Synergizing human-machine intelligence: Visualizing, labeling, and mining the electronic health record

Noah Lee

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2011

ABSTRACT

Synergizing human-machine intelligence: Visualizing, labeling, and mining the electronic health record

Noah Lee

We live in a world where data surround us in every aspect of our lives. The key challenge for humans and machines is how we can make better use of such data. Imagine what would happen if you were to have intelligent machines that could give you insight into the data. Insight that will enable you to better 1) reason about, 2) learn, and 3) understand the underlying phenomena that produced the data. The possibilities of combined human-machine intelligence are endless and will impact our lives in ways we can not even imagine today.

Synergistic human-machine intelligence aims to facilitate the analytical reasoning and inference process of humans by creating machines that maximize a human's ability to 1) reason about, 2) learn, and 3) understand large, complex, and heterogeneous data. Combined human-machine intelligence is a powerful symbiosis of mutual benefit, in which we depend on the computational capabilities of the machine for the tasks we are not good at, and the machine requires human intervention for the tasks it performs poorly on. This relationship provides a compelling alternative to either approach in isolation for solving today's and tomorrow's arising data challenges.

In his regard, this dissertation proposes a diverse analytical framework that leverages synergistic human-machine intelligence to maximize a human's ability to better 1) reason about, 2) learn, and 3) understand different biomedical imaging and health-

care data present in the patient's electronic health record (EHR). Correspondingly, we approach the data analyses problem from the 1) visualization, 2) labeling, and 3) mining perspective and demonstrate the efficacy of our analytics on specific application scenarios and various data domains.

In the first part of this dissertation we explore the question how we can build intelligent imaging analytics that are commensurate with human capabilities and constraints, specifically for optimizing data visualization and automated labeling workflows. Our journey starts with heuristic rule-based analytical models that are derived from task-specific human knowledge. From this experience, we move on to data-driven analytics, where we adapt and combine the intelligence of the model based on prior information provided by the human and synthetic knowledge learned from partial data observations. Within this realm, we propose a novel Bayesian transductive Markov random field model that requires minimal human intervention and is able to cope with scarce label information to learn and infer object shapes in complex spatial, multimodal, spatio-temporal, and longitudinal data. We then study the question how machines can learn discriminative object representations from dense human provided label information by investigating learning and inference mechanisms that make use of deep learning architectures. The developed analytics can aid visualization and labeling tasks, which enables the interpretation and quantification of clinically relevant image information.

The second part explores the question how we can build data-driven analytics for exploratory analysis in longitudinal event data that are commensurate with human capabilities and constraints. We propose human-intuitive analytics that enable the representation and discovery of interpretable event patterns to ease knowledge absorption and comprehension of the employed analytics model and the underlying data. We propose a novel doubly-constrained convolutional sparse-coding framework that learns interpretable and shift-invariant latent temporal event patterns. We apply the model to mine complex event data in EHRs. By mapping the event space to heterogeneous patient encounters in the EHR we explore the linkage between healthcare resource utilization (HRU) in relation to disease severity. This linkage may help to better understand how disease specific co-morbidities and their clinical attributes incur different HRU patterns. Such insight helps to characterize the patient's care history, which then enables the comparison against clinical practice guidelines, the discovery of prevailing practices based on common HRU group patterns, and the identification of outliers that might indicate poor patient management.

In general, we present novel approaches that exploit the synergistic aspect of human-machine intelligence by addressing problems from biomedical imaging to healthcare informatics. The generic nature and applicability of the proposed techniques, when integrated together, enable the holistic analysis of the electronic health record and its diverse data sources, which in turn can reveal hidden patterns across the different data sources.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to give thanks to my advisor Prof. Andrew Laine, who gave me the chance to pursue my doctoral studies in the Heffner Biomedical Imaging Laboratory. I still clearly remember my first encounter with Prof. Laine during the biomedical engineering department's graduate open house session. His kindness wasn't just a first impression. Throughout the years his friendliness and understanding have made my graduate school experience enjoyable. I have to thank Prof. Laine for his technical mentorship and encouragement. He introduced me to the field of Wavelets, advanced medical image analysis techniques, and exciting collaborative projects with other research institutions and the industry. Moreover, I have to deeply thank him for always being supportive and generous no matter what kind of problems I approached him with.

I would like to express my sincere gratitude towards Prof. R Theodore Smith for giving me the opportunity to work with him on challenging problems in ophthalmology. His humbleness and friendliness made graduate research work an enjoyable experience. Moreover, I am grateful for the financial support he has given me throughout all the years, for allowing me to pursue my doctoral studies, attend conferences, and other international travel. I am grateful for all his support, encouragement, patience, and especially his understanding, allowing me to widen my research experience in industry during several summer internships as well as for giving me the freedom to participate in other academic research projects.

I would like to thank the other members of my proposal and dissertation committee for their valuable feedback, support, and constructive advice. I feel grateful to Dr. Shahram Ebadollahi for giving me the opportunity to work on exiting problems in healthcare and for supporting me during my graduate studies in numerous forms. He introduced me to the world of intelligent analytics and the aspect of multimodal longitudinal data analyses. I have to say thanks for giving me the opportunity to intern at IBM research, an exciting and memorable experience. I am grateful to Dr. Fei Wang, Dr. Jianying Hu, Dr. Jimeng Sun, and Dr. Larry Stavropoulos from whom I learned a lot.

In this regard I also have to thank Dr. Gareth Funka-Lea for giving me multiple opportunities to intern at Siemens Corporate Research to work on exciting problems in diagnostic medical imaging. I have to thank Matthias Rasch and Dr. Hussein Tek for their mentorship and support. I am also thankful to Prof. Daniel Cremers, Dr. Leo Grady, and Dr. Chenyang Xu for sharing their expertise and for their kind help.

I also have to thank Prof. Arno Klein for giving me the opportunity to work on the exciting Mindboggle project and for funding my last semester of my doctoral studies. I learned and am still learning many things from him on how to become a better researcher. He has inspired and taught me a great account of using open source technology for research and introduced me to the world of Python. I truly appreciate the time he spends with me in discussing problems, giving in-depth feedback to my writing, and for sharing ideas.

I would like to thank Prof. Paul Sajda for serving on my dissertation committee and chairing the dissertation defense. He introduced me to cutting-edge machine learning techniques such as the nonnegative matrix factorization algorithm, reservoir computing, and biologically inspired machine learning models.

Thanks to all current members and alumni of the Heffner Biomedical Imaging lab, Won-Hee Lee, Ming Jack Po, Dr. Amin Katouzian, Dr. Qi Duan, Dr. Ting Song, Dr.

Jian Chen, Mustafa Selver, George Xu, Antonio Albanese, Illiyana Atanasova, Ouri Cohen, Dr. Matt Neimark, Auranuch Lorsakul, Dr. Elsa Angelini, and Yrjö Häme with whom I had fruitful discussions.

To gaeguree...

Chapter 1

Introduction

We live in a world where data are alive. When we take a cab, make a phone call, go to the bank, surf the internet, or visit the doctor, we generate a unique data footprint. Data surround us in every aspect of our lives. In the financial industry business transaction logs generate massive data amounts that capture the financial information flow of our society [1; 2; 3; 4; 5]. Advances in internet technology such as the Web 2.0 and the emergence of large social networks generate petabytes of personality profiles that provide information about people's preferences, their communications, and social interactions [6; 7; 8; 9; 10; 11]. And lastly, the healthcare industry and its initiative to bring forward the electronic health record generating enormous amounts of digital data ranging from diagnostic images, laboratory results, genetic profiles, and other healthcare specific data sources [12; 13; 14; 15; 16; 17; 18; 19].

The key challenge for humans and machines is how we can make better and more efficient use of such data. Often times, the information that may lead to actionable insight is hidden due to data scale, complexity, and data heterogeneity. Humankind will generate over one sextillion bytes of electronic data this year alone amounting to roughly one zettabyte, i.e. one trillion gigabytes. Moreover, it is predicted that the yearly amount of data humankind produces will increase by a factor of 44 over the next decade [20]. In practice the data are complex and often are incomplete, high-dimensional, noisy, and ambiguous. In addition, data heterogeneity expresses itself in multiple data sources with textual, numerical, and visual characteristics, which poses challenges to humans in coping with the diversity and scale of diverse data structures and to transform data into actionable knowledge.

Imagine what would happen if you were to have intelligent machines that could give you insight into the data. Insight provided by intelligent analytics and raw computational power that will enable you to better 1) reason about, 2) learn, and 3) understand the underlying phenomena that produced the data. Imagine having machines that could intuitively communicate the data insights in a form humans can better absorb and comprehend. Machines that are designed to exchange and combine effectively synthetic knowledge with human knowledge. The possibilities of combined synergistic human-machine intelligence are endless and will impact our lives in ways we can not even imagine today.

1.1 Motivation

While research in artificial intelligence [21; 22; 23; 24], machine learning [25; 26; 27], and data mining [28; 29; 30] show promising progress and state-of-the-art performance in many data analyses tasks, the intelligent aspect of machines is still a product of human-engineered knowledge and intelligence.

One question that has received limited attention is how one can best leverage synergistic human-machine intelligence to better cope with large, complex, and heterogeneous data? This question is not trivial since a human and a machine deal with data at multiple scales in different ways. Clear is that with increasing data scale, complexity, and data heterogeneity humans quickly become overwhelmed and have difficulties in reasoning about and making inference from data.

In this regard, synergistic human-machine intelligence aims to facilitate the analytical reasoning and inference process of humans by creating machines that maximize a human's ability to 1) reason about, 2) learn, and 3) understand large, complex, and heterogeneous data. Combined human-machine intelligence is a powerful symbiosis of mutual benefit, in which we depend on the computational capabilities of the machine for the tasks we find difficult, and the machine requires human intervention for the tasks it performs poorly on. This relationship provides a compelling alternative to either approach in isolation for solving today's and tomorrow's data challenges.

We hypothesize that in order to leverage synergistic human-machine intelligence, an analytical model (i.e. the machine) or framework must be commensurate with human capabilities. Further it is of essence that the human and the machine communicate effectively. Therefore, the input/output and the internal structure of the model should take into account the human factor. For the input, the model should be able to make optimal use of available human knowledge where possible, while at the same time cope with situations where human knowledge resources are constrained. The intermediate and output level of the model should adhere to the capabilities of humans so that the analytical model is intuitive and understandable.

Having data-driven analytics that are commensurate with human capabilities and constraints is essential for facilitating the knowledge exchange between a human and a machine. The synergistic interaction of human-machine intelligence has also immense practical importance, since it will be impossible to give machines and their learning algorithms all of the knowledge that they will need to serve useful autonomous long-term roles in our dynamic and complex environment. Rather, it is more practical to have humans control the intelligence of the machine when it is desired and enable humans to effectively make use of the synthetic knowledge the machine generates.

The healthcare industry in particular provides a challenging environment to develop intelligent analytics that leverage synergistic human-machine intelligence. In healthcare, medical practitioners deal with large, complex, and heterogeneous data sources for purposes of improved diagnosis and evidence-based decision making. Technological advances in biomedical imaging provide us with novel ways to see and analyze the functioning of the human body to understand and improve disease diagnosis. Such data can take on many forms, e.g. unimodal, multimodal, or longitudinal image data in planar or volumetric form. On another front, technical advances towards the availability of electronic health records (EHRs)[1]provide a rich collection of heterogeneous data sources including numerical, textual, visual, and time-dependent data such as time series and events.

EHRs provide a complete record of patient encounters and a detailed account of the medical history of the patient that enable improved data integration and automated access, yet the wealth and diversity of information they provide is currently underutilized. Most data analyses approaches take a short-sighted approach by focusing on a particular data and problem domain. Only recently have researchers begun to fuse multiple data sources together for combined analyses to search for individual data patterns and correlations between different data sources. Yet, a holistic approach to analyze the complete electronic patient record with its diverse data sources has not been accomplished yet. Fig. 1.1 depicts an illustrative example of the diversity of data sources that are contained in an EHR.

The scale and complexity of biomedical and healthcare data in EHRs pose challenges not only to medical practitioners, but also to the data and information analysis by machines. Often times, data that could provide important insight for decision making are hidden, making it difficult to understand and transform the data into actionable knowledge. What would be desirable to medical practitioners are generic data analyses tools that can 1) visualize, 2) label, and 3) discover meaningful infor-

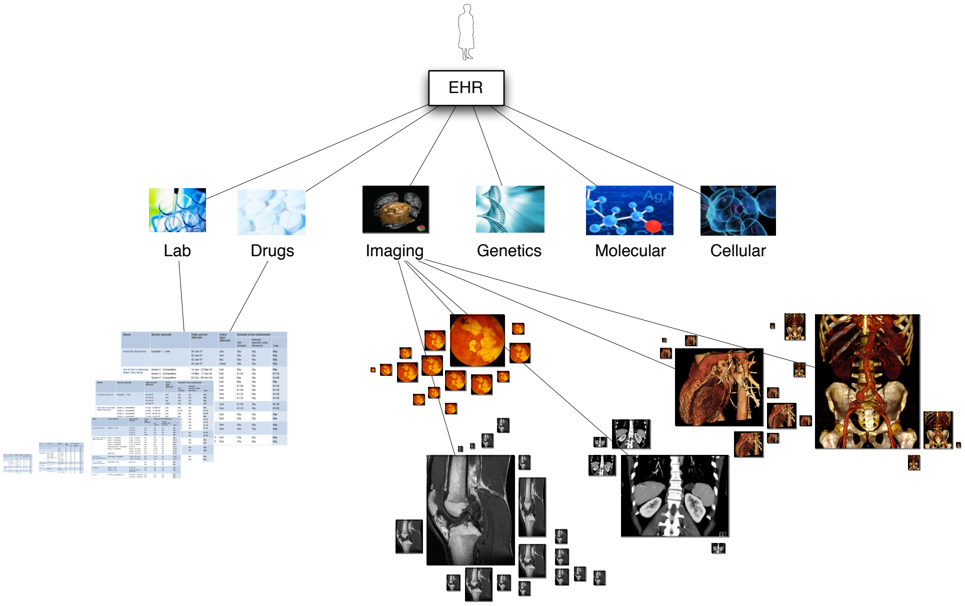


Figure 1.1: The electronic healthcare record–a large, complex, and heterogeneous data source. An electronic healthcare record may include all of the key administrative clinical data relevant to that person's care, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, test results, radiology reports, and diagnostic imaging data.

mation from EHRs in a holistic human-intuitive form.

1.2 Challenges

In this section we focus on three inter-related challenges that to us seem most important considering the increasing amount of data in the biomedical and healthcare domain and in particular the EHR.

[1] EHRs capture the medical history of a patient or patient population and may include all of the key administrative clinical data relevant to that person's care, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data, test results, radiology reports, and diagnostic imaging data.

- **Challenge 1: Data scale and complexity**. The problem of data scale arises in two contexts. First, the EHR of a single patient may consist of large amounts of diagnostic image data and other auxiliary data sources as outlined above. Second, medical practitioners often need to perform group and population analyses that involve a large number of patients. Data scale already introduces the first dimension of complexity. In addition, data heterogeneity of the medical patient record introduces another dimension of complexity. The data contain latent structural, temporal, and relational information that make the automatic extraction by machines a tremendous challenge. Problems such as data noise, artifacts, ambiguity, and data incompleteness further exacerbate this challenge.

- **Challenge 2: Missing label information**. With the popularity of machine learning and data mining comes the need for labeled data. Label information serves two purposes in this context: categorization of data into semantically meaningful concepts and supervision of the learning and inference process of data-driven analytics. Yet, the amount of available label information is a limiting factor. Large annotated medical image databases are yet to be created. In the medical domain the acquisition of label information requires expert knowledge from humans, an error-prone, time intensive, and costly process.

- **Challenge 3: Shallow intelligence**. Machine learning and data mining intend to reveal the latent structure, dynamics, and relationships of data from which the machine can learn the kind of complicated functions that imitate human-like cognitive abilities such as recognition or other high-level abstractions. However, machine learning has not yet mastered the essence of human learning, and still requires the intelligent aspects of learning to be engineered by humans [31]. It has been suggested by a number of researchers that deep architectures, which are composed of multiple levels of non-linear operations, might advance the state-of-the-art towards human-like artificial intelligence.

Above challenges motivate the need for novel analytical methods that leverage the power of synergistic human-machine intelligence. What would be desired is a diverse set of tools, which enable medical practitioners to better cope with the diversity of data sources contained in the electronic health record to perform holistic analysis. A holistic approach would offer potential to find hidden relationships between different data sources and their latent patterns.

1.3 Objectives

This dissertation proposes a diverse analytical framework to maximize a human's ability to better 1) reason about, 2) learn, and 3) understand biomedical imaging and healthcare data within the patient's electronic health record. The goal of this work is to improve the capabilities of medical practitioners to efficiently cope with the diversity of the EHR and its large, complex, and heterogeneous data sources by leveraging the synergistic aspect of human-machine intelligence.

In the first part of this dissertation we explore the question how we can build intelligent imaging analytics that are commensurate with human capabilities and constraints, specifically for optimizing data visualization and automated labeling workflows. Our journey starts with heuristic rule-based analytical models that are derived from task-specific human knowledge. From this experience, we move on to data-driven analytics, where we adapt and combine the intelligence of the model based on prior information provided by the human and synthetic knowledge learned from partial data observations. Within this realm, we propose a novel Bayesian transductive Markov random field model that requires minimal human intervention and is able to cope with scarce label information to learn and infer object shapes in complex spatial, multimodal, spatio-temporal, and longitudinal data. We then study the question how machines can learn discriminative object representations from dense human provided label information by investigating learning and inference mechanisms that make use of deep learning architectures. The developed analytics can aid visualization and labeling tasks, which enables the interpretation and quantification of clinically relevant image information.

The second part explores the question how we can build data-driven analytics for exploratory analysis in longitudinal event data that are commensurate with human capabilities and constraints. We propose human-intuitive analytics that enable the representation and discovery of interpretable event patterns to ease knowledge absorption and comprehension of the employed analytics model and the underlying data. We propose a novel doubly-constrained convolutional sparse-coding framework that learns interpretable and shift-invariant latent temporal event patterns. We apply the model to mine complex event data in EHRs. By mapping the event space to heterogeneous patient encounters in the EHR we explore the linkage between healthcare resource utilization (HRU) in relation to disease severity. This linkage may help to better understand how disease specific co-morbidities and their clinical attributes incur different HRU patterns. Such insight helps to characterize the patient's care history, which then enables the comparison against clinical practice guidelines, the discovery of prevailing practices based on common HRU group patterns, and the identification of outliers that might indicate poor patient management.

1.4 Contributions

Our contributions can be organized into two main parts. The first part (see Chapter 2- 5) presents a diverse collection of human-assisted image analytics to visualize and label meaningful information from large, complex, and heterogeneous image data. The second part (see Chapter 6) presents human-intuitive exploratory analytics for temporal event pattern mining in large collections of EHRs. Chapters 7 and 8 reiterate the contributions of our work, describe the significance of our research, outline the limitation of our framework, and point to future work.

The research proposed herein uniquely bridges different data domains, data analyses tasks, and knowledge domains by presenting a diverse set of analytics that leverage synergistic human-machine intelligence to enable medical practitioners to cope with the EHR and its different data sources. Our interdisciplinary approach draws from a variety of techniques and crosses the boundary from biomedical imaging to healthcare informatics. In addition, the generic nature and applicability of the proposed techniques, when integrated and combined together, enable the analysis of the electronic health record within a holistic perspective. Each chapter addresses a unique aspect of how one can exploit synergistic human-machine intelligence to solve specific data challenges within the EHR. In general terms we approach the data analyses task from the 1) visualization, 2) labeling, and 3) mining perspective with the following contributions outlined below.

- In **Chapter 2** we present analytics for human-assisted visualization of complex latent tree structures within large volumetric image data. We invent and implement a novel algorithm that enables the intuitive exploration of complete vascular trees and their internal volume structure. We compare our algorithm with the state-of-the-art and demonstrate superior performance in terms of visualization quality and the ability to preserve anatomical shape appearance. Our developed tools enable the interpretation of sparse information in large volumetric image data. The intuitive visualization of tree-like objects enables medical practitioners to better reason about the complex vasculature of the human body at different scales.

- **Chapter 3** presents an interactive pipeline for human-assisted automated labeling of object boundaries in unimodal image data. Our method enables labeling of objects that exhibit high variability in shape, intensity, and texture. We show as part of a large evaluation experiment that our pipeline improves the state-of-the-art for interactive labeling of geographic atrophy lesions in ophthalmic image data. The generic applicability of our pipeline enables medical practitioners to quantify and label a wide range of disease phenotypes and organ anatomies within the human body.

- **Chapter 4** proposes extensions to the naive Bayes algorithm within a transductive learning and inference paradigm. We introduce a semi-parametric form of the naive Bayes algorithm in combination with a Markov random field model. We develop an algorithm for automated object and multi-object labeling with minimal human intervention. In numerous experiments we demonstrate that the algorithm generalizes to different data sources and application domains. We show the performance of the algorithm on unimodal, multimodal, and spatio-temporal data comprising planar and volumetric image data. The ability to label multiple objects with minimal human intervention enables medical practitioners to more efficiently quantify complex disease phenotypes that occur at different locations. Labeling tools that can cope with scarce label information are important until large annotated medical image database are available.

- **Chapter 5** presents our initial investigations to employ deep learning and inference architectures to automate anatomical labeling of human brain image volumes. We present a novel application of convolutional networks to build discriminative features for brain parcellation, which are automatically learned from labels provided by human experts. Initial validation experiments show promising results for automatic brain parcellation, suggesting that the proposed approach has potential to be an alternative to template or atlas-based parcellation approaches. Moreover, the ability to learn complex functions from only human provided labels without feature engineering has important practical implications. Such tools have the potential to be easily transferable to other analysis tasks as long as rich label information is available.

- In **Chapter 6** we propose extensions to the nonnegative matrix factorization algorithm. We present a novel temporal event matrix representation and learning scheme to perform event pattern mining in longitudinal heterogeneous EHRs. We propose a doubly-constrained convolutional sparse-coding framework that learns interpretable and shift-invariant latent temporal event patterns. We apply our methods to study the linkage between healthcare resource utilization and disease severity in a pool of over 20,000 patients. The developed analytics for knowledge discovery bring the aforementioned contributions together in representing a patient within a generic event knowledge representation to enable the mining of group-specific patient characteristics that are derived from heterogeneous data sources within the EHR. This work has the potential to revolutionize the way how event data is treated within the EHR.

- **Chapters 7 and 8** discuss the limitations of our proposed framework and concludes the dissertation by reiterating the main contributions. At the end we outline the significance of our research and point to ongoing directions of future work.

A more detailed account of our contributions is outlined in each abstract and introduction of the individual chapters. Our contributions provide a diverse set of methods that enable medical practitioners to visualize, label, and discover a variety of heterogeneous data sources. Such interdisciplinary research with a focus on a diverse set of data and problem domains is unique and enables medical practitioners to take a holistic approach. The employed labeling methods can be extended or modified for a diverse set of application scenarios to quantify anatomical regions within the human body or disease phenotypes that are captured by large and complex medical imagery. The diversity of tools developed in this dissertation addresses a broad range of problem domains in analyzing the electronic health records and its large, complex, and heterogeneous data sources.

The research presented in this dissertation was published/submitted for publication in [32; 33; 34; 35; 36; 37; 38; 39; 40; 41]. Research works not included in this dissertation were published in [42; 43; 44; 45; 46]. As part of the research work several software frameworks have been implemented, which are outlined in Section IV.

Part I

Synergistic human-machine intelligence for optimizing visualization and labeling workflows in biomedical data

Chapter 2

Human-intuitive object visualization in volumetric data

2.1 Abstract

In this chapter, we address the problem of enabling humans to better access and interpret sparse information in large volumetric image data. We present analytics for human-assisted visualization of latent tubular trees and develop a novel algorithm that enables the intuitive exploration of the complete internal volume structure. Our work extends the curved planar reformation (CPR) technique and overcomes limitations of existing CPR extensions.

The shape of the medial axis of the complete tubular tree is obtained within an energy optimization framework utilizing front propagation and graph-theoretic approaches. The medial axis tree (MAT) guides a reformation process that projects the complete three-dimensional tree onto two-dimensional image planes. Each image plane slices through every compartment of the volumetric tree at certain rotation angles. We use shape properties to estimate the orientation of the tree for rotation-invariant projection. Radial sampling planes perpendicular to the MAT tangents are the basis for topological and orientation invariant visualization of the vascular lumen.

We demonstrate our algorithm within the field of diagnostic cardiology. Comparative assessment of our algorithm results in superior visualization performance with respect to the state-of-the-art.

The projective mapping from three-dimensional space to the two-dimensional space allows intuitive exploration of the vessel tree interior, which could be used as an interface to human intuitive labeling of the complete vessel tree.

2.2 Introduction

Visualization of diagnostic relevant information in large volumetric images is an important topic in biomedical imaging. Computed tomography angiography (CTA) or magnetic resonance angiography (MRA) are non-invasive high-resolution *in-vivo* image acquisition techniques that enable the examination of vascular diseases such as coronary artery disease. Heart disease is the leading cause of death for both women and men in the United States. In 2005, 652,091 people died of heart disease and accounts for 27.1% of all U.S. deaths. Coronary heart disease is caused by atherosclerosis; the narrowing of the coronary arteries due to fatty build ups of plaque, and is likely to produce angina pectoris (chest pain), heart attack or both. In 2005, coronary heart disease caused 445,687 deaths and is the single leading cause of death in America today. This year an estimated 1.26 million Americans will have a new or recurrent coronary attack. The cost of heart disease and stroke in the U.S. in 2005 was projected to exceed $394 billion: $242 billion for healthcare expenditures and $152 billion for lost productivity from death and disability.

CTA and MRA provide high-resolution image information about the vascular anatomy and pathology such as the narrowing of the artery lumen, calcification, and atherosclerotic plaque formations. The vessel interior is of great importance for characterizing the degree and extent of vascular diseases. In order to evaluate the vascular tree the whole vessel lumen must be investigated, which is tedious and time consuming [47] given large volumetric images and complex tree topologies that are hidden in the data.

Curved Planar Reformation (CPR) has proved to be a useful visualization technique for practical assessment of curved tubular structures within the human body. CPR resamples a single vessel compartment along its medial axis to produce a curved cross-section through the vessel lumen. This enables the accurate visualization of diagnostic relevant information within the vessel lumen. Extensions to the CPR technique were proposed by Kanitsar *et al.* [48; 49], to improve the visualization of the complete vascular tree and its lumen for all compartments. While these projective transformations provide enhanced visualization, they are not able to correctly visualize trees that exhibit non-planar alignment and arbitrary tree topologies.

Vascular trees in medical image volumes are not aligned to planar cross-sections of the volumetric image grid and thus aggravate simultaneous visualization of diagnostic relevant information. Furthermore, complex tree topologies require the need for an adaptive projection scheme to prevent visualization artifacts in order to preserve anatomical information. Fig. 2.1 shows volume rendering examples of two vascular tree topologies in the human body. The left images show the MAT of the peripheral artery tree and the right image shows the MAT of the coronary arteries.

In this chapter, we will present analytics for human-assisted visualization of latent tubular trees and develop a novel algorithm that enables the intuitive exploration of the complete internal volume structure. Our work extends the curved planar reformation (CPR) technique [47; 50] and overcomes limitations of existing CPR extensions [48; 49].

We obtain the shape of the complete medial axis tree (MAT) by means of an energy optimization framework utilizing front propagation and graph-theoretic approaches. The MAT guides a reformation process that projects the complete three-dimensional tree onto two-dimensional image planes. Each image plane slices through every compartment of the volumetric tree at certain rotation angles. We use shape properties

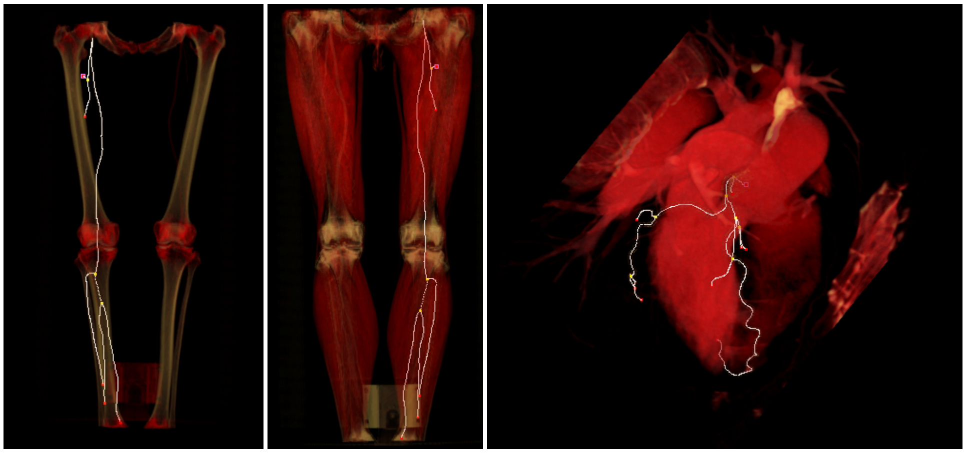


Figure 2.1: Volume rendering examples of the peripheral and coronary artery tree. Left: Medial axis tree of the peripheral vasculature rendered in white. Right: Medial axis tree of the coronary vasculature rendered in white. Both trees exhibit different tree topologies and orientations with respect to the grid structure of the image volume.

to estimate the orientation of the tree for rotation-invariant projection. Radial sampling planes perpendicular to the MAT tangents are the basis for topological and orientation invariant visualization of the vascular lumen.

We will demonstrate our algorithm within the field of diagnostic cardiology. Comparative assessment of our algorithm results in superior visualization performance with respect to the state-of-the-art [48; 49].

The projective mapping from three-dimensional space to the two-dimensional space allows intuitive exploration of the complete vessel tree lumen. This transformation aids the visualization of diagnostic relevant information and the interpretation of sparse information in large volumetric image data. By having a complete view of the vascular tree the human can quickly obtain an overview of disease characteristics that are otherwise difficult to assess using traditional visualization techniques. The projective mapping could further help to obtain diagnostic relevant annotations with the volumetric image by letting the human place seed labels on the two-dimensional view to perform three-dimensional interactive seed label placement, which could then be used by an automated interactive labeling algorithm.

2.3 Prior art

2.3.1 Curved planar reformation and extensions

Not much work has been published on technical aspects of automated techniques for visualizing the interior of vascular trees that exhibit arbitrary topology and non-planar alignment. Regarding vascular visualization a pool of techniques exist such as multi-planar reformation (MPR), shaded surface display (SSD), maximum intensity projection (MIP), curved planar reformation (CPR), or volume rendering (VR). However, diagnostic features of vascular trees with small scale diameter such as in coronary arteries are often hard to investigate simultaneously with aforementioned visualization techniques. In VR, atherosclerotic plaque accumulating on vessel walls can obstruct the view of the vascular lumen, which is important for disease diagnosis. Furthermore, other body tissues often occlude the objects of interest while traditional MPR views only visualize localized cross-sections of the vascular tree due to non-planar alignment. Manual creation of curved MPRs for whole vascular trees is tedious and time consuming especially in large patient studies. Other methods for automatic generation of curved planar reformations only consider a single vessel segment to visualize [47; 50]. Vrtovec *et al.* applied CPR on 3D spine images for automated visualization [51]. Kanitsar *et al.* proposed extensions to traditional CPR techniques consisting of "multi-path" and "rotated" CPR [48]. They also proposed two advanced CPR techniques called "helical" and "untangled" CPR enabling the visualization of complete vascular interiors for single vessel segments in one image and occlusion free reformation of the whole vascular tree using an untangling scheme [49].

However, to deal with arbitrary topology and non-planar alignment aforementioned projective transformations cannot directly be applied. Complex tree topology with high vessel curvature requires an adaptive projection scheme with tangential aligned radial sampling planes for artifact-free visualization. In this work we will describe solutions and implementation details to the problem of handling arbitrary tree topology and orientations of vascular networks and apply our projection scheme to the visualization of coronary artery and peripheral artery trees. We will name the proposed projection scheme as tangential curved planar reformation (TCPR).

2.4 Methods

2.4.1 Preliminaries

Consider a voxel $\mathbf{x} = (x_1, x_2, x_3) \in \mathbb{Z}^3$. In $\mathbb{Z}^3$, $\mathbf{x}$ can have different neighborhood structures. The most simplest setting is the 6, 18, and 26-neighborhood of $\mathbf{x}$ expressed through the following distance constraints

$$N_6(\mathbf{x}) = \{\mathbf{y} | \sum_i |\mathbf{y}_i - \mathbf{x}_i| \leq 1\} \tag{2.4.1}$$

$$N_{26}(\mathbf{x}) = \{\mathbf{y} | \max_i |\mathbf{y}_i - \mathbf{x}_i| \leq 1\} \tag{2.4.2}$$

$$N_{18}(\mathbf{x}) = \{\mathbf{y} | \sum_i |\mathbf{y}_i - \mathbf{x}_i| \leq 2\} \cap \{\mathbf{y} | \max_i |\mathbf{y}_i - \mathbf{x}_i| \leq 1\}. \tag{2.4.3}$$

The concept of a voxel neighborhood is important since its neighborhood characterizes the connectivity of a set of voxels $\Omega = \{\mathbf{x}_i\}_{i=1}^N$. Further, the topology of Ω can be assessed by analyzing the neighborhood structure of each voxel.

Besides the topology of Ω we are interested in distance properties of each voxel with respect to the boundary of Ω. We notate this boundary with $\partial\Omega$. This brings us to the concept of front propagation and the Eikonal equation by considering $\partial\Omega$ to be the starting point of an evolving surface Γ, from which one can measure movement and distance properties.

The Eikonal equation is a non-linear partial differential equation encountered in problems of wave propagation and takes the following form

$$|\nabla u(\mathbf{x})| = F(\mathbf{x}), \mathbf{x} \in \Omega, \qquad (2.4.4)$$
$$\text{s.t. } u(\mathbf{x})|_{\partial\Omega} = 0.$$

Here Ω is an open set in $\mathbb{R}^n$, u an unknown function, $F(\mathbf{x})$ the time cost at $\mathbf{x}$, and $\partial\Omega$ the boundary to $\mathbf{x}$ in side Ω. An approximate solution to Equation 2.4.4 can be obtained by the *fast marching method*. In the case of $F = 1$, $u(\mathbf{x})$ gives the signed distance from $\partial\Omega$ to each point $\mathbf{x}$.

A signed distance function $f(\mathbf{x})$ of a set Ω in a metric space determines the closest distance of a given point to $\partial\Omega$. At the boundary we have $f(\mathbf{x}) = 0$. Outside of Ω the signed distance function takes negative values $f(\mathbf{x}) < 0$. This distance function $f(\mathbf{x})$ enables the characterization of medial points in Ω.

2.4.2 Representation of the medial axis tree

We are interested in finding the medial axis tree of an object. Let's assume $\mathcal{M}_T$ represents an anatomical manifold of a tree with arbitrary shape topology and orientation in domain $\Omega \subset \mathbb{R}^3$. $\mathcal{M}_T$ is binarized such that $\mathcal{M}_T(\mathbf{x}) = 1$ corresponds to the inside and $\mathcal{M}_T(\mathbf{x}) = 0$ to the outside region. Furthermore, let's assume this manifold is topological correct, i.e. does not contain holes or handles, is closed, and differentiable. We seek the set of medial paths $\mathcal{M}_P \subset \mathcal{M}_T$, such that $\mathcal{M}_P$ fulfills the following *properties of medialness*: 1) connectedness, 2) one-voxel thickness, and 3) reconstruct-ability of $\mathcal{M}_T$ from $\mathcal{M}_P$. With medial we mean, that $\mathcal{M}_P$ is centered with respect to the boundary $\partial\mathcal{M}_T$.

First, we assign a label to each voxel in $\mathcal{M}_T$ with a distance function f from the boundary set $\partial\mathcal{M}_T = \{\mathbf{x} \in \partial\mathcal{M}_T | N_6(\mathbf{x}) < 6\}$. From $\partial\mathcal{M}_T$ we propagate a front Γ with constant travel time F in the normal direction of the boundary pointing inwards. The arrival time t of Γ labels every voxel in $\mathcal{M}_T$, such that the Eikonal equation is satisfied

$$|\nabla M_T| \; = \sum_{i=1}^{n} \left(\frac{\partial M_T}{\partial \mathbf{x}_i} \right)^2 = 1 \tag{2.4.5}$$

Since $F = const$, arrival time t is a measure for medialness. Once all voxels in $\mathcal{M}_T$ have a label associated to their medialness measure, we build an undirected graph $\mathcal{G} = (V, E)$ from $\mathcal{M}_T$. Each edge $e \in E$ is assigned a value from a weighting function $w(e) : E \to \mathbb{R}$. To obtain a set of connected and one-voxel thick medial paths $\mathcal{M}_P$ we form a minimum spanning tree $\mathcal{MST} \subset \mathcal{G}$ by minimizing the following weighting function

$$\min_e w(\mathcal{MST}) = \sum_{e \in \mathcal{MST}} w(e) = -f(\mathbf{x}_i) + |\mathbf{x}_i - \mathbf{x}_{i-1}|, \text{ with} \tag{2.4.6}$$

$$\mathbf{x}_{i-1} \in N_{26}(\mathbf{x}_i). \tag{2.4.7}$$

where we use two distance metrics: 1) negative medialness $-f(\mathbf{x})$ computed from Equation 2.4.5 and 2) the geodesic distance in the topological neighborhood $N_{26}(\mathbf{x})$. Note that the edge weights must be distinct, i.e. $w(e_i) \neq w(e_j)$ for any pair of edges e_i and e_j, so the $\mathcal{MST}$ is unique. For obtaining the $\mathcal{MST}$ from $\mathcal{G}$ we use Kruskal's algorithm. To reduce the complexity of $\mathcal{MST}$ we perform a recursive node abstraction, where the $\mathcal{MST}$ is simplified into a binary node tree such that a single node in the tree has at most 2 children.

2.4.3 Tree orientation determination and adjustment

As noted before, $\mathcal{MST}$ can have arbitrary orientation due to non-planar alignment, thus no assumption can be made regarding the *axis of projection* vector $\mathbf{p}$. To be invariant against arbitrary tree orientations, we change the viewing vector (basis) $\mathbf{v}$, such that $\mathbf{v}$ is perpendicular to $\mathbf{p}$ pointing towards the human observer. To determine $\mathbf{p}$ we first compute the center of mass $\mu_{\mathbf{n}}$ from all nodes $\{n_i\}_{i=0}^{N} \in \mathcal{MST}$

$$\mu_{n_x} = \frac{\int\int\int n_x \rho(n_x, n_y, n_z) dV}{M} \tag{2.4.8}$$

$$\mu_{n_y} = \frac{\int\int\int n_y \rho(n_x, n_y, n_z) dV}{M} \tag{2.4.9}$$

$$\mu_{n_z} = \frac{\int\int\int n_z \rho(n_x, n_y, n_z) dV}{M}, \text{ with} \tag{2.4.10}$$

$$M = \int\int\int \rho(n_x, n_y, n_z) dV, \tag{2.4.11}$$

where $\rho(\cdot)$ is a node density function. Then, $\mathbf{p}$ is defined as the direction vector pointing from $\mu_{\mathbf{n}}$ to $\mathbf{n}_0$, with $\mathbf{n}_0$ being the root node of $\mathcal{MST}$. Note that this approach is robust against different tree orientations and to ambiguous shape topologies, which cause problems when simply assuming the largest principle component to be $\mathbf{p}$. Our assumption allows to determine a correct $\mathbf{p}$ for tree topologies that are flat in depth, but wide. Recalling the property of *reconstructability*, $\mathcal{M}_T$ can be reconstructed from $\mathcal{MST}$. At last, in order to be able to project $\mathcal{MST}$ onto $\mathbb{R}^2$ we convert the discrete node representation to a continuous representation by resampling every path in $\mathcal{MST}$ using a B-Spline interpolation scheme. The result is a smooth, connected, one-voxel thick medial axis tree as shown in Fig. 2.1

2.4.4 Tree projection from three to two dimensions

We project the $\mathcal{MST}$ from $\mathbb{R}^3$ to $\mathbb{R}^2$ and perform alignment, such that the projection is oriented top-to-bottom meaning that the root of the tree is at the top of the image and the leaves of the tree at the bottom part of the image. Furthermore, we project $\mathcal{MST}$ at different viewing angles $\theta \in [0...2\pi]$, where each view is projected on a separate image plane $\mathbf{I}_\theta$.

From the tree root we cast a line segment into a predefined direction perpendicular to the global rotation axis $\mathbf{p}$. The length is determined by computing the maximum distance of all medial axis tree points to $\mathbf{p}$. For a specific viewing angle we project every node in $\mathcal{MST}$ onto $\mathbf{I}_\theta$, where the x-coordinates are determined by the shortest distance from each centerline point to the reference plane and the y-coordinates are computed by taking the absolute travel distance to the root element of the tree. This preserves the original vessel length information in the final image.

2.4.5 Topological invariant tangential resampling

We compute the Frenét trihedral for the set of vector functions $\{\mathbf{r}_i(t)\}_{i=1}^{M}$, where i indexes the individual path segments of $\mathcal{MST}$. The set of paths satisfy the following conditions: 1) $\mathbf{r}_i(t)$ is open and 2) the second derivative $\mathbf{r}_i^{''}(t)$ exists. Then we sample along each $\mathbf{r}_i(t)$ by casting a set of sample lines $L = \{l_\theta\}_{\theta=0}^{2\pi}$ for each point in $\mathbf{r}_i(t)$, such that L lies in a plane P that is perpendicular to $\mathbf{r}_i'(t)$. The plane P is spanned by $\hat{\mathbf{n}}(t)$ and $\hat{\mathbf{b}}(t)$ defined as

$$\hat{\mathbf{t}}(t) = \frac{\mathbf{r}'(t)}{\|\mathbf{r}'(t)\|} \tag{2.4.12}$$

$$\hat{\mathbf{n}}(t) = \frac{\hat{\mathbf{t}}'(t)}{\|\hat{\mathbf{t}}'(t)\|} \tag{2.4.13}$$

$$\hat{\mathbf{b}}(t) = \hat{\mathbf{t}}(t) \times \hat{\mathbf{n}}(t), \tag{2.4.14}$$

where × denotes the cross product, $\hat{\mathbf{t}}(t)$ a tangent, $\hat{\mathbf{n}}(t)$ a normal, and $\hat{\mathbf{b}}(t)$ a binormal vector. To cast corresponding sampling lines l_θ for tangential sampling we rotate $\hat{\mathbf{b}}(t)$ around $\hat{\mathbf{t}}(t)$ using

$$s = 1 - \cos(\theta), \tag{2.4.15}$$

$$\hat{\mathbf{a}} = (x, y, z) \tag{2.4.16}$$

$$R(\hat{\mathbf{a}}, \theta) = \begin{bmatrix} sx^2 + \cos(\theta) & sxy - \sin(\theta)z & sxz + \sin(\theta)y & 0 \\ sxy + \sin(\theta)z & sy^2 + \cos(\theta) & syz - \sin(\theta)x & 0 \\ sxz - \sin(\theta)y & syz + \sin(\theta)x & sz^2 + \cos(\theta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.4.17}$$

where $\hat{\mathbf{a}}$ is the axis of rotation, θ the rotation angle, and R the rotation matrix with homogeneous coordinates. Each sampling line is defined as

$$l_\theta = \langle r_x, r_y, r_z \rangle + t \langle a_\theta, b_\theta, c_\theta \rangle, \text{ with} \tag{2.4.18}$$

$$t \in [d_{nl}, -d_{nl}] \tag{2.4.19}$$

$$\langle a_\theta, b_\theta, c_\theta \rangle = R(\hat{\mathbf{t}}(t), \theta) \cdot \hat{\mathbf{b}}(t), \tag{2.4.20}$$

where t is a scale parameter, which upper and lower bound is the maximal node-to-line distance d_{nl}. Note that l_θ samples a complete cross-section of the lumen at $\mathbf{r}_i(t)$ and since L is perpendicular to $\mathbf{r}'_i(t)$, the visualization of vessel interiors is invariant to arbitrary tree topologies. The final step consists of mapping the line samples L to consecutive image regions.

2.4.6 Image partitioning and rendering

Once we map the projected tree $\mathcal{MST} \subset \mathbb{R}^2$ onto the image $\mathbf{I}_\theta$ we obtain an image partitioning R by casting horizontal rays from each point in the tree until all empty image pixels have been visited once. The locations where two rays meet mark the region boundaries between two neighboring tree segments. This procedure is repeated for all rotation angles θ. Each region is filled with the tangential line sample data obtained in Section 2.4.5. The rendering process starts with the tree segments that are most distant to the projection plane. By prioritizing the partition rendering to the segments that are closest to the observer, we can prevent rendering artifacts.

2.5 Experiments and results

We implemented an interactive user interface and performed validation experiments on two types of vascular trees: 1) coronary and 2) peripheral abdominal trees. Coronary trees are highly curved and comprise arbitrary shape topology and non-planar orientation. Peripheral trees are elongated long structures and span hundreds of image slices. Binary label masks of both types of trees were obtained from CTA and MRA volumetric image datasets using a human-assisted model-based vessel labeling algorithm. The visualization proceeds in an automatic fashion and computes within a couple of seconds. The human can interactively examine different aspects of the tree and its interior by adjusting the projection and radial sampling angle.

Visualization results obtained by our TCPR algorithm on the coronary artery tree are shown in part B and D of Fig. 2.2. Comparison between non-tangential CPR and TCPR are shown in part A and B. The visualization results on two peripheral trees are shown in part C.

2.6 Discussion

The goal was to visualize the whole vessel tree in one image plane while being able to show the whole vessel lumen throughout the vascular tree. We found that TCPR is able to produce artifact-free visualization of the vascular tree lumen. The reformation procedure is computationally efficient and runs fully automatic, but also enables user interactive manipulation. In Figure 2.2 we can see the superior visualization performance of our TCPR algorithm in comparison to the state-of-the-art when com-

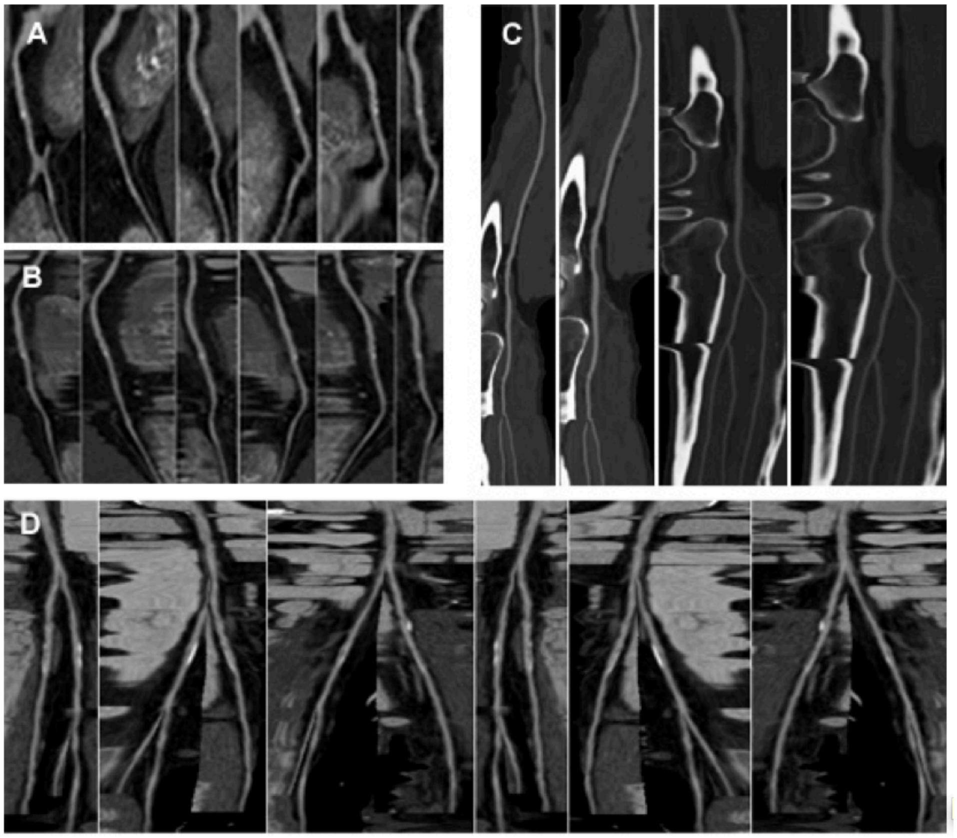


Figure 2.2: Comparison of the proposed Tangential Curved Planar Reformation (TCPR) algorithm with the state-of-the-art. (A), (B) Comparison between the non-tangential and our TCPR projection scheme in MRA data. Note the sampling artifacts of the lumen structure in A, whereas the TCPR algorithm shows correct artifact-free anatomic information of the vessel lumen in (B). (C) Complete vascular tree visualization using TCPR on two different peripheral artery trees. (D) Coronary vascular tree visualization using TCPR at different radial sampling angles.

paring the images in part A and B. Coronary vessels exhibit high curvature changes, which may cause visualization artifacts when the re-sampling direction and the projection vector coincide. It is possible that parts of the vessel tree are aligned along the sampling direction, which leads to artificial lumen scaling as shown in part A. In contrast, TCPR samples along perpendicular directions of the medial axis tangent. The projection scheme adapts to the tree topology resulting in artifact-free visualization of highly curved vessel compartments as shown in part B, which makes the method suitable for a wide range of vascular networks within the human body. We further demonstrate that TCPR can be applied to other tree topologies (e.g. abdonimal vascular trees) as shown in part C.

The projective mapping from three-dimensional space to the two-dimensional space enables intuitive exploration of the complete vessel tree lumen. This transformation aids the visualization of diagnostic relevant information and the interpretation of sparse information in large volumetric image data. By having a complete view of the vascular tree the human can quickly obtain an overview of disease characteristics that are otherwise difficult to assess using traditional visualization techniques. Also the holistic approach to provide human-intuitive visualizations of the complete vascular tree interior enables medical practitioners to quickly assess the whole tree, which may reveal hidden patterns that would not be evident when visualizing each vessel compartment in an independent manner. The projective mapping could further help to obtain diagnostic relevant annotations with the volumetric image by letting the human place seed labels on the two-dimensional view to perform three-dimensional interactive seed label placement, which could then be used by an automated interactive labeling algorithm.

One limitation of our method is the missing context information at regions that are more distant to the vessel tree. The tangential reformation scheme follows the curvature and topology of the tubular tree, which causes distortions far beyond the vessel lumen. One way to circumvent this problem is to incorporate a visualization overlay that provides contextual information of the vessel tree's surrounding environment. Alternatively, standard visualization techniques such as perpendicular cross-sections could be combined to provide additional context information about the vessel's background structure.

Failure cases of our proposed visualization method depend on the quality of the pre-processing results and the correctness of the obtained vessel tree labels. The proposed reformation method assumes a high-quality vessel labeling method. We observed that vascular pathology and low image quality required a user-interactive labeling method to correct for erroneous labeling results. Low image contrast at small vessel scales and vascular pathology were the main reasons for missing vessel compartments. Furthermore, vessel boundary labels have to be regularized and corrected for topological errors to provide a smooth medial axis tree.

2.7 Conclusion

We have presented novel analytics for human-intuitive topological and orientation invariant vascular tree visualization by exploiting intrinsic shape properties of the vascular tree for rotation-invariant projection and radial sampling planes perpendicular to the medial axis tangent. Several visualization experiments were presented to demonstrate the efficacy of our algorithm. The projective mapping from three-dimensional space to the two-dimensional space allows improved intuitive exploration of the vessel tree interior, which could be used as an interface to human intuitive labeling of the complete vessel tree.

Chapter 3

Human-assisted interactive object labeling in image data

3.1 Abstract

This chapter presents a novel image analytic pipeline for human-assisted interactive labeling of object boundaries that exhibit high variability in shape, intensity, and texture.

We compute non-linear gradient approximations using generalized Sobel kernels to deal with varying degrees of noise and to detect edge responses at multiple scales. The human interactively places seed labels near the desired object boundaries, which serve as constraints to impose local minima at the seed locations. Local minimal are enforced by a morphological geodesic reconstruction process to remove degenerate solutions when computing the watershed transform. The watershed transform produces an image partitioning along the gradient such that the boundaries of the partitioning are aligned to the object boundary. Final object boundaries are obtained in an iterative fashion through human-assisted interactive seed label refinement, where we relabel each region according to their respective seed labels.

We demonstrate our approach within the domain of ophthalmology by quantify-

ing pathologic image regions in patients with age-related macular degeneration. A large evaluation study consisting of 100 test cases shows that our interactive labeling pipeline compares favorably with respect to a state-of-the-art interactive graph-based labeling algorithm.

The integration of human knowledge into the labeling process enables the robust delineation of object boundaries in the presence of high variability in shape, intensity, and texture. Compared to fully automated labeling approaches the proposed human-assisted interactive labeling pipeline provides an optimized labeling workflow for a broad class of object appearances.

3.2 Introduction

Fundus autofluorescence (FAF) imaging is a non-invasive technique for *in vivo* ophthalmoscopic inspection of age-related macular degeneration (AMD). AMD is the leading cause of blindness in the U.S. and the developed world [52]. The macula is the central region of the retina with the highest concentration of photoreceptors responsible for sharp central vision. FAF image signals are reliable markers of lipofuscin in the retinal pigment epithelium (RPE) cell layer [53; 54], which closely interacts with the photoreceptors to maintain visual function. Lipofuscin are finely granular yellow-brown pigment granules composed of lipid-containing residues of lysosomal digestions. The accumulation of lipofuscin is a major risk factor for AMD.

AMD occurs in two forms, dry (or atrophic) AMD and wet (or neovascular) AMD. Geographic atrophy (GA) of the RPE, an advanced form of dry AMD, accounts for 12-21% of severe visual loss in this disorder [52]. GA is characterized by round or multi-lobed patches of atrophy of the RPE. Over time, atrophic patches may increase in size and number or may coalesce to form larger areas of atrophy, thus leading to high variability in shape, intensity and texture. Figure 3.1 shows examples of the GA disease phenotype.

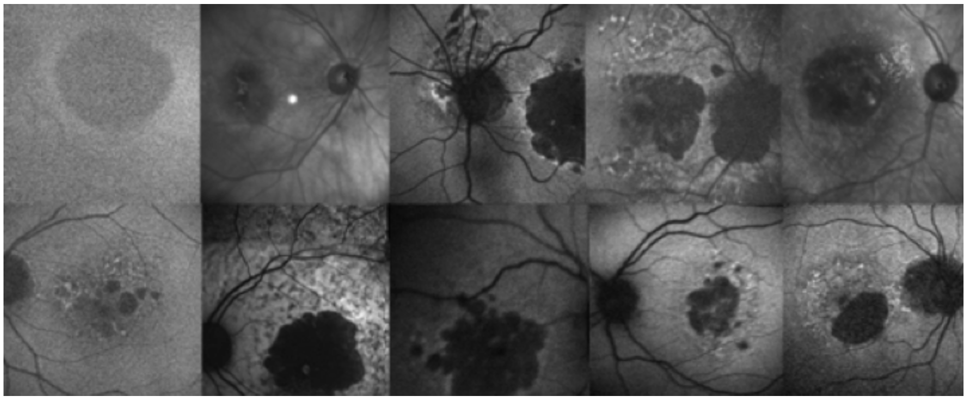


Figure 3.1: Human-assisted interactive object labeling in the presence of high variability in shape, intensity, and texture. Ophthalmic images and pathologic examples of age-related macular degeneration (AMD).

To reduce the enormous morbidity of AMD and its intermediate forms, we must be able to monitor and quantify accurately the natural history and response to treatment of the pathologic phenotype. The quantification of GA is important for determining disease progression and facilitating clinical diagnosis of AMD. Generally, GA quantification methods in the literature have typically relied on visual inspection of FAF images [55], which prevents quantification, or time-consuming manual delineation of GA boundaries [56]. Manual quantification of GA is time-consuming and prone to inter- and intra-observer variability [55]. There has been a continued interest in the use of machine vision techniques to label and quantify the pathologic phenotype of AMD.

The problem of automatic labeling of pathological regions in image data has been widely studied by the medical image analysis community, yet it still remains an unsolved problem [57]. The current state-of-the-art shows that few automated image analysis techniques can be applied fully autonomously with reliable results. Often times post-processing of the obtained labeling result is necessary to validate labeling accuracy and correct for errors.

In the realm of computer-aided diagnosis interactive labeling schemes are well received by physicians, where the combination of human and machine intelligence can provide improved labeling accuracy and efficacy. Interactive labeling schemes may be seen as an appropriate alternative to solve the problem of robust and accurate labeling for a variety of different labeling problems [57; 58; 32].

In this chapter we present a novel image analytic pipeline for human-assisted interactive labeling of object boundaries that exhibit high variability in shape, intensity, and texture. We compute non-linear gradient approximations using generalized Sobel kernels to deal with varying degrees of noise and to detect edge responses at multiple scales. The human interactively places seed labels near the desired object boundaries, which serve as constraints to impose local minima at the seed label locations. Local minimal are enforced by a morphological geodesic reconstruction process to remove degenerate solutions when computing the watershed transform. The watershed transform produces an image partitioning along the gradient such that the region boundaries are aligned to the object boundary. We choose the watershed transform due to its well-defined properties, its simplicity, and computational efficiency [59]. Final object boundaries are obtained in an iterative fashion through interactive seed label refinement, where we relabel each region according to their closest seed labels.

We demonstrate our approach within the domain of ophthalmology by quantifying pathologic image regions of GA. A large evaluation study of 100 cases shows that our algorithm compares favorably with respect to the Random Walker [57], a state-of-the-art interactive graph-based labeling algorithm. Quantitative evaluation shows a mean sensitivity (SE)/specificity (SP) of 98.3/97.7% for our pipeline approach and a mean SE/SP of 88.2/96.6% for the Random Walker respectively.

The integration of human knowledge into the labeling process allows to perform robust delineation of object boundaries in the presence of high variability in shape, intensity, and texture. Automatic labeling algorithms have difficulties in reliably labeling objects that exhibit high variability in appearance and data ambiguity. By integrating human intelligence into the labeling process a variety of labeling problems can be addressed in a robust and accurate manner.

3.3 Prior art

Fully automatic object labeling in image data is still an unsolved problem [57; 58; 32]. Interactive object labeling also known as semi-automatic or semi-supervised labeling is a practical solution to alleviate the inherent limitations of fully automatic labeling. The interactive approach to object labeling has great practical advantages since human knowledge and synthetic machine knowledge can be combined. Interaction can take various forms such as 1) direct human guidance as in the case of the intelligent scissors [60] approach, 2) object boundary initialization based approaches common in active contour models [61] or the level set framework [62], and 3) seed label placement within object regions to constrain the solution space of the final object boundaries. In this section we limit the scope to interactive object labeling approaches that make use of the third type of interaction.

3.3.1 Graph-based interactive labeling

The literature on existing interactive graph-based labeling approaches is vast. Early work that employed graph theory for the task of object labeling in image data was proposed by [63]. Graph-based approaches represent the image as a weighted graph, where each pixel in the image corresponds to a node in the graph and the edges of the graph represent neighborhood relations between the image pixels. Edge weights quantify the similarity between two nodes in the graph, where large values are given to edges that link similar looking nodes together and low edge weights to dissimilar nodes.

A popular approach for graph-based interactive labeling of object boundaries was presented by Boykov *et al.* [64; 65; 66; 67]. He proposed a combinatorial optimization technique termed Graph Cuts using a novel max-flow min-cut algorithm. So-called object/background seed labels (hard constraints) are treated as source/sink nodes in a graph that provide the initial conditions for the max-flow min-cut operation to minimize an energy objective function (soft constraints) that incorporates object region and boundary terms. This minimum cut is a global minimizer and corresponds to the set of edges in a graph with minimum total weight separating the source and sink nodes in a global optimum. While the Graph Cut algorithm and extensions of it were successfully applied to a variety of labeling problems, the Graph Cut approach has several limitations. First, the minimum cut criterion might lead to overly cautious estimated object boundaries in the presence of noise, low contrast, and limited number of human-placed seed labels. Another difficulty is the extension of the Graph Cut framework to the multi-class labeling task, an NP-hard problem requiring the use of heuristic approximations to obtain a solution.

Grady *et al.* [68; 57; 69] proposed an interactive graph-based multi-class labeling approach, where the user provides initial label information in form of seed labels indicating the object regions within an image. A seed label is a human-provided location in the image that has associated a certain label value. The edge weights in the Random Walker algorithm are treated as probabilities of a particle with the probability to first reach a certain seed label. The Random Walker algorithm labels an unseeded pixel by resolving the question: Given a random walker starting at location x, what is the probability that it first reaches each of the K seed points? Validation studies have shown state-of-the-art performance.

3.3.2 Interactive labeling based on the watershed transform

Beucher *et al.* first applied the concept of the watershed transform to the image labeling problem [70] in the late 1970s. The watershed transform is a morphological image partitioning technique and found wide use in medical image processing [59]. The labeling technique is derived directly from the topographical watershed idea whereby all points on the surface are grouped according to the concept of water falling onto the surface and flooding each local minimum until total immersion [71]. The analogy can be explained by taking the example of rain drops associated with each point in the image. Any two points are in the same region, also named a catchment basin, if they fall to the same point. The watershed lines, which divide the image, result from the catchment basins that start to meet each other as more rain falls onto the surface.

Meyer *et al.* proposed a marker-controlled watershed labeling method [72] to overcome the over-segmentation problem. The marker-based watershed transform [73; 74; 75] is a technique suitable for interactive labeling. Its properties have been studied in [76] and its robustness with respect to the marker placement has been shown. The watershed transform yields the same results for two different sets of markers as long as they are located within the same catchment basin. Many extensions to the marker-controlled watershed algorithm have been proposed such as [77; 59; 78; 79; 80]. Couprie *et al.* [81] proposed power watersheds as a new image labeling framework extending the graph cuts, random walker, and the optimal spanning forest approach.

3.4 Methods

3.4.1 Preliminaries

A key component of our pipeline approach is the watershed transform. We can define a continuous watershed with the help of distance functions. Assume that the image f is an element of the space $\mathcal{C}(D)$ of real twice continuously differentiable functions on a connected domain D with only isolated critical points. Then the *topographical distance* between points p and q in D is defined by [82]

$$T_f(p, q) = \inf_{\gamma} \int_{\gamma} \|\nabla f(\gamma(s))\| ds, \tag{3.4.1}$$

where the infimum is over all paths γ inside D with $\gamma(0) = p, \gamma(1) = p$. The *topographical distance* between a point $p \in D$ and a set $A \subseteq D$ is defined as $T_f(p, A) = \min_{a \in A} T_f(p, a)$. The path with shortest T_f-distance between p and q is a path of steepest slope on the graph of f. From this, one can define the following definition of the watershed transform.

Definition (Watershed transform) *Let $f \in \mathcal{C}(D)$ have minima $\{m_k\}_{k \in I}$,for some index set I. The catchment basin $CB(m_i)$ of a minimum m_i is defined as the set of points $x \in D$, which are topographically closer to m_i than to any other regional minimum m_j:*

$$CB(m_i) = \{x \in D | \forall j \in I \setminus \{i\} : f(m_i) + T_f(x, m_i) < f(m_j) + T_f(x, m_j)\}. \tag{3.4.2}$$

The watershed of f is the set of points, which do not belong to any catchment basin:

$$W_{shed}(f) = D \cap \left(\bigcup_{i \in I} CB(m_i) \right)^c . \tag{3.4.3}$$

Let W be some label, $W \notin I$. The watershed transform of f is a mapping $\lambda : D \to I \cup \{W\}$, such that $\lambda(p) = i$ if $p \in CB(m_i)$, and $\lambda(p) = W$ if $p \in W_{shed}(f)$.

So the watershed transform of f assigns labels to the points of D, such that: 1) different catchment basins are uniquely labeled, and 2) a special label W is assigned to all points of the watershed of f. For implementation details and other definitions of the watershed transform we refer the reader to [82].

3.4.2 A human-assisted interactive object labeling pipeline

The interactive labeling pipeline consists of several modules, i.e. 1) an optional image preprocessing module for noisy image data, 2) a module for computing non-linear gradient approximations, 3) a geodesic reconstruction module, 4) a watershed transformation module, and 5) a relabeling module. A schematic overview of the iterative labeling workflow is shown in Figure 3.2.

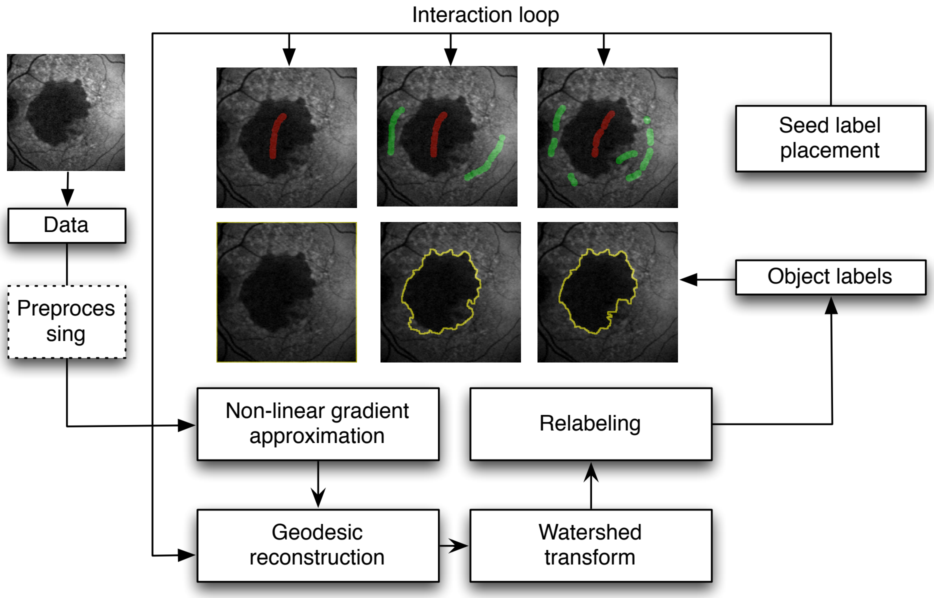


Figure 3.2: Human-assisted interactive labeling pipeline. Interactive labeling workflow for object quantification. The process starts with approximate seed label placement drawn by a human expert. The non-linear gradient approximation module computes image cues that indicate object boundaries. The geodesic reconstruction module enforces local minima on the approximated gradient map. The watershed transform module computes an image partitioning, whose labels are reassigned by the relabeling module to produce an intermediate label boundary. The labeling process iterates until the final label boundaries are obtained.

The interaction pipeline starts with human input. The human draws so-called seed labels to indicate approximate locations of the object of interests. Here we consider the case of binary object labeling, where the image is partitioned into two disjoint regions, one indicating the object and the other the image background. The seed labels serve as constraints to impose local minima regions at the seed locations.

In the next step, we perform non-linear gradient approximations using generalized separable Sobel kernels [83; 84] for fast multi-scale edge detection and smoothing. A Sobel edge detection operator in its standard form consists of two 3x3 kernels for each gradient direction

$$\mathbf{G}_x = \mathbf{K}_x \star \mathbf{I} \tag{3.4.4}$$

$$\mathbf{G}_y = \mathbf{K}_y \star \mathbf{I} \tag{3.4.5}$$

$$\mathbf{K}_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \; \mathbf{K}_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{3.4.6}$$

where $\star$ is a two-dimensional convolutional operator, $\mathbf{I}$ an image, and the kernels $\mathbf{K}_{x,y}$ itself perform a smoothing operation that is perpendicular to the direction of the derivative, which is approximated with a central difference scheme. The integration of a smoothing step within the Sobel kernel enables noise robustness, the reduction of aliasing artifacts, and regularization along edge responses for smoother object boundaries. From Equation 3.4.4- 3.4.5 one can compute the gradient magnitude and phase as follows

$$|\mathbf{G}| = \sqrt{\mathbf{G}_x^2 + \mathbf{G}_y^2}, \tag{3.4.7}$$

$$\theta = atan2\left(\mathbf{G}_y, \mathbf{G}_x\right). \tag{3.4.8}$$

To account for multi-scale edge responses higher order Sobel kernels can be computed with a polynomial transform representation of the form

$$H(p,q) = \mathcal{P}_2[f(x,y)] = \sum p^x q^y f(x,y), \tag{3.4.9}$$

where p and q are filter indices for the x and y direction with the origin being in the left bottom corner of the filter H. Thus the Sobel kernel in Equation 3.4.6 can be written as

$$\mathbf{K}_x = H(p,q) = -1 - 2q - q^2 + p^2 + 2p^q + p^2q^2 \tag{3.4.10}$$

$$= (1+q)^2(-1+p^2) \tag{3.4.11}$$

$$= (-1+p)(1+p)(1+q)(1+q). \tag{3.4.12}$$

Here each factor represents the atomatic element of the decomposed Sobel kernel into four separable two element convolution kernels. The extension to higher order kernels of size $k \times k$ can be computed with

$$\frac{\partial}{\partial x} \rightarrow (-1+p)^{n-1}(1+q)^n(1+p) \tag{3.4.13}$$

$$\frac{\partial}{\partial y} \rightarrow (-1+p)^n(1+q)^{n-1}(1+q), \tag{3.4.14}$$

where $n = k - 1$.

Given multiple Sobel kernels of size $k \times k, k \in \{3, 5, 7, 9\}$ we compute the gradient magnitude of the image f using Equation 3.4.7. On $|\mathbf{G}|$ we impose local minima at the seed locations that are provided by the human expert. A seed s consists of an input output pair $s = \{\mathbf{x}, y\}$, were $\mathbf{x}$ can be the seed location and y the seed label. To enforce local minima we compute the morphological reconstruction to prevent degenerative solutions.

To obtain an image partitioning we choose the watershed transform (WT) algorithm defined in Equation 3.4.3 due to its well-defined properties and computational efficiency [59]. We compute the WT to obtain an image partitioning and relabel each region according to their respective seed labels. Final object boundaries are obtained in an iterative manner through seed label refinement. Through this interactive process the human obtains direct feedback as object and background seed labels are drawn onto the image. An example of the intermediate processing results are shown in Fig. 3.3.

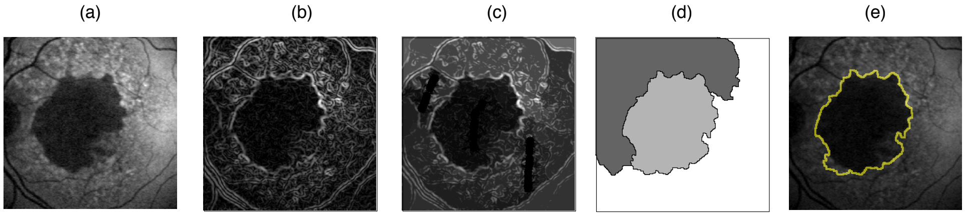


Figure 3.3: Intermediate processing results of the interactive object labeling pipeline. (a) The original image, (b) the gradient magnitude image, (c) the local minima image, (d), the watershed catchment basins, and (e) the object boundary labels.

3.5 Experiments and results

Fundus autofluorescence (FAF) images have been recorded using the Heidelberg model HRA confocal SLO (Heidelberg Inc, Heidelberg, DE). This instrument uses blue laser light at 488nm for illumination and a barrier filter at 500nm to limit the captured light to auto fluorescent structures. The FAF images consisted of bit-mapped laser scans of varying image resolution ranging from 256x256 to 870x870 pixels in size. Each image was an average of 3 to 6 scans composed by the SLO software. A human expert provided ground truth information for all 100 FAF images by manually drawing the object boundaries within each image. Figure 3.4 shows an example of the pool of images that were used for the evaluation study. From this pool one can see the high variability in shape, intensity, and texture.

We have evaluated our interactive labeling pipeline on 100 FAF images and compared the interactive labeling performance with the Random Walker [57], a state-of-the-art interactive labeling algorithm. The Random Walker algorithm was executed

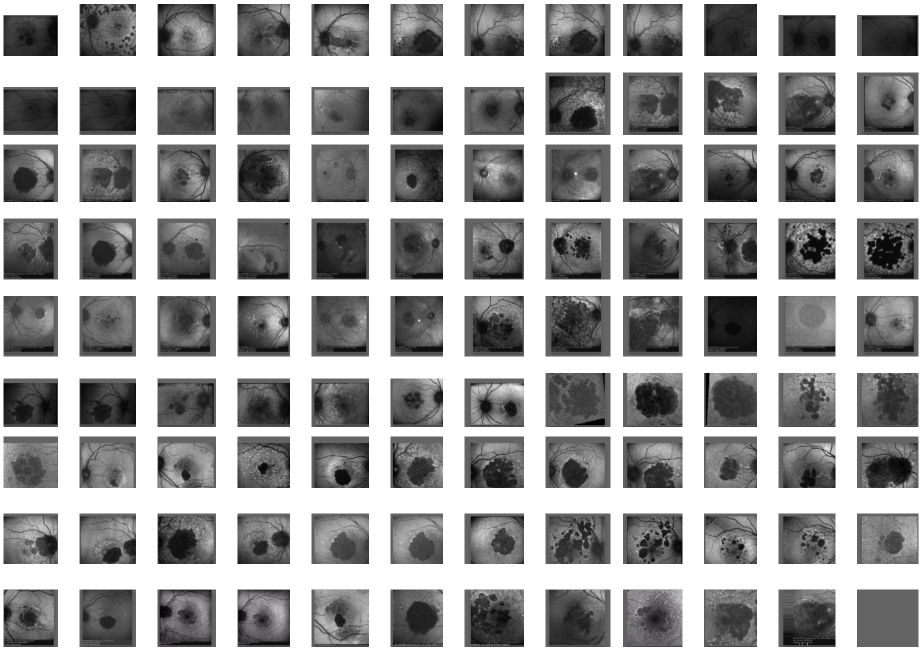


Figure 3.4: A diverse collection of geographic atrophy cases comprising high variability in intensity, shape, and texture.

with the default parameter value of $\beta = 90$ for all test cases. Implementation was obtained from available source code referenced in [57]. Both algorithms used the same seed labels that the human provided in an iterative manner. The interactive labeling pipeline was run with a Sobel kernel size of 5x5. No other preprocessing was performed on the image data. The labeling process was stopped as soon as one of the algorithm produced a high-quality labeling of the desired object boundaries. Receiver operating characteristic (ROC) analysis was performed on a pixel-by-pixel basis with respect to human expert manual gradings and compared for both algorithms. Quantitative evaluation experiments on 100 FAF images show a mean sensitivity/specificity of 98.3/97.7% for our interactive labeling pipeline and a mean sensitivity/specificity of 88.2/96.6% for the Random Walker algorithm.

Figure 3.5 shows the ROC curve where the red curve belongs to the interactive labeling approach and the green curve to the labeling performance of the Random

Walker algorithm.

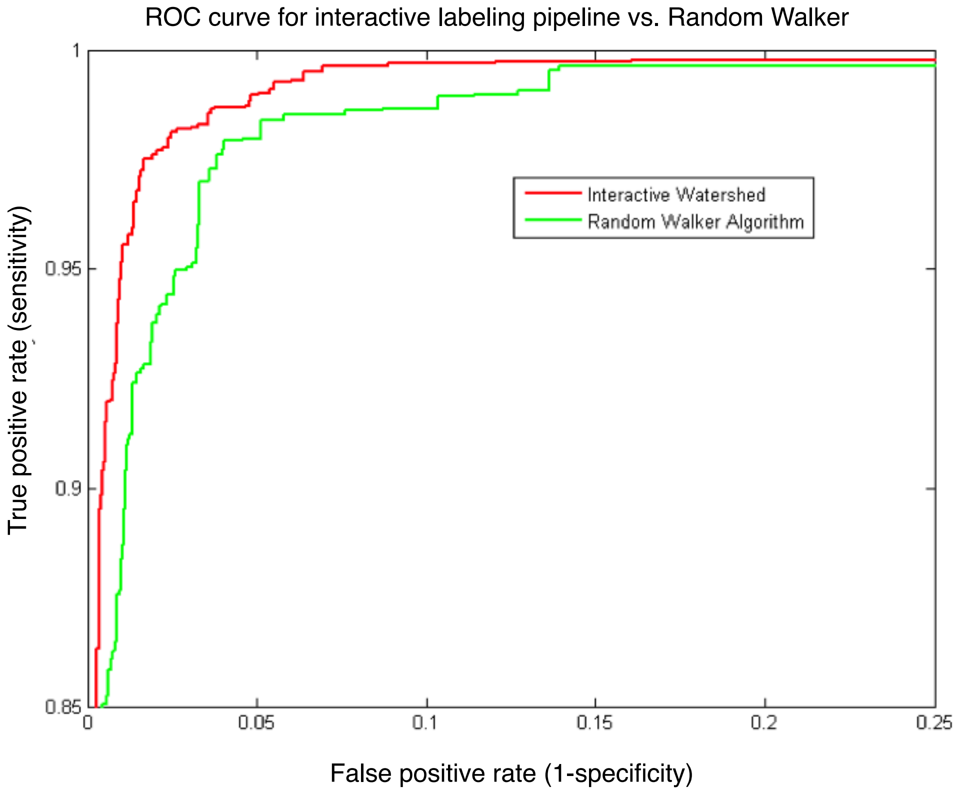


Figure 3.5: ROC curve for the two interactive labeling approaches. The upper left curve (red) shows the performance of our interactive labeling pipeline employing the watershed transform whereas the lower right curve (green) shows the Random Walker performance.

Table 3.1 shows the quantitative performance results of the ROC validation study. We report mean and standard deviations of the sensitivity and specificity.

Fig. 3.6 shows qualitative labeling results of GA boundaries obtained with our interactive labeling pipeline.

ROC Statistic	Watershed Transform	Random Walker
$\mu_{sensitivity}$	98.3%	88.2%
$\mu_{specificity}$	97.7%	96.6%
$\sigma_{sensitivity}$	2.3%	10.8%
$\sigma_{specificity}$	2.1%	8%

Table 3.1: Comparison of receiver operating characteristics (ROC) against the random walker algorithm.

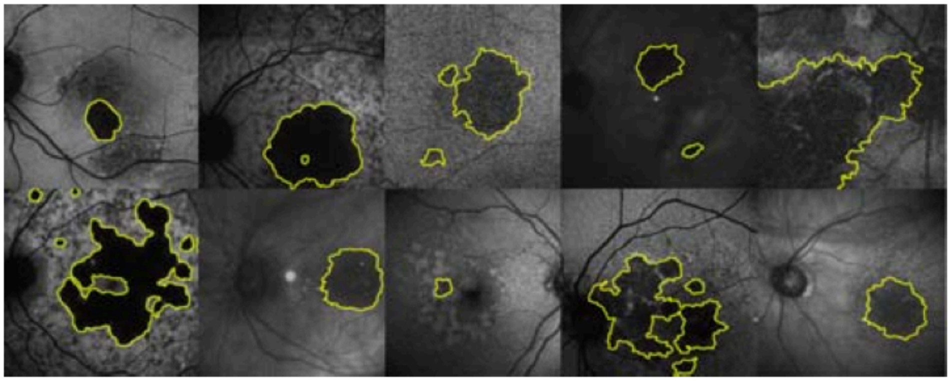


Figure 3.6: Qualitative interactive labeling performance for geographic atrophy (GA) quantification. The yellow contour shows the computed object boundaries by our interactive labeling pipeline. Examples showing different labeling complexity are shown.

3.6 Discussion

The proposed interactive labeling approach is well suited for the task of GA labeling and quantification. The intuitive interface and interaction with the image data was well perceived by human clinical experts and involved short learning curves to get familiar with the graphical user interface and interactive labeling workflow. After short training times high-quality object label boundaries could be obtained for a variety of different object appearances. The proposed interactive labeling pipeline outperformed the Random Walker (RW) algorithm in terms of the sensitivity and specificity by 10.2%/1.1%. This result is surprising given the simplicity of the labeling pipeline.

During the validation study we observed that the RW algorithm has improved noise resistance compared to the WT algorithm. To account for very noisy image data an optional preprocessing step could be employed before the non-linear gradient approximation step to increase the noise robustness of the interactive labeling pipeline. Initial experiments not reported here using a hybrid combination of total variation norm and the bilateral filter to regularize for noise showed further performance improvements. Images that contained high degree of noise could still be robustly segmented with combined with a noise removal step in our pipeline.

One advantage of the WT algorithm compared to the RW algorithm is the robustness to a noisy seed label placement. As long as the placed seed labels were within the local catchment basins of the watershed transform the same labeling result could be obtained, which leads to more reproducible object boundaries in settings where the images are graded multiple times by a single grader or by multiple graders.

We note that both approaches are designed to perform single object labeling. In cases where multiple objects are present the human needs to place many seed labels in order to obtain high-quality label boundaries, which is time-consuming and error prone. Our interactive labeling pipeline failed in cases where the objects of interest exhibited elongated thin shapes (e.g. vessels). In such cases user intervention required more iterations until the desired labeling accuracy was achieved.

Nevertheless, the developed interactive object labeling pipeline can be used to quickly generate ground truth label information for the creation of large annotated image databases. The interaction pipeline is intuitive and commensurate with human capabilities and constraints.

3.7 Conclusion

We have presented a simple and intuitive interactive labeling approach for object boundary labeling using the watershed transform. We demonstrated our approach for the task of quantifying geographic atrophy a wet form of age-related macular degeneration. We validated our approach with quantitative comparison to the Random Walker algorithm using ROC statistics. Our approach has potential to perform well for other retinal disorders and application areas for generic object labeling. The interactive labeling procedure iterates to the desired labeling result that is in conformance to the perception and knowledge of the human. Furthermore, since only approximate label information is required the labeling process across humans is more coherent, reproducible, and time-efficient when compared to manual labeling. Especially in pathological cases, where medical expert knowledge is crucial to distinguish ambiguous region boundaries this approach directly integrates expert a priori information, which would be difficult to robustly model mathematically. Future research is intended towards the integration of interaction tools that allow the labeling of multiple spot-like GA manifestations.

Chapter 4

Learning object and multi-object labeling with minimal human intervention

4.1 Abstract

In this chapter, we present analytics for human-assisted automated object and multi-object labeling with minimal human intervention.

Within this realm, we propose extensions to the naive Bayes algorithm within a transductive learning and inference paradigm. We introduce a semi-parametric form of the transductive naive Bayes algorithm in combination with a Markov random field model. Our extensions impose a multidimensional mixture assumption on each covariate feature dimension to explain more complex distributions. Thus, the complexity of the model grows with the data dimensionality. We combine this model with Markov random fields to impose spatial regularization constraints.

We demonstrate our algorithm within the field of diagnostic ophthalmology and neuro-oncology. In numerous experiments we report on the automated labeling performance of the algorithm in unimodal, multimodal, hyperspectral, and spatio-temporal

data comprising volumetric and planar images.

The developed analytics have practical value by leveraging the interplay of synergistic human-intelligence, enabling humans to better reason about, perceive, and understand heterogeneous image data.

4.2 Introduction

The lack of labeled medical images is a major obstacle for devising data-driven analytics for automated interactive object labeling. In the biomedical imaging domain labeling objects in images requires human expert knowledge and often time-consuming editing to obtain accurate label information. The lack of label information is also known as the scarce label problem.

To cope with the scarce label problem, transductive learning (TL) and semi-supervised learning (SSL) offer a workaround by either simplifying the inference problem or by leveraging unlabeled and labeled data to perform label inferences. TL and SSL are suitable learning paradigms for designing interactive labeling tools that require minimal human expert intervention.

The inductive learning formulation considers a function $f : X \rightarrow Y$ that maps instances from the entire input space $\mathbf{x} \in X$ to output labels $y = f(\mathbf{x})$. In inductive function learning we seek to form a hypothesis that can recover f given a training set of example pairs $\{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$. However, inductive learning and inference assumes the availability of sufficient training data, which clearly in the medical domain is expensive to obtain. It is realistic to assume that the training set at learning stage is incomplete or insufficient to guarantee reliable generalization performance of the learning algorithm.

An alternative learning formulation is the transductive learning and inference setting motivated by Vapnik in the 90's. In transductive inference the learning algorithm is given a labeled training set and an unlabeled test set with the goal to learn a function f that only needs to learn the mapping to the given test set. Consequently, the transductive learning algorithm can explore the labeled training and unlabeled test set. The usefulness of the unlabeled test set in transduction has also been advocated in the context of co-training and SSL.

In this chapter, we present analytics for human-assisted automated object and multi-object labeling with minimal human intervention. Within this realm, we propose extensions to the naive Bayes algorithm within a transductive learning and inference paradigm. We introduce a semi-parametric form of the transductive naive Bayes algorithm in combination with a Markov random field model. We demonstrate our algorithm within the field of diagnostic ophthalmology and neuro-oncology. The algorithm is applicable to different data sources and application domains. In numerous experiments we report the automated labeling performance in unimodal, multimodal, hyperspectral, and spatio-temporal data comprising volumetric and planar images.

We impose unconditional and conditional Gaussian mixture models on each covariate feature dimension to learn and infer the relationships between the input and the output space using naive Bayesian transduction. The naive conditional independence assumption allows efficient inference of marginal and conditional distributions for large-scale learning and inference. The transductive generative formalism allows us to provide 1) predictive confidence of the classification and 2) performance guarantees of the inference. In a probabilistic formulation and using the framework of graphical models we consider a bounded probability measure $\mathbb{P}_{XY}$ describing the joint distribution of the given input and output label space. The generative graphical model formalism provides a unifying framework for capturing complex dependencies between random variables and allows the design of large-scale multivariate statistical models to account for uncertainty and missing data. The objective is to minimize the conditional expected error rate of a classification rule through the conditional $p(y|x)$ with hypothesis $\mathcal{H}_t$ given the observed training sample and the test set. The posterior provides the basis for building $\mathcal{H}_t$ to recover f. Since the goal is to obtain label information only for the test set we allow the posterior distribution to depend on the test set with spatial regularization constraints to exploit the smoothness- and cluster assumption between $p(x)$ and $p(y|x)$. Our approach has the following advantages: i) the classification result supports a reject option and confidence bounds for risk-sensitive applications, ii) has the ability to handle class imbalance through scaled likelihoods, and iii) the conditional independence assumptions allow separate model learning and model combination and sample complexity reduction.

4.3 Prior art

4.3.1 The naive Bayes model

Here we review the naive Bayes model for learning and inference in discrete and continuous data domains. The naive Bayes model is a probabilistic model and makes use of the Bayes' theorem. The model can be efficiently trained within a supervised learning setting given the independence assumption of each covariate feature dimension. Instead of learning a complete covariance matrix the naive Bayes model only requires the learning of individual variances. The independence constraint reduces the parameter space, which in turn minimizes the amount of training data. While the naive independence assumption is not realistic the naive Bayes model performs remarkably well in many real-world problem domains.

Let $p(\mathbf{y}, \mathbf{X})$ denote the joint distribution of the class labels and the input samples, where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ and $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^d$. Each $\mathbf{x}_i$ is a d-dimensional feature vector. The naive conditional independence assumption allows us to factorize the joint distribution as a product of class prior $p(\mathbf{y})$ and independent conditional probability distributions $\prod_{j=1}^d p(\mathbf{x}_j|\mathbf{y})$. In graphical model notation the naive Bayes model has for each $\mathbf{x}_j$ node the parent node $\mathbf{y}$, where j indexes the covariate feature dimension. For the discrete case we assume each $\mathbf{x}_j$ to be sampled from a multinomial probability model of the form

$$p(\mathbf{x}_j) = \prod_{m=1}^{M} \theta_m^{\delta(\mathbf{x}_j = m)}, \text{with} \sum_m \theta_m = 1, \text{ and } \theta_m \geq 0, \tag{4.3.1}$$

where δ is an indicator function and $m = 1, .., M$ indexes the discrete states of the multinomial. Consider the discrete multinomial case by letting $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_{j-1}, \mathbf{x}_j, \mathbf{x}_{j+1}, ..., \mathbf{x}_d)$ be the individual feature vectors and each $\mathbf{x}_j$ a multinomial random variable with components $\mathbf{x}_j^m$. The joint distribution $p(\mathbf{y}, \mathbf{X})$ factorizes into

$$p(\mathbf{x}, \mathbf{y}|\Theta) = p(\mathbf{y}|\pi) \prod_{j=1}^{d} p(\mathbf{x}_j|\mathbf{y}, \theta_j) \text{ with} \tag{4.3.2}$$

$$\Theta = (\pi, \theta_1, \theta, 2, ..., \theta_d). \tag{4.3.3}$$

The class-conditional density $p(\mathbf{x}|\mathbf{y})$ takes the form

$$p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d|y_k = 1, \theta) = \prod_j^d \prod_m \theta_{kjm}^{\mathbf{x}_j^m}, \tag{4.3.4}$$

with $\theta_{kjm}^{\mathbf{x}_j^m} = p(\mathbf{x}_j^m = 1|y_k = 1, \theta)$ being the probability that the j-th feature takes on its m-th value for the k-th class label. Taking the log-likelihood over a labeled training set $\mathcal{D}_l = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ we obtain the following objective function subject to nonnegative constraints, which is solvable by forming the Lagrangian and maximizing over Θ

$$\max_{\Theta} \mathcal{L}(\Theta|\mathcal{D}_l) = \sum_{i=1}^{n} \log p(y_i|\pi) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log p(\mathbf{x}_{ji}|y_i, \theta_j) \tag{4.3.5}$$

$$\text{subject to } \sum_m \theta_{kjm} = 1.$$

The class-conditional probability for each $\mathbf{x}_j$ for the continuous case takes the form of a Gaussian

$$p(\mathbf{x}_j) \sim \mathcal{N}(\mu_j|\sigma_j), \quad \text{with} \tag{4.3.6}$$

$$\mu_{jk} = E[\mathbf{x}_j|y_k] \quad \text{and} \tag{4.3.7}$$

$$\sigma_{jk}^2 = E[(\mathbf{x}_j - \mu_{jk})^2|y_k], \tag{4.3.8}$$

where μ_{jk} denotes the class-conditional mean and σ_{jk}^2 the class-conditional variance. The class-conditional densities for $\mathbf{x}_j$ are

$$p(\mathbf{x}_j|y = k, \theta_j) = \frac{1}{\left(2\pi\sigma_j^2\right)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma_j^2}(\mathbf{x}_j - \mu_{kj})^2\right\}, \tag{4.3.9}$$

with $\mu_k = (\mu_{k1}, \mu_{k2}, ..., \mu_{kd})^T$ and $\sum = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$. The joint probability with conditional independent covariates and a Gaussian class-conditional likelihood factorizes into

$$p(\mathbf{x}, \mathbf{y}|\Theta) = p(\mathbf{y}|\pi) \prod_{j=1}^{d} p(\mathbf{x}_j|\mathbf{y}, \theta_j), \tag{4.3.10}$$

with $\Theta = (\pi, \theta_1, \theta_2, ..., \theta_d)$. Similar to the discrete case the log-likelihood over the input data $\mathcal{D}_l$ can be obtained as in Equation 4.3.5

$$\max_{\Theta} \mathcal{L}(\Theta|\mathcal{D}_l) = \sum_{i=1}^{n} \log p(y_i|\pi) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log p(\mathbf{x}_{ji}|y_i, \theta_j) \tag{4.3.11}$$

$$\text{subject to } \sum_m \theta_{kjm} = 1.$$

4.3.2 The transductive naive Bayes model

The transductive naive Bayes classifier [85] was introduced for the application of text classification. The classifier uses both the training documents and the distribution of the test documents to learn a function f that maps the input space to the output space. This model is similar to the discrete naive Bayes model outlined in Section 4.3.1 with the extension to perform transductive inference. The algorithm classifies the test documents using a two-step procedure. First, [85] learns a multinomial naive Bayes model from the labeled training documents to predict the label distribution on the test set. Second, the test set labels are then used to train a new classifier to predict the final class distribution. This two-step procedure iterates until the distribution of the test set labels converges to a fixed test label distribution.

As opposed to the discrete multinomial, the continuous single Gaussian, or the transductive discrete naive Bayes model we allow $p(X)$ to be continuous and non-uniformly distributed with a multimodal cluster and smoothness assumption. In real-world applications often times the single Gaussian assumption is too limited to fully explain the complexity of $p(X)$. In non-negative data domains the uniform Gaussian assumption may produce incorrect model behavior near zero due to model symmetry or insufficient descriptive power. Previously outlined multinomial probability models in Section 4.3.1 are often employed in text classification [86] and assume discrete data domains with a fixed set of values. In the case of multimodal image data it is impractical to learn a discrete multinomial model as each image value would require a separate model parameter. Given those limitations we extend the transductive naive Bayes model to account for continuous and multimodal image data.

4.4 Methods

4.4.1 Preliminaries

Consider a dataset $\mathcal{D} = [\mathcal{D}_L, \mathcal{D}_U]$, where $\mathcal{D}_L = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{l}$ denotes the labeled training set and $\mathcal{D}_U = \{(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}_{i=l+1}^{l+u}$ the unlabeled test set with $\hat{\mathbf{y}}_i$ unknown. The usual case is $l << u$, i.e., labeled data is scarce and unlabeled data is abundant. Our goal is to learn a function $f : X \rightarrow Y$, where $X \in \mathbb{R}^d$ and $Y = [1, ..., K]$ by taking into account the training and unlabeled test data to find $\hat{\mathbf{y}}_i$.

4.4.2 A semi-parametric model for naive Bayesian transduction

We propose a novel transductive learning algorithm for generative classification casted as an interactive labeling problem with minimal expert intervention. In particular we present a conditional mixture naive Bayes model (TCMNB) with spatial regularization constraints in a transductive learning and inference setting. Compared to [85] and [86] our model assumes for the class-conditional likelihood a semi non-parametric Gaussian mixture model on each covariate feature dimension allowing us to represent and describe more complex distributions. To simplify the estimation we reduce the parameter space by assuming naive conditional independence between the feature space and the class label. The naive conditional independence assumption allows efficient inference of marginal and conditional distributions [87] suitable for large-scale learning and inference. The posterior is formed by learning class-conditional mixture models and priors for each class in each covariate feature dimension exploiting labeled and unlabeled data. Another extension is that we allow the posterior distribution to depend on the unlabeled test set with spatial regularization constraints to exploit the smoothness- and cluster assumption between $p(\mathbf{X})$ and $p(\mathbf{y}|\mathbf{X})$.

Our modeling problem consists of two latent variables one for the conditional

$p(\mathbf{y}|\mathbf{X})$ and the other for approximating the marginal $p(\mathbf{X})$. For the latter, we build an unconditional mixture density on each $p(\mathbf{x}_j)$ to account for multimodal densities by considering a sub-probability model $f_c(\mathbf{x}_j|\theta_c)$ for each component $c = [1, ..., C]$

$$p(\mathbf{x}_j|\Theta = \{\alpha_c, \theta_c\}) = \sum_{c=1}^{C} p(\mathbf{x}_j, z^c = 1|\Theta) \tag{4.4.1}$$

$$= \sum_{c=1}^{C} \alpha_c f_c(\mathbf{x}_j|\theta_c) \tag{4.4.2}$$

$$\text{subject to } \sum_{c=1}^{C} \alpha_c = 1, \text{ and } \alpha_c \geq 0. \tag{4.4.3}$$

Here $\Theta = (\alpha_c, \theta_c)_{c=1}^{C}$ denotes the parameter space of the complete mixture model, $f_c(\mathbf{x}_j|\theta_c)$ denotes the *mixture components* obtained by marginalizing and conditioning over a latent or hidden multinomial variable z having $c = [1, ..., C]$ values. The non-negativity constraints α_c are the *mixing proportions* and θ_c the model parameters for the sub-probability models. In generative graphical model language the latent variable z forms the parent node over the data node $\mathbf{x}_j$ with the arrow pointing from parent to child. This form of graphical model corresponds to the problem of density estimation.

For the former, we consider a probabilistic model $p(\mathbf{y}|\mathbf{X}, \Theta)$ to infer the class labels $\mathbf{y}$ using a conditional mixture model on the data. The Bayes rule inverts the graphical model of the density estimation problem such that the observed data is the parent with the arrow pointing to the latent variable $\mathbf{y}$. Conditioning on $p(\mathbf{x}_j)$ the conditional probability of the latent class variable $\mathbf{y}$ is

$$p(\mathbf{y}|\mathbf{X}, \Theta = \{\alpha_{k,c}, \theta_{k,c}\}_{c=1}^{C}) = p(\mathbf{y}|\pi) \sum_{j} p(\mathbf{x}_j|\mathbf{y} = k, \Theta), \text{ where} \tag{4.4.4}$$

$$= p(\mathbf{y}|\pi) \sum_{j} \sum_{c=1}^{C} \alpha_{k,c} f_{k,c}(\mathbf{x}_j|\theta_{k,c}), \tag{4.4.5}$$

where $p(\mathbf{y}|\pi)$ is the class piror, and $p(\mathbf{x}_j|\mathbf{y}=k,\Theta)$ the covariate conditional mixture model.

4.4.3 Transductive learning and inference

The conditional model is learned using training data $\mathcal{D}$. Given $\mathcal{D}_L$ we learn the class-conditional and unconditional mixture densities of each class by maximizing the log-likelihood of $p(\mathbf{X}|\mathbf{y})$ and $p(\mathbf{y})$. To learn the marginal $p(\mathbf{X})$ for a given class label we assume $p(\mathbf{X})$ to be distributed as a Gaussian mixture on each covariate feature dimension. To approximate both latent variables $\mathbf{z}$ and $\mathbf{y}$ we build the following maximum a posteriori (MAP) model on $\mathcal{D}_L$ and $\mathcal{D}_U$

$$\mathcal{L}(\mathcal{D}_L|\Theta_L) = \prod_{i=1}^{l} p(\mathbf{y}_i|\pi_L) \sum_{j=1}^{d} p(\mathbf{x}_{ji}|\mathbf{y}_i,\theta_L) \tag{4.4.6}$$

$$\mathcal{L}(\mathcal{D}_U|\Theta_U) = \prod_{i=l+1}^{l+u} p(\mathbf{y}_i|\pi_U) \sum_{j=1}^{d} p(\mathbf{x}_{ji}|\mathbf{y}_i,\theta_U). \tag{4.4.7}$$

The MAP estimates of Θ_L and Θ_U for $\mathcal{D} = [\mathcal{D}_L, \mathcal{D}_U]$ have no closed form solution. Maximizing the log-likelihood of Equation 4.4.6- 4.4.7 gives

$$\max_{\hat{\Theta}_L=\{\hat{\pi}_L,\hat{\theta}_L\}} \mathcal{L}(\hat{\Theta}_L|\mathcal{D}_L) = \sum_{i=1}^{l} \log p(\mathbf{y}_i|\pi_L) + \sum_{i=1}^{l} \log \sum_{j=1}^{d} p(\mathbf{x}_{ji}|\mathbf{y}_i,\theta_L) \tag{4.4.8}$$

$$\max_{\hat{\Theta}_U=\{\hat{\pi}_U,\hat{\theta}_U\}} \mathcal{L}(\hat{\Theta}_U|\mathcal{D}_U) = \sum_{i=l+1}^{l+u} \log p(\mathbf{y}_i|\hat{\pi}_L) + \sum_{i=l+1}^{l+u} \log \sum_{j=1}^{d} p(\mathbf{x}_{ji}|\mathbf{y}_i,\hat{\theta}_L), \tag{4.4.9}$$

where we can independently solve for the prior and likelihoods terms. The log-sum terms of above log-likelihood in Equation 4.4.8- 4.4.9 are marginal probabilities with respect to each covariate feature dimension and require a non-linear optimization scheme. We use the expectation-maximization (EM) algorithm [88] to learn the model parameters within a maximum likelihood (ML) framework. Lower bounding the log-sum term with an auxiliary function $\mathcal{L}(q, \theta)$ we can obtain a local solution by iteratively computing the E-step and M-step equations

$$\text{E-Step: } q^{t+1} = \max_{q} \mathcal{L}(q, \theta^{(i)}) \tag{4.4.10}$$

$$\text{M-Step: } \theta^{(i+1)} = \max_{\theta} \mathcal{L}(q^{(i+1)}, \theta). \tag{4.4.11}$$

A proof that the update Equations in 4.4.10 indeed maximize the log-likelihood can be found in [88]. The maximum likelihood estimate of the prior terms of Equations 4.4.8- 4.4.9 are much simpler. Maximizing the log-likelihood with respect to π_L and π_U the solution to the constraint optimization problem for $\mathcal{D}_L$ and $\mathcal{D}_U$ is

$$\hat{\pi}_L = \max_{\pi_L} \sum_{i=1}^{l} \log p(\mathbf{y}_i|\pi_L) = \sum_{i=1}^{l} \mathbf{y}_i/l \tag{4.4.12}$$

$$\hat{\pi}_U = \max_{\pi_U} \sum_{i=l+1}^{l+u} \log p(\mathbf{y}_i|\pi_U) = \sum_{i=l+1}^{l+u} \mathbf{y}_i/(l+u). \tag{4.4.13}$$

The knowledge of $\mathbf{x}_j$ and $\hat{\Theta}_L$ and $\hat{\Theta}_U$ enables us to obtain the probability of the model given the data for each class label $\mathbf{y}_k$. From this probability we can classify and predict the class distribution to perform labeling using

$$\hat{\mathbf{y}} = \max_{\Theta} p(\mathbf{y} = k|\mathbf{X}) = \frac{p(\mathbf{y})p(\mathbf{X}|\mathbf{y})}{\sum_k^K p(\mathbf{y} = k)p(\mathbf{X}|\mathbf{y} = k)}. \tag{4.4.14}$$

4.4.4 Bayesian transductive random fields

The Markov random field model is an undirected graphical model that consists of a set of nodes and edges. Consider a graph $\mathcal{G} = (V, E)$ with V nodes and E edges. The nodes of $\mathcal{G}$ can represent the pixels or voxels in an n-dimensional image, with a neighborhood system N encoded by the edges of $\mathcal{G}$. The neighborhood system can be a 4-, 8-, 6-, 18-, or 26-connected neighborhood. A Gibbs distribution with respect to N has the following form

$$\begin{aligned} p(Y = y) &= \frac{1}{Z} \exp\left\{-\beta U(y)\right\}, \qquad (4.4.15) \\ Z &= \sum_{y} \exp\left\{U(y)\right\}, \\ U(y) &= \sum_{c \in \mathcal{C}} \psi_c(y_c), \end{aligned}$$

where Z is a normalization factor (i.e., the partition function), β a weighting factor for the Gibbs energy function $U(y)$. The Gibbs energy sums over all clique configurations c, where $\mathcal{C}$ denotes the set of cliques for N. A clique in $\mathcal{G}$ is a subset of nodes where each node is connected to each other. Each clique in the graph has associated a nonnegative potential function $\psi_C(\cdot)$. The potential function maps a clique to a real number based on some homogeneity criterion on the values of each nodes within the clique. A popular choice is the Ising model where $\psi_C(y_c) = -1$ for $y_i = y_j$ and $\psi_C(y_c) = 1$ for $y_i \neq y_j$, where the index $j \in N_i$ is a neighbor of node i.

We present a Markov random field (MRF) model using conditional mixture naive Bayesian assumptions within a transductive learning and inference setting. We allow the posterior distribution to depend on the unlabeled test set $\mathcal{D}_U$ with spatial MRF regularization constraints to exploit the smoothness and cluster assumption between the marginal $p(\mathbf{X})$ and the conditional $p(\mathbf{y}|\mathbf{X})$ to perform labeling.

We model the prior term $p(\mathbf{y})$ with a Markov random field model as described in Equation 4.4.15. Conditional independence assumptions allow for local factorizations of the joint distribution. To maximize our objective function constrained by the Markov random field model we use a deterministic approach. We approximate the MAP estimate with the iterated conditional modes (ICM) algorithm [89], which takes the following form

$$\hat{\mathbf{y}}_i = \max_{\mathbf{y}_i \in [1,\ldots,K]} p(\mathbf{y}_i | \mathbf{y}_{N_i}, \mathbf{X}). \tag{4.4.16}$$

4.5 Experiments and results

4.5.1 Application to multimodal data

We applied our algorithm to the task of interactive brain tumor (edema) labeling and evaluated our method with quantitative comparisons to human expert ground truth labels. We performed experimental evaluation on a real-world multimodal magnetic resonance (MR) brain dataset comprising nine modalities. The dataset had a resolution of 256x256x30 voxels per modality and anisotropic voxel dimensions of 0.4mm x 0.4mm x 5mm. Columbia University Medical Hospital provided the dataset after de-identification of patient information. We performed intensity normalization and multimodal registration to bring the data sources into a common coordinate system. A combination of landmark-based and affine registration was used. The landmarks were manually selected by an expert. We started with the most simplistic configuration of multimodal features by looking at FLAIR and DWI ASSET B voxel intensities. We assumed a bag-of-voxels representation of our multimodal features analogous to the bag-of-words representation often used in natural language processing (NLP) applications. The human expert provided approximate object and background seed label information on the FLAIR volume using a custom-developed interactive labeling environment. The interaction step continued until the desired labeling result and accuracy was achieved. The performance of our labeling technique was evaluated with human expert provided manual ground truths using voxel-by-voxel based receiver operating characteristic (ROC) statistics. We further compared our algorithm with the continuous single Gaussian naive Bayes model outlined in Section 4.3.1.

Figure 4.1 shows the input labels on a single slice with the corresponding classified labels for edema in yellow.

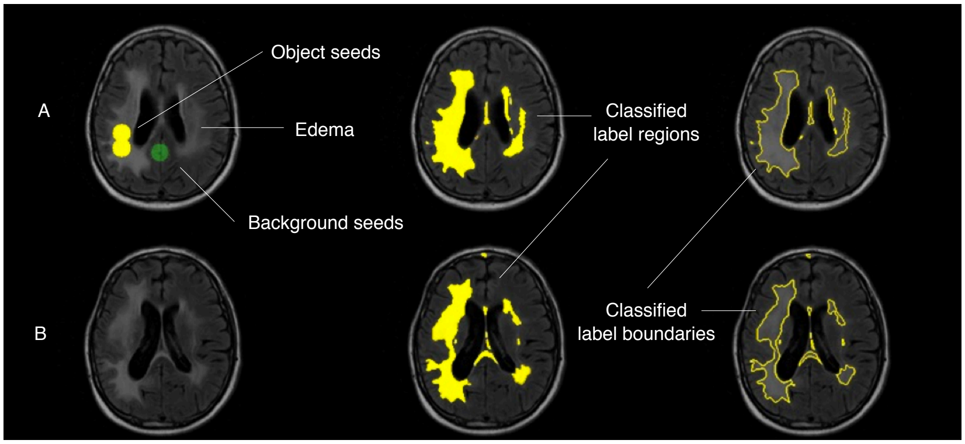


Figure 4.1: Example of our labeling technique on multimodal data. (A) Expert drawn object and background seed labels on single slice (left). Classified label regions (center) and label boundaries (right) produced by our TCMNB algorithm. (B) No labels were provided for other slices in the image volume (left). Classified label regions (center) and label boundaries (right) in yellow for the different slice.

Figure 4.2 shows a comparison of the continuous single Gaussian naive Bayes model (row A) and our TCMNB algorithm (row B).

Figure 4.3 shows the nine modalities registered into a common coordinate system and the classified label boundaries in yellow overlaid on each modality.

Our TCMNB algorithm showed a labeling performance with a sensitivity and specificity of 90.37% and 99.74%. The technique is computationally efficient and takes about 1-3 seconds on a 256x256x30x2 multimodal dataset using a dual core 2.4GHz machine showing that the algorithm is applicable to an interactive labeling environment.

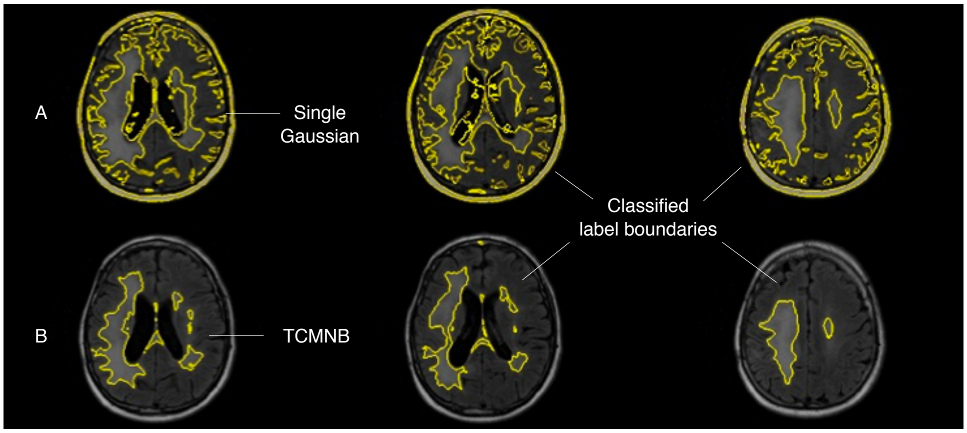


Figure 4.2: Comparison between continuous single Gaussian naive Bayes model and our TCMNB algorithm. Row A shows the classified label boundaries in yellow for three different slices of the FLAIR dataset. Row B shows the classified label boundaries that were produced by the TCMNB algorithm for the corresponding slices in row A.

4.5.2 Application to multimodal spatio-temporal data

We applied our algorithm to the task of interactive multimodal spatio-temporal brain tumor (edema) labeling and evaluated our method with quantitative comparison to human expert provided ground truth labels. Multimodal temporal registration was applied to bring the longitudinal data sources into a common coordinate system. We used the same multimodal feature configuration and preprocessing steps as outlined in Section 4.5.1. The human expert provided approximate object and background seed label information on the FLAIR volume using a custom-developed interactive labeling environment. This time the seed labels were placed on a single time point volume with the goal to predict edema regions on subsequent time points in the multimodal longitudinal dataset. The interaction step continued until the desired labeling result and accuracy was achieved. To apply the TCMNB algorithm to the task of multimodal spatio-temporal labeling we casted the transductive learning and inference setting with respect to the time domain, where the labeled training set was

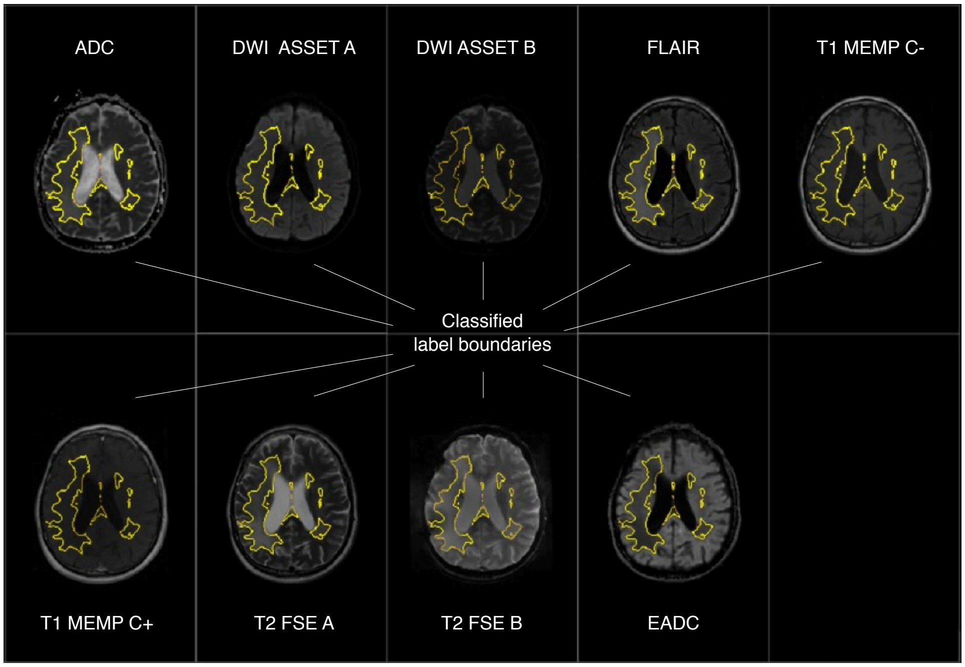


Figure 4.3: Multimodal labeling for nine different modalities using the TCMNB algorithm. Example images of multimodal data sources aligned to a common coordinate frame. The labeling result is overlaid in yellow on all nine image modalities. The DWI ASSET B and FLAIR modality are shown in the 3rd and 4th window from top left to right bottom ordering.

drawn from a single time point volume $\mathcal{X}_t$ and the unlabeled test set from a future time point volume $\mathcal{X}_{t+1}$. The performance of our labeling technique was evaluated with human expert provided manual ground truths using voxel-by-voxel based receiver operating characteristic (ROC) statistics.

Fig. 4.4 shows examples of multimodal spatio-temporal labeling results produced by our TCMNB algorithm. Discarding the yellow label boundaries for now row A shows different image slices of the brain captured on 02/27/2008. Correspondingly row B shows the same image slices at a later time point captured after 70 days on 05/07/2008. The human expert provided object and background seed labels (not shown) at a single time point volume as depicted in row A. Row B shows the predicted label boundaries overlaid in yellow for the time point 05/07/2008. The same label boundaries are transferred to the image volume at time point 02/27/2008 to show edema change. The white arrows in row A show the shrinked edema regions that existed 70 days before.

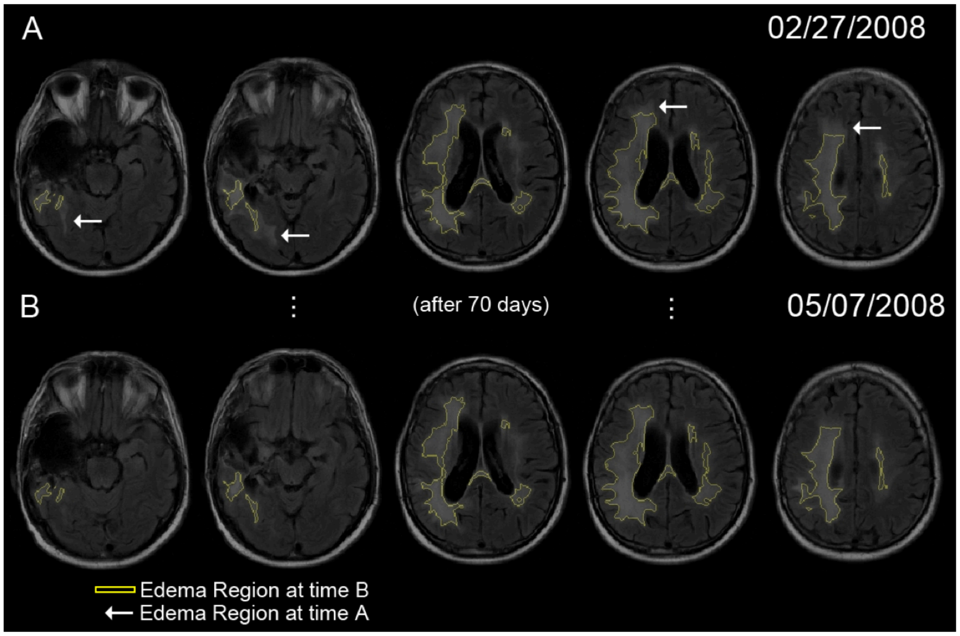


Figure 4.4: Qualitative multimodal spatio-temporal labeling results. A) Predicted label contours for time point B are overlaid on the dataset for time point A to show edema change. The white arrows point out existing edema regions at time point A that disappeared 70 days later in time point B. From left to right different image slices of the brain are shown.

The algorithm showed a labeling performance with a sensitivity and specificity of up to 93.16% and 99.91% respectively. By labeling a single slice at time t accurate predictions for future observations at time $t+1$ could be obtained. We employed our algorithm for the generation of pseudo ground truth data for assessing automated multimodal longitudinal edema prediction. For a detailed account we refer to Caban *et al.* [34].

4.5.3 Application to unimodal data

We have applied our TCMNB labeling algorithm on real-world multimodal ophthalmologic images for the task of multi object labeling with minimal human expert intervention. We evaluated our algorithm on four different modalities including fluorescein angiography (FA), color fundus (CF), infrared (IR), and autofluorescence (AF) images to assess to performance of our algorithm on different labeling tasks. The multimodal images contained retinal disorders such as drusen, geographic atrophy (GA), and choroidal neovascularisation (CNV) with exudates and sub-retinal fibrosis. The motivation for the experiments was to assess the Markov random field model in combination with the TCMNB model within a Bayesian transductive learning and inference framework.

As in Section 4.5.1 and 4.5.2 the expert provided approximate seed label information marking object (red) and background (green) regions for the objects of interest. Images from 16 patients were centered around the macula and cropped to 324x324 regions to reduce image size and improve visualization. The human expert continued the labeling process until the desired labeling accuracy was achieved.

In Fig. 4.5 to 4.8 we show model comparisons for two sets of parameters $\beta = \{0, 1.5\}$ to show the effect on our model with and without the Markov random field constraint. The left column shows the provided seed labels by the human expert. The middle column shows the classified label boundaries in yellow without the MRF constraint $\beta = 0$. The right column shows the classified label boundaries in yellow with the MRF constraint $\beta = 1.5$.

4.6 Discussion

The validation experiments on multimodal brain MR image data showed that the TCMNB algorithm could label volumetric edema regions with a sensitivity and specificity of 90.37% and 99.74%. The algorithm required on average approximately 250

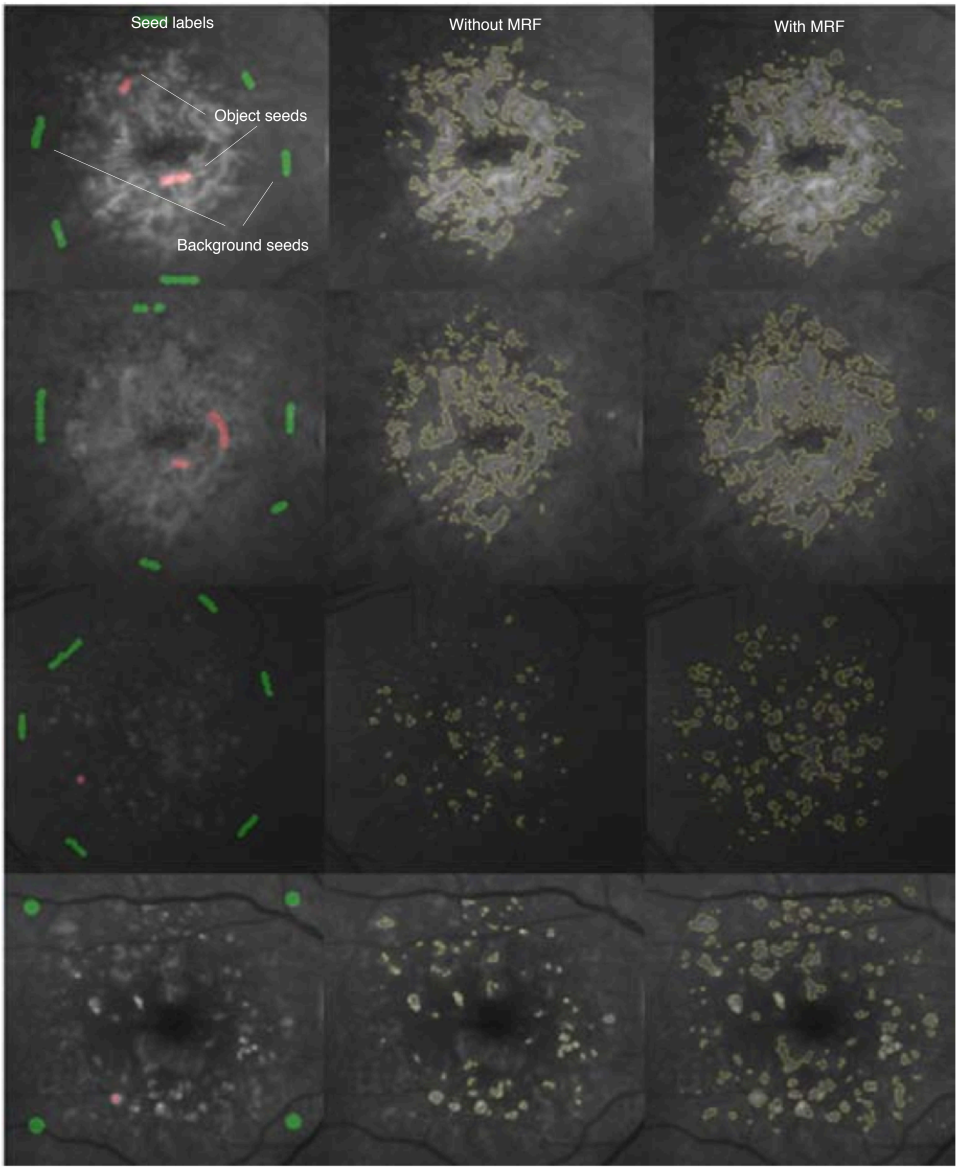


Figure 4.5: Automated multi-object labeling with minimal human intervention on ophthalmology FA image data.

seed labels. When compared to the sample cardinality of the unlabeled test set 0.013% labeled training data was required, which shows that our algorithm is able to accurately label edema regions with minimal human expert intervention. The algorithm is applicable to problems involving single and multi object labeling. Transductive

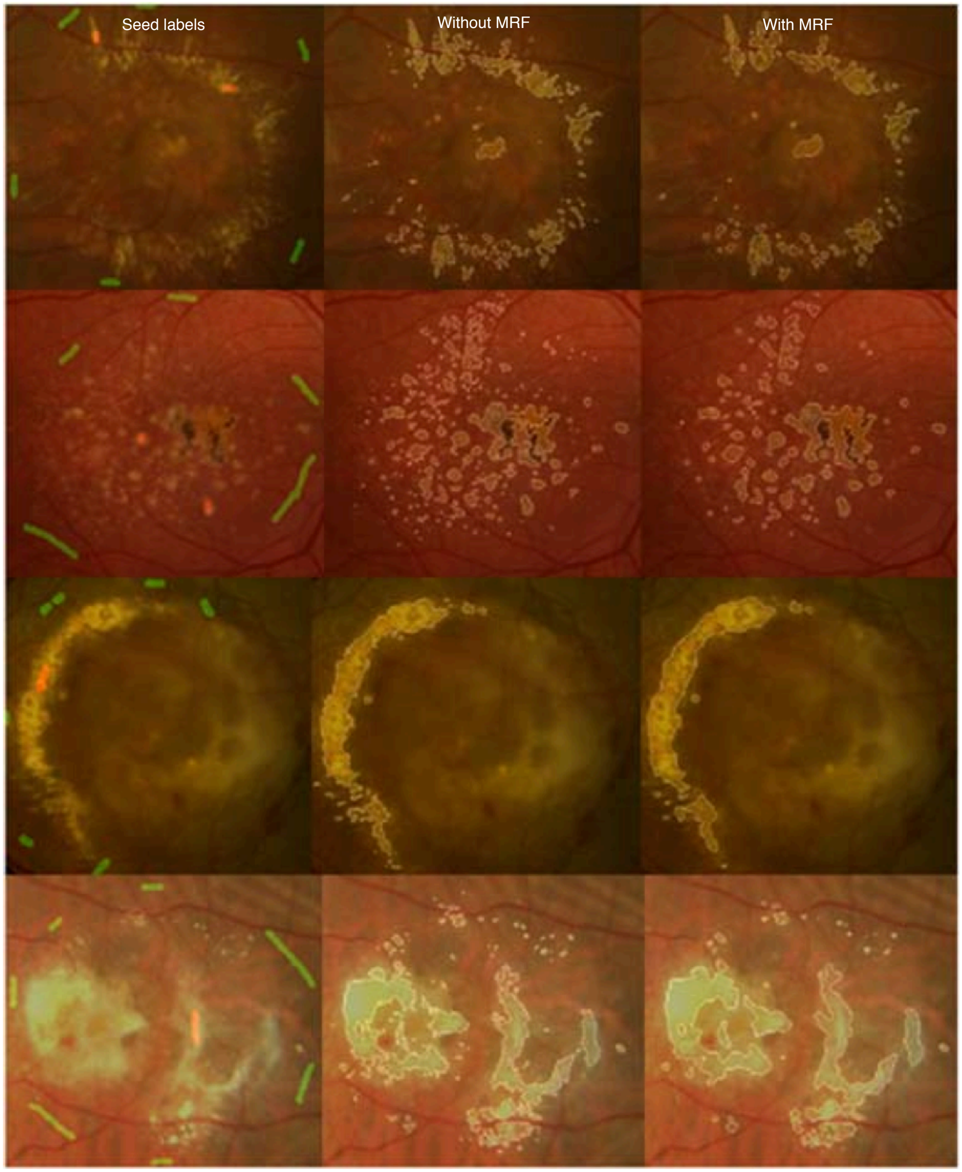


Figure 4.6: Automated multi-object labeling with minimal human intervention on ophthalmology CF image data.

learning and inference is computationally efficient given the naive Bayes assumption. Our algorithm is suitable for large-scale learning and inference as demonstrated in our experiments. Automated labeling of approximately two million data samples took only a couple of seconds. Further the computational efficiency of our algorithm en-

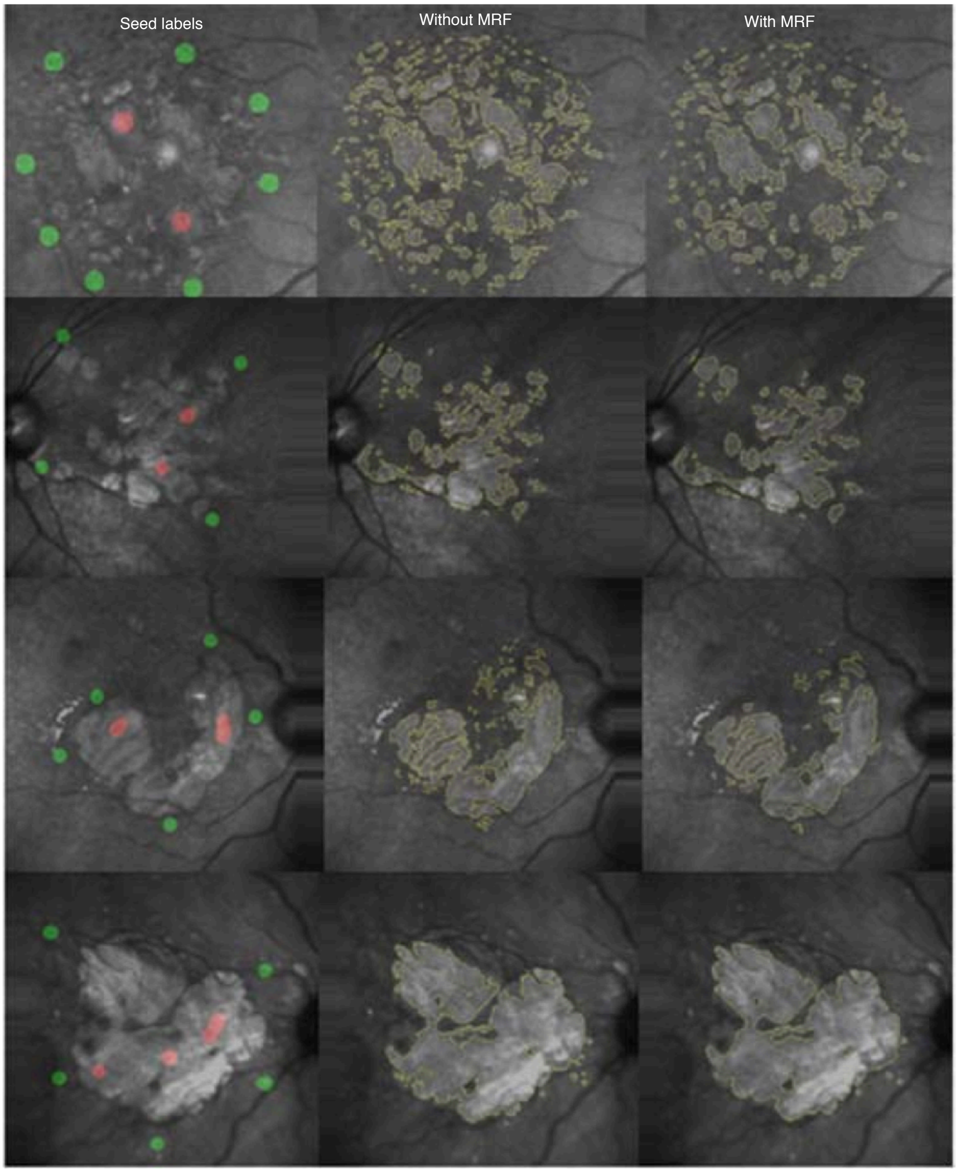


Figure 4.7: Automated multi-object labeling with minimal human intervention on ophthalmology IR image data.

ables the extensions to higher dimensional labeling problems involving temporal data or the incorporation of application-specific constraints as shown in Section 4.5.2 and 4.5.3.

The comparative results shown in Fig. 4.2 demonstrate the superior performance

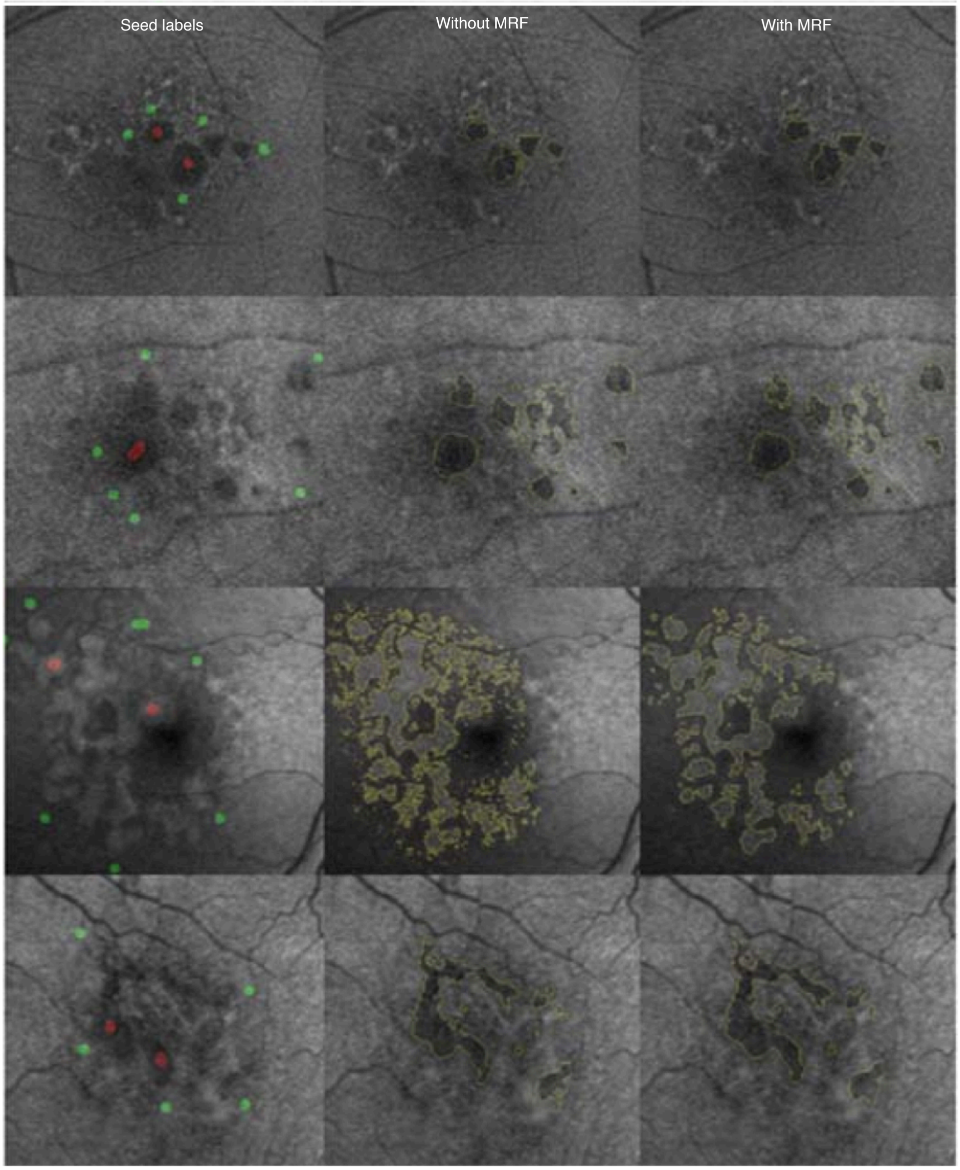


Figure 4.8: Automated multi-object labeling with minimal human intervention on ophthalmology AF image data.

of our TCMNB algorithm in comparison to the continuous single Gaussian naive Bayes model. Clearly the single Gaussian assumption is not able to distinguish between edema regions and the brain skull given their similar intensity ranges. However, the transductive conditional mixture assumption of the TCMNB algorithm can explain more complex multimodal distributions. From the qualitative labeling results we see that false positive labels in the cerebrospinal fluid (CSF) are correctly discarded by the TCMNB algorithm.

Fig. 4.3 shows the automated labeling results on all nine modalities. Combined with an affine and landmark-based registration we are able to transfer the classified label boundaries to the other modalities, which could either be used as the final label boundaries in each respective modality or as initializations to a labeling refinement step to adjust for modality-dependent local variations. From the experiments we observed that while the usage of multimodal data aids the labeling performance the naive approach in using all available image modalities led to inferior label performance. Empirical results suggest that high labeling performance can be achieved when combining FLAIR and DWI ASSET B modalities. By taking a look at the appearance characteristics of each modality one can see why TCMNB can successfully discard the skull regions and why the continuous single Gaussian naive Bayes erroneously detects CSF regions. The CSF regions in the DWI ASSET B modality have similar intensity regions as the edema regions in the FLAIR modality. In the DWI ASSET B modality the skull regions exhibit dark image intensities. The conditional mixture model in combination with the transductive learning and inference scheme can successfully learn discriminative probability distributions of the human provided object and background seed labels. We have to mention that due to the anisotropic voxel dimensions the registration accuracy is a crucial factor and affects the results of our labeling algorithm.

Fig. 4.4 shows the results on multimodal spatio-temporal edema labeling. We can cast the TCMNB model with respect to the time domain to perform spatio-temporal labeling. By comparing the yellow predicted edema boundaries one can observe that the algorithm could successfully predict the edema regions 70 days later. In this demonstrative example we showed that the algorithm can successfully detect edema shrinkage. Only minimal human expert intervention was required to achieve a sensitivity and specificity of up to 93.16% and 99.91% respectively.

In Fig. 4.5 to 4.8 we show automated labeling results of our TCMNB algorithm in combination with a Markov Random field constraint. The spatial regularization constraint acts as a smoothness prior allowing to perform labeling in noisy datasets. In our experiments we considered simple neighborhood relations of 4-connected neighbors. The results show that our algorithm is applicable to a wide range of different labeling tasks. With minimal human expert intervention complex object boundaries can be learned and inferred by the TCMNB model in combination with the Markov random field constraint. This enables the rapid generation of pseudo ground truth labels or the creation of annotated image databases without tedious manual delineation of object boundaries.

By exploiting unlabeled data we can achieve good labeling performance for the task of interactive brain tumor (edema) labeling in anisotropic multimodal medical image volumes–both in cross sectional and temporal labeling tasks. Using a Bayesian transductive learning and inference scheme enables to link human expert provided knowledge with the synthetic knowledge obtained by the probabilistic model to adapt to changing model assumptions over time. This is especially desirable in quantifying unpredictable tumor growth. In general the model with spatial regularization, as expected, provided smoother label outlines and improved the labeling results confirmed by the expert. Bayesian transductive Markov random fields fuses concepts from probabilistic generative learning and spatially constrained Bayesian inference. The combined model allows efficient learning and inference in a semi-supervised setting given only minimal approximate label information.

Failure cases included images where the object boundaries in the image data contained ambiguous appearance information with respect to the learned model. Especially in cases where the object and background had the same appearance led to erroneous label boundaries. For known object shapes one could introduce a shape prior model into the feature space to account for such failure cases. Another alternative is to led the user define shape priors in an online fashion to complete and refine ambiguous label boundaries during the interactive labeling procedure. We also observed that in real data image modalities can be inconsistent and missing, which caused our learned model to fail. Finally, we note that our approach assumes that the object boundaries across modalities follow the same appearance model.

4.7 Conclusion

We have presented a novel interactive labeling approach for single and multi object boundary labeling using a naive Bayesian transductive conditional mixture model in combination with a Markov random field constraint. We demonstrated our approach for the task of quantifying edema regions within multimodal and spatio-temporal brain images and pathologic regions in multimodal retinal images. We validated our approach with quantitative comparison to the continuous single Gaussian native Bayes model using ROC statistics.

Our approach has potential to perform well for other retinal disorders and application areas for generic object labeling. The labeling adaptively iterates to the desired labeling result that is in conformance to the perception and knowledge of the human. Furthermore, since only approximate label information is required the labeling process across humans is more coherent, reproducible, and time efficient when compared to manual labeling. Especially in pathological cases where higher medical expert knowledge is crucial to distinguish similar looking object regions this approach directly integrates expert a priori information, which would be hard or even impossible to robustly model mathematically. The model allows efficient learning and inference in a transductive setting given only minimal approximate label information. This is especially desirable in unpredictable tumor growth and other degenerative diseases.

Future research is intended towards the integration of other interaction constraints to tune the algorithm to specific application contexts.

Chapter 5

Learning discriminative object representations from human labels

5.1 Abstract

In this chapter we present our initial investigations of employing deep learning and inference schemes to automate anatomical labeling of volumetric human brain images.

We present a novel application of deep convolutional networks to autonomously build discriminative object representations from human expert-provided parcellation labels. Different network architectures in combination with context-sensitive feature configurations are studied. A feature consists of a local image patch without further manual engineering effort towards a particular task.

Initial validation experiments show promising results for automatic brain labeling and parcellation. Preliminary results suggest that deep learning and inference schemes can learn complicated object representations that humans find difficult.

The deep learning approach has immense practical value by leveraging human expert intelligence in form of manual provided label regions from which the machine autonomously learns complex discriminative object representations. This form of synergistic human-machine intelligence enables humans to leverage synthetic generated

knowledge in situations where human intuition is limited due to data complexity.

5.2 Introduction

The human brain is a complicated object and is still today poorly understood. Delineation of structural and functional regions (”parcellation”) of the human brain is an important and challenging task for neuroscience and cognitive psychology. Accurate and precise parcellation enables quantification of normal and abnormal changes in the brain as well as analysis of relationships between brain function and structural appearance. Such information is crucial for clinical diagnosis, prediction of treatment outcome in neurodegenerative and pychiatric disorders, and more profoundly for brain-machine interfaces.

However, there still does not exist a widely accepted standard (protocol) for brain image parcellation [90]. The choice of parcellation units is usually dictated by software packages that make use of a labeled atlas brain volume, in which a parcellation protocol has been applied to a single individual. Only recently have large-scale efforts come about to establish and manually apply a standard brain parcellation protocol to many volumetric brain images[1]. However, because manual parcellation is a tedious, time-consuming, and inconsistent endeavor that requires human expertise, many researchers rely on automatic brain parcellation methods. The challenge for both humans and machines is the intrinsic variability of the human brain and its complexity, which makes it extremely difficult to define consistent correspondences across brains.

To establish correspondences, researchers ubiquitously co-register brain images to each other, commonly with a template or labeled atlas brain of the same imaging modality [91]. However, such registration methods typically assume image similarity as a surrogate for anatomical similarity, continuous mapping between corresponding

[1] http://www.braincolor.org/protocols

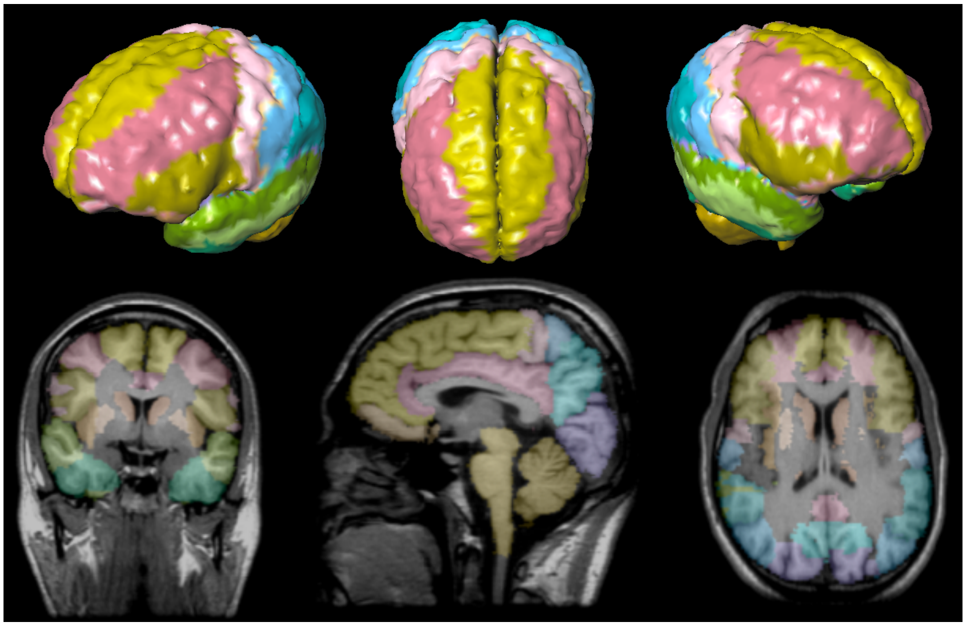


Figure 5.1: Example of a brain parcellation. The top row shows a surface rendering of the different brain regions defined by the parcellation protocol. The bottom row shows three multi-planar reformation views (axial, sagittal, coronal) of the brain with translucent color overlays from human-provided brain parcellation labels.

features, and representativeness of the template or atlas.

On a lower level a main drawback with existing automatic brain parcellation approaches is that they 1) employ algorithms with shallow architectures, 2) are based on heuristic manual feature engineering, and 3) assume the validity of the underlying feature engineered model. In [31] the authors have demonstrated that shallow architectures are limited and non-optimal when learning complex high-dimensional functions. Examples of learning algorithms with shallow architectures are kernel machines or single-layer neural networks. In comparison to deep architectures, shallow learning algorithms are limited in efficiently representing complex function families to learn high-level learning tasks. Many learning algorithms rely on human knowledge to handcraft features, which requires a complete understanding of the problem domain. Such feature engineering approaches limit the generalizability of the model, which may lead to feature redesign and validation, a costly, error prone, and impractical process.

Our research is driven by two main questions. First, can we truly automate brain parcellation in a realistic clinical setting? Second, can artificial cortical network models, which possess deep learning and inference architectures, provide the computational intelligence for this challenging task?

In this chapter we report on a novel application of convolutional networks (CNs) to build discriminative object representations (features) for brain parcellation, which are automatically learned from labels provided by human experts. The idea we would like to pursue is a structured hierarchical approach using context-aware feature learning to perform parcellation without resorting to an atlas or a template-based registration approach. Moreover, our approach does not require the engineering design of handcrafted features, reducing human expert intervention and the need for prior knowledge. The employed analytics model is intuitive to humans, since the architecture consists of visual feature images that the human can assess.

We present initial validation experiments that show promising results for automatic brain labeling and parcellation, suggesting that the proposed approach has potential as an alternative to existing template or atlas-based parcellation approaches.

5.3 Prior art

5.3.1 Brain parcellation approaches

One of the earliest brain parcellation works was proposed by [92]. The authors described a method for building an attributed relational graph (ARG) from T1-weighted MRI to represent the cortical structure of the brain. They proposed a segmentation algorithm based on topology-preserving deformations to segment the gray matter (GM) and the cerebral spinal fluid (CSF). A 3D skeleton was extracted from the union of GM and CSF, which was then used to build the ARG. The nodes of the

ARG consisted of cortex folds (sulci) and edges in the graph encoded pairwise topological connectivity between folds and pairs of gyrus-enclosing folds. The authors used the GM/WM interface instead of the GM/CSF interface for easier segmentation of a topologically correct and smooth surface that served as an input to noise-sensitive 3D skeletonization. The main idea of the proposed method was the notion of discrete object homotopy and homotopic deformations derived from a relaxed topological equivalence class. A homotopic deformation of a discrete binary object is a transformation that preserves the homotopy of the object. The concept of simple points was used to perform homotopic deformations by adding or deleting simple points in an iterative manner. From an initial binary image with known topological properties, the proposed algorithm minimizes a Gibbs energy objective that consisted of a data driven and a regularization term. To segment the GM/CSF interface the following pipeline was employed: i) brain segmentation (binarization, erosions, marker selection, reconstruction, closing, and cavity elimination), and ii) GM/CSF segmentation using homotopic deformations. From the union of GM/CSF, the authors used homotopic thinning based on simple points to obtain a 3D skeleton surface. The surface was further partitioned into simple surfaces, i.e. external brain surface and hemispheric fissure.

The authors in [93] presented a nonparametric generative mixture model for supervised image parcellation. The key idea was to allow global and locally weighted label fusion within a probabilistic maximum a posteriori (MAP) framework. Special cases of the model are i) global, ii) local (voxel-wise independence assumption + uniform prior of training volumes), and iii) semi-local (neighborhood-wise independence assumption + MRF prior) mixture models. Label fusion accounts for multiple training subjects in contrast to the STAPLE algorithm []. The algorithm was validated on 39 (9:training, 30:testing) brain T1-weigthed MR image volumes (256x256x256, 1mm isotropic voxel resolution) involving nine regions of interest, the left and right White Matter (WM), Cerebral Cortex (CT), Lateral Ventricle (LV), Hippocampus

(HP), Thalamus (TH), Caudate (CA), Putamen (PU), Pallidum (PA) and Amygdala (AM).

The authors in [94] proposed a nonparametric hierarchical Bayesian model for functional brain parcellation. Group-level patterns of functional brain images were learned in an unsupervised manner using a two-layer generative model. The first layer consisted of binary activation variables that model functional brain responses induced by visual stimuli. The second layer contained a group-based prior over all binary activation variables modeled as a hierarchical Dirichlet process. Group-specific response patterns as well as the number of such patterns were learned. The authors applied their method on functional brain imaging data of the visual cortex.

An in-depth review on existing registration-based brain parcellation approaches was performed by Klein et al. [95], where they evaluated 14 non-linear deformation algorithms applied to human brain MRI registration. Fourteen algorithms from laboratories around the world were evaluated using 8 different error measures. More than 45,000 registrations between 80 manually labeled brains were performed by algorithms including: AIR, ANIMAL, ART, Diffeomorphic Demons, FNIRT, IRTK, JRD-fluid, ROMEO, SICLE, SyN, and four different SPM5 algorithms (SPM2-type and regular Normalization, Unified Segmentation, and the DARTEL Toolbox). One of the most significant findings of their study is that the relative performances of the registration methods under comparison appear to be little affected by the choice of subject population, and labeling protocol, suggesting that the findings are generalizable to new subject populations that are labeled or evaluated using different labeling protocols. Furthermore, they ranked the 14 methods according to three completely independent analyses (permutation tests, one-way ANOVA tests, and indifference-zone ranking). They further derived three almost identical top rankings of the methods. ART, SyN, IRTK, and SPMs DARTEL Toolbox gave the best results according to overlap and distance measures, with ART and SyN delivering the most consistently high accuracy across subjects and label sets.

In 2010, Klein *et al.* provided a validation study of volume-based and surface-based brain image registration methods [96]. In [96] the authors presented a study that directly compared some state-of-the-art volume registration with surface registration methods. They also compared registrations of whole-head versus brain-only (de-skulled) images. They used permutation tests to compare the overlap or Hausdorff distance for more than 16,000 registrations between 80 manually labeled brain images. The authors primary findings were the following: 1. de-skulling aids volume registration methods; 2. custom-made optimal average templates improve registration over direct pairwise registration; and 3. resampling volume labels on surfaces or converting surface labels to volumes introduces distortions that preclude a fair comparison between the highest ranking volume and surface registration methods using present resampling methods. From the results of this study, they recommend constructing a custom template from a limited sample drawn from the same or a similar representative population, using the same algorithm used for registering brains to the template.

In contrast to existing methods, we would like to advocate a deep learning [31] approach to automate brain image parcellation. We are motivated by models from biologically inspired artificial intelligence, in particular artificial cortical network models such as deep convolutional networks [97] (CNs).

5.4 Methods

5.4.1 Preliminaries

Convolutional networks are extensions to classical multi-layer back-propagation neural networks. CNs involve the use of convolutions and the concept of the multi-layer back-propagation algorithm. First, consider the one dimensional convolution of two continuous functions f and g written $f \star g$

$$\begin{aligned}(f \star g)(t) &= \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau.\end{aligned} \tag{5.4.1}$$

Here t denotes the time domain, τ a free variable, $g(-\tau)$ the reverse of $g(\tau)$ centered at $t = 0$. The $\star$ operator integrates or sums the product of f and g, where either f or g is reversed and shifted along τ. The result is a weighted average of $f(\tau)$ at time t, where the weighting coefficients are the values in $g(t - \tau)$. The discrete case is analogous to the above definition except that the integrals are replaced with summation signs. The two dimensional or n-dimensional case follows the above definition by adding another integral/summation sign to each dimension.

In standard CNs, the architecture often contains a fully connected layer at the end of the network hierarchy. The fully connected layer is equivalent to the layer type found in artificial neural networks (ANNs). Fully connected layers are one dimensional layers in contrast to CN-specific layer types. In ANNs the classical learning algorithm performs error back-propagation through the network to learn the parameters of the model. The parameters of the fully connected layer can be learned by solving the following optimization problem

$$\min_{\theta} \mathcal{O} := \mathcal{L}(\mathcal{X}, \mathbf{y}, \mathcal{H}_\theta) \tag{5.4.2}$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{C} (\mathbf{y}_i^c - \mathcal{H}_\theta(\mathcal{X})_i^c)^2 \text{, with} \tag{5.4.3}$$

$$\mathcal{H}_\theta = \begin{cases} u_p^k = \left(\sum_q I_q^{k-1} \star w_{p,q}^k\right) + b_p^k \\ I_p^k = f(u_p^k) \end{cases} \tag{5.4.4}$$

where $\mathcal{L}$ is a squared error loss, C the number of output label classes, N the number of input samples, $\mathbf{y}$ the output class label vector, $\mathcal{X}$ the input sample, and $\mathcal{H}_\theta$ the

CN architecture and its model parameters. Standard gradient descent can be used to solve Equation 5.4.2

$$\frac{\partial \mathcal{L}}{\partial w^k_{p,q}} = \sum_{i=1}^{N}\sum_{c=1}^{C} \frac{\partial \mathcal{L}}{\partial \left(u^k_p\right)^c_i} \frac{\partial \left(u^k_p\right)^c_i}{w^k_{p,q}} \tag{5.4.5}$$

$$\frac{\partial \mathcal{L}}{\partial b^k_p} = \sum_{i=1}^{N}\sum_{c=1}^{C} \frac{\partial \mathcal{L}}{\partial \left(u^k_p\right)^c_i} \frac{\partial \left(u^k_p\right)^c_i}{b^k_p}, \tag{5.4.6}$$

which leads to

$$\frac{\partial \mathcal{L}}{\partial w^k_{p,q}} = \sum_{i=1}^{N}\sum_{c=1}^{C} \frac{\partial \mathcal{L}}{\partial \left(u^k_p\right)^c_i} \left(I^{k-1}_q\right)^c_i = \left(\delta^k_p I^{k-1}_q\right)^c_i \tag{5.4.7}$$

$$\frac{\partial \mathcal{L}}{\partial b^k_p} = \sum_{i=1}^{N}\sum_{c=1}^{C} \left(\delta^k_p\right)^c_i, \tag{5.4.8}$$

where

$$\delta^{k=L}_p = \frac{\partial \mathcal{L}}{\partial u^L_p} = \frac{\partial \mathcal{L}}{\partial I^L_p}\frac{\partial I^L_p}{\partial u^L_p} = \left(y_p - I^L_p\right) f'(u^L_p) \tag{5.4.9}$$

$$\delta^{k=l}_p = \frac{\partial \mathcal{L}}{\partial u^l_p} = \frac{\partial \mathcal{L}}{\partial I^l_q}\frac{\partial I^l_q}{\partial u^l_p} = \sum_q w^{l+1}_{q,p}\delta^{l+1}_q f'(u^l_p). \tag{5.4.10}$$

The intuition of the last term is that the error δ^l_p at an inner layer l is a function of the feature map errors δ^{l+1}_q in the proceeding layer $l+1$.

5.4.2 Problem formulation

Consider the problem of finding a function $f : X \rightarrow Y$ that maps an input space to an output space. Here X refers to the brain image data and Y to a multi-class label space of a brain parcellation. We are given a dataset $\mathcal{D}$ as a collection of N images $\{\mathcal{I}\} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_N\}$. The dataset is further partitioned into $\mathcal{D} = \{\mathcal{D}_l, \mathcal{D}_u\}$, where

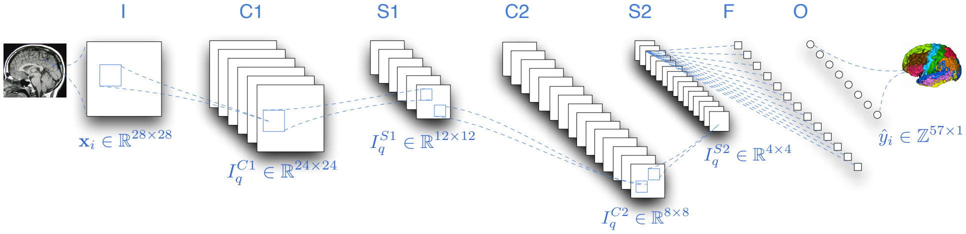

I
C1
S1
C2
S2
F
O
$\mathbf{x}_i \in \mathbb{R}^{28\times 28}$
$I_q^{C1} \in \mathbb{R}^{24\times 24}$
$I_q^{S1} \in \mathbb{R}^{12\times 12}$
$I_q^{C2} \in \mathbb{R}^{8\times 8}$
$I_q^{S2} \in \mathbb{R}^{4\times 4}$
$\hat{y}_i \in \mathbb{Z}^{57\times 1}$

Figure 5.2: Deep learning and inference with a convolutional network model. From left to right, the deep architecture consists of several layers starting with the input layer (I). In an alternating manner the CN consists of a hierarchical architecture of convolutional (C1, C2) and subsampling (S1, S2) layers followed by a full-connection layer (F), and finally the output layer (O).

$\mathcal{D}_l = \{s_i, y_i\}_{i=1}^{l}$ denotes the labeled training set and $\mathcal{D}_u = \{s_i, \hat{y}_i\}_{i=l+1}^{n}$ the unlabeled test set. Each pair consists of an image site s_i (e.g., voxel) and a label y_i, which assumes values in a finite set $y = \{0, ..., C\}$. The index n refers to the number of sites within each image. For each site in the training set we form a d-dimensional patch $\mathbf{x}_i \in \mathbb{R}^d$. A detailed description of $\mathbf{x}_i$ can be found in Section 5.5.1. The input-output pairs in $\mathcal{D}_l$ are drawn in an independent and identically distributed manner from some unknown probability distribution $\mathbb{P}(X, Y)$ defined jointly over X and Y. Our goal is, given $\mathcal{D}_l$, to predict $\hat{y}$ for the unlabeled test set $\mathcal{D}_u$ such that the learned approximation to f has low probability of error $\mathbb{P}(f(X) \neq Y)$.

5.4.3 The convolutional network architecture

Convolutional networks (CN) belong to the class of artificial cortical network models and are an extension to the classical multilayer perceptrons (MLPs) model. CNs consist of a hierarchical multilayer architecture of maps as depicted in Fig. 5.2. The learned model $\theta = \{\mathcal{W}, \mathbf{b}\}$ includes convolutional operators $\mathcal{W}$ and bias terms $\mathbf{b}$, which in combination with a nonlinear activation function γ (e.g. sigmoid or hyperbolic tangent), form so-called "activity feature" maps $\mathbf{I}_p^k$, where $k \in [1, ..., K]$

indexes a CN layer and p a particular feature map of layer k. Each layer can have different numbers of feature maps. CNs enables the sharing of weights between the nodes of the network.

The layers (S1, S2) are simple subsampling layers (also called max-pooling layers) to reduce the computational load of the model and to introduce some degree of scale invariance. The full-connection layer (F) is a hidden layer as in standard MLPs to reduce the dimensionality of the last subsampling layer and to aggregate the information to each output node of the CN. The output nodes of layer (O) represent the individual class labels of the CN. The O-layer compares the forward propagated class labels from the network with manual ground truth labels using an application-dependent loss function.

5.4.4 Deep learning and inference

Given a CN model θ and training data $(\mathcal{X}_i, \mathbf{y}_i)$, the first step is to perform a forward propagation of an input patch through the CN architecture shown in Fig. 5.2. The feature maps in each convolutional layer (C1, C2) are computed through a recursive forward dynamic of the form

$$\mathbf{I}_q^k = \gamma(\mathbf{u}_q^k) \tag{5.4.11}$$

$$\mathbf{u}_q^k = b_q^k + (\sum_p \mathbf{w}_{q,p}^k \otimes \mathbf{I}_p^{k-1}), \tag{5.4.12}$$

where γ denotes a smooth differentiable nonlinearity to ensure differentiability across layers, $\mathbf{u}_q^k$ a pre-activation image, $\mathbf{I}_p^{k-1}$ the feature image at layer $k-1$, $\mathbf{w}_{q,p}^k$ a directed convolution kernel from map p to q, and b_q^k a bias term for layer k.

The errors are then back-propagated through the network to refine and learn the CN model in an iterative fashion. By learning θ we basically learn convolutional filters that represent discriminative object representations of the provided parcellation units.

Learning the model can be achieved by solving the following optimization problem

$$\min_{\theta} \mathcal{O}(\mathcal{X}_i, \mathbf{y}_i, \theta) := \theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta} \mathcal{O}(\mathbf{X}_i, \mathbf{y}_i, \theta). \tag{5.4.13}$$

Equation 5.4.13 can be solved with a recursive error back-propagation scheme within an online learning setting using stochastic gradient descent. We maximize the likelihood model by minimizing the negative log-likelihood

$$\mathbb{P}(Y = c|\mathcal{X}, \theta) = \frac{e^{W_i x + b_i}}{\sum_j e^{W_j x + b_j}} \tag{5.4.14}$$

$$\ell(\theta|\mathcal{X}, \mathbf{y}) = \sum_i = \log(\mathbb{P}(Y = \mathbf{y}_i|\mathcal{X}_i, \theta)) \tag{5.4.15}$$

$$\min_{\theta} \mathcal{O}(\mathbf{X}, \mathbf{y}, \theta) = -\ell(\theta|\mathcal{X}, \mathbf{y}) \tag{5.4.16}$$

$$\tag{5.4.17}$$

Solving the Equation 5.4.14 can be done by following the procedure outlined in Section 5.4.1. In the O-layer the final output label can be inferred by taking the class with the maximum probability given the data.

$$\hat{y} = \max_i \mathbb{P}(Y = i|\mathbf{X}, \theta). \tag{5.4.18}$$

5.5 Experiments and results

5.5.1 Context-aware feature learning

We have built two context-aware feature configurations to examine labeling performance in relation to varying degrees of available contextual information around a particular voxel location. We performed experiments on a single planar slice to assess the labeling performance of the deep convolutional network architecture. Fig. 5.3 shows examples of feature configuration C1 and C2 that were used to assess labeling performance of the CN architecture. The white dot denotes a voxel location and the white rectangle the feature patch. Feature configuration C1 consisted of a single patch, whereas feature configuration C2 comprised a cross configuration of four patches (i.e., north, south, west, east). Both feature configurations were obtained through randomized sampling to build the training and validation set $(\mathbf{x}_i, y_i)$. Linear sampling was used to generate the test set. To perform cortical and subcortical parcellation, we constrained the context area for feature configuration C1 to a 28 x 28 dimensional patch. For C2, the four-element patches had dimensions of 14 x 14 pixels, which were concatenated back to the final 28 x 28 patch dimension as in C1. Fig. 5.4 shows an example of 5000 randomly selected C1 patches from a pool of 50,000 patch samples.

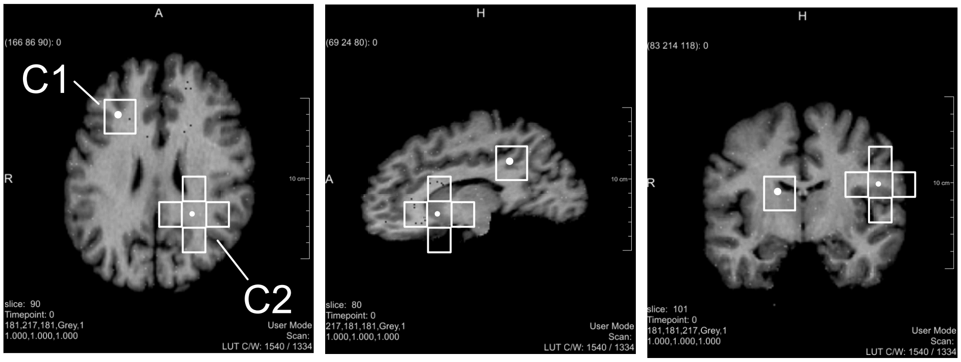


Figure 5.3: Context-aware feature configurations for the deep learning and inference scheme. Shown are the orthogonal slices of subject one from the LBPA40 dataset. The white dot denotes a voxel location and the white rectangle the feature patch. By adjusting the size of the patch surrounding context information can be incorporated.

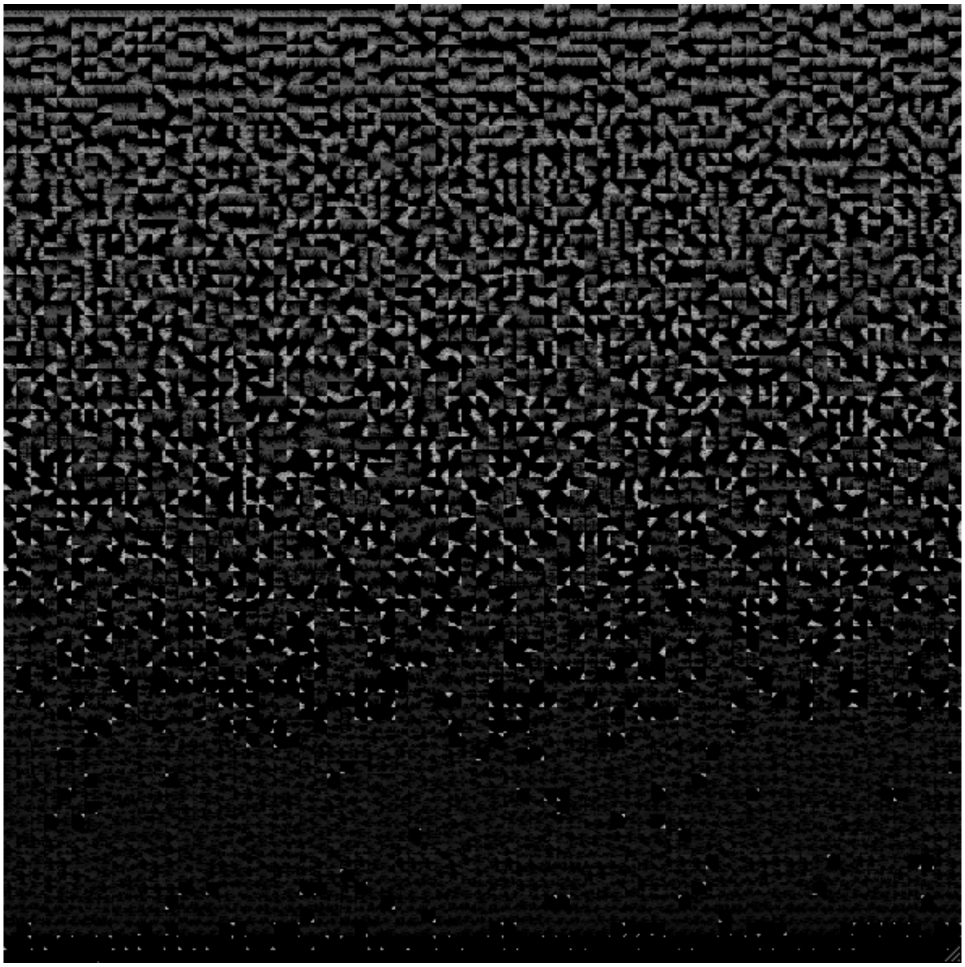


Figure 5.4: 5000 random voxel locations and their associated raw image patches. Each patch captures local information at a certain voxel position in the brain.

5.5.2 Deep learning performance in resource constraint settings

For our experiments we used 40 brain images and their labels (56 structures + background) from the LONI Probabilistic Brain Atlas (LPBA40) at the Laboratory of Neuro Imaging (LONI) at UCLA [?]. We performed two sets of experiments on the LPBA40 dataset to assess the performance of our approach using feature configuration C1 and C2. The motivation of our experiments was to assess whether deep learning and inference can learn complex discriminative object representations from human expert-provided labels.

For both configurations we have used the following settings (learning rate $\eta = 0.1$, batch size = 25, number of training epochs = 50, number of randomized patches = 25000, number of feature maps in each layer $(C1, S1) = 20$, $(C2, S2) = 50$). We split the training and validation set with a ratio of 80:20 from a single central slice of subject 1. The validation set of subject 1 was used to determine the best performing model during online stochastic gradient descent learning. After the best model was obtained test performance was assessed on single central slices of all other remaining LPBA40 subjects. For testing, we performed linear patch sampling in order to learn a site-wise class probability. Invalid patch samples near the border were ignored. Parcellation labels were then obtained by choosing the class label with the highest probability given the patch information and the learned CN model.

For quantitative validation we computed sensitivity, specificity, positive predictive value, negative predictive value, and the Dice coefficient D_c for the overall cortical structure and for individual subcortical structures. The Dice coefficient is defined as $D_c = \frac{2|A \cap B|}{|A|+|B|}$, where D_c measures the set agreement between the ground truth labels and the predicted brain parcels. The D_c score ranges from (0-1), where 1 means perfect agreement.

For experiment C1, the complete cortical structure had a mean D_c of 0.73 ($\pm$ 0.05), whereas for C2, the same structure had a mean D_c of 0.65 ($\pm$ 0.04) over all 39 test subjects.

Fig. 5.5 shows the manual ground truth (LPBA40) labels provided by the human expert for all 40 subjects of the LPBA40 dataset.

Fig. 5.6 shows the computed parcellation results by the convolutional network using feature configuration C1 for all 40 subjects of the LPBA40 dataset.

Table 5.1 and 5.2 show the quantitative labeling results for all 40 subjects in terms of the sensitivity, specificity, positive predictive value, and the negative predictive value.

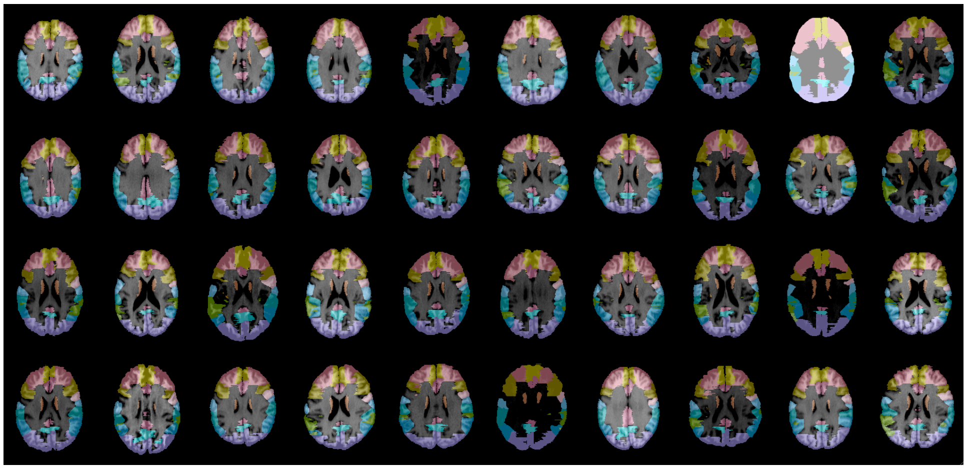


Figure 5.5: Ground truth labels provided by the human expert.

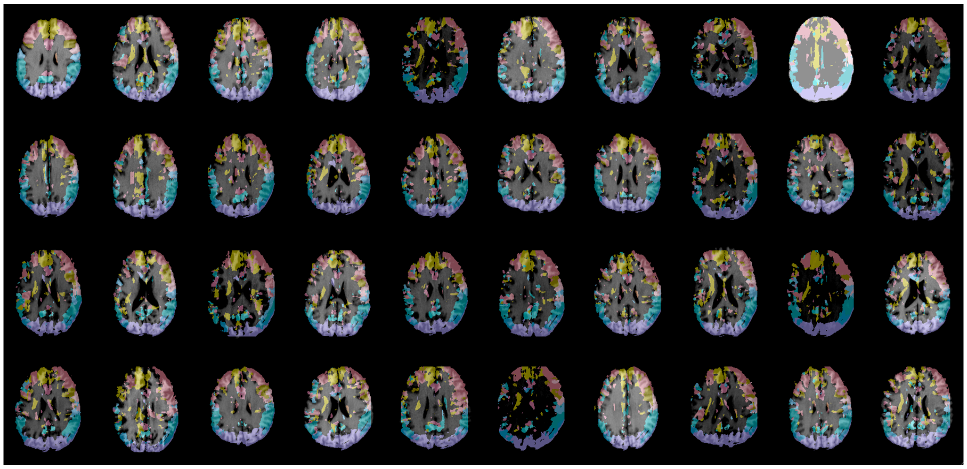


Figure 5.6: Predicted parcellations labels by the CN.

5.6 Discussion

Given the limited training set the overall performance for labeling the complete cortex was surprisingly good in terms of the Dice coefficient and the other four performance measures outlined in Table 5.1 and 5.2. From the qualitative parcellation results in

Subject Id	Sensitivity	Specificity	PPV	NPV
1	0.9850	0.9223	0.9777	0.9466
2	0.9347	0.6076	0.8804	0.7507
3	0.9443	0.7857	0.9350	0.8120
4	0.9419	0.7638	0.9325	0.7913
5	0.9510	0.6666	0.8967	0.8172
6	0.9560	0.6231	0.8706	0.8423
7	0.9571	0.6863	0.9039	0.8384
8	0.9516	0.6357	0.9002	0.7919
9	0.9492	0.6795	0.9006	0.8140
10	0.9447	0.6249	0.8823	0.7916
11	0.9640	0.7040	0.9228	0.8420
12	0.9383	0.7110	0.9094	0.7883
13	0.9439	0.7815	0.9354	0.8060
14	0.9632	0.7420	0.9321	0.8459
15	0.9445	0.7309	0.9236	0.7925
16	0.9485	0.6623	0.9081	0.7854
17	0.9657	0.7060	0.9110	0.8685
18	0.9458	0.6812	0.9010	0.8039
19	0.9445	0.6798	0.9100	0.7811
20	0.9226	0.4621	0.8269	0.6820

Table 5.1: Quantitative labeling results produced by the convolutional network for subjects 1-20. PPV stands for Positive Predictive Value and NPV for Negative Predictive Value.

Fig. 5.6 we can observe that the CN provides better parcellation results for larger regions than for smaller ones. In general the D_c was low for small brain parcels in comparison to larger sub-cortical structures. This can be explained by the primitive training dataset and unbalanced class sample distributions. However, the initial experiments look promising given the limited amount of training data. To improve upon the current results several options are available. We believe that an affine registration in combination with location information into the feature vector will improve the current parcellation performance. Instead of one large CN architecture, training an ensemble of multiple CNs for each parcellation region would further improve the re-

Subject Id	Sensitivity	Specificity	PPV	NPV
21	0.9458	0.6890	0.9092	0.7941
22	0.9433	0.6114	0.9020	0.7399
23	0.9422	0.5331	0.8409	0.7788
24	0.9493	0.6370	0.8849	0.8102
25	0.9585	0.7374	0.9243	0.8417
26	0.9422	0.7305	0.9212	0.7907
27	0.9443	0.6702	0.9083	0.7768
28	0.9333	0.6259	0.8842	0.7539
29	0.9358	0.6779	0.9042	0.7648
30	0.9352	0.6512	0.9056	0.7376
31	0.9539	0.7155	0.9116	0.8346
32	0.9360	0.6207	0.8873	0.7525
33	0.9517	0.7475	0.9356	0.8008
34	0.9304	0.6046	0.8885	0.7195
35	0.9296	0.6842	0.9026	0.7555
36	0.9420	0.6248	0.8924	0.7654
37	0.9612	0.6953	0.9134	0.8429
38	0.9344	0.6135	0.8940	0.7282
39	0.9543	0.7468	0.9291	0.8246
40	0.9306	0.6078	0.8799	0.7392

Table 5.2: Quantitative labeling results produced by the convolutional network for subjects 21-40. PPV stands for Positive Predictive Value and NPV for Negative Predictive Value.

sults. The overall parcellation performance showed high variability and no significant difference between feature configuration C1 and C2. We have used the convolutional network implementation provided by Theano [98].

Initial experiments showed that in cases where image quality is poor, i.e. low image contrast and high noise levels, the learned model failed in producing good labeling results. Furthermore, bringing the multiple brain image volumes into a common affine coordinate system could circumvent failure cases where individual brain volumes had high variability in appearance such as scale and rotation. In general a single CN model failed in producing robust and accurate labelings for small parcellation regions.

5.7 Conclusion

In this chapter we have presented a novel application of biologically inspired cortical network models to automate brain image parcellation using a deep convolutional network architecture. We were able to demonstrate parcellation of the cerebral cortex, without human intervention to build handcrafted features or to provide other prior knowledge. The feature configurations were able to correctly reject the detection of the main white matter regions. We attribute the low parcellation performance and high inter-subject variability to the very limited training set that we used. Another factor that affected the performance was the crude registration of the dataset causing the central slices that were used for training and testing to be misaligned. Misalignment caused by registration errors however can be accounted for by enriching the training set samples from a slab of slices. In future work we plan to improve upon the results obtained by these initial experiments and to extend our current approach to three-dimensional, context-aware feature learning and in-depth validation of the model in a clinical setting.

Part II

Synergistic human-machine intelligence for exploratory event analysis in healthcare data

Chapter 6

Learning human intuitive spatio-temporal event patterns

6.1 Abstract

In this chapter we address the problem of how we can build data-driven analytics for exploratory analysis in longitudinal event data that are commensurate with human capabilities and constraints.

We propose human-intuitive analytics that enable the representation and discovery of interpretable event patterns to ease knowledge absorption and comprehension of the employed analytics model. We develop a novel temporal event matrix representation and learning algorithm to perform large-scale temporal pattern mining of longitudinal heterogeneous event data. The algorithm enables the representation, extraction, and mining of complex latent event relationships within single and multiple event entities to perform exploratory group analysis.

We demonstrate the developed analytics within the healthcare domain by exploring healthcare resource utilization (HRU) in relation to diabetic disease severity. Experimental results demonstrate that we can learn meaningful interpretable patterns.

The developed analytics have practical value by leveraging the interplay of synergistic human-intelligence, enabling medical practitioners to better reason about, perceive, and understand complex temporal event relationships in longitudinal EHRs.

6.2 Introduction

Temporal event data are ubiquitous in nature and all aspects of our everyday life. Examples include the 1) neural firing pattern of individual neurons in our brains [99], 2) business transactions in the financial sector [100], 3) external event stimuli a robot interacts with [101], 4) event-related data from sensor measurements [102; 103], or 5) the healthcare industry [12].

Finding latent temporal patterns is important in many domains as they encode temporal concepts such as event trends, episodes, cycles, and abnormalities. For example, in the healthcare domain latent event patterns facilitate decision support for patient diagnosis, prognosis, and management. Of particular interest is the temporal aspect of information hidden in event data that may be used to perform intelligent reasoning and inference about the latent relationships between event entities over time. An event entity can be a person, an object, or a location in time. In our case the entity we refer to is the electronic health record.

Recent efforts towards the implementation of electronic health records show promise towards better data integration, automated access, and improved care delivery, yet their full potential is still underutilized. The vast amount of data contained in an EHR pose challenges not only to medical practitioners, but also to the information analysis by machines. Mining electronic health records is a difficult task due to data complexity and scale. More specifically, we face the following challenges:

- *Complexity.* Medical treatment is a complex multi-faceted endeavor involving nested hierarchies of temporal event relationships that involve multiple covariates and heterogeneous events composed of a mixture of single events, intervals,

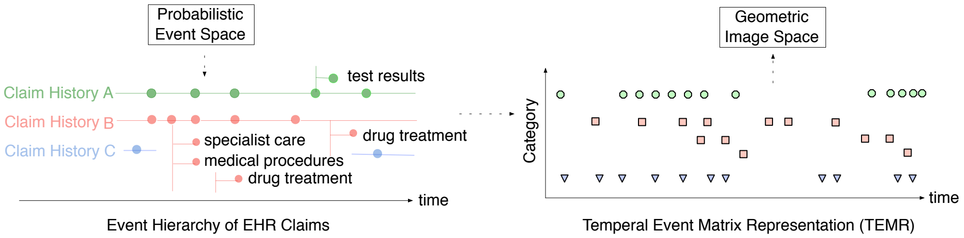


Figure 6.1: The proposed temporal event matrix representation (TEMR). The space-time dimension of a probabilistic event space $\mathbb{R}^3 \times \mathbb{R}$ is mapped onto a measurable geometric space of $\mathbb{R}^2$ by encoding events within a structured sparse image matrix. Left: The EHR can be abstracted as a nested heterogeneous event hierarchy. Event realizations can comprise multiple event groups (e.g. A, B, and C). Each event-group itself consists of an event category hierarchy. Right: In TEMR event types and groups are represented with a visual symbol system to ease interpretation. The temporal (time) dimension of events is mapped along the x-dimension and different event categories are stacked up along the y-dimension. Visual shift-invariant latent patterns can then be mined.

and sequences. Multivariate *nested* and *heterogeneous* event patterns over time are difficult to make sense of.

- *Incompleteness.* In reality we often face the missing data problem and data irregularity, where patients may face sudden death, do not follow recommended treatment guidelines, or physician dependent differences in care delivery.

- *Interpretability.* Clinical decision making relies on precise knowledge of the patient's medical history in context to group specific patient characteristics. In order to make optimal use of human and machine knowledge, the employed analytics should be commensurate with human capabilities and constraints, while at the same time provide deep insight into the patient record. Thus the mined patterns should be interpretable.

- *Shift-invariance.* Medical histories of patients differ in their absolute time stamps (i.e. they are not time aligned). Making inferences on common and individual patient characteristics within populations requires a time (shift)-invariant representation of the patient.

- *Scalability.* The patient's medical history captured within EHRs contains massive collections of heterogeneous data sources and modalities. Especially for diabetic patients with chronic diseases these records are long and complex. The employed analytics should efficiently cope with big data to support deep analysis.

This chapter proposes human-intuitive analytics that enable the representation and discovery of interpretable event patterns to ease knowledge absorption and comprehension of the employed analytics model. We develop a novel temporal event matrix representation, which we name TEMR, and learning algorithm to perform large-scale temporal pattern mining of longitudinal heterogeneous event data. The algorithm enables the representation, extraction, and mining of complex latent event relationships within single and multiple event entities to perform exploratory group analysis. We outline the contributions of this chapter at three levels.

- *Algorithmic level.* We propose 1) a novel stochastic optimization scheme for large-scale longitudinal event pattern mining of multiple event entities in a group and 2) a doubly constrained sparse coding framework that learns over-complete, shift-invariant, and sparse temporal event patterns for improved interpretability. We show how to cope with the sparsity in the data as well as in the latent factor model by inducing ℓ_1-norm constraints on the latent factors and its basis coefficients. We demonstrate that appropriate normalization constraints on the sparse latent factor model allow for automatic rank determination.

- *Representation level.* TEMR supports 1) the representation of hierarchical event data composed of single events, event intervals, and higher-order event structure, 2) the combination of semantically invariant shape and probabilistic metrics to facilitate knowledge reasoning and inference, 3) the joint representation of continuous and discrete categorical event data, and 4) the application of our representation to the challenging task of mining heterogeneous longitudinal event data from electronic health records. TEMR maps the space-time dimensionality of a probabilistic event space onto a measurable geometric space of $\mathbb{R}^2$ by encoding events as a structured spatial-temporal shape or point process. This projective mapping is achieved by using a rich geometric visual symbol system forming a structured two-dimensional sparse matrix. We address the missing data problem by representing the probabilistic event space with a geometric temporal event matrix representation, where missing events are handled implicitly as empty elements in the matrix.

- *Experimental level.* We validated our approach with extensive large-scale experiments on synthetic data and on a real-world electronic health records (EHRs) dataset. In total over 70,000 latent factor models were computed. We employ a special case of the β-divergence and show that this parameterization optimally copes with binary sparse data. The β-divergence is a parameterized family of cost functions that measures the difference between two probability distributions [104]. The β-divergence has been mainly used for continuous, non-binary, and non-sparse music spectrogram data within a nonnegative matrix factorization framework by considering special cases of β, (e.g. $\beta = 0, 1,$ and 2) [105]. We report on optimal model parameters (β, rank, window size, sparsity), convergence behavior, and the reconstruction accuracy. We link temporal patient encounter patterns against a diabetic complications severity index to explore the relationship between HRU and diabetic disease severity.

6.3 Prior art

This section is divided into two parts. The first part provides related work on the topic of knowledge representations for temporal data mining. We address time series knowledge representations as the continuous counterpart to our proposed geometric event matrix representation for discrete temporal event data. The second part outlines related work on nonnegative matrix factorization and its various extensions. In each section we contrast our contributions with the state of the art.

6.3.1 Time series knowledge representations

Most of the relevant prior research in temporal data mining transforms multivariate continuous time series into discrete symbolic representations (string, nominal, categorical, and item sets). Keogh *et al.* summarized existing time series representations as data adaptive and non-data adaptive representations such as the standard discrete Fourier transform (DFT), the Wavelet transform (DWT), piecewise linear approximation (PLA), adaptive piecewise constant approximation (APCA), the singular value decomposition (SVD), and symbolic aggregate approximation (SAX) [106].

A multitude of temporal knowledge representations in the form of symbolic languages and grammars have been formulated as a means to perform intelligent reasoning and inference from time-dependent data. Mörchen *et al.* (2006 [107], 2007 [108], 2010 [109]) proposed a novel *Time Series Knowledge Representation (TSKR)* as a pattern language (grammar) for temporal knowledge discovery from multivariate time series and symbolic interval data. A main drawback is that symbolic languages and temporal grammars specify time-dependent structure explicitly. They assume a fixed event structure such as event intervals, which limits the flexibility of learning unknown patterns that are not part of the specified language or grammar model.

In contrast to [106; 107; 108; 109] our knowledge representation does not use a symbolic language or grammar to represent knowledge, but rather uses a geometric approach to visually encode event data. In our approach, single events, event intervals, and high-order event structure (trends, episodes, cycles, etc.) can be represented jointly. Whereas other languages address the missing value problem by modeling event intervals instead of single events our representation can jointly encode both types of event structure. Another advantage is that our knowledge representation condenses complex multivariate temporal event relationships into an intuitive, interpretable, simple visual form that can be easily absorbed and understood by humans for synergistic human-machine intelligence. Moreover our representation supports rich analysis by employing methods from the image processing community in combination with standard probabilistic models for statistical event analysis.

6.3.2 Nonnegative matrix factorization and extensions

Early nonnegative matrix factorization (NMF) algorithms include the work from Paatero (1994) [110] and Lee and Seung (1997, 1999, 2001) [111], [112], [113]. Since then, many extensions have been proposed. Hoyer (2002) [114] and Eggert (2004) [115] introduced sparse NMF by adding a sparsity inducing regularizer to the standard NMF objective. The concept of sparsity is important for model interpretability, improved algorithm performance, and efficient data representation. To address the dynamic nature of the data, convolutional NMF models have been proposed by Eggert (2004) [116], Smaragdis (2004) [117], and O'Grady and Pearlmutter (2007) [105]. Recently, several forms of online NMF have been proposed as in Cao *et al.* (2007) [118], Mairal *et al.* (2010) [119], and Wang *et al.* (2011) [120].

[0] In what follows we will adopt the notation from the standard matrix factorization literature. The literature uses several notations for the *bases* and *coefficient* matrix e.g. $\{(\mathbf{W}, \mathbf{H}), (\mathbf{F}, \mathbf{G}), (\mathbf{A}, \mathbf{S})\}$. We will follow Lee and Seung's notation.

Our contributions differ with the work in [105],[117],[118],[119], and [120] in several ways. We perform a stochastic optimization scheme within a doubly constrained convolutional sparse coding framework to support: 1) large-scale factorization of single and multiple spatial-temporal point processes in a group and 2) an over-complete sparse latent factor model for solving the rank selection problem to learn common and individual temporal patterns within a group. In [114],[105],[117],[116] it was mentioned that $\mathcal{W}$ requires having unit norm bases. However, basis-wise normalization in conjunction with a sparsity constraint on $\mathcal{W}$ leads to a latent factor model that is non-sparse with respect to the bases set. In contrast, an element-wise normalization enables a sparse over-complete latent factor model, where the majority of basis atoms are zero. Whereas [105] used the β-divergence for continuous music spectrogram analysis with the standard setting of $\beta = 1$, we employ a special case ($\beta = 0.5$) of the parameterized β-divergence and show that it outperforms the generalized Kullback-Liebler divergence for $\beta = 1$. Our update rules for $\mathbf{W}$ and $\mathbf{H}$ differ as does the normalization constraint on $\mathbf{W}$ to allow for an over-complete sparse bases representation accounting for sparsity in the latent factor model $\mathbf{W}$ as well as in $\mathbf{H}$. For $\mathbf{H}$ we use a true convolutional update rule that does not average the weighting coefficients across multiple shifted windows. The update equations of our stochastic optimization scheme employ multiplicative update rules, which lead to simplified implementations without the need to optimize the learning rate of the gradient descent step.

6.4 Methods

6.4.1 The temporal event matrix representation

The proposed temporal event matrix representation is composed of a rich set of geometric shape primitives that symbolize multivariate event data. TEMR maps the space-time dimensionality of a probabilistic event space $\mathbb{R}^3 \times \mathbb{R}$ onto a measurable

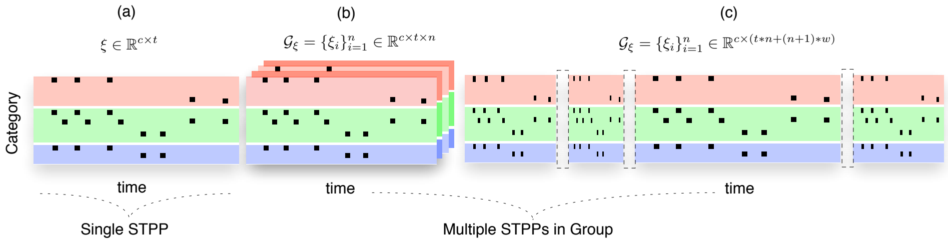

(a)
$\xi \in \mathbb{R}^{c\times t}$
(b)
$\mathcal{G}_\xi = \{\xi_i\}_{i=1}^n \in \mathbb{R}^{c\times t\times n}$
(c)
$\mathcal{G}_\xi = \{\xi_i\}_{i=1}^n \in \mathbb{R}^{c\times (t*n+(n+1)*w)}$
Category
time
time
time
Single STPP
Multiple STPPs in Group

Figure 6.2: TEMR as a matrix and tensor for single and multiple spatial-temporal point processes. (a) Multiple event sequences $\mathcal{E}_{p,q}$ of a group G_p (rows within a colored block) are vertically stacked together. At a second level, multiple event sequence groups G_p (red, green, blue) are vertically stacked on top of each other. The shown spatial-temporal process ξ on the left consists of three groups G1 (red), G2 (green), and G3 (blue). Black dots denote active events. (b) Multiple such STPPs in a group form a three-way tensor $\mathcal{G}_\xi$. (c) A tensor unfolded view of $\mathcal{G}_\xi$ is shown on the right, where n denotes the sample size, c the number of categories, t the number of time points, and w the temporal window size. Note the different temporal dimensions each STPP can have.

geometric space of $\mathbb{R}^2$ by encoding events with a rich visual symbol system. The visual symbol system can use color, shape, texture, position, value, and orientation to encode information. This mapping produces a structured sparse image matrix (see Fig. 6.1), which is flexible in terms of visually representing different event data types in a form intuitive to humans.

TEMR can be used to represent single or multiple event entities, where an event entity would correspond to all the events that are captured within a patient's EHR. Multiple event entities would thus correspond to multiple TEMR matrices forming a 3-order tensor or alternatively an unfolded 2-order tensor. Each event belongs to a labeled *event-type (ET)* and an *event-group (EG)* category (red, green, blue boxes). Multiple event sequences can be encoded with TEMR forming a multivariate representation, where the rows of TEMR encode ET/EG categories and the columns encode the time domain. These concepts are illustrated in Fig. 6.2.

TEMR can also be used to build a rich patient signature, which is composed of multiple temporal pattern windows of different lengths (e.g., week, month, quarter). Each pattern window captures specific latent aspects of the patient's medical history. This visual patient representation is efficient, intuitive, and is commensurate with human capabilities and constraints. Fig. 6.3 shows an example of a TEMR patient signature composed of multiple weekly, monthly, and quarterly temporal event pattern windows. The fundamental unit of analysis is a *single event* (a black square) from which a *pair of events* (two black squares) can be used to define an *event interval* or a *set of ordered events* to define an *event sequence* (a multiple black squares).

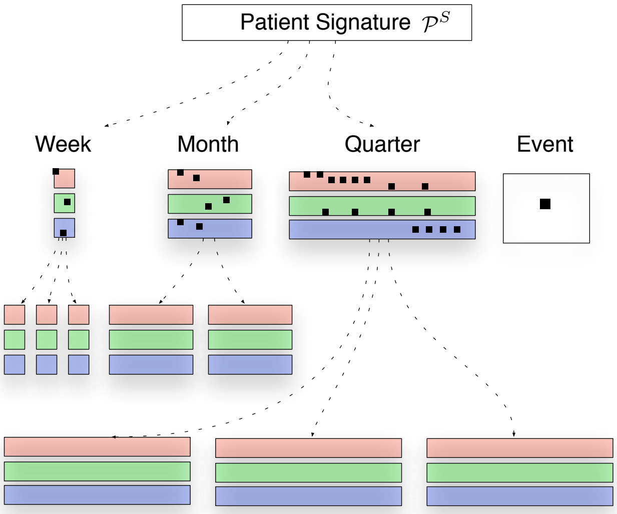


Figure 6.3: The TEMR patient signature. $\mathcal{P}^S$ is composed of multiple temporal pattern windows of different lengths (e.g. week, month, and quarter). One could also imagine a more generic hierarchical structure of the temporal pattern windows, where weekly windows could be extracted from monthly windows.

TEMR enables the encoding of the temporal concepts of *order, duration, coincidence, concurrency, synchronicity, periodicity, and trends* of time patterns. Temporal operators for qualitative temporal reasoning have quantitative meaning in the measurable geometric space. For example, temporal operators such as *before, meets, overlaps,*

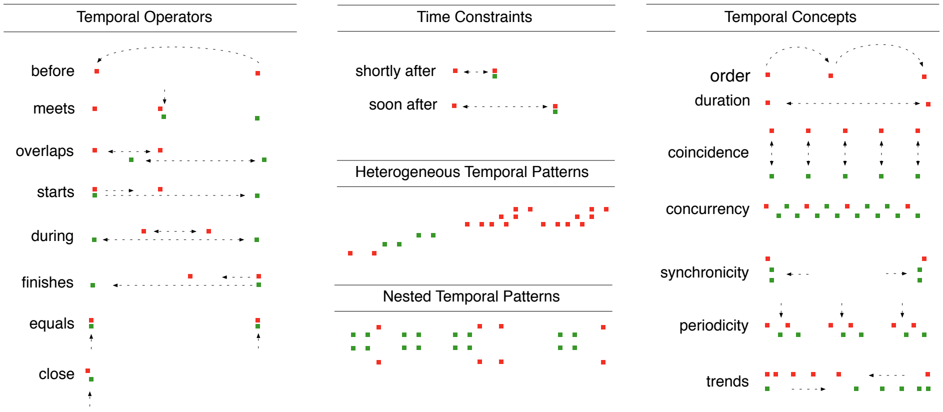


Figure 6.4: Temporal operators, constraints, and concepts. Left: The temporal operators describe common temporal event interval operators as proposed by Allen's interval logic. Middle: TEMR can represent qualitative temporal reasoning of temporal constraints such as *shortly after, soon after.* In addition, TEMR also enables the representation of event trends, intervals as well as *heterogeneous and nested events.* Right: The list of temporal concepts describe the event interval concepts as proposed by Mörchen *et al.* [109; 107; 108]. The red and green colors were chosen for improved visualization purposes and do not have a particular meaning.

starts, during, finishes, after, close, equals, or in combination with time constraints *shortly after and soon after* can be expressed in terms of geometric shape distances. We refer to a mixture of single events, event pairs, and event sequences as *high-order heterogeneous temporal events.* These concepts are illustrated in Fig. 6.4.

By using shape invariant metrics semantic invariances can be modeled and included into the mining framework. We note that the chosen geometric representational space offers a wide set of tools to be applied from the signal and image processing community.

6.4.2 Preliminaries

Suppose we have a TEMR matrix $\mathbf{X} \in \mathbb{R}^{c \times t}$, where c is the number of event categories and t the length of the patient's medical history. In order to learn a shift-invariant representation of $\mathbf{X}$ we consider the following convolutional latent factor model

$$\mathbf{X} = \mathcal{W} \star \mathcal{H} + \mathbf{N} \tag{6.4.1}$$

where $\mathcal{W} = \{\mathbf{W}^{(d)}\}_{d=1}^{r} \in \mathbb{R}^{u \times v \times r}$ and $\mathcal{H} = \{\mathbf{H}^{(d)}\}_{d=1}^{r} \in \mathbb{R}^{r \times c \times t}$ are 3-order tensors, $\star$ a shift-invariant convolutional operator, and $\mathbf{N}$ a noise model. Here u and v refer to the dimensions of a single basis element and r to the size of the basis set. The $\star$ operator in equation 6.4.1 can be expanded into

$$(\mathcal{W} \star \mathcal{H})_{ij} = \sum_{d} \sum_{m,n} \mathbf{W}^{(d)}(i-m, j-n)\mathbf{H}^{(d)}(m,n), \tag{6.4.2}$$

where m, n denote the shift indices and d a basis index. In the case of ($m = 0$ and $n > 0$) we obtain a *horizontal shift-invariant* latent factor model in which case $\mathcal{H}$ reduces to $\mathbf{H} \in \mathbb{R}^{r \times t}$. Similarly, for ($m > 0$ and $n = 0$) we obtain *vertical shift invariance*. Equation 6.4.2 consists of a linear superposition of a basis set convolved with its coefficient matrix.

The *horizontal shift-invariant* latent factor model takes the form

$$\mathbf{X} \approx \widetilde{\mathbf{X}} = \sum_{n} \sum_{d} \mathbf{W}^{(d)}(n)\mathbf{h}^{(d)}(j-n) + \mathbf{N}, \tag{6.4.3}$$

where $\mathbf{h}^{(d)} \in \mathbb{R}^{1 \times t}$ corresponds to the weighting coefficients of a single basis atom $\mathbf{W}^{(d)}$, n the shift index, and r the number of basis atoms (i.e., the rank of the latent factor model). In matrix form we can write

$$\mathbf{X} \approx \widetilde{\mathbf{X}} = (\mathcal{W} \star \mathbf{H})_{ij} = \sum_{\tau=1}^{T} (\mathcal{W})_{i,j=\tau,k} (\mathbf{H})_{i,j-\tau}. \tag{6.4.4}$$

where we assume $\mathbf{N} = 0$. Here $\mathcal{W} \in \mathbb{R}^{c \times w \times r}$ is a three order tensor and $\mathbf{H} \in \mathbb{R}^{c \times t}$ the basis coefficient matrix. One can see that the convolutional form also returns a $c \times t$ matrix.

Next, we define the matrix β-divergence $d_\beta(\mathbf{X}|\widetilde{\mathbf{X}})$ [121],[104][1]between the original matrix $\mathbf{X}$ and the approximated matrix $\widetilde{\mathbf{X}}$ as a general divergence measure

$$d_\beta(\mathbf{X}|\widetilde{\mathbf{X}}) = \frac{1}{\beta(\beta-1)} \sum_{ij} (\mathbf{X}_{ij}^\beta + (\beta-1)\widetilde{\mathbf{X}}_{ij}^\beta - \beta \mathbf{X}_{ij} \widetilde{\mathbf{X}}_{ij}^{\beta-1}), \tag{6.4.5}$$

where $\beta \geq 0$ is a nonnegative constant. Taking the limit of equation 6.4.5 by letting β approach $0, 1,$ and 2 gives the Itakura-Saito divergence (D_{IS}), generalized Kullback-Liebler divergence (D_{KL}), and the Euclidean distance (D_E) respectively.

$$\lim_{\beta \to 0} d_\beta(\mathbf{X}|\widetilde{\mathbf{X}}) = d_{\beta=0}^{D_{IS}}(\mathbf{X}|\widetilde{\mathbf{X}}) = \sum_{ij} \frac{\mathbf{X}_{ij}}{\widetilde{\mathbf{X}}_{ij}} - \log \frac{\mathbf{X}_{ij}}{\widetilde{\mathbf{X}}_{ij}} - 1 \tag{6.4.6}$$

$$\lim_{\beta \to 1} d_\beta(\mathbf{X}|\widetilde{\mathbf{X}}) = d_{\beta=1}^{D_{KL}}(\mathbf{X}|\widetilde{\mathbf{X}}) = \sum_{ij} \mathbf{X}_{ij} \log \frac{\mathbf{X}_{ij}}{\widetilde{\mathbf{X}}_{ij}} + \widetilde{\mathbf{X}}_{ij} - \mathbf{X}_{ij} \tag{6.4.7}$$

$$\lim_{\beta \to 2} d_\beta(\mathbf{X}|\widetilde{\mathbf{X}}) = d_{\beta=2}^{D_E}(\mathbf{X}|\widetilde{\mathbf{X}}) = \sum_{ij} \frac{1}{2}(\mathbf{X}_{ij} - \widetilde{\mathbf{X}}_{ij})^2 \tag{6.4.8}$$

Finally, we introduce a shifting matrix $\mathbf{S}$. $\mathbf{S}$ is a $n \times n$ matrix of the form

$$\mathbf{S}_{\tau=1} = \begin{bmatrix} 0 & & & & \\ 1 & \ddots & & & 0 \\ & \ddots & \ddots & & \\ & 0 & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \mathbf{I}_{n-1} & 0 \end{bmatrix}, \tag{6.4.9}$$

[1] Besides the β-divergence one could also use the α-divergence [121] or γ-divergence [**?**].

which when multiplied with a matrix $\mathbf{X}_{ij}$ causes $\mathbf{x}_j$ to shift by $j = \tau$ columns. Shifting matrices are zero-one matrices that have useful properties for the analysis of time series models. Here $\mathbf{S}_{\tau=1}$ is the first associated shifting matrix out of n shifting matrices. $\mathbf{S}_{\tau=1}$ has off-diagonal entries of value one and zero elsewhere. The pre- or post-multiplication of a matrix $\mathbf{X}$ with $_{\tau=1}$ shifts the entries of $\mathbf{X}$ by one row or column respectively. The operator fills up the empty space that is created due to the shift with zero values. Shifting can be performed in a bidirectional manner. To right and left-shift $\mathbf{X}$ by n shifts we can use $\mathbf{X}_{n\rightarrow} = \mathbf{X}\mathbf{S}_n^T$ and $\mathbf{X}_{\leftarrow n} = \mathbf{X}\mathbf{S}_n$.

6.4.3 Learning a single spatial-temporal point process

We are interested in two aspects. First, to learn a human-intuitive *horizontal shift-invariant sparse bases* (i.e., a double sparse dictionary) $\mathcal{W}$ and the associated *sparse weighting coefficients* (sparse code) $\mathbf{H}$ of $\mathcal{W}$. The emphasis is on learning an efficient, minimalistic, and interpretable representation $\mathcal{R}_\theta$, with $\theta = \{\mathcal{W}, \mathbf{H}\}$. Learning $\mathcal{R}_\theta$ can be achieved by coupling an approximation error objective with a sparsity constraint on $\mathcal{W}$ and $\mathbf{H}$

Given a sparse data matrix $\mathbf{X} \in \mathbb{R}^{c\times t}$ we learn a *horizontal shift-invariant latent factor model* by minimizing the following constrained optimization problem

$$\min_{\mathcal{W},\mathbf{H}} \mathcal{O}(\mathbf{X}, \mathcal{W}, \mathbf{H}) := d_\beta \left(\mathbf{X}, \sum_{\tau=1}^{T} (\mathcal{W})_{i,j=\tau,k} (\mathbf{H})_{i,j-\tau} \right) \tag{6.4.10}$$
$$\text{s.t.} \forall \tau := 1, 2, ..., T, \mathcal{W} \geq 0, \mathbf{H} \geq 0,$$

Taking the partial derivate of 6.4.10 with respect to the latent factor model we have

$$\frac{\partial \mathcal{O}}{\partial \mathcal{W}_{jk}} = \sum_{\tau=1}^{t} \sum_{d=1}^{r} \left(\widetilde{\mathbf{X}}^{\beta-1} - \mathbf{X} \odot \widetilde{\mathbf{X}}^{\beta-2} \right) \left(\mathbf{H}\mathbf{S}_\tau^T \right)_i^T \tag{6.4.11}$$

and

$$\frac{\partial \mathcal{O}}{\partial \mathbf{H}} = \sum_{\tau=1}^{t} \left(\widetilde{\mathbf{X}}^{\beta-1} - \mathbf{X} \odot \widetilde{\mathbf{X}}^{\beta-2} \right) \mathcal{W}_j^T, \tag{6.4.12}$$

where $\odot$ and $/$ are element-wise operations. From equation 6.4.11 and 6.4.12 we can obtain multiplicative update rules by diagonally rescaling the gradient to cancel out the learning rate

$$\mathcal{W}_{jk} \leftarrow \sum_{\tau=1}^{t} \sum_{d=1}^{r} \mathcal{W}_{jk} \odot \frac{(\mathbf{X} \odot \widetilde{\mathbf{X}}^{\beta-2}) \left(\mathbf{H}\mathbf{S}_\tau^T \right)_i^T + \epsilon}{(\widetilde{\mathbf{X}}^{\beta-1}) \left(\mathbf{H}\mathbf{S}_\tau^T \right)_i^T + \epsilon}, \tag{6.4.13}$$

$$\mathbf{H} \leftarrow \sum_{\tau=1}^{t} \mathbf{H} \odot \frac{\mathcal{W}_j^T \left(\mathbf{X} \odot \widetilde{X}^{\beta-2} \mathbf{S}_\tau \right) + \epsilon}{\mathcal{W}_j^T \left(\widetilde{\mathbf{X}}^{\beta-1} \mathbf{S}_\tau \right) + \epsilon}. \tag{6.4.14}$$

To account for double sparsity, i.e. the sparsity in the data as well as in the latent factor model, we regularize the β-divergence by inducing a double sparsity constraint on $\mathcal{W}$ and $\mathbf{H}$. The double sparsity constraint requires a different normalization constraint to support an over-complete bases set, where the factorization rank r can be defined such that r is larger than the actual bases elements in the data. The over-complete bases representation addresses the rank selection problem, where irrelevant basis elements are squashed to be zero and only a few basis elements that are supported in the data are retained. The regularized β-divergence with the double sparsity constraint can be formulated by the following constraint optimization problem

$$\begin{aligned}
&\min_{\mathcal{W}, \mathbf{H}} \mathcal{O}_1, \text{where} \\
&\mathcal{O}_1(\mathbf{X}, \mathcal{W}, \mathbf{H}) := d_\beta \left(\mathbf{X}, \sum_{\tau=1}^{T} (\mathcal{W})_{i,j=\tau,k} (\mathbf{H})_{i,j-\tau} \right) + \lambda_1 \sum \left| \mathcal{W}_{i,j,k} \right| + \lambda_2 \sum \left| \mathbf{H}_{i,j} \right|, \qquad (6.4.15) \\
&\text{s.t.} \forall \tau := 1, 2, ..., T, \mathcal{W} \geq 0, \mathbf{H} \geq 0.
\end{aligned}$$

The joint objective of equation 6.4.15 is non-convex, but convex with respect to $\mathcal{W}$ and $\mathbf{H}$ individually. The problem can be solved with an alternative optimization (block coordinate descent), where each factor is optimized in an alternate fashion. By inducing a ℓ_1-norm constraint on the β-divergence, sparsity can be enforced in a trade-off with the approximation error of the factorization. As in equation 6.4.11- 6.4.14 we obtain the multiplicative update rules by following the same procedure

$$\frac{\partial \mathcal{O}_1}{\partial \mathcal{W}_{jk}} = \sum_{\tau=1}^{t}\sum_{d=1}^{r}\left(\widetilde{\mathbf{X}}^{\beta-1} - \mathbf{X}\odot\widetilde{\mathbf{X}}^{\beta-2}\right)\left(\mathbf{H}\mathbf{S}_{\tau}^{T}\right)_{i}^{T} + \lambda_1 \tag{6.4.16}$$

$$\frac{\partial \mathcal{O}_1}{\partial \mathbf{H}} = \sum_{\tau=1}^{t}\left(\widetilde{\mathbf{X}}^{\beta-1} - \mathbf{X}\odot\widetilde{\mathbf{X}}^{\beta-2}\right)\mathcal{W}_{j}^{T} + \lambda_2, \tag{6.4.17}$$

$$\mathcal{W}_{jk} \leftarrow \sum_{\tau=1}^{t}\sum_{d=1}^{r}\mathcal{W}_{jk}\odot\frac{(\mathbf{X}\odot\widetilde{\mathbf{X}}^{\beta-2})\left(\mathbf{H}\mathbf{S}_{\tau}^{T}\right)_{i}^{T} + \epsilon}{(\widetilde{\mathbf{X}}^{\beta-1})\left(\mathbf{H}\mathbf{S}_{\tau}^{T}\right)_{i}^{T} + \lambda 1 + \epsilon}, \tag{6.4.18}$$

$$\mathbf{H} \leftarrow \sum_{\tau=1}^{t}\mathbf{H}\odot\frac{\mathcal{W}_{j}^{T}\left(\mathbf{X}\odot\widetilde{X}^{\beta-2}\mathbf{S}_{\tau}\right) + \epsilon}{\mathcal{W}_{j}^{T}\left(\widetilde{\mathbf{X}}^{\beta-1}\mathbf{S}_{\tau}\right) + \lambda 2 + \epsilon}. \tag{6.4.19}$$

Considering normalization-invariant update rules the optimization problem from equation 6.4.15 becomes

$$\begin{aligned}
&\min_{\widehat{\mathcal{W}},\mathbf{H}} \mathcal{O}_2, \text{where}\\
&\mathcal{O}_2(\mathbf{X},\widehat{\mathcal{W}},\mathbf{H}) := d_\beta\left(\mathbf{X}, \sum_{\tau=1}^{T}\left(\widehat{\mathcal{W}}\right)_{i,j=\tau,k}(\mathbf{H})_{i,j-\tau}\right) + \lambda_1\sum\left|\widehat{\mathcal{W}}_{i,j,k}\right| + \lambda_2\sum\left|\mathbf{H}_{i,j}\right|,\\
&\text{s.t.}\forall\tau := 1,2,...,T, \widehat{\mathcal{W}} \geq 0, \mathbf{H} \geq 0,
\end{aligned} \tag{6.4.20}$$

where different types of normalization constraints can be employed. Here we consider *total normalization* and *individual normalization* of $\mathcal{W}$

$$\widehat{\mathcal{W}}_T = \frac{\mathcal{W}}{\|\mathcal{W}\|_F} = \frac{\mathcal{W}}{\sqrt{\sum_{ijk} |\mathcal{W}_{ijk}|^2}}, \text{ or} \tag{6.4.21}$$

$$\widehat{\mathcal{W}}_I = \sum_{d=1}^{r} \frac{\mathbf{W}^{(d)}}{\|\mathbf{W}^{(d)}\|_F} = \sum_{d=1}^{r} \frac{\mathbf{W}^{(d)}}{\sqrt{\sum_{ij} |\mathbf{W^{(d)}}_{ij}|^2}}. \tag{6.4.22}$$

For *total normalization* we normalize the complete tensor (each element) $\mathcal{W}_{ijk}$ by its ℓ_2-norm. By inserting equation 6.4.21 or 6.4.22 into equation 6.4.20 we can solve the partial derivatives of the latent factor model and its multiplicative update rules by using the quotient rule

$$\frac{\partial \mathcal{O}_2}{\partial \widehat{\mathcal{W}}_{ij}} = \sum_{\tau=1}^{t} \sum_{d=1}^{r} \left(\widetilde{\mathbf{X}}^{\beta-1} - \mathbf{X} \odot \widetilde{\mathbf{X}}^{\beta-2} \right) \frac{\partial \widetilde{X}}{\partial \widehat{\mathcal{W}}_{ij}} + \lambda_1 \tag{6.4.23}$$

$$= \sum_{\tau=1}^{t} \sum_{d=1}^{r} \left(\widetilde{\mathbf{X}}^{\beta-1} - \mathbf{X} \odot \widetilde{\mathbf{X}}^{\beta-2} \right) \frac{\|\mathcal{W}\|_F - \mathcal{W}_{jk} \widehat{\mathcal{W}}_{jk}^T}{\|\mathcal{W}\|_F^2} \left(\mathbf{HS}_\tau^T \right)_i^T + \lambda_1 \tag{6.4.24}$$

$$\frac{\partial \mathcal{O}_2}{\partial \mathbf{H}} = \sum_{\tau=1}^{t} \left(\widetilde{\mathbf{X}}^{\beta-1} - \mathbf{X} \odot \widetilde{\mathbf{X}}^{\beta-2} \right) \frac{\mathcal{W}_j^T}{\|\mathcal{W}\|_F} + \lambda_2, \tag{6.4.25}$$

which then results in

$$\widetilde{\mathcal{W}}_{jk} \leftarrow \sum_{\tau=1}^{t} \sum_{d=1}^{r} \mathcal{W}_{jk} \odot \frac{\left(\mathbf{X} + \mathcal{W}_{jk} \widetilde{\mathcal{W}}_{jk}^T \widetilde{\mathbf{X}} \right) \odot \widetilde{\mathbf{X}}^{\beta-2} \left(\mathbf{HS}_\tau^T \right)_i^T + \epsilon}{\left(\widetilde{\mathbf{X}} + \mathcal{W}_{jk} \widetilde{\mathcal{W}}_{jk}^T \mathbf{X} \right) \odot \widetilde{\mathbf{X}}^{\beta-2} \left(\mathbf{HS}_\tau^T \right)_i^T + \lambda_1 + \epsilon}, \tag{6.4.26}$$

$$\mathbf{H} \leftarrow \sum_{\tau=1}^{t} \mathbf{H} \odot \frac{\widetilde{\mathcal{W}}_j^T \left(\mathbf{X} \odot \widetilde{X}^{(\beta-2)} \mathbf{S}_\tau \right) + \epsilon}{\widetilde{\mathcal{W}}_j^T \left(\widetilde{\mathbf{X}}^{\beta-1} \mathbf{S}_\tau \right) + \lambda_2 + \epsilon}. \tag{6.4.27}$$

where the indices for $\widetilde{\mathcal{W}}_{jk}$ and $\mathbf{H}_i$ run over $j = \tau, k = d, i = d$. This factorization allows us to learn an over-complete minimalistic and interpretable representation that is intuitive to humans.

The general algorithm for learning an over-complete sparse latent factor model of a single spatial-temporal point processes can be implemented as outlined in algorithm 6.1.

Algorithm 6.1 Learning a single spatial-temporal point process

```
Require: $\mathbf{X}, \mathcal{W}, \mathbf{H}, r, T, \beta, \lambda, iter$
Ensure: $\mathcal{W} \geq 0, \mathbf{H} \geq 0$
  Initialize $\mathcal{W}, \mathbf{H}$
  Normalize $\mathcal{W}$ via equation 6.4.21
  for $i = 1$ to $iter$ do
    Update $\mathbf{H}$ via equation 6.4.27
    Update $\mathcal{W}$ via equation 6.4.26
    Normalize $\mathcal{W}$ via equation 6.4.21
    Compute $d_\beta(\mathbf{X}|\widetilde{\mathbf{X}})$ via equation 6.4.5
    if (converged) then
      break
    end if
  end for
  return $\mathcal{R}_\Theta = \{\mathcal{W}, \mathbf{H}\}$
```

6.4.4 Learning multiple spatial-temporal point processes in a group

Whereas one could now use the learned latent factor model to perform group-based analysis using a clustering scheme, the non-convex and non-unique objective pose problems regarding the reproducibility of the learned representation (i.e., its latent bases). Each individual factorization will result in arbitrary orderings of the latent basis atoms. Thus any clustering based on the individual latent basis atoms will give meaningless results. In this regard our goal is to learn a hidden group structure for multiple spatial-temporal point processes in a joint fashion.

Multiple spatial-temporal point processes form a 3-way tensor $\mathcal{X} = [\mathbf{X}_1, ..., \mathbf{X}_n]$ where individual TEMR's are stacked up on top of each other along the 3rd dimension. To learn a group structure of multiple spatial-temporal point processes several strategies can be employed.

A straightforward way is to adopt the learning process from the previous section on an unfolded zero-padded two-way tensor of size $\mathbb{R}^{c\times(t*n+(n+1)*w)}$ (see Fig. 6.2).

For small n, factorizing the unfolded tensor enables to learn a minimalistic and interpretable group representation that jointly learns the common temporal patterns. However, this approach has several limitations. If c, t, or n is large, the individual learning scheme quickly becomes computationally expensive and intractable. For large w zero-padding individual slices of the original three-way tensor is wasteful.

Another alternative is to keep the three-way tensor structure and employ a stochastic optimization scheme. This prevents the storage of zero-padded elements as well as computationally efficient optimization. Within the stochastic optimization scheme large values of w and n do not incur an overhead in space complexity. Large values of c and t could be efficiently treated with parallel optimizations. Within the stochastic optimization scheme data samples arrive in a sequential manner and $\mathcal{R}_\Theta$ is adaptively learned in an online fashion allowing to learn the latent group structure for large n.

We consider single group and multiple group structure learning. Given $\mathcal{X}_c = [\mathbf{X}_{c1}, ..., \mathbf{X}_{cn}]$ and assuming $c \in \mathcal{C} = \{1\}$ learning the group structure of $\mathcal{X}_c$ amounts to solving the following constrained optimization problem

$$\min_{\widehat{\mathcal{W}}} \mathcal{O}_3, \text{where}$$

$$\mathcal{O}_3\left(\mathcal{X}_c, \widehat{\mathcal{W}}_c, \mathcal{H}_c\right) := \sum_{l=1}^{n} d_\beta \left(\mathbf{X}_l, \sum_{\tau=1}^{T} \left(\widehat{\mathcal{W}}\right)_{i,j=\tau,k} (\mathbf{H}_l)_{i,j-\tau}\right) + \lambda_1 \sum \left|\widehat{\mathcal{W}}_{i,j,k}\right| + \lambda_2 \sum \left|(\mathbf{H}_l)_{i,j}\right|, \tag{6.4.28}$$

$$\text{s.t.} \forall \tau := 1, 2, ..., T;\ l = 1, ..., n;\ \widehat{\mathcal{W}} \geq 0, \mathbf{H}_l \geq 0,$$

where $\mathcal{H}_c := \{\mathbf{H}_l\}_{l=1}^{n}$, $\mathbf{X}_l \in \mathbb{R}^{c \times t}, \mathcal{W} \in \mathbb{R}^{u \times v \times r}$, and $\mathbf{H}_l \in \mathbb{R}^{r \times t}$. Following the same approach as in equation 6.4.16 and 6.4.25- 6.4.27 we obtain

$$\widehat{\mathcal{W}}_{jk} \leftarrow \sum_{l=1}^{n} \sum_{\tau=1}^{t} \sum_{d=1}^{r} \mathcal{W}_{jk} \odot \frac{\left(\mathbf{X}_l + \mathcal{W}_{jk} \widehat{\mathcal{W}}_{jk}^T \widetilde{\mathbf{X}}_l\right) \odot \widetilde{\mathbf{X}}_l^{\beta-2} \left(\mathbf{H}_l \mathbf{S}_\tau^T\right)^{T_i} + \epsilon}{\left(\widetilde{\mathbf{X}}_l + \mathcal{W}_{jk} \widehat{\mathcal{W}}_{jk}^T \mathbf{X}_l\right) \odot \widetilde{\mathbf{X}}_l^{\beta-2} \left(\mathbf{H}_l \mathbf{S}_\tau^T\right)_i^T + \lambda_1 + \epsilon}, \tag{6.4.29}$$

$$\mathbf{H}_l \leftarrow \sum_{\tau=1}^{t} \mathbf{H}_l \odot \frac{\widehat{\mathcal{W}}_j^T \left(\mathbf{X}_l \odot \widetilde{\mathbf{X}}_l^{\beta-2} \mathbf{S}_\tau\right) + \epsilon}{\widehat{\mathcal{W}}_j^T \left(\widetilde{\mathbf{X}}_l^{\beta-1} \mathbf{S}_\tau\right) + \lambda_2 + \epsilon}. \tag{6.4.30}$$

Note that $\mathbf{H}_l$ does not have any meaning after the stochastic optimization pass. To obtain individual basis coefficients a second pass through $\mathcal{X}_c$ is required.

In case of $c \in \mathcal{C} = \{1, ..., C\}$ we have

$$\min_{\widehat{\mathcal{W}}} \mathcal{O}_4, \text{ where}$$

$$\mathcal{O}_4\left(\mathcal{X}_c, \widehat{\mathcal{W}}_c, \mathcal{H}_c\right) := \sum_{c=1}^{C} \left[\sum_{l=1}^{n} d_\beta \left(\mathbf{X}_l, \sum_{\tau=1}^{T} \left(\widehat{\mathcal{W}}\right)_{i,j=\tau,k} (\mathbf{H}_l)_{i,j-\tau}\right)\right] \tag{6.4.31}$$

$$+ \lambda_1 \sum \left|\widehat{\mathcal{W}}_{i,j,k}\right| + \lambda_2 \sum \left|(\mathbf{H}_l)_{i,j}\right|, \tag{6.4.32}$$

$$\text{s.t.} \forall \tau := 1, 2, ..., T;\ l = 1, ..., n; c = 1, ..., C\ \widehat{\mathcal{W}} \geq 0, \mathbf{H}_l \geq 0,$$

where

$$\widehat{\mathcal{W}} = \left[\widehat{\mathcal{W}}^S \quad \widehat{\mathcal{W}}^{I_{c=1}}, ..., \quad \widehat{\mathcal{W}}^{I_{c=C}}\right] \tag{6.4.33}$$

$$\widehat{\mathcal{W}}_{jk}^S \leftarrow \sum_{c=1}^{C} \sum_{l=1}^{n} \sum_{\tau=1}^{t} \sum_{d=1}^{r} \mathcal{W}_{jk}^S \odot \frac{\left(\mathbf{X}_{cl} + \mathcal{W}_{jk}^S \widehat{\mathcal{W}}_{jk}^{T^S} \widetilde{\mathbf{X}}_{cl}\right) \odot \widetilde{\mathbf{X}}_{cl}^{\beta-2} \left(\mathbf{H}_{cl} \mathbf{S}_\tau^T\right)^{T_i} + \epsilon}{\left(\widetilde{\mathbf{X}}_{cl} + \mathcal{W}_{jk}^S \widehat{\mathcal{W}}_{jk}^{T^S} \mathbf{X}_{cl}\right) \odot \widetilde{\mathbf{X}}_{cl}^{\beta-2} \left(\mathbf{H}_{cl} \mathbf{S}_\tau^T\right)_i^T + \lambda_1 + \epsilon}, \tag{6.4.34}$$

$$\widehat{\mathcal{W}}_{jk}^I \leftarrow \sum_{l=1}^{n} \sum_{\tau=1}^{t} \sum_{d=1}^{r} \mathcal{W}_{jk}^I \odot \frac{\left(\mathbf{X}_{cl} + \mathcal{W}_{jk}^I \widehat{\mathcal{W}}_{jk}^{T^I} \widetilde{\mathbf{X}}_{cl}\right) \odot \widetilde{\mathbf{X}}_{cl}^{\beta-2} \left(\mathbf{H}_{cl} \mathbf{S}_\tau^T\right)_i^T + \epsilon}{\left(\widetilde{\mathbf{X}}_{cl} + \mathcal{W}_{jk}^I \widehat{\mathcal{W}}_{jk}^{T^I} \mathbf{X}_{cl}\right) \odot \widetilde{\mathbf{X}}_{cl}^{\beta-2} \left(\mathbf{H}_{cl} \mathbf{S}_\tau^T\right)_i^T + \lambda_1 + \epsilon}, \tag{6.4.35}$$

$$\mathbf{H}_{cl} \leftarrow \sum_{\tau=1}^{t} \mathbf{H}_{cl} \odot \frac{\widehat{\mathcal{W}}_j^T \left(\mathbf{X}_{cl} \odot \widetilde{\mathbf{X}}_{cl}^{\beta-2} \mathbf{S}_\tau\right) + \epsilon}{\widehat{\mathcal{W}}_j^T \left(\widetilde{\mathbf{X}}_{cl}^{\beta-1} \mathbf{S}_\tau\right) + \lambda_2 + \epsilon}. \tag{6.4.36}$$

6.5 Experiments and results

6.5.1 Datasets

6.5.1.1 Real-world data

The real-world dataset consisted of an EHR data model. In conjunction with medical experts we have selected a diabetic patient pool (n=21,384) that was stratified into three groups A, B, and C. Group A consisted of patients (n=16,205) with no disease complications, group B consisted of patients (n=4,925) with chronic disease complications, and group C of patients (n=254) with acute complications. For all three groups we generated temporal event matrix representations (TEMRs) for each patient using *event-group* and *event-type level* criteria that defined general out-patient encounters specific to diabetes care (see 8.1 and 8.2 in the Appendix for a detailed list). The chosen criteria consisted of 30 different conditions that were grouped into four groups over a time period of 365 days: medical procedures (G_1 = CPTs), lab results (G_2 = LABS), primary care physician visits (G_3 = PCP), and specialty visits (G_4 = SPEC). Fig. 6.5 shows an example of a temporal event matrix from a patient in the diabetic patient pool.

6.5.1.2 Synthetic data

We have created four sets of synthetic datasets. The synthetic data matrices for all four sets encoded events as binary activation units in the form of a single 1-or-0 valued pixel, where a value of 1 (black) denoted an event realization and 0 (white) no event activity. Each row of the matrix referred to a particular event-type-level category and each column to a single time unit scale (e.g. days).

- *Set 1.* We created three TEMR data matrices $\mathcal{X} = [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3]$ to simulate a variety of different temporal event patterns.

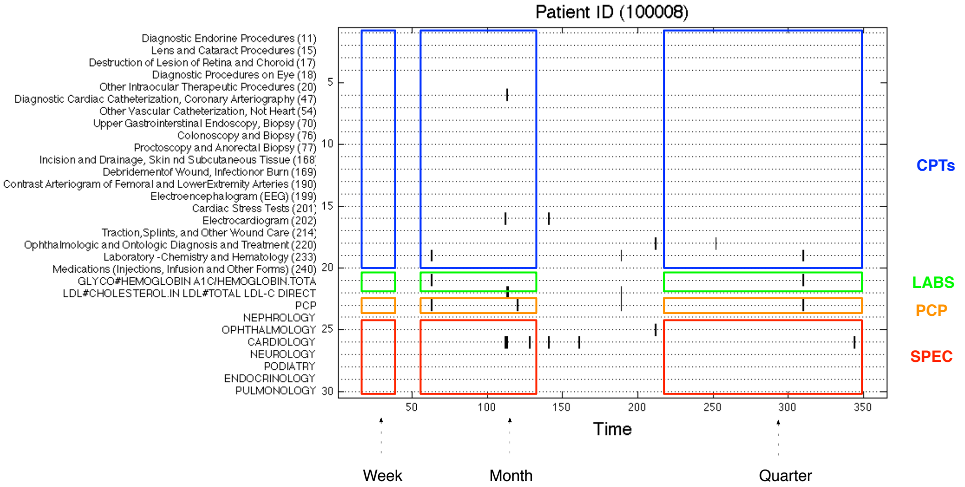


Figure 6.5: Example of a real-world TEMR extracted from a EHR data model. A real-world dataset that contains repeating and non-repeating event patterns.

According to Fig. 6.8, $\mathbf{X}_1$ and $\mathbf{X}_2$ consisted of Mörchens's time series knowledge representation's (TSKR) *event interval test pattern* [107],[108],[109]. The pattern comprises a trivariate interval event sequence (e.g., A, B, C), where so called *Tones* represent different *event interval durations*, *Chords* represent *coincidences of Tones*, and *Phrases* represent a *partial ordering of the Chords*. The red box corresponds to the partially ordered *Phrase* (AB-ABC-AC) and the green box to (AB-BC-AC) accordingly. An example of this TSKR pattern can be seen in Fig. 6.6 together with an equivalent TEMR representation.

Subsequently, $\mathbf{X}_3$ consisted of various temporal concepts and operators as introduced in Fig. 6.4. The red box corresponds to *synchronicity*, the green boxes to a *trend of decreasing coincidences* and a *trend of increasing coincidences*, and the blue box to *concurrency*. The four temporal concepts implicitly included the temporal operators: *order, duration, close, far, before, meets, overlaps, equals, during*. The remaining temporal operators: *starts* and *finishes* were not considered as they can be easily represented within TEMR. All synthetic data matrices in $\mathcal{X}$ had dimensions of 30 rows x 120 columns.

- *Set 2.* We have created two TEMR data matrices $\mathcal{X} = [\mathbf{X}_1, \mathbf{X}_2]$ to simulate *heterogeneous* and *nested* event patterns. $\mathbf{X}_1$ comprised a synthetic example of *heterogeneous event patterns.* $\mathbf{X}_2$ consisted of a synthetic example that simulated *nested event patterns.* All synthetic data matrices had dimensions of 10 rows x 60 columns. Examples can be seen in Fig. 6.9.

- *Set 3.* In Fig. 6.10 we created three TEMR data matrices $\mathcal{X} = [\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3]$ to simulate common and individual group patterns for a single group. The red and green box shows a hypothetical temporal event pattern, which occurs in $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ with the difference that the green box has multiple occurrences. The blue box shows a temporal event pattern, which only occurs in $\mathbf{X}_1$ and $\mathbf{X}_2$. The orange box shows a temporal event pattern, which only occurs in $\mathbf{X}_1$. All event patterns span a time window of seven days. All synthetic data matrices had dimensions of 30 rows x 120 columns.

- *Set 4.* We created nine TEMR data matrices $\mathcal{X} = \{\mathbf{X}_{c1}, \mathbf{X}_{c2}, \mathbf{X}_{c3}\}_{c=1}^{3}$ to simulate *common* and *individual group patterns* for three groups. Each group has associated three TEMR data matrices. The red box indicates the shared event pattern that occurs in all three groups and their individual matrices. The green, blue, and olive colored boxes indicate group specific event patterns for group $c = 1, 2, 3$ accordingly. All synthetic data matrices had dimensions of 10 rows x 60 columns. Examples can be seen in Fig. 6.11.

6.5.2 Performance metrics

To assess the quantitative performance of our framework we have chosen multiple validation metrics, as each of them has practical implications. The metrics consist of the average number of iterations until convergence I_{conv}, the mean reconstruction error R_{err}, and the mean Dice coefficient D_c over T trials

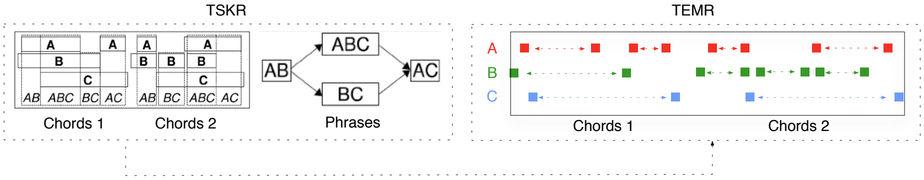


Figure 6.6: TSKR and TEMR examples of Mörchen's event interval test pattern. Left: TSKR enables one to distinguish between the partial ordering of so-called *Chords*. Such partial orderings form *Phrases*. Two *Chord* configurations are contained (i.e., AB-ABC-AC and AB-BC-AC). Right: TEMR can be used to emulate the same test pattern by representing an event interval with two consecutive events. The dotted arrows indicate the event interval that is marked by the colored solid squares, which denote the start and end of the interval.

$$I_{conv} = \frac{1}{T}\sum_{t=1}^{T} I_t \tag{6.5.1}$$

$$R_{err} = \frac{1}{T}\sum_{t=1}^{T} ||\mathbf{X}_t - \mathbf{R}_t||_F^2 \tag{6.5.2}$$

$$D_c = \frac{1}{T}\sum_{t=1}^{T} D_c(\mathbf{X}_t, \mathbf{R}_t) \tag{6.5.3}$$

The Dice coefficient is defined as $D_c = \frac{2|A \cap B|}{|A|+|B|}$, where D_c measures the set agreement between the original temporal event matrix and the reconstruction. The D_c score ranges from [0, 1], where 1 means perfect agreement. For all three metrics we computed 95% confidence intervals to assess the true mean of each performance measure.

6.5.3 Cross-validation of the proposed model

We validated the model on real-world data. We analyzed the reconstruction performance and convergence behavior of the learned representation for a single TEMR data matrix by performing a permutation test to cross-validate for the optimal model parameters. A representative dataset was selected from the patient pool that included multiple repeating patterns and non repeating patterns. The cross-validation involved 1,225 factorizations over 25 trials. In total 30,625 factorizations were computed. We examined the approximation error of the factorization as a function of different parameterizations of the

- β-divergence $\beta = \{0, 0.1, 0.25, 0.5, 1, 1.5, 2\}$
- degree of sparsity $\lambda = \{0, 0.5, 1, 2, 10\}$
- temporal window size $w = \{7, 14, 30, 60, 90\}$
- rank of the factorization $r = \{1, 5, 10, 15, 20, 25, 50\}$

From this pool we examined the optimal model with respect to β, λ, w, r by computing I_{conv}, R_{err}, and D_c and their 95% confidence intervals. Fig. 6.7 shows the cross validation results and table 6.1 summarizes the optimal model parameters.

Performance Measures	Optimal λ	Optimal w	Optimal β	Optimal r
Mean convergence	2.0	7	0.5	1
Mean Dice coefficient	0.5	30	0.5	50
Mean ℓ_2-norm	0.5	30	0.5	50

Table 6.1: Permutation test for cross-validation on a real-world dataset. Cross-validation results with respect to mean convergence (number of iterations), the reconstruction error (ℓ_2)-norm, and the Dice coefficient.

6.5.4 Learning complex event patterns

The second set of experiments was performed on synthetic data. We examined the efficacy of the regularized β-divergence with the double sparsity constraint. We demonstrate that our proposed framework is able to 1) learn shift-invariant, interpretable, and high-order latent temporal patterns, 2) cope with the double sparsity problem,

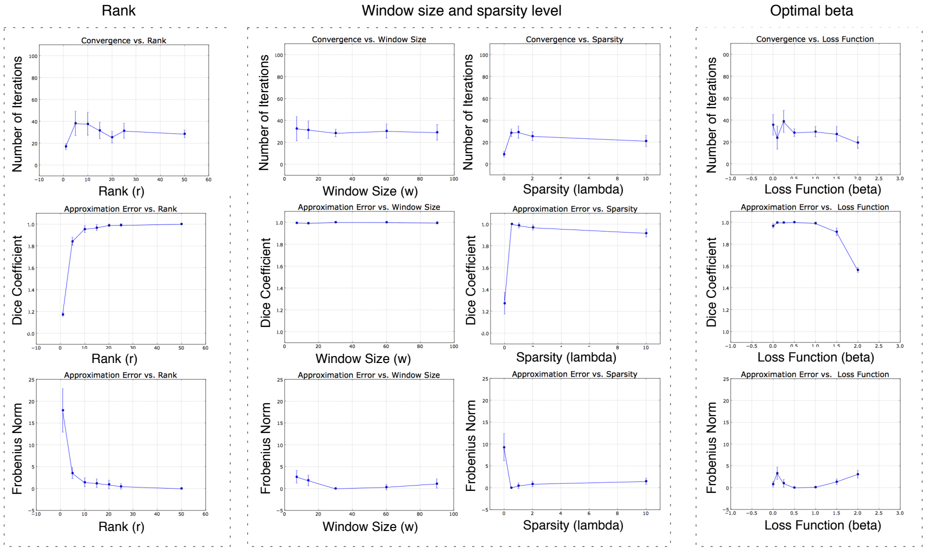


Figure 6.7: Permutation test for cross-validation on a real-world dataset. Shown are mean performance measures with 95% confidence intervals. Row 1: Mean convergence vs. the model parameters (rank, window size, sparsity, and different loss functions of the β-divergence). Row 2: Mean Dice coefficient vs. the model parameters. Row 3: Mean Frobenius norm vs. the model parameters.

and 3) estimate the rank of the latent factor model through the induced double sparsity constraint on the β-divergence.

Our motivation for the experiments was to examine three questions. First, can our geometric learning framework learn shift-invariant interpretable latent temporal patterns that are sparse in the data and in the latent factor model? Second, is the model sensitive to an optimally chosen rank? And third, does the framework handle binary data entries in the data matrix as well as complex high-order event structure?

In Fig. 6.8 we show that the algorithm can learn *Tones, Chords, and Phrases* and a diverse set of temporal concepts and operators.

For all factorizations we have used the following parameter settings: $\lambda = 0.5, \beta = 0.5$ as determined by our cross-validation study outlined in Section 6.5.3. The number of iterations were set to $iter = 100$ and the convergence threshold to 1e-9 to ensure that the algorithm converges. For the $\mathbf{X}_1$ and $\mathbf{X}_2$ we have used a $w = 35$ and a rank of $r = 2$ and $r = 10$ to account for the number of true patterns in the data and their pattern duration. For $\mathbf{X}_3$ we have used a window size of $w = 3$ and a rank of $r = 4$ and $r = 11$ accordingly, where $r = 11$ is an over-complete specified rank. For each case, the first rank (i.e., $r = 2, r = 4$) was chosen based on the known number of distinct temporal patterns in the data. The second rank (i.e., $r = 10, r = 11$) was chosen as an over-complete rank, where the pre-specified number of basis elements exceeds the number of true latent factors in the data. We ran 25 trials to assess the mean performance, standard error, and 95% confidence intervals. Fig. 6.8 shows the results for the second set of experiments.

In Fig. 6.9 we show how to learn *heterogeneous event patterns* and *nested event patterns*. For each set of patterns we demonstrate the effect on using different normalization schemes during the learning procedure. The left box shows the results when using the *total normalization* scheme and the right box shows the learned bases for the *individual normalization* scheme. The colored boxes indicate the temporal event patterns in the data and the learned latent factor model $\mathcal{W}$.

6.5.5 Learning event patterns within groups

We tested the stochastic optimization scheme on synthetic (i.e., set 3 and 4) and real-world data to assess the efficacy of the algorithm to learn latent event structures within single and multiple groups.

We examined the performance for incremental group learning on two different synthetic datasets as shown in Fig. 6.10 and Fig. 6.11. For the synthetic experiments the motivation was to assess whether we can learn common and individual latent event patterns from multiple data matrices in single or multiple groups. Another objective was to assess the convergence behavior and the sensitivity to a pre-defined rank (see Fig. 6.10 and 6.11).

Real-world data experiments were performed on the electronic health record dataset

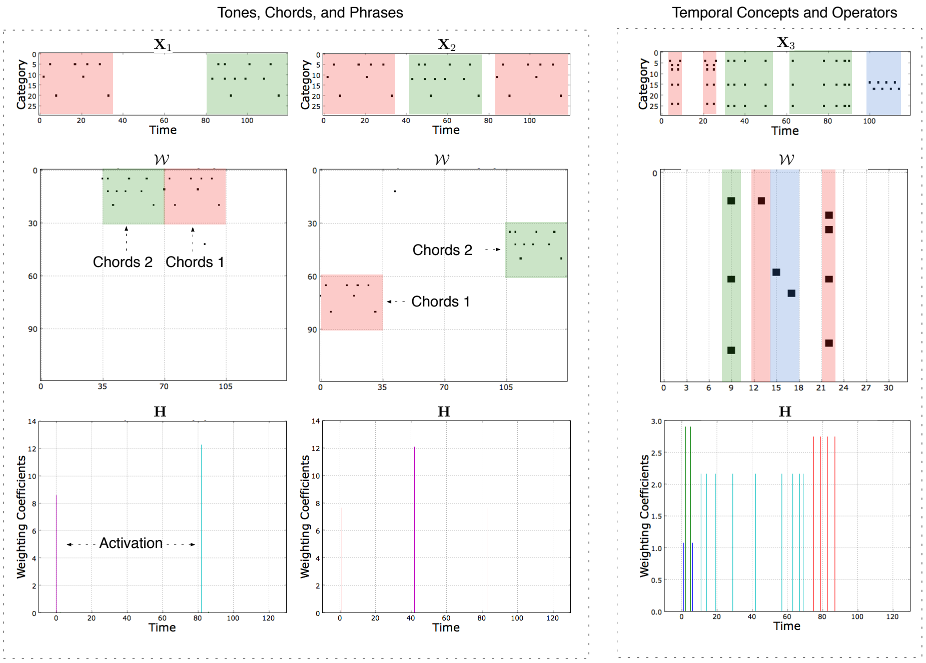


Figure 6.8: Learning Chords, Tones, Phrases and temporal concepts/operators. The first row shows the original dataset, the second row shows the sparse bases $\mathcal{W}$, and the last row the sparse code $\mathbf{H}$. The weighting coefficients in $\mathbf{H}$ were colored based on an arbitrary random color map. They do not have an explicit meaning. The event patterns of interest one should pay attention to are color coded in the original datasets and the learned bases.

outlined in Section 6.5.1.1. The motivation for the real-world data experiments were two-fold. First, to investigate whether we can learn meaningful latent event patterns within multiple groups (see Fig. 6.12). Second, to examine the approximation error, in terms of the mean Dice coefficient, as a function of r and w for two variations of the stochastic optimization scheme. We have studied two cases. In algorithm type I we set the maximum update iterations for both latent factors to 1, whereas in algorithm type II (green) we set let the $\mathbf{H}$ update converge in each iteration of the stochastic gradient descent walk (see Fig. 6.13).

We have used the following parameter settings

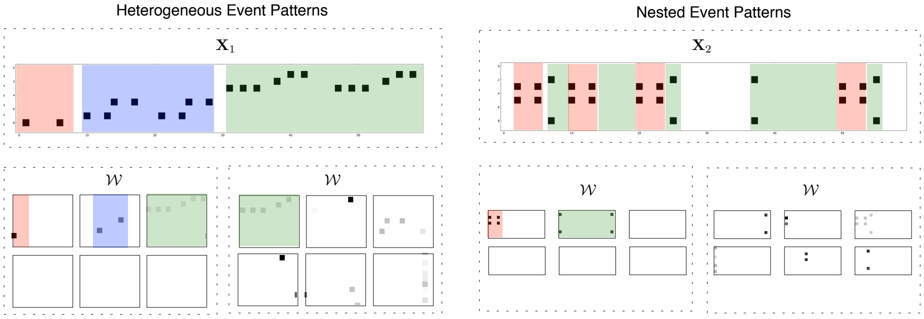


Figure 6.9: Learning heterogeneous and nested event patterns. The left and right box show examples of learning *heterogeneous* and *nested event patterns*. The bottom part of the left and right box show the learned latent patterns $\mathcal{W}$. We show two cases where the left bottom box shows the result of our over-complete latent factor model employing *total normalization*, and the right bottom box the case of *individual normalization*.

- $\beta = 0.5$ and $\lambda = 0.5$
- the random sampling parameter $iter = 100$
- convergence threshold of 1e-9
- the number of $\mathbf{H}$ updates were set to $iterH = \{1, 50\}$
- the number of $\mathbf{W}$ updates was $iterW = 1$
- $r = \{1, 5, 10, 50, 100, 200, 500, 1000, 5000, 10000\}$
- $w = \{3, 7, 14, 30\}$

6.5.6 Diabetic Complication Severity Index (DCSI)

The DCSI is a discrete 13-point scale scored from automated diagnostic, pharmacy, and laboratory data to quantify the severity of complications and to potentially bet-

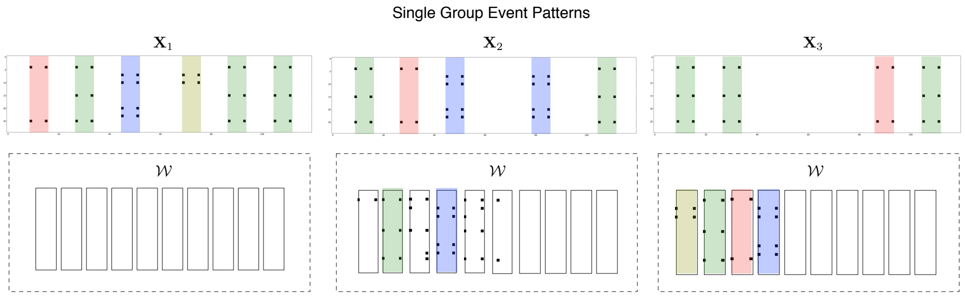


Figure 6.10: Stochastic optimization scheme for learning single group patterns on synthetic data.

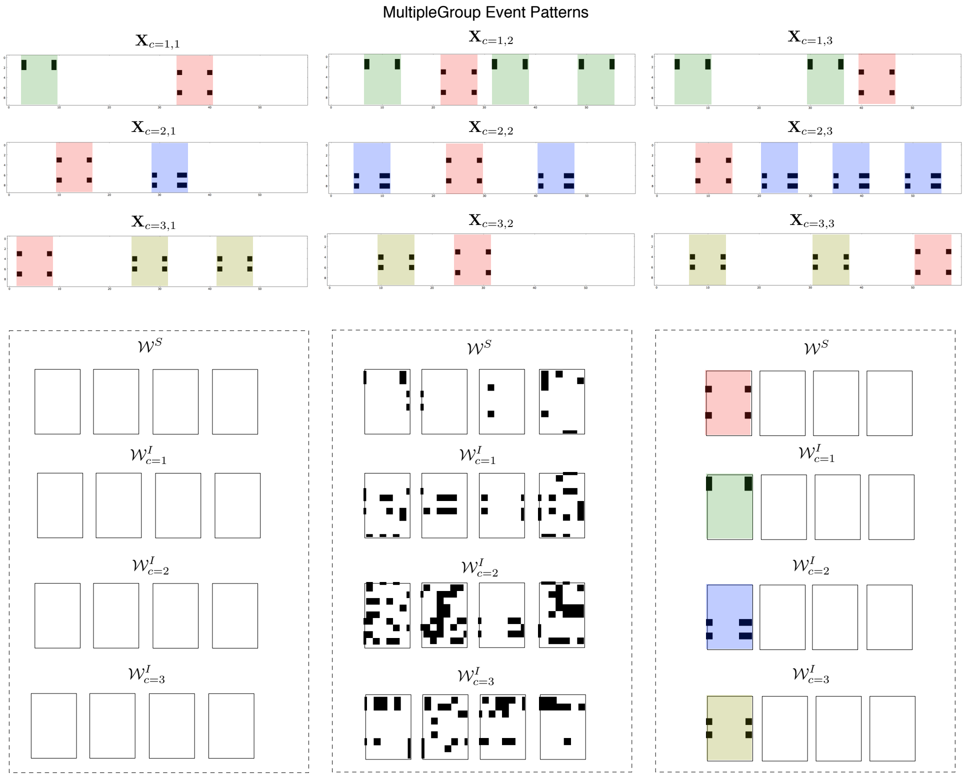


Figure 6.11: Stochastic optimization scheme for learning multiple group patterns on synthetic data.

ter predict the risk of adverse outcomes. Since number of diabetic complications and its severity are associated with greater risk of mortality and hospitalizations, DCSI can be used as a tool for adjusting for baseline severity of disease in populations

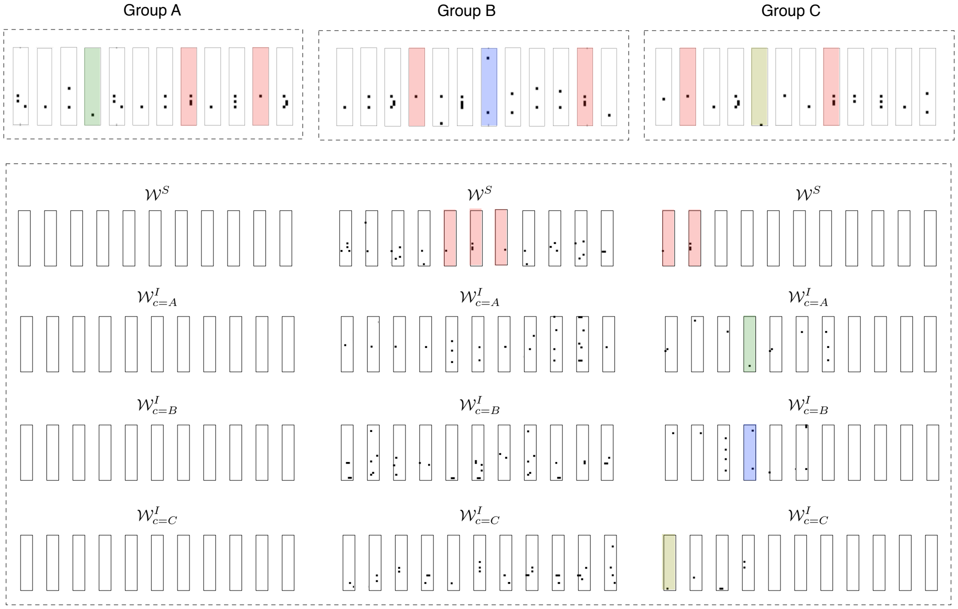


Figure 6.12: Stochastic optimization scheme for learning multiple group patterns on real-world data.

with diabetes. Young *et al.* (2008) [122] proposed DCSI as a better predictor for mortality and risk of hospitalization. The objective was to determine whether the number and severity of diabetes complications are associated with increased risk of mortality and hospitalizations. DCSI was developed from automated clinical baseline data of a primary care diabetes cohort and compared with a simple count of complications to predict mortality and hospitalizations. Cox proportional hazard and Poisson regression models were used to predict mortality and hospitalizations, respectively. Replacing a simple disease complications count index with the DCSI showed a similar mortality risk. Each level of the DCSI was associated with a 1.34-fold (95% CI = 1.28, 1.41) greater risk of death. Similar results were obtained for the association of the DCSI with risk of hospitalization. Comparison of receiver operating characteristic curves verified that the DCSI was a slightly better predictor of mortality than a count of complications. Compared with DCCI, DCSI performed slightly better and appears to be a useful tool for prediction of mortality and risk of hospitalization.

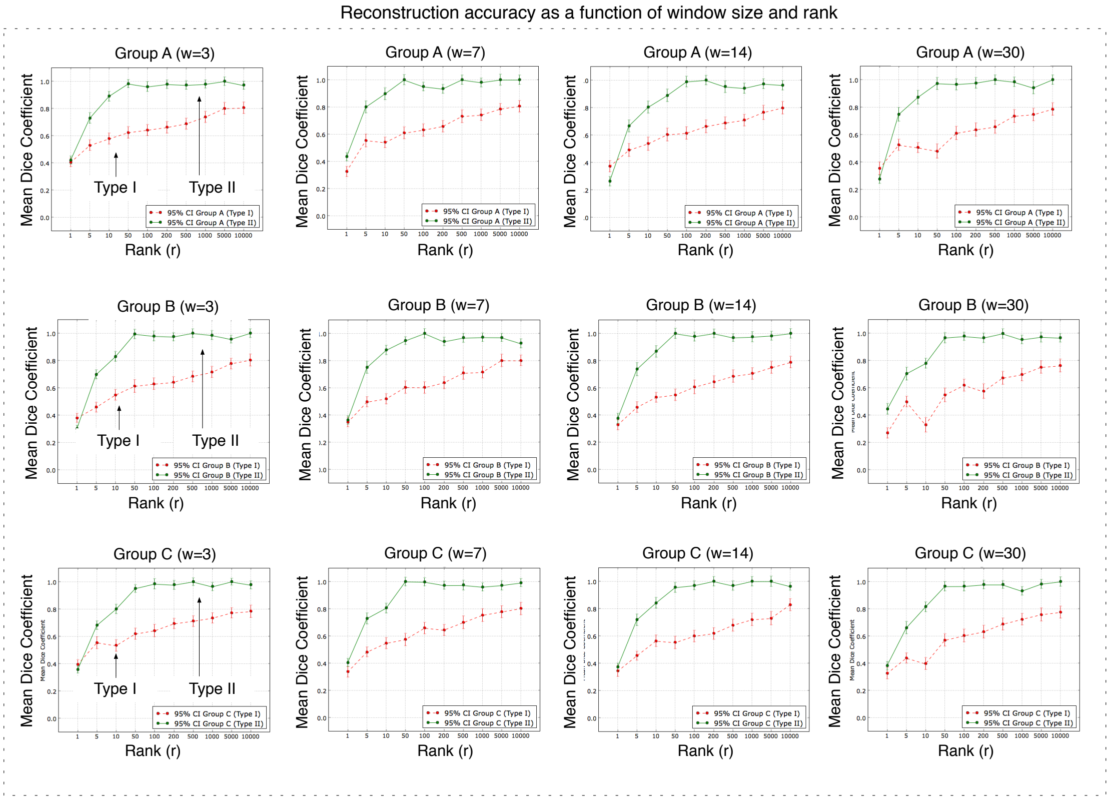


Figure 6.13: Algorithm comparison of the stochastic optimization scheme. Mean reconstruction performance and 95% confidence intervals for algorithm I (red) and II (green). Row 1: Mean Dice coefficient vs. different window sizes for group A. Row 2: Mean Dice coefficient vs. different window sizes for group B. Row 3: Mean Dice coefficient vs. different window sizes for group C.

We use the DCSI to stratify the three groups of our diabetic patient pool outlined in Section 6.5.1.1 and to use the obtained severity score as group labels to correlate against HRU patterns.

6.5.7 Linking HRU patterns to diabetic disease severity

We performed an exploratory analysis of the diabetic patient pool outlined in Section 6.5.1.1 to assess how patterns of healthcare resource utilization (HRU) relate to the severity of diabetic complications. We started by learning the latent event structure of each patient in the diabetic patient pool using our framework outlined in Section 6.4. We learned 30 weekly, bi-weekly, monthly, and quarterly temporal patterns for all patients. Then we performed pairwise feature matching within the latent temporal event pattern pool to identify the closest pattern matches between each pair of patients. We computed pairwise distances that were then weighted by the difference of their associated convolution coefficients to account for the different number of pattern occurrences within TEMR. Then, we computed a KNN-graph to examine the latent cluster structure of the mined latent patterns by looking at the Fiedler vector. Different cluster groups were computed together with the DSCI score for each patient. Finally, we generated a histogram that captured the patient distribution in each cluster. We performed visual examination of the patient distribution based on their severity level to look for group specific differences. In Fig. 6.14 we show four computed clusters and their patient distribution based on the DCSI severity index.

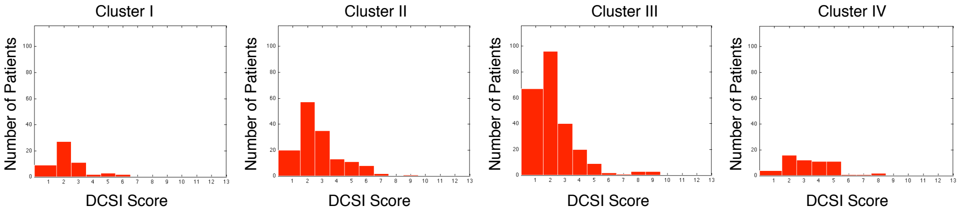


Figure 6.14: HRU pattern groups vs diabetic disease severity level. Histograms that show the number of patients within each computed cluster vs. their DCSI severity score.

6.6 Discussion

In Fig. 6.7 one can observe that the algorithm converges within 50 iterations for all different model parameters. The approximation error measured in terms of the Dice coefficient and the Frobenius norm exponentially increased and decreased as the rank was increased. For $k > 10$, different rank sizes had an overall approximation error above $D_c > 0.9$ and $|| \cdot ||_F < 2.5$. The reconstruction performance showed that the algorithm is robust against varying window sizes and the sparsity parameter for $\lambda > 0$. The effect on different sparsity constraints showed that the mean convergence is not indicative of a low approximation error. The best model was achieved with a sparsity constraint of $\lambda = 0.5$ and a $\beta = 0.5$. Setting the sparsity constraint to $\lambda = 0$ led to a very low Dice coefficient. Also setting $\beta = 2$ gave the lowest Dice coefficient showing that the Frobenius loss is not able to cope with double sparsity. We summarize the optimal model parameters with respect to the computed performance metrics in Table 1. One can see that the convergence criterion should not be considered as a cross-validation measure. The optimal mean Dice coefficient and mean ℓ_2-norm both gave the same optimal model parameters, whereas the parameters for the convergence criterion disagreed. In general, the framework shows robustness with respect to the chosen window size and the sparsity parameter. This is encouraging, since learning patterns of different window sizes is important for extracting a rich event structure within TEMR. Also the optimal parameterization of the β-divergence with $\beta = 0.5$ shows that it outperforms the Itakura-Saito and generalized KL divergence.

Fig. 6.8 shows the results for the second set of experiments. One can observe that the algorithm successfully learned the correct bases set even though the rank was specified to be over-complete. The sparse code ($\mathbf{H}$) and the sparse bases ($\mathbf{W}$) that were learned from synthetic datasets showed interpretable shift-invariant sparse activation patterns. By looking at the activation codes one exactly knows when a particular latent temporal pattern occurred in the data. Also the induced sparsity constraints on the latent factor model in conjunction with the non-negativity constraints enable easy interpretation of the model. The experimental results demonstrate that our framework is able to learn shift-invariant latent event patterns of different complexity. Note that the patterns implicitly encode missing event values as no event activity is simply encoded with zeros within TEMR. Where as other languages address the missing value problem by modeling event intervals instead of single events our representation can jointly encode both types of event structure. In general complex high-order event structure can be represented and learned within our framework. One can simply imagine to add additional rows to TEMR that encode single events, event intervals, and event sequences.

In Fig. 6.9 we can observe that our framework can successfully learn the heterogeneous event structure of the synthetic dataset. The top box shows the synthetic dataset, the bottom left(right) box the learned heterogeneous temporal event patterns of our algorithm with total(individual) normalization respectively. In this synthetic example, the patient received three types of diabetic treatments A (red), B (blue), and C (green), where treatment A consisted of single patient encounters, treatment B of two consecutive encounters separated by a time interval, and treatment C of a more complicated structure. We have used an over-complete specified rank (r=6) to simulate the situation where the known latent event structure is not known in advance. The bottom left box shows that our framework is able to learn the true heterogeneous event structure, whereas the bottom right box contained small artifacts and a hard-to-interpret pattern structure. As in the first synthetic example the bottom left/right boxes show the learned nested event patterns. We show results on learning *nested event patterns.* The pattern consists of a two level nested structure, where the green area is the first level, and the red area the second level. The top box shows the synthetic dataset that simulates the following scenario. The patient received two consecutive types of treatment A (first green box) and treatment B (second green box). Within each treatment another treatment type C (red boxes) happened outside (first red box) and within (red boxes encapsulated by the green boxes) treatment A and B.

In Fig. 6.10 we show that the stochastic optimization scheme is able to learn an over-complete sparse latent factor model from a group of multiple synthetic spatial-temporal point processes. Different factorization ranks (r=5,10) resulted in the same learned latent basis elements demonstrating that the framework can learn interpretable shift-invariant latent temporal patterns even though the true latent event structure is not known in advance. This property is particular advantageous for real data as one does not know how many true latent patterns exist in the data. The convergence plot shows that our algorithm converges.

In Fig. 6.11 we show that we can learn the synthetic multi group patterns. The upper three blocks show the synthetic datasets for the three groups. The lower blocks show the learned group patterns for three different parameter settings. The left block shows empty event group patterns, which shows that the Euclidean norm without sparsity constraints is not able to learn meaningful event group patterns. The middle block shows the learned group patterns when using the beta-divergence with $\beta = 0.5$ and an individual normalization scheme. One can see that the learned patterns contain noise artifacts. The right block shows the learned group patterns when using the total normalization scheme. One can observe that the algorithm successfully learned the shared group pattern, which is marked with a red translucent box. Accordingly the other three group specific event patterns marked with green, blue, and olive boxes, could also be discovered by the algorithm. Learning the group event structure is important for doing patient group analysis to assess patient similarity and to perform automated patient group stratification.

In Fig. 6.12 we show that we can also learn meaningful multi group pattern in real-world data. The top boxes show examples of true real-world event patterns. In total three groups were considered as explained in Section 6.5.1.1. The bottom boxes from left to right show the discovered patterns by our algorithm. The parameter settings are the same as previously described Fig. 6.11. Note the red box indicating the shared event pattern that exists in all three groups. The green, blue, and olive box show individual group specific event patterns. One can also observe that a single run did not discover all latent event patterns, which is expected as the solution to our algorithm is local. To circumvent this, one could perform multiple runs of the algorithm to search for shared and individual event group patterns.

In Fig. 6.13 we show the reconstruction behavior of the stochastic optimization scheme for two different algorithmic configurations. For all three groups A, B, and C, algorithm type II (in green) outperformed algorithm type I (in red). Algorithm type I showed a linear increase of the mean Dice coefficient whereas algorithm type II an exponential increase as the rank was increased. This observation also holds for different window sizes showing that the reconstruction performance is robust against the window size and the number of basis elements (rank). For all three groups the stochastic optimization scheme could learn a latent factor model that led to a mean Dice coefficient close to 1. The 95% confidence interval showed that the computed means were representative of the three population groups. Visual examination of the learned patterns also confirmed that the algorithm could learn interpretable latent event patterns for all three groups.

Fig. 6.14 shows an example of a four-cluster partitioning of a random subset of our diabetic patient population. One can infer that the identified patterns in cluster IV mostly occur in groups of patients with a high DCSI score. Taking a closer look to cluster IV one can see the low number of patients with a low severity score (i.e., 1) in contrast to the overall histogram shape. The majority of patients in cluster IV exhibit a higher DCSI score and thus have higher risk of hospitalization and mortality. Cluster II and III show similar shapes of the overall histogram indicating that the learned patterns within these patient groups mainly consist of common HRU patterns that are not indicative of disease severity. The longer right tail of the histogram can be explained by the rarity of patients who have a very high DCSI score. We note that one can go back to the individual patterns to investigate what kind of care the patients received.

One drawback of our approach is the problem of permutation invariance across the rows of TEMR. Permuting the rows of TEMR may lead to different latent event patterns and thus different event relationships. Though we note that we can circumvent the problem by using a temporal pattern window that spans all the rows of the TEMR matrix as shown in the experiments section. In this case learning latent temporal patterns on different row permutations would visually result in different patterns, but the shift-invariant model would still be able to learn their latent relationships.

We note that the chosen temporal pattern window is an important factor in obtaining semantically meaningful latent event patterns. In cases where the event pattern window length was incorrectly chosen the true latent event patterns could not be recovered though the learned latent factor model was able to reconstruct the original data matrix. Also the double sparsity constraint is important for obtaining meaningful and interpretable latent event patterns. When using the sparsity constraint in combination with the standard Euclidean norm our model failed in learning the latent event structure of the data matrix. In some cases it was necessary to repeat the learning process to obtain meaningful patterns due to the non-convex optimization problem.

6.7 Conclusion

In this chapter we have presented a novel temporal event matrix representation and learning framework in conjunction with an in-depth validation of over 40,000 learned latent factor models.

We have demonstrated that our proposed framework is able to cope with the double sparsity problem and that the induced double sparsity constraint on the β-divergence enables automatic relevance determination for solving the optimal rank selection problem via an over-complete sparse latent factor model. Further, the framework is able to learn shift-invariant high-order latent event patterns in large-scale data such as latent temporal *event operators, concepts, and time constraints*, individual and common event group structure, as well as *heterogeneous* and *nested temporal event patterns*. We empirically showed that our stochastic optimization scheme converges to a fixed point. We applied our framework to build patient-specific signatures that capture individual characteristics of the patient and a fingerprint of the medical care history they received.

Our representation and learning framework is commensurate with human capabilities and constraints, since the latent temporal patterns are interpretable and easy to comprehend. The developed analytics have wide applicability to a variety of data and application domains that involve large-scale longitudinal event data. Future work will be devoted to a thorough clinical assessment for visual interactive knowledge discovery in large electronic health record databases.

Part III

Summary and conclusion

Chapter 7

Summary

7.1 Significance

The goal of this work is to improve the capabilities of medical practitioners to efficiently cope with large, complex, and heterogeneous data sources–a diverse and challenging problem. To this end, the dissertation proposes a diverse analytical framework that leverages synergistic human-machine intelligence to maximize a human's ability to better 1) reason about, 2) learn, and 3) understand biomedical imaging and healthcare specific event data within the patient's EHR. Within this scope we have targeted a range of specific application scenarios where the analysis by machines or humans alone is difficult, time-consuming, and error-prone. By combining human and machine intelligence the developed analytics in this dissertation focus and exploit the synergistic aspect of combined human-machine intelligence to better 1) visualize, 2) label, and 3) discover knowledge from large, complex, and heterogeneous data.

Chapter 2 presents analytics for human-assisted visualization of complex latent tree structures within large volumetric images. We develop a novel algorithm that enables the intuitive exploration of complete vascular trees and their internal volume structure. Comparative validation of our algorithm with the state-of-the-art demonstrates superior performance in terms of visualization quality and preservation of anatomical shape appearance. The ability to visualize complex vascular networks within the human body not only enables us to better reason about large complex images and their sparse information content, but is also important for a variety of diagnostic procedures that involve the examination of the vascular network. Branching tree-like networks are found throughout the human body at multiple scales and locations ranging from the micro-to-macro scale. The developed analytics could be transferred and applied to such data and problem domains. Diseases with complex co-morbidities might be related to the health condition of the vascular system. Not to mention diseases that directly affect the vascular system, such as diabetes, require efficient tools to access and visualize diagnostic relevant information for correct data interpretation. Whereas visualization enables the subjective assessment to better interpret and reason about complex data we also need quantitative information that help us to better learn about the hidden relationships of diagnostic relevant patterns.

In this regard, we describe a variety of interactive analytics for human-assisted automated labeling of object boundaries in unimodal, multimodal, and spatio-temporal image data. Our methods described in Chapter 3 enable the labeling of objects that exhibit high variability in shape, intensity, and texture–a scenario where fully automatic methods fail to perform in a robust and accurate manner. We show as part of a large evaluation study to quantify geographic atrophy, that our algorithm improves upon a state-of-the-art graph-based interactive labeling algorithm. In Chapter 4 we propose extensions to the naive Bayes algorithm within a transductive learning and inference paradigm. We introduce a novel semi-parametric form of the naive Bayes algorithm in combination with a Markov random field model. The algorithm enables automated object and multi-object labeling with minimal human intervention. In numerous experiments we show the performance of the algorithm on unimodal, multimodal, and spatio-temporal data comprising images and volumes. We also demonstrate that the algorithm generalizes to different data sources and application domains. In Chapter 5 we present our initial investigations to employ deep learning and inference algorithms to automate anatomical labeling of human brain image volumes based on manually labeled data provided by a human expert.

The developed analytics in Chapter 3 - 5 aim to address two possible scenarios that arise in clinical practice. The first case where label information is scarce and the second case where we have rich human expert provided labels to exploit. In the former case, our developed analytics based on Bayesian transduction provide a practical solution to quickly generate ground truth label information to build annotated medical image databases for a variety of disease phenotypes with minimal human intervention while exploiting human-expert knowledge for a variety of different disease phenotypes. In the latter case our preliminary studies on deep learning and inference explore technical avenues that allow us to go one step further in developing truly automatic labeling analytics that better generalize to different data sources and existing label information. Moreover, the integration of our visual analytics described in Chapter 2 with the labeling analytics described in Chapter 3 - 5 provide an analytical framework that is commensurate to humans. Though we demonstrate our framework on specific applications the general nature of the developed analytics and their hybrid combination can be easily extended or adopted to different application scenarios. This provides a rich set of tools to visualize and label a diverse set of data sources within the electronic health record.

Finally, to integrate the previously described image analytics with exploratory analytics that make use of ancillary healthcare specific data sources we propose extensions to the nonnegative matrix factorization algorithm in Chapter 6. We present a novel temporal event matrix representation and learning scheme to perform event pattern mining in longitudinal heterogeneous EHRs. Specifically, we propose a double-constrained convolutional sparse coding framework to learn latent event patterns within a single and group data setting. This work is one of its first kind and enables the straightforward representation and mining of event data that could be derived from image data and heterogeneous healthcare specific data sources for groups of patients to perform mining at a population level.

Analyzing the EHR and its diverse data sources within and across patients ultimately requires a holistic approach. Holistic analysis integrates analytics for visualizing, labeling, and mining EHR data, which enables the discovery of hidden patterns at a scale and depth not possible when taking an individualistic approach. Especially complex disease that result in multiple co-morbidities would benefit from holistic analysis. This dissertation provides a diverse analytical framework that would allow the such analysis by addressing challenging data analysis issues that involve large, complex, and heterogeneous data sources.

7.2 Limitations

We would like to make some general remarks regarding the limitations of our research in a wider context, which was not discussed in the individual chapters. The contributions of this dissertation along with the developed software tools by no means represent a generic solution to the problem of making better use of the diverse heterogeneous data sources within the electronic health record. The scope of this problem is beyond this dissertation. In this work we have just scratched the surface of this problem. Also the topic of synergistic human-machine intelligence was studied from a practical point of view within specific clinical applications and from the perspective of single human-machine intelligence. While our methods are applicable and transferrable to other clinical problems and application domains, they do not provide an all-in-one solution. Our focus was to build data-driven analytics that are intuitive to humans while leveraging human intelligence, which in the medical domain is of utmost importance.

Another limitation of our work is the partially limited validation in terms of clinically relevant performance measures. The validation of algorithms in the medical domain is a time and resource intensive challenging endeavor. In the ideal scenario each component of our diverse analytical framework could have been validated in a more rigorous manner. The human-intuitive analytics for visualization of complex vascular tree structures could have been validated in a real clinical scenario in conjunction with user studies involving multiple experts to assess the efficacy of our methods. However, such studies are costly to perform. An alternative strategy to further validate our visualization methods is the combination with successive labeling tasks. Due to the complex shape topology of vascular networks in volumetric space the proposed visualization methods serve as an ideal interface to enable medical practitioners to quickly identify and mark regions of interest. One could use our developed system in combination with interactive labeling analytics to compare the efficacy of interactive human-assisted labeling of vascular networks based on usability performance measures. However, given the many possibilities to improve upon more rigorous validation procedures one should not forget that our initial validation results shown in Chapter 2 directly show the advantages of our method. Also our interactive labeling pipeline described in Chapter 3 could have been validated with respect to multiple graders using multi ROC studies as well as comparative performance experiments on non-medical data. While we have investigated such applications with positive results we kept the focus of our validation study to the medical domain. Another extension of our validation study could have been a more detailed comparative analysis with respect to different object types, noise levels, and other data ambiguities. The quantitative validation in Chapter 4 showed impressive labeling results on a variety of image modalities and application problems. As in Chapter 3, a more detailed account on label prediction with respect to the feature dimensionality, the effect on cross-modality dependences, and the comparison to existing inductive approaches would provide useful information about the limitations of our proposed labeling analytics. Due to the preliminary nature of Chapter 5 we skip our commentary on how a more rigorous validation strategy may look like. The discovery analytics in Chapter 6 were validated with extensive experiments. To further validate the efficacy and usefulness of our analytics one could implement a visual analytic prototype that allows medical practitioners to slice and dice the individual discrete event matrices to search for clinically relevant event patterns. As mentioned in the previous Chapters more rigorous user studies involving multiple experts with different levels of expertise would provide a more detailed account of the benefits of our proposed analytics. Ideally patient relevant validation studies should be designed right at the point of care to measure and assess the efficacy with respect to improved patient care, reduction of costs, and the minimization of mortality and the risk of hospitalization. On a more holistic perspective our proposed diverse analytic framework as a whole would be interesting. Yet, under the given resource constraints this aim was out of the scope of this dissertation work.

7.3 Remarks

Over the course of this research I have learned important lessons that I would like to share. First of all, through collaborations with medical practitioners I came to realize the importance of closely working with physicians and medical domain experts in order to work on problems that have true clinical significance and value for the patient. This collaborative exchange has taught me that any developed method should be an aid to the physician and should intuitively fit into the current workflows of clinical practice. When devising a novel visualization technique such as the one outlined in Chapter 2 close interaction with medical practitioners is required to ensure that the developed techniques will be positively received by the physician. At the end what matters is the practicality and usefulness of the developed tools in clinical practice.

Second of all, the task of automatic labeling is a tremendous challenge and I came to realize that in a practical clinical setting fully automatic analyses are dangerous to perform and require careful quality and testing. Especially in the medical domain and in situations where data characteristics change over time. In degenerative disease often times the disease process is not fully understood and the disease phenotype progresses in non-deterministic ways, which makes the generalization task of automatic methods difficult. Instead of tweaking the parameters of a fixed algorithm to perform labeling of objects contained in image data I realized that interactive human-intuitive approaches provide a robust and attractive alternative. Until true artificial intelligence reaches a performance that closely resembles human intelligence the combination of human-machine intelligence seems to be the right approach to address the challenges we have outlined in the introduction of this dissertation.

Chapter 8

Conclusion

This dissertation explores the question how human-machine intelligence can be leveraged for optimal protocols of visualization, labeling, and knowledge discovery in large, complex, and heterogeneous electronic health record data. We propose a diverse analytical framework to maximize a human's ability to better 1) reason about, 2) learn, and 3) understand biomedical imaging and healthcare data within the patient's electronic health record. By combining human and machine intelligence in a synergistic form we develop analytics that are commensurate to humans. This synergism improves the capabilities of medical practitioners to efficiently cope with the EHR and its large, complex, and heterogeneous data sources.

Potential future work can take on different venues. We are most intrigued by taking the proposed work to the next level. Instead of just looking at individual synergistic human-machine intelligence a more interesting challenge is the aspect of synergistic human-machine intelligence within a collective crowd setting. What we mean by that is instead of considering an individual human and his/her intelligence it is more interesting to exploit the synergism of collective intelligence from a group of people or groups of people and a cluster of machines. Such an approach would require interdisciplinary research that bridges and integrates concepts from research that studies collective social intelligence with distributed parallel computational learning machinery. Within this realm, designing and developing analytics that are commensurate to a group of people is a non-trivial problem both in terms of how to exploit social intelligence analytically and how to actually compute and communicate synthetic knowledge generated from a cluster of machines in a collective human-intuitive form. How can one communicate effectively collective human-machine intelligence to individuals and groups of people? What computational mechanisms can address the scale and complexity of such knowledge? How can one adapt the intelligence of the model to account for knowledge bias within the crowd? Before one can leverage the potential of social intelligence married with large-scale parallel distributed machine intelligence much ground work has to be performed both in terms of research and development. Novel ways of visualizing, labeling, and mining are required that support the paradigm of collective synergistic human-machine intelligence. In return, the possibilities of collective synergistic human-machine intelligence are endless and will impact our lives in ways we can hardly imagine today.

Bibliography

[1] R. Hochreiter, C. Wiesinger, and D. Wozabal, “Large-scale computational finance applications on the open grid service environment,” in *Proceedings of the Advances in Grid Computing - EGC 2005*, vol. 3470, 2005, pp. 433–433.

[2] D. K. Bhattacharyya and A. Das, “A new distributed algorithm for large data clustering,” in *Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents*, 2000, pp. 29–34.

[3] N. Aleksandrov and B. M. Hambly, “A dual approach to multiple exercise option problems under constraints,” *Mathematical Methods of Operations Research*, vol. 71, pp. 503–533, 2010.

[4] V. Schetinin, J. E. Fieldsend, D. Partridge, W. J. Krzanowski, R. M. Everson, T. C. Bailey, and A. Hernandez, “Estimating classification uncertainty of bayesian decision treetechnique on financial data,” in *Perception-based Data Mining and Decision Making in Economics and Finance*, 2007, pp. 155–179.

[5] A. Brabazon and M. O’Neil, “Evolving technical trading rules for spot foreign-exchange markets using grammatical evolution,” *Computational Management Science*, vol. 1, pp. 311–327, 2004.

[6] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? large scale analytics on factors impacting retweet in twitter network." IEEE Computer Society, 2010, pp. 177–184.

[7] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. Sen Sarma, R. Murthy, and H. Liu, "Data warehousing and analytics infrastructure at facebook," in *Proceedings of the International Conference on Management of Data*, 2010, pp. 1013–1020.

[8] D. Obradovic, S. Baumann, and A. Dengel, "A social network analysis and mining methodology for the monitoring of specific domains in the blogosphere," in *Proceedings of the ASONAM*, 2010, pp. 1–8.

[9] H. Gao, X. Wang, G. Barbier, and H. Liu, "Promoting coordination for disaster relief: from crowdsourcing to coordination," in *Proceedings of the 4th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, 2011, pp. 197–204.

[10] S. Peters, L. Denoyer, and P. Gallinari, "Iterative annotation of multi-relational social networks," in *Proceedings of the ASONAM*, 2010, pp. 96–103.

[11] N. Memon and R. Alhajj, "Introduction to the first issue of social network analysis and mining journal," *Social Network Analysis Mining*, vol. 1, no. 1, pp. 1–2, 2011.

[12] D. C. J. Ruben, "Data mining in healthcare: Current applications and issues," Master's thesis, Carnegie Mellon University, 2009.

[13] M. R. Kraft, K. C. Desouza, and I. Androwich, "Data mining in healthcare information systems: case study of a veterans' administration spinal cord injury population," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, 2003, p. 9.

[14] A. Olinsky and P. A. Schumacher, “Data mining for health care professionals: Mba course projects resulting in hospital improvements,” *International Journal of Business Intelligence Research*, vol. 1, no. 2, pp. 30–41, 2010.

[15] M. K. Obenshain, “Application of data mining techniques to healthcare data,” *Infection control and hospital epidemiology the official journal of the Society of Hospital Epidemiologists of America*, vol. 25, no. 8, pp. 690–695, 2004.

[16] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, “Data mining in healthcare and biomedicine: A survey of the literature,” *Journal of Medical Systems*, pp. 1–18, 2011.

[17] H. Kaur and S. K. Wasan, “Empirical study on applications of data mining techniques in healthcare,” *Journal of Computer Science*, vol. 2, no. 2, pp. 194–200, 2006.

[18] D. H. Pari, B. Z. Arkady, K. Shonali, and M. G. Mohamed, “Mobile data mining for intelligent healthcare support,” in *Proceedings of the 42nd Hawaii International Conference on System Sciences*, 2009, pp. 1–10.

[19] P. Lucas, “Bayesian analysis, pattern analysis, and data mining in health care,” in *Proceedings of the Current Opinion Critical Care*, 2004, pp. 399–403.

[20] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva, “The diverse and exploding digital universe - an updated forecast of worldwide information growth through 2011,” in *An IDC White Paper - Sponsered by EMC*, 2008.

[21] T. Bosse, Z. A. Memon, and J. Treur, “A recursive bdi agent model for theory of mind and its applications,” *Applied Artificial Intelligence*, vol. 25, no. 1, pp. 1–44, 2011.

[22] W. Chung, S. Kim, M. Choi, J. Choi, H. Kim, C.-b. Moon, and J.-B. Song, “Safe navigation of a mobile robot considering visibility of environment,” *IEEE Transactions on Industrial Electronics*, vol. 56, no. 10, pp. 3941–3950, 2009.

[23] A. Ryan, D. George, and M. Iain, “Incorporating side information in probabilistic matrix factorization with gaussian processes,” in *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, 2010, pp. 1–9.

[24] Q. Liu and A. Ihler, “Negative tree-reweighted belief propagation,” in *Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, CA, USA, 2010.

[25] M. Tan, L. Wang, and I. W. Tsang, “Learning sparse svm for feature selection on very high dimensional datasets,” in *Proceedings of the International Conference on Machine Learning ICML’10*, 2010, pp. 1047–1054.

[26] M. Cummins and P. Newman, “Fab-map: Probabilistic localization and mapping in the space of appearance,” *Int. J. Rob. Res.*, vol. 27, pp. 647–665, 2008.

[27] R. Salakhutdinov, “Learning deep boltzmann machines using adaptive mcmc,” in *Proceedings of the International Conference on Machine Learning ICML’10*, 2010, pp. 943–950.

[28] J. Chen, J. Liu, and J. Ye, “Learning incoherent sparse and low-rank patterns from multiple tasks.” in *Proceedings of KDD’10*, 2010, pp. 1179–1188.

[29] H. Fei and J. Huan, “Boosting with structure information in the functional space: an application to graph classification.” in *Proceedings of KDD’10*, 2010, pp. 643–652.

[30] V. Leroy, B. B. Cambazoglu, and F. Bonchi, “Cold start link prediction.” in *Proceedings of KDD’10*, 2010, pp. 393–402.

[31] Y. Bengio and Y. LeCun, “Scaling learning algorithms towards AI,” in *Large-Scale Kernel Machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, Eds. MIT Press, 2007.

[32] H. Lee and S. Choi, “Group nonnegative matrix factorization for eeg classification,” *Journal of Machine Learning Research*, vol. 5, pp. 320–327, 2009.

[33] N. Lee and M. Rasch, “Tangential curved planar reformation for topological and orientation invariant visualization of vascular trees.” in *Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, 2006, pp. 1073–1076.

[34] J. J. Caban, N. Lee, S. Ebadollahi, A. F. Laine, and J. R. Kender, “Concept detection in longitudinal brain mr images using multi-modal cues,” in *Proceedings of the 6th IEEE international conference on Symposium on Biomedical Imaging: From Nano to Macro*, 2009, pp. 418–421.

[35] N. Lee, A. F. Laine, S. Ebadollahi, and R. L. DeLaPaz, “Bayesian transduction and markov conditional mixtures for spatiotemporal interactive segmentation,” in *Proceedings of the 4th International IEEE EMBS Conference on Neural Engineering*, 2009, pp. 226–229.

[36] N. Lee, R. T. Smith, and A. F. Laine, “Interactive segmentation for geographic atrophy in retinal fundus images,” in *Proceedings of the Asilomar Conference on Signals, Systems, and Computers*, 2008, pp. 655–658.

[37] N. Lee, A. F. Laine, and R. T. Smith, “Bayesian transductive markov random fields for interactive segmentation in retinal disorders,” in *Proceedings of the World Congress on Medical Physics and Biomedical Engineering*, vol. 25/11, 2009, pp. 227–230.

[38] N. Lee and A. Klein, “Towards a deep learning approach to brain parcellation,” in *Proceedings of the 8th International Symposium on Biomedical Imaging (ISBI11). To Appear.*, 2011.

[39] N. Lee, J. Hu, F. Wang, S. Jimeng, S. Ebadollahi, and A. Laine, “A temporal event matrix representation and learning framework for event pattern mining.” *IEEE Transactions on Pattern Analysis and Machine Intelligence. Submitted.*, 2011.

[40] F. Wang, N. Lee, S. Jimeng, J. Hu, and S. Ebadollahi, “One-side convolutional nonnegative matrix factorization for pattern discovery on longitudinal clinical record. submitted.” in *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[41] N. Lee, A. Laine, J. Hu, F. Wang, S. Jimeng, and S. Ebadollahi, “Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients. accepted.” in *Proceedings of the First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology. To Appear.*, 2011.

[42] N. Lee, A. F. Laine, and R. T. Smith, “A hybrid segmentation approach for geographic atrophy in fundus auto-fluorescence images for diagnosis of age-related macular degeneration.” in *Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, 2007, pp. 4965–4968.

[43] N. Lee, A. F. Laine, and R. Smith, “Coarse to fine segmentation of stargardt rings using an expert guided dual ellipse model,” in *Proceedings of the International Conference of IEEE Engineering in Medicine and Biology Society*, 2008, pp. 2250–2253.

[44] N. Lee, A. F. Laine, and T. R. Smith, “Learning non-homogenous textures and the unlearning problem with application to drusen detection in retinal images,”

in *Proceedings of the 5th IEEE International Symposium on Biomedical Imaging From Nano to Macro (ISBI)*, 2008, pp. 1215–1218.

[45] N. Lee, J. Wielaard, A. A. Fawzi, P. Sajda, A. F. Laine, G. Martin, M. S. Humayun, and R. T. Smith, "In vivo snapshot hyperspectral image analysis of age-related macular degeneration," in *Proceedings of the Annual International Conference of the IEEE*, 2010, pp. 5363–5366.

[46] N. Lee, A. F. Laine, G. Marquez, J. M. Levsky, and J. K. Gohagan, "Potential of computer-aided diagnosis to improve ct lung cancer screening," *IEEE Reviews in Biomedical Engineering (RBME)*, vol. 2, pp. 136–146, 2009.

[47] R. Raman, S. Napel, C. F. Beaulieu, E. S. Bain, R. B. Jeffrey, and G. D. Rubin, "Automated generation of curved planar reformations from volume data: method and evaluation," *Radiology*, vol. 223, no. 1, pp. 275–80, 2002.

[48] A. Kanitsar, D. Fleischmann, R. Wegenkittl, P. Felkel, and E. Groeller, "Cpr-curved planar reformation," in *Proceedings of the 13th IEEE Visualization 2002 (VIS'02)*, 2002, pp. 37–44.

[49] A. Kanitsar, R. Wegenkittl, D. Fleischmann, and M. E. Groeller, "Advanced curved planar reformation: flattening of vascular structures," in *Proceedings of the 14th IEEE Visualization 2003 (VIS'03)*, 2003, pp. 43–50.

[50] L. Saroul, S. Gerlach, and R. D. Hersch, "Exploring curved anatomic structures with surface sections," in *Proceedings of the 14th IEEE Visualization 2003 (VIS'03)*, 2003, pp. 27–34.

[51] T. Vrtovec, B. Likar, and F. Pernus, "Automated curved planar reformation of 3d spine images," vol. 50, no. 19, pp. 4527–40, 2005.

[52] J. S. Sunness, "The natural history of geographic atrophy, the advanced atrophic form of age-related macular degeneration," *Molecular Vision*, vol. 5, pp. 25–9, 1999.

[53] F. C. Delori, C. K. Dorey, G. Staurenghi, O. Arend, D. G. Goger, and J. J. Weiter, "In-vivo fluorescence of the ocular fundus exhibits retinal pigment epithelium lipofuscin characteristics," *Investigative Ophthalmology and Visual Science*, vol. 36, no. 3, pp. 718–29, 1995.

[54] F. C. Delori, M. R. Fleckner, D. G. Goger, J. J. Weiter, and C. K. Dorey, "Autofluorescence distribution associated with drusen in age-related macular degeneration," *Investigative Ophthalmology and Visual Science*, vol. 41, no. 2, pp. 496–504, 2000.

[55] A. Bindewald, A. C. Bird, S. S. Dandekar, J. Dolar-Szczasny, J. Dreyhaupt, F. W. Fitzke, W. Einbock, F. G. Holz, J. J. Jorzik, C. Keilhauer, N. Lois, J. Mlynski, D. Pauleikhoff, G. Staurenghi, and S. Wolf, "Classification of fundus autofluorescence patterns in early age-related macular disease," *Investigative Ophthalmology and Visual Science*, vol. 46, no. 9, pp. 3309–14, 2005.

[56] J. C. Hwang, J. W. K. Chan, S. Chang, and R. T. Smith, "Predictive value of fundus autofluorescence for development of geographic atrophy in age-related macular degeneration," *Investigative Ophthalmology and Visual Science*, vol. 47, no. 6, pp. 2655–61, 2006.

[57] L. Grady, "Random walks for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1768–83, 2006.

[58] N. Lee, A. F. Laine, and R. Smith, "Coarse to fine segmentation of stargardt rings using an expert guided dual ellipse model," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society*, 2008, pp. 2250–3.

[59] V. Grau, a. U. J. Mewes, M. Alcañiz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 447–58, 2004.

[60] E. N. Mortensen and W. A. Barrett, "Interactive segmentation with intelligent scissors," *Graphical Models and Image Processing*, vol. 60, no. 5, pp. 349–384, 1998.

[61] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[62] S. Osher and N. Paragios, *Geometric level set methods in imaging, vision, and graphics.* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2003.

[63] D. M. Greig, B. T. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society Series B*, vol. 51, no. 2, pp. 271–279, 1989.

[64] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images," in *Proceedings 8th IEEE International Conference on Computer Vision (ICCV)*, 2001, pp. 105–112.

[65] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[66] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[67] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient n-d image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.

[68] L. Grady, “Multilabel random walker image segmentation using prior models,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 763–770.

[69] L. Grady and A. K. Sinop, “Fast approximate random walker segmentation using eigenvector precomputation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[70] S. Beucher and C. Lantuejoul, “Use of watersheds in contour detection,” in *Proceedings of the International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation*, 1979, pp. 2.1–2.12.

[71] L. Vincent and P. Soille, “Watersheds in digital spaces: An efficient algorithm based on immersion simulations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

[72] F. Meyer and S. Beucher, “Morphological segmentation,” *Journal of Visual Communication and Image Representation*, vol. 1, no. 1, pp. 21–46, 1990.

[73] J.-F. Rivest, “Marker-controlled segmentation: An application to electrical borehole imaging,” *Journal of Electronic Imaging*, vol. 1, no. 2, p. 136, 1992.

[74] F. Meyer, *Minimum spanning forest for morphological segmentation*, J. Serra and P. Soille, Eds. Kluwer Academic Publisher, 1994.

[75] R. Lotufo and W. Silva, “Minimal set of markers for the watershed transform,” in *Proceedings of ISMM 2002*, 2002, pp. 359–368.

[76] R. Lotufo and A. Falcao, “The ordered queue and the optimality of the watershed approaches,” in *Proceedings of the Mathematical Morphology and its Applications to Image and Signal Processing*, 2000, pp. 341–350.

[77] W. E. Higgins and E. J. Ojard, “Interactive morphological watershed analysis for 3d medical images,” *Computerized Medical Imaging and Graphics*, vol. 17, no. 4-5, pp. 387–395, 1993.

[78] H. K. Hahn and H.-O. Peitgen, “Interactive watershed transform: A hierarchical method for efficient interactive and automated segmentation of multidimensional gray-scale images,” in *Proceedings of Medical Imaging, SPIE*, vol. 5032, 2003, pp. 643–653.

[79] F. C. Flores and R. D. A. Lotufo, “Watershed from propagated markers: An interactive method to morphological object segmentation in image sequences,” *Image and Vision Computing*, vol. 28, no. 11, pp. 1491–1514, 2010.

[80] B. Klava, N. Sumiko, and T. Hirata, “Interactive image segmentation with integrated use of the markers and the hierarchical watershed approaches,” in *Proceedings of the International Conference on Computer Vision Theory and Applications - VISAPP 2009*, 2009, pp. 1–6.

[81] C. Couprie, L. Grady, L. Najman, and H. Talbot, “Power watersheds: A new image segmentation framework extending graph cuts, random walker and optimal spanning forest,” in *Proceedings of the IEEE 12th International Conference on Computer Vision*, 2009, pp. 731–738.

[82] J. B. T. M. Roerdink and A. Meijster, “The watershed transform : definitions , algorithms and parallelization strategies,” *Fundamenta Informaticae*, vol. 41, pp. 1–40, 2001.

[83] P.-E. Danielsson, *Generalized and separable Sobel operators: Machine vision for three-dimensional scenes*, H. Freeman, Ed. Academic Press, 1990.

[84] M. Gokstorp and P.-E. Danielsson, “Velocity tuned generalized sobel operators for multiresolution computation of optical flow,” in *Proceedings of 1st International Conference on Image Processing*, 1994, pp. 765–769.

[85] K. Branson, “A naive bayes classifier using transductive inference for text classification. unpublished.” 2001.

[86] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, “Text classification from labeled and unlabeled documents using em,” *Machine Learning*, vol. 39, no. 2, pp. 103–134, 2000.

[87] D. Lowd and P. Domingos, “Naive bayes models for probability estimation,” in *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 529–536.

[88] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society Series B Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[89] S. Geman and D. Geman, “Stochastic relaxation, gibbs distributions and the bayesian restoration of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 6, pp. 721–741, 1984.

[90] J. W. Bohland, H. Bokil, C. B. Allen, and P. P. Mitra, “The brain atlas concordance problem: quantitative comparison of anatomical parcellations,” *PloS One*, vol. 4, no. 9, p. e7200, 2009.

[91] A. Gholipour, N. Kehtarnavaz, R. Briggs, M. Devous, and K. Gopinath, “Brain functional localization: A survey of image registration techniques,” *Medical Imaging, IEEE Transactions on*, vol. 26, no. 4, pp. 427 –451, 2007.

[92] J.-F. Mangin, V. Frouin, I. Bloch, J. Régis, and J. López-Krahe, “From 3d magnetic resonance images to structural representations of the cortex topography using topology preserving deformations,” *J. Math. Imaging Vis.*, vol. 5, pp. 297–318, 1995.

[93] M. R. Sabuncu, B. T. T. Yeo, K. Van Leemput, B. Fischl, and P. Golland, "Nonparametric mixture models for supervised image parcellation," in *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, vol. 12, no. WS, 2009, pp. 301–313.

[94] D. Lashkari, R. Sridharan, E. Vul, P.-J. Hsieh, N. Kanwisher, and P. Golland, "Nonparametric hierarchical bayesian model for functional brain parcellation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 15 –22.

[95] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren, R. P. Woods, J. J. Mann, and R. V. Parsey, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.

[96] A. Klein, S. S. Ghosh, B. Avants, B. T. T. Yeo, B. Fischl, B. Ardekani, J. C. Gee, J. J. Mann, and R. V. Parsey, "Evaluation of volume-based and surface-based brain image registration methods," *NeuroImage*, vol. 51, no. 1, pp. 214–220, 2010.

[97] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, 1998, pp. 2278–2324.

[98] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, and Y. Bengio, "Theano: a cpu and gpu math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

[99] J. Shlens, G. D. Field, J. L. Gauthier, M. Greschner, A. Sher, A. M. Litke, and E. J. Chichilnisky, "The structure of large-scale synchronized firing in primate

retina," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 29, no. 15, pp. 5022–31, 2009.

[100] Y. R. R. Kumar and A. Govardhan, "Stock market predictions - integrating user perception for extracting better prediction - a framework," *International Journal of Engineering Science*, vol. 2, no. 7, pp. 3305–3310, 2010.

[101] R. A. Russell, "Mobile robot learning by self-observation," *Autonomous Robots*, vol. 16, no. 1, pp. 81–93, 2004.

[102] L. Xie, H. Sundaram, and M. Campbell, "Event mining in multimedia streams," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 623–647, 2008.

[103] M. Dong, "A tutorial on nonlinear time-series data mining in engineering asset health and reliability prediction: Concepts, models, and algorithms," *Mathematical Problems in Engineering*, vol. 2010, pp. 1–23, 2010.

[104] C. Fevotte and J. Idier, "Algorithms for non-negative matrix factorization with the beta-divergence," pp. 1–20, 2010.

[105] P. D. O'Grady and B. A. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity," in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 520–527.

[106] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03*, 2003, pp. 2–11.

[107] F. Mörchen, "Time series knowledge mining," Ph.D. dissertation, Philipps-University Marburg, 2006.

[108] F. Mörchen and A. Ultsch, "Efficient mining of understandable patterns from multivariate interval time series," in *Proceedings of Data Mining and Knowledge Discovery*, 2007, pp. 181–215.

[109] F. Mörchen and D. Fradkin, "Robust mining of time intervals with semi-interval partial order patterns," in *Proceedings SIAM Conference on Data Mining*, 2010, pp. 315–326.

[110] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no. 1, pp. 23–35, 1997.

[111] D. D. Lee and S. H. Seung, "Unsupervised learning by convex and conic coding," in *Proceedings of Advances in Neural Information Processing Systems*, 1997, pp. 515–521.

[112] ——, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[113] ——, "Algorithms for non-negative matrix factorization," in *Proceedings of Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[114] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of the IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557–565.

[115] J. Eggert, "Sparse coding and nmf," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 2, no. 4, 2004, pp. 2529–2533.

[116] J. Eggert, H. Wersing, and E. Korner, "Transformation-invariant representation and nmf," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, vol. 4, 2004, pp. 2535–2539.

[117] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proceedings of the 4th International*

Conference on Independent Component Analysis and Blind Signal Separation, vol. 3195, 2004, pp. 494–499.

[118] B. Cao, D. Shen, J. Sun, X. Wang, Q. Yang, and Z. Chen, "Detect and track latent factors with online nonnegative matrix factorization," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 2007, pp. 2689–2694.

[119] J. Mairal, F. Bach, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[120] W. Fei, L. Ping, and K. Christian, "Online nonnegative matrix factorization for document clustering," in *Proceedings of the 11th SIAM International Conference on Data Mining (SDM). To Appear.*, 2011.

[121] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S.-i. Amari, "Non-negative tensor factorization using alpha and beta divergences," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007, pp. 1393–1396.

[122] B. A. Young, E. Lin, M. Von Korff, G. Simon, P. Ciechanowski, E. J. Ludman, S. Everson-Stewart, L. Kinder, M. Oliver, E. J. Boyko, and W. J. Katon, "Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization," *The American Journal of Managed Care*, vol. 14, no. 1, pp. 15–23, 2008.

[123] F. Heckel, M. Schwier, and H.-O. Peitgen, "Object-oriented application development with mevislab and python," in *Proceedings of the INFORMATIK: Im Focus das Leben*, vol. 154, 2009, pp. 1338–1351.

Part IV

Appendix

CPT Code	G_1 Description
11	Diagnostic Endorine Procedures
15	Lens and Cataract Procedures
17	Destruction of Lesion of Retina and Choroid
18	Diagnostic Procedures on Eye
20	Other Intraocular Therapeutic Procedures
47	Diagnostic Cardiac Catheterization, Coronary Arteriography
54	Other Vascular Catheterization, Not Heart
70	Upper Gastrointerstinal Endoscopy, Biopsy
76	Colonoscopy and Biopsy
77	Proctoscopy and Anorectal Biopsy
168	Incision and Drainage, Skin nd Subcutaneous Tissue
169	Debridementof Wound, Infectionor Burn
190	Contrast Arteriogram of Femoral and LowerExtremity Arteries
199	Electroencephalogram (EEG)
201	Cardiac Stress Tests
202	Electrocardiogram
214	Traction,Splints, and Other Wound Care
220	Ophthalmologic and Ontologic Diagnosis and Treatment
233	Laboratory -Chemistry and Hematology
240	Medications (Injections, Infusion and Other Forms)

Table 8.1: Clinical conditions for diabetic patient encounters of event group G_1. The table shows one of the four event-group level categories G_1 and their respective event-type levels.

We have developed a novel image analytics environment that consists of an interactive graphical user interface. The software tool supports the loading and saving of DICOM and other medical image formats. A custom multimodal longitudinal DICOM browser enables a holistic access to the complete image record. The tool supports the integration of all major image analysis softwares for labeling, tracking, visualization, and registration. Large collections of volumetric image data can be assessed within a multi-monitor setting. The following screenshots exemplify the main functionality of our tool. The tool was implemented within the MeVisLab framework [123].

LABS	G_2 Description
	GLYCO and HEMOGLOBIN A1C/HEMOGLOBIN.TOTA LDL, CHOLESTEROL.IN LDL, and TOTAL LDL-C DIRECT
PCP	G_3 Description
	General Primary Care Physician Visits
SPECIALTY	G_4 Description
	NEPHROLOGY OPHTHALMOLOGY CARDIOLOGY NEUROLOGY PODIATRY ENDOCRINOLOGY PULMONOLOGY

Table 8.2: Clinical conditions for diabetic patient encounters of event group $G_2 - G_4$. The table shows the last three out of four event-group level categories $G_2, G_3,$ and G_4 and their respective event-type levels.

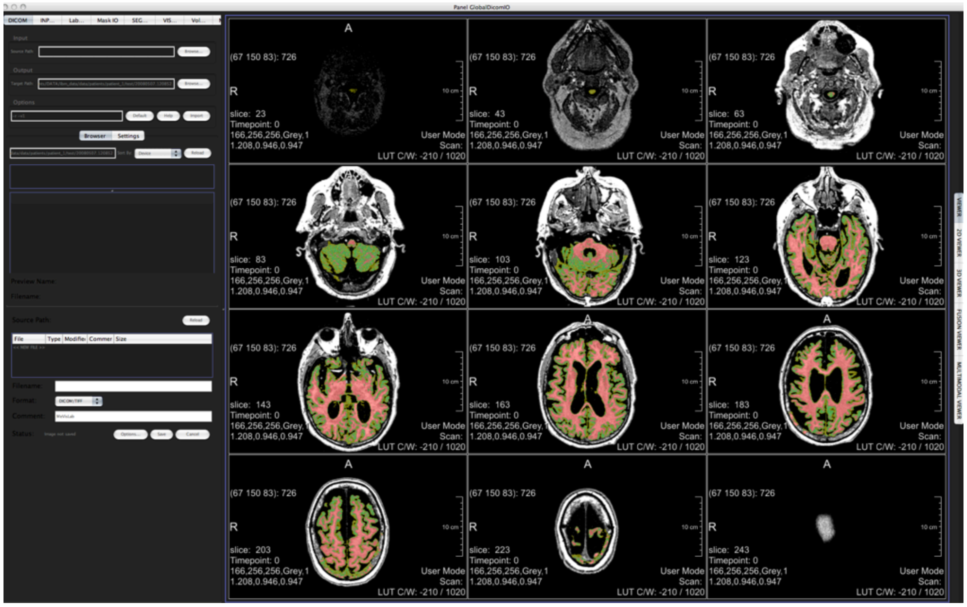


Figure 8.1: Unimodal image viewer.

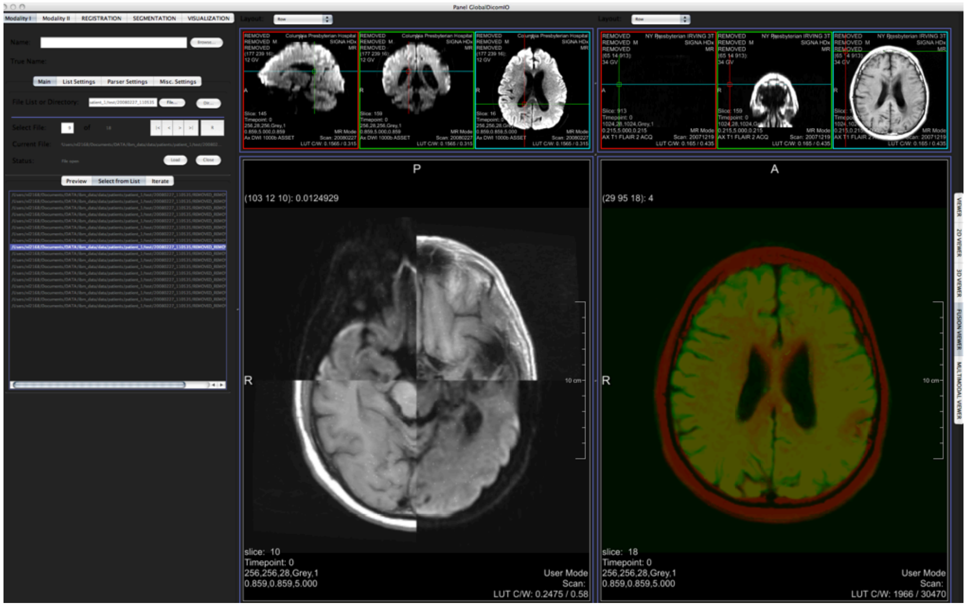


Figure 8.2: Multimodal registration environment.

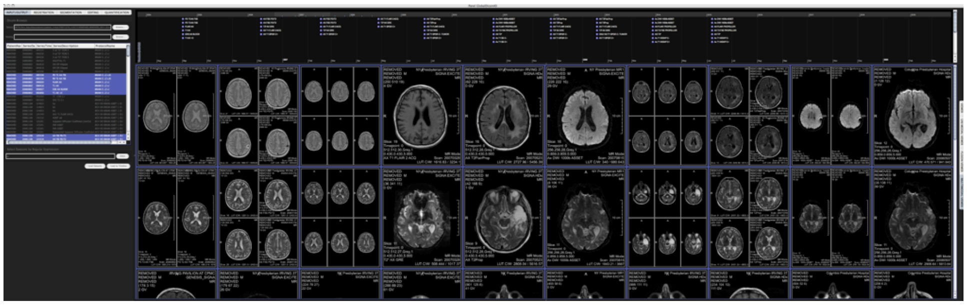


Figure 8.3: Multimodal longitudinal image viewer.

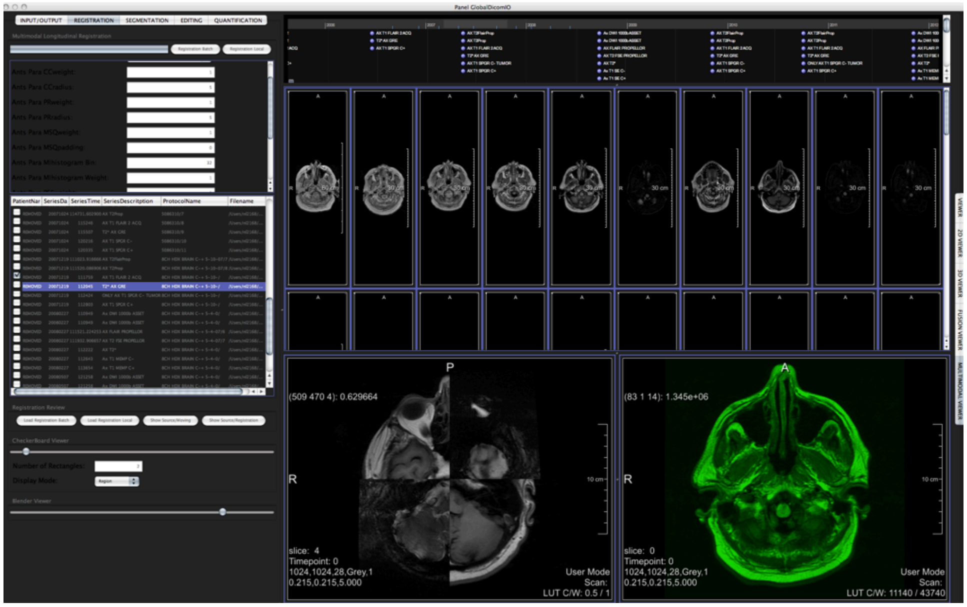


Figure 8.4: Multimodal longitudinal registration environment with holistic view to the complete patient dataset.

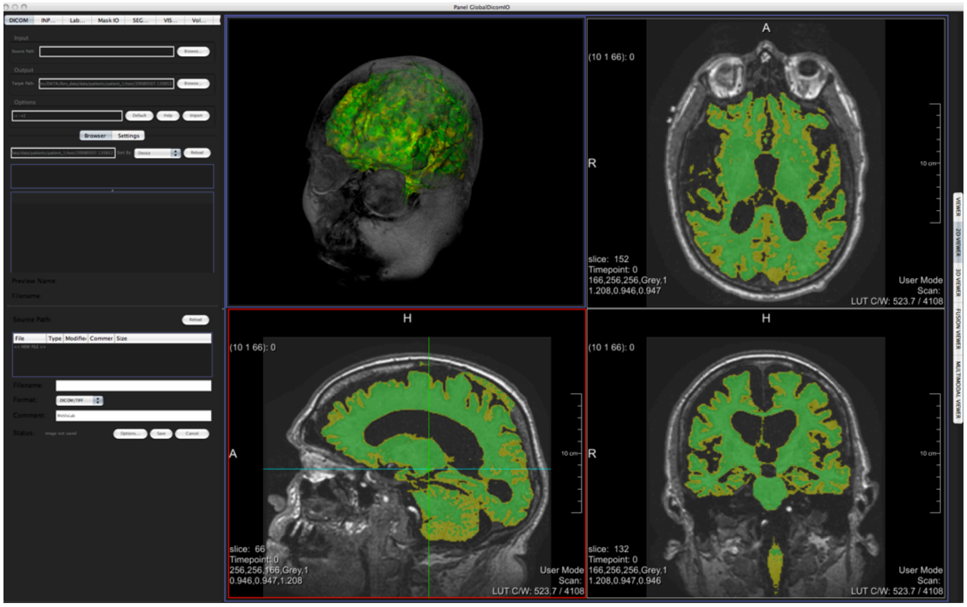


Figure 8.5: Unimodal brain labeling viewer with volume rendering, surface view, and multi-planar reformation views with mask overlays.

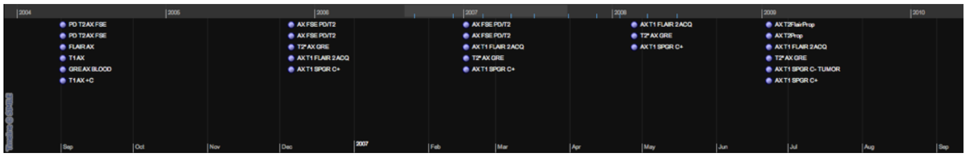


Figure 8.6: Multimodal longitudinal interactive timeline viewer.

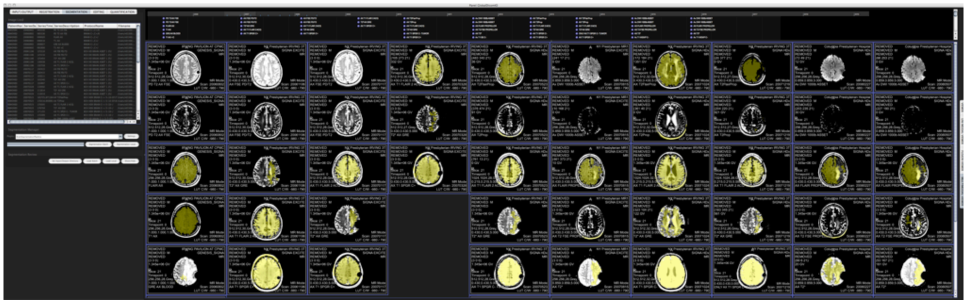


Figure 8.7: Multimodal longitudinal labeling environment with holistic view of complete dataset.

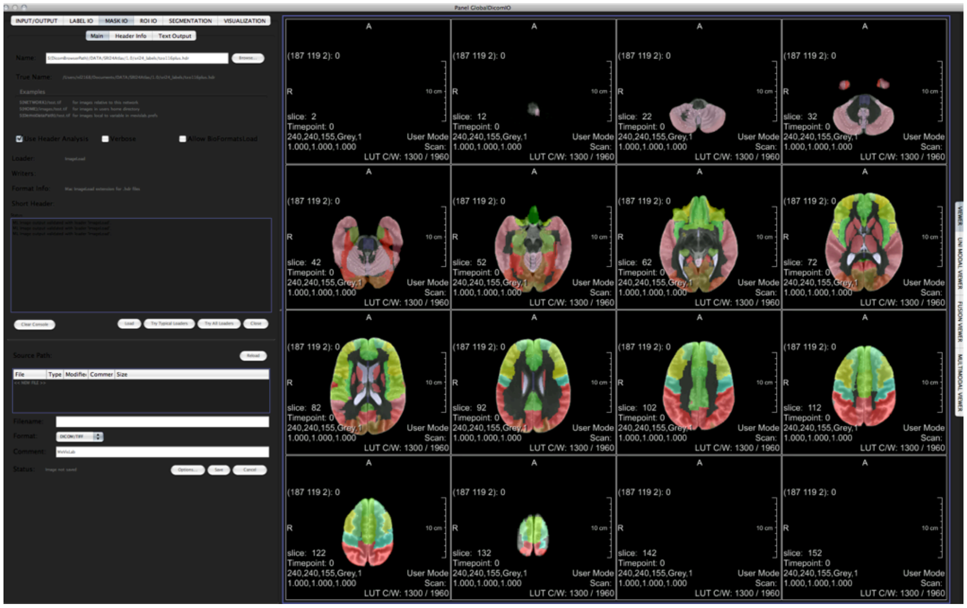
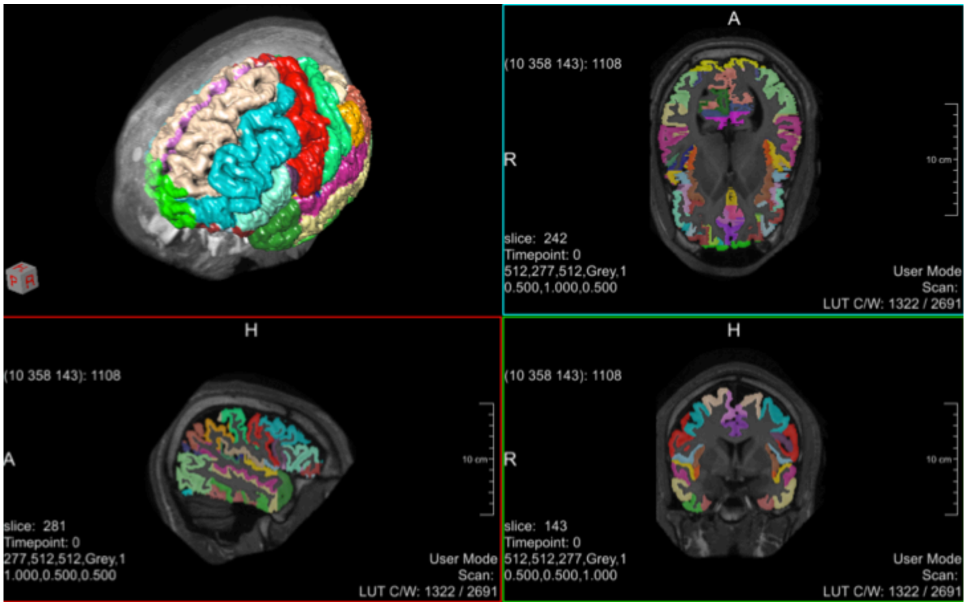


Figure 8.8: Brain parcellation viewer with multi-cross-sectional parcellation overlays.



(10 358 143): 1108

slice: 242
Timepoint: 0
512,277,512,Grey,1
0.500,1.000,0.500

User Mode
Scan:
LUT C/W: 1322 / 2691

(10 358 143): 1108

slice: 281
Timepoint: 0
277,512,512,Grey,1
1.000,0.500,0.500

User Mode
Scan:
LUT C/W: 1322 / 2691

(10 358 143): 1108

slice: 143
Timepoint: 0
512,512,277,Grey,1
0.500,0.500,1.000

User Mode
Scan:
LUT C/W: 1322 / 2691

Figure 8.9: Brain parcellation viewer with volume rendering, surface view, and multiplanar reformation views with colored mask overlays.