 (cit. on p. 21).

[20] Scott Doyle, Michael Feldman, Natalie Shih, et al. "Cascaded discrimination of normal, abnormal, and confounder classes in histopathology: Gleason grading of prostate cancer". In: *BMC Bioinformatics* 13.1 (2012), p. 282 (cit. on pp. 23, 25).

[21] Trafton Drew, Melissa L-H Võ, and Jeremy M Wolfe. "The invisible gorilla strikes again: sustained inattentional blindness in expert observers." In: *Psychological science* 24 (2013), pp. 1848–53 (cit. on p. 16).

[22] Jonathan I Epstein, Lars L Egevad, Mahul B. Amin, et al. "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma". In: *The American journal of surgical pathology* 40.2 (2016), pp. 244–52 (cit. on p. 23).

[23] K Anders Ericsson and Herbert A Simon. "Introduction and Summary". In: *Protocol Analysis: Verbal Reports as Data*. 1993. Chap. 1, pp. 1–61 (cit. on pp. 11, 42).

[24] Paul M. Fitts. "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement". In: *Journal of Experimental Psychology* 47 (1954), pp. 381–391 (cit. on p. 18).

[25] André Frank. "Mapquery: Data Base Query Language for Retrieval of Geometric Data and Their Graphical Representation". In: *ACM SIGGRAPH Computer Graphics* 16.3 (1982), pp. 199–207 (cit. on p. 17).

[26] G. W. Furnas. "Generalized fisheye views". In: *ACM SIGCHI Bulletin* 17 (1986), pp. 16–23 (cit. on p. 17).

[27] GW Furnas and BB Bederson. "Space-scale diagrams: Understanding multiscale interfaces". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (1995), pp. 234–241 (cit. on p. 17).

[28] Arkadiusz Gertych, Nathan Ing, Zhaoxuan Ma, et al. "Machine learning approaches to analyze histological images of tissues from radical prostatectomies". In: *Computerized Medical Imaging and Graphics* 46 (2015), pp. 197–208 (cit. on p. 46).

[29] John R Gilbertson, Jonhan Ho, Leslie Anthony, et al. "Primary histologic diagnosis using automated whole slide imaging: a validation study." In: *BMC Clinical Pathology* 6.4 (2006) (cit. on p. 46).

[30] Edward Goacher, Rebecca Randell, Bethany Williams, et al. "The Diagnostic Concordance of Whole Slide Imaging and Light Microscopy". In: *Archives of pathology & laboratory medicine* In Press (2016) (cit. on pp. 25, 26).

[31] Fabian Göbel, Konstantin Klamka, Andreas Siegel, et al. "Gaze-supported foot interaction in zoomable information spaces". In: *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. 2013, pp. 3059–62. ISBN: 9781450319522 (cit. on p. 18).

[32] Rafael C Gonzalez and Richard E Woods. *Digital Image Processing (3rd Edition)*. 2007, p. 976. ISBN: 013168728X (cit. on p. 20).

[33] Lena Gorelick, Olga Veksler, Mena Gaed, et al. "Prostate histopathology: Learning tissue component histograms for cancer detection and classification". In: *IEEE Transactions on Medical Imaging* 32.10 (2013), pp. 1804–1818 (cit. on pp. 23, 25).

[34] T. Greenhalgh, J. Howick, and N. Maskrey. "Evidence based medicine: a movement in crisis?" In: *BMJ* 348 (2014), g3725 (cit. on p. 25).

[35] Einar Gudlaugsson, Ivar Skaland, Emiel a M Janssen, et al. "Comparison of the effect of different techniques for measurement of Ki67 proliferation on reproducibility and prognosis prediction accuracy in breast cancer". In: *Histopathology* 61.6 (2012), pp. 1134–44 (cit. on p. 22).

[36] Yves Guiard and Michel Beaudouin-Lafon. "Target acquisition in multiscale electronic worlds". In: *International Journal of Human-Computer Studies* 61.6 (2004), pp. 875–905 (cit. on p. 18).

[37] M N Gurcan, L E Boucheron, A Can, et al. "Histopathological image analysis: a review". In: *IEEE Rev Biomed Eng* 2 (2009), pp. 147–171 (cit. on p. 21).

[38] Carl Gutwin and Amy Skopik. "Fisheye Views are Good for Large Steering Tasks". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2003, pp. 201–208. ISBN: 1581136307 (cit. on p. 18).

[39] Peter W. Hamilton, Peter Bankhead, Yinhai Wang, et al. “Digital pathology and image analysis in tissue biomarker research”. In: *Methods* 70.1 (2014), pp. 59–73 (cit. on p. 21).

[40] Peter Hamilton and Daniel Wigdor. “Conductor: Enabling and Understanding Cross-Device Interaction”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2014, pp. 2773–2782 (cit. on p. 27).

[41] Frank Hassebrock and Michael J. Prietula. “A protocol-based coding scheme for the analysis of medical reasoning”. In: *International Journal of Man-Machine Studies* 37.5 (1992), pp. 613–652 (cit. on p. 12).

[42] Bradley M. Hemminger, Anne Bauers, and Jian Yang. “Comparison of navigation techniques for large digital images”. In: *Journal of Digital Imaging* 21 (2008), pp. 13–38 (cit. on p. 19).

[43] Kasper Hornbæk, BB Bederson, and Catherine Plaisant. “Navigation patterns and usability of zoomable user interfaces with and without an overview”. In: *ACM Transactions on Computer-Human Interaction* 9.4 (2002), pp. 362–389 (cit. on pp. 18, 19).

[44] Kasper Hornbæk and Erik Frøkjær. “Reading patterns and usability in visualizations of electronic documents”. In: *ACM Transactions on Computer-Human Interaction* 10.2 (2003), pp. 119–149 (cit. on pp. 19, 35, 36).

[45] Kasper Hornbak and Erik Frokjar. “Reading of electronic documents: the usability of linear, fisheye, and overview+ detail interfaces”. In: *Proceedings of the 2001 ACM annual conference on Human Factors in Computing Systems*. 2001, pp. 293–300 (cit. on p. 19).

[46] Takeo Igarashi and Ken Hinckley. “Speed-dependent automatic zooming for browsing large documents”. In: *Proceedings of the 13th annual ACM symposium on User interface software and technology - UIST ’00* (2000), pp. 139–148 (cit. on p. 18).

[47] Humayun Irshad, Antoine Veillard, Ludovic Roux, et al. “Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential”. In: *IEEE Reviews in Biomedical Engineering* 7 (2014), pp. 97–114 (cit. on p. 21).

[48] Antoine Italiano. “Prognostic or predictive? It’s time to get back to definitions!” In: *Journal of Clinical Oncology* 29.35 (2011), p. 4718 (cit. on p. 25).

[49] Vilppu J, Sanna RuotTuominenoistenmäki, Arttu Viitanen, et al. “ImmunoRatio: a publicly available web application for quantitative image analysis of estrogen receptor (ER), progesterone receptor (PR), and Ki-67”. In: *Breast cancer research* 12.4 (2010), R56 (cit. on p. 22).

[50] Thomas Jaarsma, Henny P A Boshuizen, Halszka Jarodzka, et al. “Tracks to a Medical Diagnosis: Expertise Differences in Visual Problem Solving”. In: *Applied Cognitive Psychology* 322 (2016), pp. 314–322 (cit. on p. 16).

[51] Thomas Jaarsma, Halszka Jarodzka, Marius Nap, et al. "Expertise in clinical pathology: combining the visual and cognitive perspective". In: *Advances in Health Sciences Education* (2015), pp. 1089–1106 (cit. on p. 16).

[52] Thomas Jaarsma, Halszka Jarodzka, Marius Nap, et al. "Expertise under the microscope: processing histopathological slides." In: *Medical education* 48.3 (2014), pp. 292–300 (cit. on p. 16).

[53] Joseph G Jacobs, Eleftheria Panagiotaki, and Daniel C Alexander. "Gleason Grading of Prostate Tumours with Max-Margin Conditional Random Fields". In: *Machine Learning in Medical Imaging* (2014), pp. 85–92 (cit. on p. 23).

[54] Lisa Jacobson, Olle Bergman, Lotta Fredholm, et al. *Cancerfondsrapporten 2016*. Tech. rep. Cancerfonden, 2016 (cit. on p. 3).

[55] Waqas Javed, Sohaib Ghani, and Niklas Elmqvist. "GravNav: using a gravity model for multi-scale navigation". In: *Proceedings of the International Working Conference on Advanced Visual Interfaces* (2012), pp. 217–224 (cit. on p. 20).

[56] Waqas Javed, Sohaib Ghani, and Niklas Elmqvist. "Polyzoom: multiscale and multifocus exploration in 2d visual spaces". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, pp. 287–296 (cit. on p. 18).

[57] *Klinisk patologi - Utbildningsbok*. 1.0. Svensk Förening för Patologi Cytologi, Svensk Förening för Klinisk, 2015 (cit. on p. 8).

[58] Kristina Knaving, Paweł Woźniak, Morten Fjeld, et al. "Flow is Not Enough: Understanding the Needs of Advanced Amateur Runners to Design Motivation Technology". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (2015), pp. 2013–2022 (cit. on p. 27).

[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances In Neural Information Processing Systems* (2012), pp. 1–9. arXiv: 1102.0183 (cit. on p. 21).

[60] Elizabeth A Krupinski, Anna R Graham, and Ronald S Weinstein. "Characterizing the development of visual search expertise in pathology residents viewing whole slide images." In: *Human pathology* 44.3 (2013), pp. 357–364 (cit. on p. 15).

[61] Elizabeth A Krupinski, Allison A Tillack, Lynne Richter, et al. "Eye-movement study and human performance using telepathology virtual slides: implications for medical education and differences with experience." In: *Human pathology* 37.12 (2006), pp. 1543–56 (cit. on p. 15).

[62] Arvydas Laurinavicius, Benoit Plancoulaine, Allan Rasmusson, et al. "Bimodality of intratumor Ki67 expression is an independent prognostic factor of overall survival in patients with invasive breast carcinoma". In: *Virchows Archiv* 468.4 (2016), pp. 493–502 (cit. on pp. 22, 24, 39, 40).

[63] Maria-Jesus Lobo, Emmanuel Pietriga, and Caroline Appert. "An Evaluation of Interactive Map Comparison Techniques". In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 2015, pp. 3573–3582 (cit. on pp. 18, 39).

[64] S J Luck and E K Vogel. "The capacity of visual working memory for features and conjunctions." In: *Nature* 390.6657 (1997), pp. 279–81 (cit. on p. 14).

[65] Anant Madabhushi and George Lee. "Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities". In: *Medical Image Analysis* 33 (2016), pp. 170–175 (cit. on pp. 21, 22).

[66] Claudia Mello-Thoms, Carlos A B Mello, Olga Medvedeva, et al. "Perceptual analysis of the reading of dermatopathology virtual slides by pathology residents." In: *Archives of pathology & laboratory medicine* 136 (2012), pp. 551–62 (cit. on pp. 13, 16).

[67] Ezgi Mercan, Selim Aksoy, Linda G Shapiro, et al. "Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images". In: *International Conference on Pattern Recognition*. 2014, pp. 1179–1184 (cit. on p. 16).

[68] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, et al. "Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels". In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*. 2016. arXiv: 1512.06974v1 (cit. on p. 39).

[69] Diana Weedman Molavi. *The practice of surgical pathology: a beginner's guide to the diagnostic process*. Springer Science & Business Media, 2008 (cit. on p. 8).

[70] Steven B Most, Daniel J Simons, Brian J Scholl, et al. "Sustained Inattentional Blindness : Dynamic Events". In: 28 (2000), pp. 1059–1074 (cit. on p. 16).

[71] Allen Newell and Herbert A Simon. *Human Problem Solving*. Englewood Cliffs: Prentice-Hall, 1972 (cit. on p. 11).

[72] M. K. K. Niazi, Douglas J. Hartman, Liron Pantanowitz, et al. "Hotspot detection in pancreatic neuroendocrine tumors: Density approximation by α-shape maps". In: *SPIE Proceedings, Medical Imaging*. Vol. 9791. 2016, 97910B (cit. on p. 22).

[73] Jakob Nielsen and Thomas K. Landauer. "A mathematical model of the finding of usability problems". In: *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (1993), pp. 206–213 (cit. on p. 26).

[74] Željko Obrenović. "Design-based research". In: *Interactions* 18 (2011), p. 56 (cit. on p. 33).

[75] Liron Pantanowitz. "Digital images and the future of digital pathology." In: *Journal of pathology informatics* 1.15 (2010) (cit. on p. 6).

[76] Liron Pantanowitz, Anil Parwani, Eugene Tseytlin, et al. "Analysis of slide exploration strategy of cytologists when reading digital slides". In: *SPIE Proceedings, Medical Imaging*. Vol. 8318. 2012, 83180O–1 (cit. on p. 16).

[77] Liron Pantanowitz, Clayton A. Wiley, Anthony Demetris, et al. "Experience with multimodality telepathology at the University of Pittsburgh Medical Center". In: *Journal of Pathology Informatics* 3.1 (2012), p. 45 (cit. on p. 29).

[78] Ken Perlin, Ken Perlin, David Fox, et al. "Pad - An Alternative Approach to the Computer Interface". In: *In Proc. ACM SIGGRAPH* (1993), pp. 57–64 (cit. on p. 18).

[79] Emmanuel Pietriga, Caroline Appert, and M Beaudouin-Lafon. "Pointing and beyond: an operationalization and preliminary evaluation of multi-scale searching". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, pp. 1215–1224 (cit. on p. 19).

[80] Cyprien Pindat, Emmanuel Pietriga, Olivier Chapuis, et al. "JellyLens : Content-Aware Adaptive Lenses". In: *Proceedings of the 25th annual ACM symposium on User interface software and technology* (2012), pp. 261–270 (cit. on p. 20).

[81] Peter Pirolli and Stuart Card. "The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis". In: *Proceedings of international conference on intelligence analysis*. 2005, pp. 2–4 (cit. on pp. 11, 13, 14).

[82] MD Plumlee and Colin Ware. "Zooming versus multiple window interfaces: Cognitive costs of visual comparisons". In: *ACM Transactions on Computer-Human Interaction* 12.2 (2006), pp. 179–209 (cit. on pp. 10, 18, 47).

[83] Xin Qi, Fuyong Xing, David J. Foran, et al. "Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set". In: *IEEE Transactions on Biomedical Engineering* 59.3 (2012), pp. 754–765 (cit. on p. 22).

[84] Rebecca Randell, Thilina Ambepitiya, Claudia Mello-Thoms, et al. "Effect of Display Resolution on Time to Diagnosis with Virtual Pathology Slides in a Systematic Search Task." In: *Journal of digital imaging* 28.1 (2014), pp. 68–76 (cit. on p. 17).

[85] Rebecca Randell, RA Ruddle, and Phil Quirke. "Working at the microscope: analysis of the activities involved in diagnostic pathology". In: *Histopathology* 60.3 (2012), pp. 504–510 (cit. on pp. 9, 27).

[86] Rebecca Randell, Roy A Ruddle, Claudia Mello-Thoms, et al. "Virtual reality microscope versus conventional microscope regarding time to diagnosis: an experimental study". In: *Histopathology* 62.2 (2013), pp. 351–358 (cit. on p. 17).

[87] Rebecca Randell, Roy a. Ruddle, Rhys G. Thomas, et al. "Diagnosis of major cancer resection specimens with virtual slides: impact of a novel digital pathology workstation". In: *Human Pathology* 45.10 (2014), pp. 2101–2106 (cit. on pp. 17, 20, 27, 36, 41).

[88] Rebecca Randell, Roy a. Ruddle, Rhys Thomas, et al. "Diagnosis at the microscope: a workplace study of histopathology". In: *Cognition, Technology & Work* 14.4 (2011), pp. 319–335 (cit. on pp. 9, 11, 14, 27).

[89] Rikke Riber-Hansen, Ben Vainer, and Torben Steiniche. "Digital image analysis: a review of reproducibility, stability and basic requirements for optimal results". In: *APMIS : Acta Pathologica, Microbiologica, et Immunologica Scandinavica* 120.4 (2012), pp. 276–289 (cit. on p. 21).

[90] Ludovic Roux, Daniel Racoceanu, Nicolas Loménie, et al. "Mitosis detection in breast cancer histological images An ICPR 2012 contest." In: *Journal of pathology informatics* 4.1 (2013), p. 8 (cit. on p. 23).

[91] Roy A Ruddle, Rhys G Thomas, Rebecca Randell, et al. "The Design and Evaluation of Interfaces for Navigating Gigapixel Images in Digital Pathology". In: *ACM Transactions on Computer-Human Interaction* 23.5 (2016) (cit. on pp. 41, 42, 47).

[92] Arnout C Ruifrok and Dennis A Johnston. "Quantification of histochemical staining by color deconvolution." In: *Analytical and Quantitative Cytology and Histology* 23.4 (2001), pp. 291–299 (cit. on p. 38).

[93] David L Sackett, William M C Rosenberg, J A Muir Gray, et al. "Evidence based medicine: What it is and what it isn't - It's about integrating individual clinical expertise and the best external evidence". In: *British Medical Journal* 312.7023 (1996), pp. 71–72 (cit. on p. 24).

[94] Stephanie Santosa and Daniel Wigdor. "A field study of multi-device workflows in distributed workspaces". In: *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing - UbiComp '13*. 2013, pp. 63–72 (cit. on p. 27).

[95] Ben Shneiderman. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations". In: *Proceedings of the IEEE Symposium on Visual Languages*. 1996, pp. 336–343 (cit. on p. 14).

[96] Garth Shoemaker and Carl Gutwin. "Supporting multi-point interaction in visual workspaces". In: *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*. 2007, pp. 999–1008. ISBN: 9781595935939 (cit. on p. 18).

[97] Herbert A Simon. *The sciences of the artificial, (third edition)*. Vol. 33. 5. 1997, p. 130. ISBN: 9780262691918 (cit. on p. 26).

[98] Korsuk Sirinukunwattana, Shan Raza, Yee-wah Tsang, et al. "Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images". In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1196–1206 (cit. on p. 22).

[99] Gustav Stålhammar, Nelson Fuentes Martinez, Michael Lippert, et al. "Digital image analysis outperforms manual biomarker assessment in breast cancer". In: *Modern Pathology* 29 (2016), pp. 318–329 (cit. on pp. 22, 24).

[100] Sophie Stellmach and Raimund Dachselt. "Designing Gaze-based User Interfaces for Steering in Virtual Environments". In: *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. 2012, pp. 131–138 (cit. on p. 18).

[101] Sophie Stellmach and Raimund Dachselt. "Investigating gaze-supported multimodal pan and zoom". In: *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12* (2012), pp. 357–60 (cit. on p. 18).

[102] Sophie Stellmach and Raimund Dachselt. "Still Looking: Investigating Seamless Gaze-supported Selection, Positioning, and Manipulation of Distant Targets". In: *Proceedings of the 2013 ACM annual conference on Human Factors in Computing Systems - CHI '13*. 2013, pp. 285–294 (cit. on p. 18).

[103] Sophie Stellmach, Markus Jüttner, Christian Nywelt, et al. "Investigating Freehand Pan and Zoom". In: *Mensch & computer*. 2012, pp. 303–312 (cit. on p. 18).

[104] Erik Stolterman. "The Nature of Design Practice and Implications for Interaction Design Research". In: *International Journal of Design* 2.1 (2008), pp. 55–65 (cit. on p. 26).

[105] Marie B Suurküla and Ingvar Karlberg. *Svensk patologi – en översyn och förslag till åtgärder*. Tech. rep. 2012 (cit. on p. 3).

[106] Richard Szeliski. *Computer Vision: Algorithms and Applications*. September. Springer, 2010. ISBN: 1848829345 (cit. on p. 20).

[107] E S M Tiersma, A A W Peters, H A Mooij, et al. "Visualising scanning patterns of pathologists in the grading of cervical intraepithelial neoplasia." In: *Journal of clinical pathology* 56 (2003), pp. 677–680 (cit. on pp. 11, 15).

[108] Darren Treanor, Naomi Jordan-Owers, John Hodrien, et al. "Virtual reality Powerwall versus conventional microscope for viewing pathology slides: an experimental comparison". In: *Histopathology* 55 (2009), pp. 294–300 (cit. on p. 17).

[109] Darren Treanor, Chee Hooi Lim, Derek Magee, et al. "Tracking with virtual slides: a tool to study diagnostic error in histopathology." In: *Histopathology* 55 (2009), pp. 37–45 (cit. on pp. 11, 13, 16).

[110] Mitko Veta, Paul J. van Diest, Stefan M. Willems, et al. "Assessment of algorithms for mitosis detection in breast cancer histopathology images". In: *Medical Image Analysis* 20.1 (2015), pp. 237–248. arXiv: 1411.5825 (cit. on pp. 23, 46).

[111] Mitko Veta, Josien P W Pluim, Paul J. Van Diest, et al. "Breast cancer histopathology image analysis: A review". In: *IEEE Transactions on Biomedical Engineering* 61.5 (2014), pp. 1400–1411 (cit. on p. 21).

[112] Slawomir Walkowski, Mikael Lundin, Janusz Szymas, et al. "Students' performance during practical examination on whole slide images using view path tracking". In: *Diagnostic pathology* 9.208 (2014) (cit. on p. 16).

[113] Colin Ware. *Information Visualization: Perception for Design*. Third edit. Morgan Kaufmann, 2012 (cit. on p. 15).

[114] June Wei and Gavriel Salvendy. "The cognitive task analysis methods for job and task design: review and reappraisal". In: *Behaviour Information Technology* 23.4 (2004), pp. 273–299 (cit. on p. 12).

[115] Jarke J Van Wijk and Wim A A Nuij. "Smooth and efficient zooming and panning". In: *Information Visualization*. 2003, pp. 15–23 (cit. on p. 18).

[116] Geoffrey F Woodman and Steven J Luck. "Visual search is slowed when visuospatial working memory is occupied." In: *Psychonomic bulletin & review* 11.2 (2004), pp. 269–74 (cit. on p. 14).

[117] Paweł Woźniak, Kristina Knaving, Staffan Björk, et al. "RUFUS: Remote Supporter Feedback for Long-Distance Runners". In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2015), pp. 115–124 (cit. on p. 27).

[118] Jun Xu, Lei Xiang, Renlong Hang, et al. “Stacked Sparse Autoencoder (SSAE) based framework for nuclei patch classification on breast cancer histopathology”. In: *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* 35.1 (2014), pp. 999–1002 (cit. on p. 46).