Automatic Layout Analysis and Visual Exploration of Multidimensional Datasets with Applications in the Digital Humanities

Zur Erlangung des akademischen Grades eines

Doktor der Ingenieurwissenschaften

von der Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Swati Chandna

aus New Delhi, India

Tag der mündlichen Prüfung:	16.05.2018
Erster Gutachter:	Prof. Dr.-Ing. Carsten Dachsbacher
Zweiter Gutachter:	Prof. Dr. Marc Weber

Abstract

The rapid developments of computer technologies have led to the advancements in almost every research discipline. Researchers from various disciplines rely on the power of computers be it computing power, storage size, or advanced algorithms to extract information from their scientific data. Digital humanities, which make use of computer-aided methods in the humanities, such as literature studies and history, are gaining in importance due to a widely increasing amount of handwritten historical document images for analysis and for gaining insights.

Document layout analysis is essential for the identification of physical regions enclosed in the document images. It is utilized to determine the precise information about the physical regions. Previous research has focused on various methods to identify different physical regions of such document images that provide significant improvements regarding speed and accuracy. However, traditional methods are limited to a specific set of document layout structures, produce results in proprietary data formats, and do not allow exploration of the identified physical regions and the derived information.

The scope of this thesis is the research and development of a generic method that can be applied to a variety of documents with overlapping layout, generates reproducible and deterministic results, and enables humanities researchers to explore their data and gain valuable insights.

The first component of this method is a generic and a fully automated approach for the identification of physical regions, such as text regions and picture regions enclosed in the document images. This approach is also capable of extracting various layout features of the identified physical regions. Due to its fully automatic nature, the results produced by this approach are also deterministic and reproducible and adhere to a standard document representation format that records information about the document image characteristics, layout structure, and page content. Moreover, the ground truth evaluation shows that the results produced by the approach in this thesis are comparable to the results produced by traditional methods or tools.

The second component of this method is the application of the proposed layout analysis approach on the large and heterogeneous set of document images to identify the physical regions enclosed in them and also to extract their corresponding layout features. The proposed approach is applied to 150,000 handwritten document images digitized within the scope of the project "Virtual Scriptorium St. Matthias". The proof of generality is shown by application of the layout analysis approach to the printed Spanish magazines, PDF documents, Aristoteles documents, Parzival, and Saint Gall database.

The third component of this thesis is the generic design strategy to aid information visualization designers for efficiently choosing suitable information visualization techniques and their combinations that can be applied to a particular application. For instance, in this thesis it is applied to multidimensional and text documents data.

The fourth component is the multiple-coordinated information visualization design. This component enables researchers in the domain of digital humanities, firstly, to explore their data to determine valuable information. Secondly, to view the complete physical structure of the multiple documents at a single glance and thirdly, to determine correlations, outliers, clusters, and a range of values. The qualitative evaluation and feedback of the visualization design from the humanities researchers show that the design is capable of exploring different information of handwritten historical document images and providing beneficial information which may contribute to a more precise physical layout analysis.

As a result, this research work has enabled domain experts in the field of digital humanities to explore the identified physical regions, and their corresponding layout features more engagingly to gain better insights and discover hidden knowledge in their data.

Zusammenfassung

Die rasanten Entwicklungen der letzten Jahre in den Bereichen Speicherkapazität, Rechenleistung und komplexen Algorithmen werden von Wissenschaftlern nahezu aller Disziplinen genutzt, Informationen aus ihren wissenschaftlichen Daten zu gewinnen. Auch den Digital Humanities, die computergestützte Methoden in geisteswissenschaftlichen Disziplinen anwenden, stehen vermehrt handschriftliche historische Dokumente zur Analyse und auf diese Weise zum Erkenntnisgewinn zur Verfügung.

Durch eine Dokumentlayoutanalyse werden die physischen Regionen in Bildern des Dokuments identifiziert und zur Bestimmung präziser Informationen über diese Regionen verwendet. Traditionelle Methoden sind jedoch auf eine eingeschränkte Menge von Dokumentstrukturen festgelegt, produzieren proprietäre Datenformate und bieten keine Möglichkeit, die identifizierten physischen Regionen zu erkunden und Informationen abzuleiten.

Gegenstand der vorliegenden Dissertation ist daher die Erforschung und Entwicklung einer generischen Methode, die auf eine Vielzahl von Dokumenten angewendet werden kann, reproduzierbare und deterministische Ergebnisse erzeugt und geisteswissenschaftlichen Forschen die Datenerkundung und das Ableiten wertvoller Erkenntnisse ermöglicht.

Die erste Komponente der Methode ist ein generischer und vollautomatischer Ansatz zur Identifizierung physischer Regionen wie Text- und Bildregionen auf Dokumentenbildern sowie zur Extraktion vielfältiger Layoutmerkmale der Regionen. Die Ergebnisse sind auf Grund der Charakteristik des Ansatzes sowohl deterministisch als auch reproduzierbar und im Standformat der Dokumentenrepräsentation gespeichert, das Informationen über die Eigenschaften des Dokumentenbildes, die Layoutstruktur sowie den Seiteninhalt bereitstellt. Die Evaluation an Hand von Ground Truth Daten belegt qualitative Vergleichbarkeit von traditionellen Methoden mit dem vorgestellten Ansatz.

Die zweite Komponente ist die Anwendung der Layoutanalyse und Merkmalsextraktion auf den großen und heterogenen Datensatz des „Virtuellen Skriptoriums St.

Matthias“ mit 150.000 handgeschriebenen Manuskriptseiten. Die Anwendung bei gedruckten, spanischen Magazinen, PDF Dokumenten, Aristoteles Dokumenten, dem Parzival sowie Dokumenten der Sankt Gallen Datenbank zeigt die Übertragbarkeit und Allgemeingültigkeit des Ansatzes.

Die dritte Komponente der Methode ist eine generische Designstrategie, die Entwicklern die effiziente Auswahl und Kombination von Techniken der Informationsvisualisierung abgestimmt auf den jeweiligen Anwendungsfall ermöglicht. In dieser Arbeit wird die Strategie verwendet, passende Techniken der Informationsvisualisierung für multidimensionale Textdokumentdaten abzuleiten.

Die vierte Komponente ist das entwickelte Informationsvisualisierungsdesign, dessen vielfältige Elemente aufeinander abgestimmt sind und sich gegenseitig beeinflussen. Diese Komponente ermöglicht es Wissenschaftlern, ihre Daten zu erkunden und wertvolle Informationen abzuleiten, die äußerliche Struktur zahlreicher Dokumente auf einen Blick zu erfassen sowie Korrelationen, Ausreißer, Cluster und Wertebereiche zu bestimmen. Die qualitative Evaluierung und die Rückmeldungen der geisteswissenschaftlichen Forscher belegen, dass das Visualisierungsdesign die Untersuchung heterogener Informationen der handschriftlichen historischen Dokumente ermöglicht und wertvolle Informationen für eine präzisere physische Layoutanalyse bereitstellen kann.

Zusammengefasst ermöglicht es diese Dissertation Fachwissenschaftlern aus dem Gebiet der Digital Humanities, die identifizierten physischen Regionen und Informationen zu erforschen, neuartige Erkenntnisse abzuleiten und bisher verborgene Zusammenhänge in ihren Daten zu entdecken.

Contents

Chapter 1

Introduction

During the last years, development of various computer technologies, such as the Internet, e-publishing, and database systems has led to the advancements in almost every discipline be it in physics or in neuroscience. Researchers from most of the disciplines rely heavily on the power of computers regarding computation speed, storage size, mathematical models, and advanced algorithms to extract information from their scientific data in order to gain insights and to make better decisions.

However, the development of such computer technologies in the field of humanities is still undergoing developments [1]. The main challenge lies in the transfer of research questions from the humanities world to the digital world. To solve this challenge, digital humanities offer a close connection between computer science and humanities [2], [3]. Since the field of humanities consists of diverse disciplines, such as history, literature studies, archeology, and musicology each of these disciplines has a different view with respect to the term digital humanities, there exists no clear and precise definition of digital humanities [4]. Even the website "https://whatisdigitalhumanities.com/" [1] itself consists of approximately 800 definitions of digital humanities.

For the optimal exploitation of the interdisciplinary connection between computer science and humanities, DARIAH (Digital Research Infrastructure for the Arts and Humanities) [5] aims at developing a research infrastructure which offers various tools, services, and access to research data for further research. This infrastructure allows researchers from different discipline, for instance, from computer science and humanities to work together seamlessly in a collaborative research environment. The European Commission started the European Strategy Forum on Research Infrastructures (ESFRI) [6] in 2002 to support systematic approach towards research infrastructures in Europe as well as on the global level. DARIAH is one of the 48 projects on the roadmap of

[1] https://whatisdigitalhumanities.com/

ESFRI for supporting digitally enabled research in the arts and humanities. DARIAH-DE [7] is the German contribution to DARIAH project with the main aim of providing sustainable research infrastructure, supporting the use of digital methods in humanities and especially facilitating access to European research data. One of the research endeavors where DARIAH-DE plays an important role is the digitization projects. These digitization projects establish a digital reconstruction of a virtual library of historical documents which are scattered all over the world. The research data from these digitization projects have heterogeneous characteristics; for instance, they differ in size (from kilobytes to gigabytes), quantity (from a small number of the digitized document to massive amounts of digitized documents) and format (documents, audio, or video).

A significant amount of data (in the form of documents, audio, and video) has also been digitized at an incredible speed worldwide to provide accessibility and to achieve preservation. For instance, the British Library[2] has one of the most extensive collections of digitized material in the world. This collection includes approximately 14 million books as well as a considerable number of medieval and Renaissance manuscripts. It provides the possibility of searching, exploring, and downloading the digital images of the books or manuscripts. Similarly, the New York Public Library[3] contains a wide variety of digital data in the form of books, videos, illuminated manuscripts, and historical maps. This library also allows users to browse, search, and download its digital collections. Moreover, the online portal of the European Library[4] offers quick access to 48 national libraries across Europe to researchers and practitioners worldwide. The primary objectives of such digital libraries are to provide quick and open access to digitized data for further analysis and to allow exploration to gain knowledge about this data and discover relationships in the data.

Therefore, the vision of this thesis is to assist humanities researchers and practitioners in transferring their research questions into the digital world and providing them a possibility to analyze further and explore a digitized collection of the documents.

1.1 Motivation

Due to an increasing amount of digitized data in the form of handwritten historical documents, the humanities researchers need to analyze and explore these documents to find valuable insights and draw conclusions. The physical layout[5] of these digitized handwritten historical documents is irregular in comparison to the printed documents,

[2] `https://www.bl.uk/`

[3] `https://www.nypl.org`

[4] `http://www.theeuropeanlibrary.org/`

[5] Physical layout —Physical location and boundaries of component regions in the document

which are quite regular with respect to their layout style, column size, the spacing between different regions, font sizes, and orientation of the text. Digitized images of handwritten documents not only carry textual information but are also composed of a variety of physical regions[6] containing text in the margins or text spread out over columns, and decorative elements like illustrations or drop capitals as shown in Figure 1.2 [8].

However, despite these irregularities, the appearance of these digitized handwritten historical documents is so aesthetic that it is hard to believe that medieval artisans designed them through individual visual judgment. The question then arises of whether medieval artisans followed any geometric rules or proportions to write these documents. This question has been the basis of layout studies concerning these handwritten historical document images[7]. It is also the case that, over the course of centuries, historical documents were torn apart, destroyed, and often scattered all over the world. Precise information about physical layout characteristics of these documents can help to trace the relationships between the documents which are scattered across the globe.

The need to obtain such physical layout information leads to the discipline of document layout analysis, which tries to identify and categorize the component regions contained in the document image. Previous research has focused on various methods to identify different physical regions of such document images that provide significant improvements regarding speed and accuracy. However, these methods may only work with a specific kind of document images that are similar to the patterns accepted by these methods. Furthermore, exploring the layout of the document images and gaining knowledge from them becomes more and more complicated with vast and complex datasets where analysis problems are ill-defined. Here, the ill-defined problem means that the user does not know where to start exploring or analyzing the data.

The approach of *exploratory data analysis* developed by John Tukey [9] can help to deal with such exploration issues. This type of analysis aims to gain insights on the basis of hypothesis or formal models. It employs numerous graphical techniques to explore the data and gain insights, such as box plots, pie charts, and histograms. However, with the increasing volume of data, exploration using classical exploratory data analysis techniques and gaining insights just by looking at numbers becomes difficult. Fortunately, there exists a possibility of *visual exploration* of data interactively with the help of various visualization techniques.

[6] Physical regions —Component regions of the document, such as text region, picture region, or text lines

[7] Handwritten historical document images —Digitized handwritten historical documents

Therefore, this research is motivated by the need to facilitate identification of the physical regions enclosed in document images, extraction of their corresponding layout features, and the tools to aid visual exploration of the identified physical regions, together with their extracted layout features.

1.2 Scope of thesis research

This thesis focuses mainly on the layout analysis and visual exploration of multidimensional datasets with applications in the digital humanities. Therefore, to define the scope of this thesis, the domains of digital humanities, document layout analysis, and information visualization are shown in Figure 1.1 and briefly described in this section.

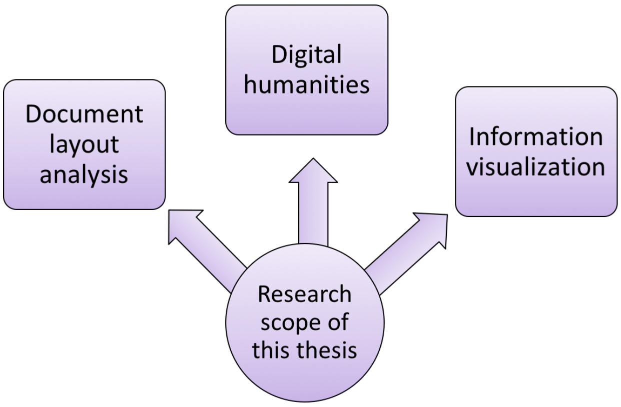


Figure 1.1 – Research scope of this thesis.

Digital humanities is defined as the use of computer-aided methods in the disciplines of humanities to improve scholarly activity, such as automating a repetitive process, such as searching a word or character [1]. Digitization is one of the most important prerequisites of any digital humanities research project. This process transforms analog material into digital images. In recent years, digital libraries have played a significant role in preserving cultural heritage, including audio, painting, or handwritten historical documents. This thesis mainly focuses on the images of such type of handwritten historical documents.

Document layout analysis breaks down a document image into a hierarchy of physical regions, such as picture regions, background, text regions, text lines. [10]. Identification and categorization of different physical regions enclosed in the document image is called *physical* (or *geometric layout) analysis*. The semantic labeling of the identified physical regions enclosed in the document image, such as title or footnote is called *logical layout analysis*. This thesis research focuses on document layout analysis because one of the primary goals of this thesis is to identify the physical regions enclosed in the document images and also extract the layout features of the identified physical regions.

Information visualization, as defined by Card et al. [11] is the “use of computer-supported, interactive visual representations of abstract data to amplify cognition.” In other words, the use of visual representations helps humans carry out tasks more efficiently [12]. Ware [13] listed various advantages of information visualization: it provides an ability to comprehend massive amounts of data, allows the perception of emergent properties, facilitates the understanding of large-scale and small-scale features of data, and helps in the formulation of hypotheses. *Visual Exploration* is defined by Tominski [14] as “an undirected search for relevant information within the data.” He also stated that the primary goal of visual exploration is, first, to provide an overview of complete data set and then to allow the user to explore the different levels of the data interactively. Visual exploration typically starts with an ill-defined hypothesis, or no hypothesis at all, and seeks to find an unusual pattern or useful information in the data. The research in this thesis focuses on information visualization in support of visual exploration of the layout features extracted from the document layout analysis.

1.3 Problem and approach

The primary goal of this thesis is to investigate various aspects of layout analysis and facilitate the visual exploration of historical document images for knowledge discovery. Although the domains of document layout analysis and information visualization may help to achieve the primary goal of this thesis, there are still some problems which are yet to be solved. These problems are described below:

- ***Heterogeneous and irregular document images***
  With the advancement of digitization technologies and the rapid growth of digital libraries around the world, large numbers of physical documents, especially historical documents, are being converted into digital format. Such improvements have led to a massive amount of high-resolution data (i.e., 300 or 400 dots per inch (DPI)), which is now being generated in large volumes, measured in terabytes or petabytes. Each of the handwritten historical document images is unique with respect to its color and size and has an irregular layout style (see Figure 1.2). Because of the heterogeneity and irregularities in such document images, identifying the physical regions is a challenging task. Such document images also include bibliographic information describing the century of production, the material of the document, and the number of pages. Current methods or frameworks for identifying the layout may only work for particular kind of document images that are similar to patterns accepted by these methods.

Furthermore, less research has been done in the area of extracting layout features describing the physical regions enclosed in the historical document images and finding valuable information from them through visual interfaces, due to their heterogeneity and layout irregularity.

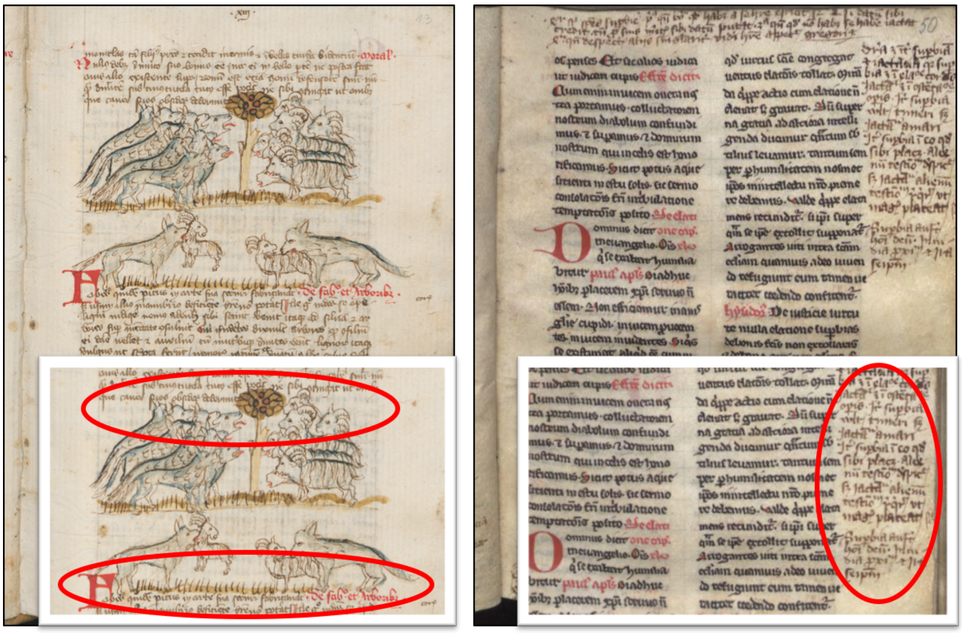


Figure 1.2 – Heterogeneous and irregular historical handwritten document images.

- ***Exploration of multidimensional datasets***
  Visualization of multidimensional datasets on a single screen leads to a vast amount of information to be explored. This makes it challenging to find valuable information and make correct decisions. Additionally, there exists a plethora of information visualization techniques which helps to visualize multidimensional information, but it is challenging to decide which of these information visualization techniques is most suitable to support exploration of multidimensional datasets especially when visualization designers are not aware of the tasks of domain experts. This makes the task of determining which information visualization technique is suitable for multidimensional information extracted from the physical layout analysis methods even more challenging.

Therefore, this thesis aims to answer the following research questions:

1. How to identify the physical regions from the heterogeneous and irregular handwritten historical document images, and also how to extract their corresponding layout features?

2. How to design and to enable interactive visual exploration of multidimensional datasets with real-world applications in digital humanities?

To answer the first question, the identification of physical regions from the heterogeneous and irregular document images can be approached in following three ways:

- manually, by identifying and measuring the physical regions of the document (e.g., measuring the size of a text region that is changing throughout a book);
- semi-automatically, by outlining the physical region present in the document image;
- automatically, by identifying and categorizing the physical regions without any user intervention.

The automatic approach is supported in this thesis because a fully automatic approach does not require any user intervention as compared to the manual approach which is very tedious and labor-intensive. Moreover, the results obtained with an automatic approach are also reproducible and deterministic. The semi-automatic approach is problematic; mainly if the tools are used by different experts, then inter-observer variability is often observed, where results differ for the same data set. Intra-observer variability can also be seen, which is when the same expert performs the analysis two or more times, and the results differ [15].

In this research, one of the main goals is to analyze the physical layout of a large number of heterogeneous and irregular datasets to enable the researchers to explore the documents easily than the overall accuracy of the extracted layout features. A generic and automated approach for identifying the physical regions without any user interaction is expected to be of significant importance to humanities researchers.

To answer the second question, i.e., to design and to enable interactive visual exploration of multidimensional datasets with real-world applications in digital humanities, the following approaches can be followed:

- summarizing, reducing, or aggregating the data and concentrating only on parts of the data.
- enabling interactive exploration of the data using information visualization techniques.

In this case, the use of information visualization techniques can facilitate the exploration of data easily by providing different interactive views of the data at different levels. This approach is much easier to learn and is better than summarizing, reducing, or aggregating the data, which leads to misinterpretations or overlooking of potential anomalies or outliers. A famous example is Anscombe's quartet [16],[12] as shown in

Figure 1.3. It shows that the set of small data designed by a statistician having the same descriptive statistics, i.e., mean, variance, correlation, and linear regression line, but in reality, they have different structures. The basic principle, which is illustrated by Anscombe's quartet is that a single summary is most often the oversimplification of the dataset that hides its true structure. This is applied even more to large and complex datasets.

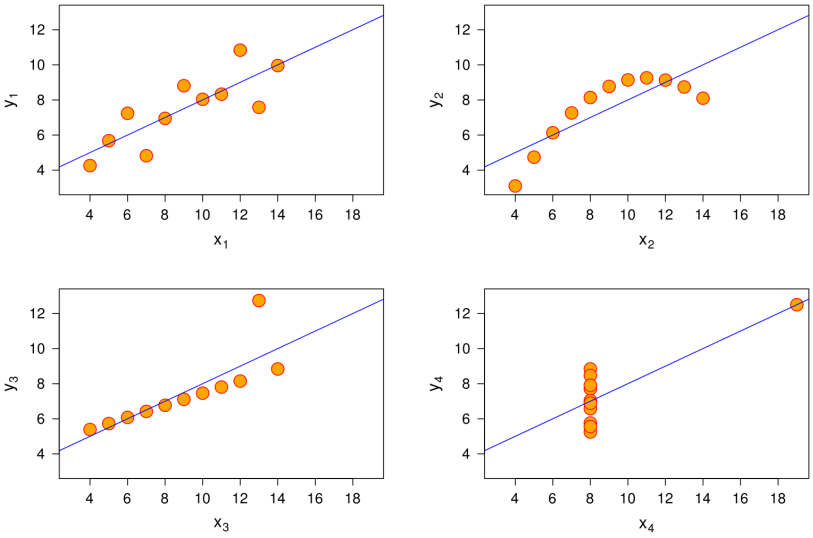


Figure 1.3 – Anscombe's quartet: Four datasets with identical statistics, i.e., mean, variance, correlation, and linear regression line. However, in reality, their structures are entirely different. (Image from Wikimedia Commons)

1.4 Contributions

The main contributions of this thesis allow automatic layout analysis of heterogeneous and irregular document images, a generic design strategy for designing effective visual exploration approach, and creating a multiple-coordinated exploration approach for exploring the multidimensional data.

1. This thesis designs and implements a generic and automatic approach for the identification of physical regions, such as page region, text regions, and picture regions enclosed in the handwritten historical document images having an irregular and overlapping layout. Moreover, it also allows extraction of features of the identified physical regions, such as color, size, count, and other derived features in order to obtain the precise information of the physical characteristics of the

handwritten historical document images. The results produced by this automated approach are deterministic and reproducible.

2. For the first time the automatic layout analysis approach is applied to a significantly large and heterogeneous dataset: 386 manuscripts with 150,000 historical handwritten documents digitized within the scope of the project "Virtual Scriptorium St. Matthias". It can also be adapted to other datasets which are shown by applying to printed Spanish Magazines, PDF documents, Parzival, Aristoteles, and Saint Gall database. The approach is currently in use by humanities scholars studying the physical layouts of historical documents and collaborating within the scope of DARIAH-DE. Currently, it is also being adopted for the additional Aristoteles handwritten documents for Collaborative Research Center 980 (CRC) "Episteme in Motion" [17] which will serve as a pilot for DARIAH-DE in future.

3. To aid visualization designers this thesis presents a generic design strategy for efficiently choosing an information visualization technique and its combinations which can be applied for a particular kind of application. The design strategy is used to select the combination of best possible information visualization techniques for visualizing textual and documents data. As a result, it is found that the parallel coordinate plot and the radial tree are most suitable for visualizing the textual and documents data and superimposition and juxtaposition are suitable for creating a composite visualization for such data.

4. Based on the third contribution, this thesis presents a multiple-coordinated information visualization design to support visual exploration of 386 handwritten historical documents with 150,000 documents pages and approximately 162,000,000 layout features. The main objective of this visual exploration approach is to aid humanities researchers in exploring their research data more effectively by presenting the complete structure of the physical layout from a bird's eye view. Furthermore, enabling humanities researchers to determine correlations, detect clusters, anomalies, and outliers. The first exploration results revealed:

   - The presence of a fixed page height for the manuscripts written on paper between the 14^{th} and 18^{th} centuries, which provides an insight of paper production techniques used in historical times.
   - The distinction between the historical documents written on the hair side and the flesh side of the parchment based on the clustering of the brightness of the pages.

- The historical documents written before 12^{th} century consisted of single columns, while after the 12^{th} century, the historical documents consists of text regions in one or two columns. This indicates the use of a ruling pattern followed after the 12^{th} century.

1.5 Thesis overview

This section briefly describes the contents of the thesis chapters.

Chapter 2 presents the theoretical background for this research and reviews the related literature. This chapter includes the background and previous work related to document layout analysis and information visualization in general and describes layout analysis, text visualization, individually as related to historical document images.

Chapter 3 describes the approach to automatically identify the physical regions and extract their corresponding layout features. It is organized according to the processing steps that are needed to be carried out. After providing the details of the document images, the preprocessing of the dataset is described. Afterwards, the region segmentation and feature extraction approaches are presented. This approach was published in:

S. Chandna, D. Tonne, T. Jejkal, R. Stotzka, C. Krause, P. Vanscheidt, H. Busch, and A. Prabhune, "Software workflow for the automatic tagging of medieval manuscript images (SWATI).," in *In Proceedings of SPIE, Document Recognition and Retrieval XXII*, vol. 9402, p. 940206, 2015

The automatic layout analysis approach presented in this thesis is generic and can be applied to a large variety of document images. This was presented for the printed Spanish magazines in:

N. Rißler-Pipka, S. Chandna, and D. Tonne, "Automatische bild-textanalyze: Chancen für die zeitschriftenforschung jenseits von reinen textdaten," in *Proceedings of Digital Humanities im Deutschsprachigen Raum*, pp. 94–99, DHd, 2017

Chapter 4 examines existing information visualization taxonomies in detail for choosing an appropriate information visualization technique and describes a generic design strategy. It explains the application of this generic design strategy to a real-world application and presents possible techniques to visualize the multidimensional information generated from the automatic layout analysis method.

In **chapter 5** designs and implementations of the visual exploration framework are described. In the elementary design (CodiViz-I) for visual exploration, the two visualization techniques, i.e., a parallel coordinate plot and a radial tree are combined using juxtaposition composition where these two techniques are displayed next to each other. In the advanced design (CodiViz-II), exploration is carried out hierarchically with the fusion of a parallel coordinate plot, a superimposition plot and a document layout montage plot using superimposition and juxtaposition composition. The application of the generic design strategy to generate CodiViz-I design is published in:

S. Chandna, D. Tonne, R. Stotzka, H. Busch, P. Vanscheidt, and C. Krause, "An effective visualization technique for determining co-relations in high-dimensional medieval manuscripts data," in *In Proceedings of Electronic Imaging*, vol. 2016, pp. 1–6, Society for Imaging Science and Technology, 2016

Whereas, the CodiViz-II is published in:

S. Chandna, F. Rindone, C. Dachsbacher, and R. Stotzka, "Quantitative exploration of large medieval manuscripts data for the codicological research," in *Large Data Analysis and Visualization (LDAV), 2016 IEEE 6th Symposium on*, pp. 20–28, IEEE, 2016

Finally, **Chapter 6** provides a discussion of the covered topics and the conclusion.

Chapter 2

Theoretical background and related work

The main aim of document layout analysis is to obtain the precise information about the physical regions enclosed in the document image. Additionally, information visualization aims to use the interactive visual interfaces to represent the data. In this chapter, the fundamentals and related literature of document layout analysis and information visualization are introduced from the document image to document layout analysis and from information visualization to visualization of textual documents.

2.1 Document image

An Image I is defined as a two-dimensional function, *f(x, y)*, where x and y are the spatial coordinates in horizontal and vertical direction and the amplitude of f at any pair of coordinates x, y is called gray level of the image [21]. The finite, discrete values of f and x, y defines the image as digital image. This digital image consists of a finite number of elements called pixels where each pixel has a defined location and a value.

A digital image that includes color information for each pixel is called colored digital image. A color image contains three values for each pixel. They measure the intensity and chrominance of light. The three values (channels) for each pixel are interpreted as coordinates in a defined color model. Most commonly used color model is the RGB color model. Other color models include HSV, YCbCr.

Document images is a class of such images which is used in a variety of applications and acquired through digitization process from the physical documents using scanners or digital cameras. Most common applications include handwritten character recognition or extraction of information present in the document images. There exist different types of document images namely books, newspapers, or magazines in various forms, such as handwritten or printed. The layout of these document images can be classified into two categories as described below:

- **Physical layout (or geometrical layout)**
  Physical layout refers to the physical locations or the boundaries of various regions present in the document images, such as background, text regions, or picture regions.
- **Logical layout**
  Logical layout refers to the logical units present in the document, such as title, headings, charts, or tables.

This thesis work mainly focusses on analyzing the physical or geometrical layout. Kise [22] further classified the physical layout of the document images into four classes as described below and shown in Figure 2.1

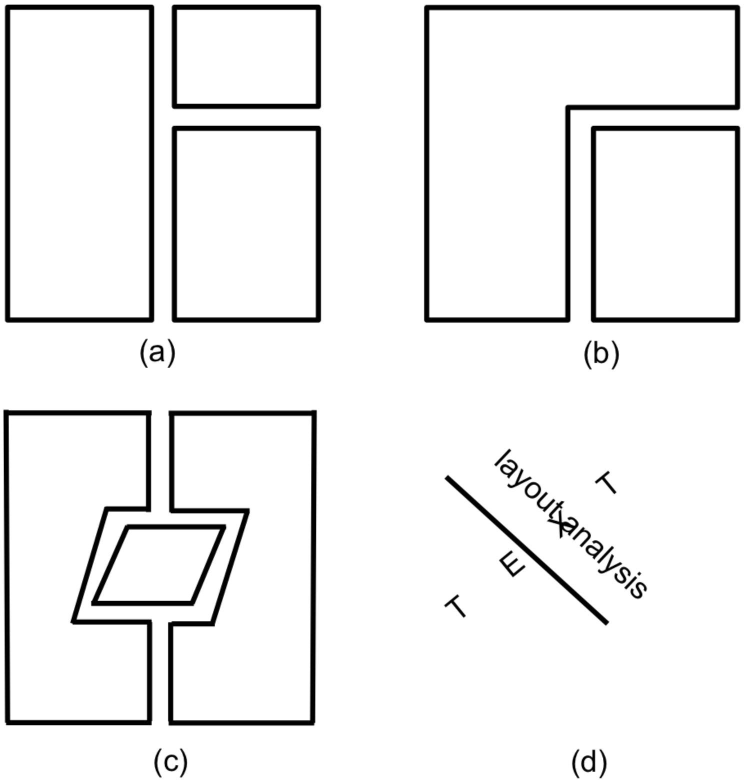


Figure 2.1 – Different classes of document layout a) Rectangular, b) Manhattan, c) non-Manhattan, d) Overlapping layout. (Images are adapted from [22]).

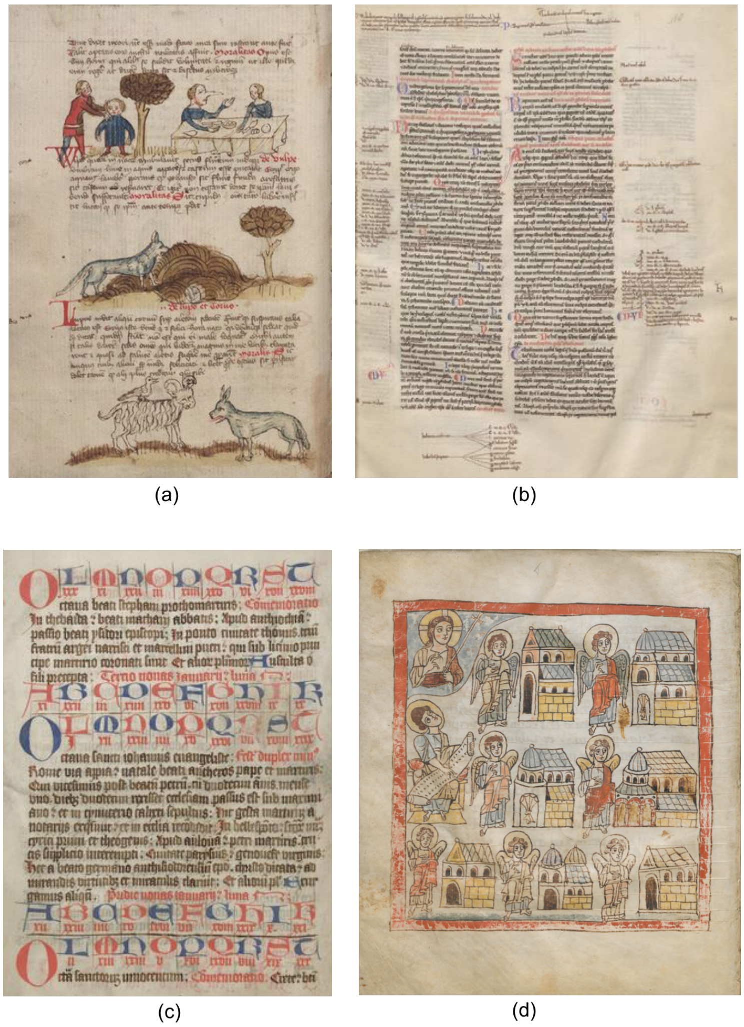
- ***Rectangular layout***
  When the non-overlapping rectangle bounds all the regions present in the document image and when sides of the regions are parallel or perpendicular to the border of the document image as shown in Figure 2.1a, the layout is called rectangular layout. Books or scientific research papers have rectangular layouts.

- ***Manhattan layout***
  When some regions present in the document image having concave shape are represented by sides parallel or perpendicular to one another as shown in Figure 2.1b, the layout is called Manhattan layout. Manhattan layout also includes rectangular layout as its subclass. Most documents with multiple columns, such as magazines or newspapers have Manhattan layout.

- ***Non-Manhattan layout***
  When non-overlapping regions bound only some of the regions present in the document image and when sides of the regions are neither parallel nor perpendicular to the borders of the document image as shown in Figure 2.1c, the layout is called non-Manhattan layout. Magazines or newspapers with larger figures and picture normally have the non-Manhattan layout.

- ***Overlapping layout***
  When regions of the document image intersect with each other as shown in Figure 2.1d, where text regions overlay on picture regions, the layout is called overlapping layout. Handwritten historical documents, modern advertisements, maps have an overlapping layout.

As mentioned above, the handwritten historical document images have the overlapping type of physical layout. Figure 2.2 shows some examples of handwritten historical document images. The physical layout of such document images is very irregular; their regions overlap with one another, the text is spread out over columns, and the picture regions and the background of document images share same background color. Such documents sometimes enclose decorative elements, illustrations, comments on the margins, or text between lines as shown in Figure 2.2. Moreover, the physical layout continuously changes throughout the whole document collection. This research work mainly focuses on the layout analysis of such type of documents which directly leads to the approach of document layout analysis described in the next section.



(a)

(b)

(c)

(d)

Figure 2.2 – Examples of historical document images from St. Matthias (City Library of Trier, Germany) a) Overlapping text and picture regions, b) Comments in the margin, c) Various initials in the text, d) Picture region and document image with same background color.

2.2 Document layout analysis

Document layout analysis decomposes a given document image into its constituent regions, such as text regions or picture regions as shown in Figure 2.3. Identifying and categorizing these constituent regions present in the document image is called *physical layout analysis* whereas assigning logical labels to these identified physical regions, such as titles, paragraphs is called *logical layout analysis*. The algorithms for the *document layout analysis* are classified into three classes: *top-down algorithms*, *bottom-up algorithms*, and *hybrid algorithms*.

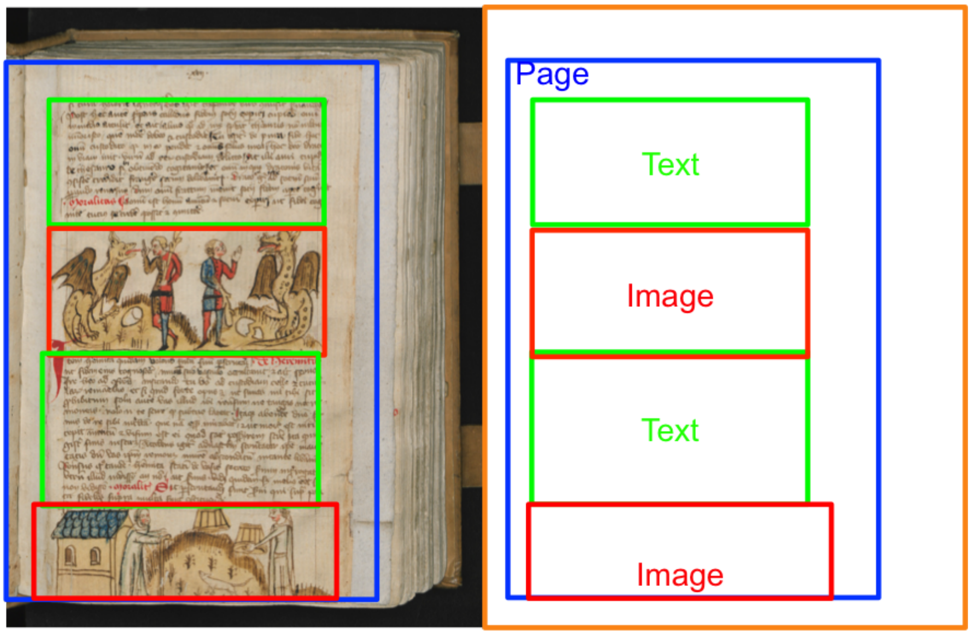


Figure 2.3 – Physical regions enclosed in the handwritten historical document image T Hs 1108/55 4° 37r, from St. Matthias (City Library of Trier, Germany).

Top-down algorithms start with the complete document image and repeatedly decompose it to form smaller homogeneous regions. The decomposition terminates when the resulting regions correspond to the primitives described by the document. Whereas, *bottom-up algorithms* start from the smallest component of the document image and group them to form larger regions. There also exists hybrid approach which works with the combination of top-down and bottom-up algorithms. Some of these algorithms are described below:

X-Y Cut algorithm [23] is also called recursive x-y cut algorithm. It is a tree-based top-down algorithm which starts by dividing a document image into rectangular blocks and repeatedly splits the document image by projecting the regions of the current block on the horizontal and vertical axes. Once the criteria which decide the atomicity of the region are met, the further splitting of the document image stops. This algorithm can only segment the document images having rectangular or Manhattan layout. Also, it is extremely dependent on the skew of the document image, for instance, a small amount of skew in the document image can lead to incorrect projection profiles.

Whitespace analysis algorithm given by Baird et al. [24] divides the document image into smaller regions based on the structure of the white background in the document image. The basic idea in the first step is to detect the maximal set of white rectangles covering the complete background using maximal empty space algorithm. These white rectangles are then sorted by a sort key, K(c) where c is the white rectangle, and W(.) is the weighting function. The purpose of this weighting function is to assign the higher weight to long rectangles because they are supposed to be separators for text regions.

In the second step, the white rectangles are combined one by one to generate a sequence of segmentations. Here, segmentation is the open area left by the union of the white rectangles combined so far. This unification continues until the stopping rule is satisfied. The connected components which remain in the end within the remaining uncovered parts are called candidate text regions. Since the uncovered regions thus obtained are not necessarily rectangular, bounding boxes of these uncovered regions are taken as representative of the text segments.

Run-length smearing algorithm [25] is an example of the bottom-up approach. It works on the binary images where white pixels are represented by 0's and black pixels by 1's. The binary sequence x is transformed into y according to the following rules:

- 0's in x are changed to 1's in y if the number of adjacent 0's is less than or equal to a predefined threshold T.
- 1's in x are unchanged in y.

This is applied row-wise to the document using a threshold t_{sh} and column-wise using t_{sv}, yielding two distinct images. These two images are combined in a logical AND operation. Then, the connected component analysis is performed on the resultant combined image to obtain the document regions.

The *Docstrum algorithm* [26] is a bottom-up approach based on the nearest neighbor clustering of connected documents. The connected components are divided into two categories, i.e., first category is with characters with most dominant font size and second category is with characters present in the titles or headings. Afterwards, for each component, the K-nearest neighbors are identified and then histogram of the distance and angle is computed for each connected component. After this, the transitive closures of adjacent components are computed using the within-line-distance and between-line-distance to determine the text lines. Lastly, text lines are merged using parallel and perpendicular distance threshold to form text blocks.

The *Voronoi-diagram based segmentation algorithm* given by Kise et al. [27] is also a bottom-up approach, and it also functions by grouping the connected components enclosed in the document image. At first, the borders of the connected components compute a Voronoi diagram of the document image. After that, the Voronoi edges are deleted to obtain the boundaries of the document components. This algorithm is capable of handling very complex layouts and gives accurate and reliable results.

Detailed performance evaluation and benchmarking of these popular top-down and bottom-up algorithms on the UW-III dataset is shown by [28].

This dataset contains 1,600 English document images, scanned from various archival journals and manually edited ground truth. Based on the performance evaluation following recommendations are made:

- the X-Y cut algorithm is preferred for well-defined documents with no or little skew.
- the Docstrum and Voronoi algorithms can be used for homogeneous collection of documents, i.e., with same resolution or similar font sizes. However, these algorithms need parameter tuning to obtain good results for a given set of document collections.
- the Constrained Text-line finding algorithm can be used best for documents having different font sizes and styles.
- the Voronoi algorithm can be used for the documents having non-Manhattan layout.

Based on the top-down and the bottom-up algorithms described above, several document layout analysis methods have been designed and implemented during the last years. These methods have been used for the document images having the Manhattan, non-Manhattan, rectangular, and the overlapping layouts. These methods are described below:

A modified version of Constrained Run-Length Algorithm (CRLA) [29] called selective CRLA [30] has been presented to perform layout analysis for documents with Manhattan and non-Manhattan layouts. It is performed two times with different set of parameters on the label image which is derived from the document image itself. Here, the label image is obtained from the connected component labeling algorithm [21].

The Voronoi++ approach presented by Agrawal and Doermann [31] is based on the Voronoi diagram based segmentation algorithm. They proposed a combination of Docstrum and Voronoi algorithms. Where the Voronoi algorithm associates only distance metric to each Voronoi edge and angular, the neighbor features are added from the Docstrum algorithm. A comparison is also performed between the Voronoi and the Voronoi++. As a result, it is shown that the Voronoi++ approach improves the accuracy for analyzing the layout by 33%.

Saqib Bukhari et al. [32] presented a method for text and non-text segmentation. Here, a machine learning approach is used to classify connected components to the relevant class of text. The features like normalized height, foreground area, relative distance, orientation, and neighborhood information of the connected components are considered to train a Multilayer Perceptron (MLP) classifier. The authors extended this approach in [33] to segment side-notes from the main-body text in ancient Arabic documents.

Cohen et al. [34] proposed a two-stage bottom-up approach. In the first stage, the Laplacian of Gaussian is applied to the binarized image to extract the connected components. In the second stage, features, such as bounding box size, area, stroke width, estimated text lines distance are used to label each connected component into text or non-text.

Kai Chen et al. [35] presented an unsupervised learning approach for page segmentation of historical documents. Here, a page segmentation is considered as a pixel labeling problem and each pixel is classified as the background, text block, or decoration. Convolutional encoders are applied to learn features directly from the pixel intensity values. After that these features are used to train a Support Vector Machine (SVM) classifier.

Mehri et al. [8] proposed texture based segmentation method in which firstly, some random foreground pixels are selected, and then textual features like autocorrelation, Grey Level Co-occurrence Matrix, and Gabor features are computed based on these foreground pixels. Afterwards, the Consensus Clustering method is applied to estimate a similar cluster and labeling is done using a nearest neighbor algorithm with Mahalanobis distance.

Document physical layout analysis process also have performance uncertainties and so may provide uncertain input to the logical layout analysis process. Stochastic models [36], represented by stochastic grammars and related parsing techniques, could be used to address these problems. The input to the parser could be regarded as probabilistic to reflect uncertainty due to erroneous physical layout analysis results [37].

However, the approaches described above are suitable for a specific set of the document images having regular layout style. Moreover, it is also stated by Tran et al. [38] that the state-of-the-art document layout methods are limited to particular document language. The methods are not adaptable to handwritten historical document images with irregular or overlapping layouts. During the last years, numerous research projects have also been set up with the goal of providing automatic layout analysis and indexing of historical document images. Some research projects deal with the whole document images while other research projects deal with analyzing the handwriting of the document images or the graphics/images present in the images. The next section describes some of these research projects.

2.2.1 Research projects related to layout analysis of historical documents

During the last years, several research projects started to analyze and explore the layout of handwritten historical documents to support enrichment and exploitation of digital libraries. These projects are described below:

The European project DEBORA (Digital AccEss to BOoks of the RenAissance) [39] proposed a complete system for analysis, indexing, retrieval, and compression of digitized Renaissance books. It extracts the metadata related to physical layout mainly using bottom-up algorithm and afterward compresses the document images. Here, at first all the connected components are localized, and then these components are merged to form higher interpreted elements, such as text or graphics. The text regions (handwritten area) and graphics (pictorial area) are then separated by comparing the size of each connected component to the average size of book characters. It also assists users to perform the manual transcription of printed documents with any arbitrary typography or language using computer-aided transcription (CAT). Moreover, it stores the document image and its physical layout and logical layout in a single annotation file using a proprietary data format.

The METAe project [40] developed a set of tools to digitize and perform automatic analysis of books and journals. In this context, this project modified a commercial OCR tool called DocWorks [41] to read the ancient books printed in the eighteenth century in Fraktur fonts. Docworks is used to automatically identify and describe the physical and logical layout of the documents. It then generates the metadata of the image and recognizes the characters from the printed documents using OCR. It can also identify page numbers, titles, and fonts.

AGORA [42], [43] is a user-driven annotation tool which performs layout analysis of printed historical document images and indexes them. AGORA also includes a new segmentation algorithm based on the connected component analysis technique. This segmentation algorithm builds two maps, a shape map for foreground information analysis based on the connected component analysis technique, and a background map for extracting blocks by performing white area analysis. Based on the simple descriptors, such as spatial position of the extracted blocks in the analyzed document image or the neighborhood relationships between the identified blocks, the user can define many indexing scenarios for the document images. After these indexing scenarios have been validated, they can be applied to the remaining of the document images. The most important use of AGORA is to extract and label the graphical regions. Its other uses include, the extraction of the table of contents and the transcription of the text blocks.

DMOS (Description and MOdification of Segmentation) [44] deals with the extraction of the layout structure from damaged military forms from French archives written in the

19th century. It recognizes 2D structures present in the images of military forms. For example, it detects each cell which is present in the military form. This method proposed a new grammatical formalism EPF (Enhanced Position Formalism) which is a description language for the structured documents.

STATE [45] is the computer assisted transcription system for ancient Spanish documents. This system consists of two main components, i.e. *StateTA* and *StateRE*.

- *StateTA* is a modular interactive application. It consists of four stages: the Project Manager, the Image Conditioner, the Layout Manager, and the Line Transcriber. The Project Manager is used for transcribing the set of pages. The Image Conditioner offers a set of tools for restoring the torn or faded out ancient documents. As a result, an enhanced document is obtained which can be directly fed into an OCR system. The Layout Manager offers a set of tools for automatically identifying the layout of the page and an interactive environment to edit the false detected layout. Moreover, lastly, the Line Transcriber helps to transcribe all the lines present in the text of the page.

- *StateRE* is a recognition engine. It can be easily adapted to other documents because it learns from the transcription of the pages obtained from *StateTA*.

Transkribus[1] is a platform to perform automated recognition, transcription, and searching of historical documents. It consists of three components a) an expert tool b) a web interface and c) cloud services. These services offer a set of tools for automatic analysis of the historical documents. It includes Handwritten Text Recognition (HTR), layout analysis, document understanding, writer identification and optical character recognition using ABBYY Finereader Engine [46].

Larex [47] is a semi-automatic tool for the segmentation and classification of regions enclosed in the early printed books. It is a rule-based connected components approach. It also allows intuitive manual correction from users if necessary. The results are stored in the PageXML format for further integration into OCR workflows. It works on two assumptions: a) Related characters, words, and text lines are closer to each other than other unrelated ones. b) it assumes that the pages within the same book have similar layouts. However, to generate a generic approach, a more extensive set of rules needs to be added to program code or during the experiment.

The HisDoc [48] research project mainly deals with the analysis, recognition, indexation, and retrieval of the historical documents. It is organized into three modules described below:

[1] `https://transkribus.eu/Transkribus/`

- ***Layout analysis***
  This module deals with the detection of different layout elements present on a digitized historical document, such as ornaments, illustrations, and text elements. Machine learning techniques are mainly used to extract the layout elements at different levels of resolution automatically.

- ***Handwriting recognition***
  This module deals with the transcription of historical document images into machine-readable text. It receives the text lines images from the layout analysis module as input and automatically transcribes it into a machine-readable format.

- ***Information retrieval***
  This module creates a search engine for transcribed manuscripts. It receives machine-readable text as input from module handwriting recognition module and performs a full-text search on them.

This system is applied to the datasets present in IAM-HistDB [49], which contains freely available handwritten historical document images, such as the Parzival database containing 47 images, the Saint Gall database containing 60 historical document images, and the George Washington database consisting of 20 pages from the George Washington papers.

DivaDia [50] is a research project related to HisDoc 2.0 [48]. The main goal of this project is to develop a tool for semi-automatic layout analysis. It helps users in labeling parts of the document, such as text, images, and initials. Based on the annotations made by the user the tools learn a model and enables to predict the labeling of the unseen page. The user then validates the result by accepting or rejecting the anticipated solution.

Alberti et al. [51] presented a tool for evaluating the layout analysis task of the document images at pixel level, and also support multi-labeled pixel groundtruth. This tool provides many metrics to investigate the layout analysis predictions, such as *Hamming score, Precision, Recall, and Jaccard Index*. Once the document image is evaluated, a visualization method is provided which allows users to look at the quality of the prediction through colors. Here, the visualization can also be overlapped with the original image. Such a visualization is capable of interpreting the segmentation mistakes by displaying exactly which pixels have been misclassified.

From the above examples of research projects, it is clear that a lot of research work has been done related to the historical document images in the direction of image preprocessing, layout analysis and graphics recognition, etc. However, to the best of author's knowledge and also stated by Cheriet et al. [52] there is no generic method which can be applied to large and heterogeneous historical document images. He also indicated that the combination of various document analysis algorithms and design of a multi-level framework could lead to optimized exploitation of historical document images.

Moreover, approaches to discover knowledge and making informed decisions from the vast, heterogeneous, and irregular document images are still missing. Only the automatic analysis approach proposed by Grana et al. [53] enables retrieval and exploration of a particular set of picture regions present in the digital collection of Borso d's Este Holy Bible using Sammon Mapping [54]. It is used to visualize the correlation between data, similarities, and clusters. A query image is set at the center, and other images are positioned randomly around this query image, but they are proportional to the distance between each pair of images in the two-dimensional space. Therefore, the domain of information visualization and visual exploration may help to solve the issues related to knowledge discovery and decision making. These two domains are described in detail in the following sections.

2.3 Information visualization and visual exploration

Availability of a significant amount of data presents a challenge in gaining knowledge and making informed decisions. Information visualization is a research field which can help researchers and practitioners to tackle this challenge. The information visualization tools help to visualize a large amount of information in detail as well as the summary of it. As described in the previous chapter, information visualization is the use of computer-supported, interactive, and visual representations of data to amplify cognition [11]. Card et al. [11] proposed the theory of creating effective information visualization techniques by using the reference data model for visualization as shown in Figure 2.4.

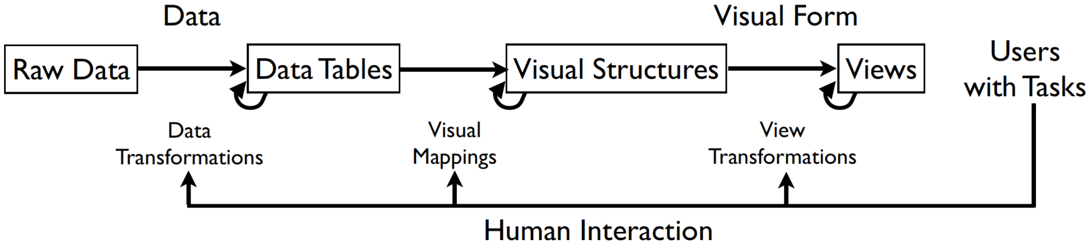
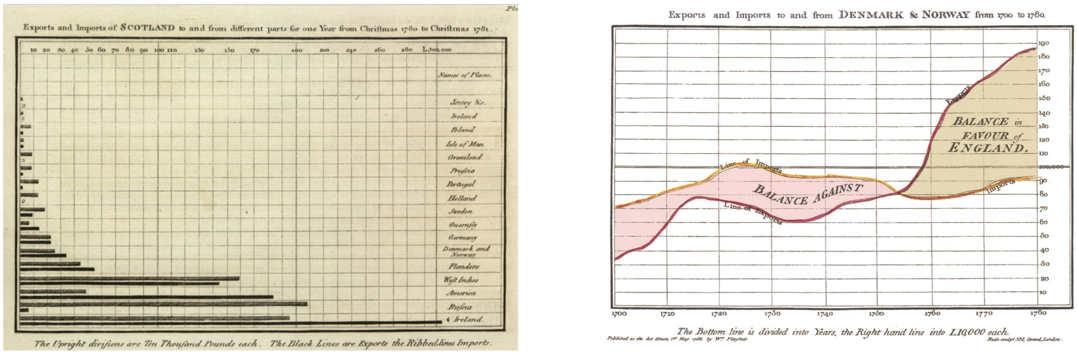


Figure 2.4 – Card et al. [11] reference model for visualization.

It starts with the processing of raw data in order to convert the original data format into the format which is suitable for the visualization techniques. The most commonly used data format suitable for information visualization techniques is a data table. It represents a set of relations, also called tuples. Tuples are represented by the rows of the data table and attributes are represented by the columns of the data table. These attributes can be of many types, such as nominal or ordinal. After the transformation of the raw data into data tables, the abstract data presented in the data tables are mapped to a visual structure to help users identify clusters, patterns or errors in the data. Here, the choice of the visual structure to represent the data matters most. After that, the identified visual structure is transformed to generate a visual view, and the users can interact with this view to reveal new information which is not known before. The beginning of the most intensive research in the field of information visualization started with the invention of the line chart, area chart and bar chart in 1786 and pie chart, circle graph in 1801 by William Playfair [55]. Figure 2.5 shows his early visualizations where (a) shows exports and imports of Scotland to and from different parts for one year from Christmas in 1780 to Christmas in 1781 (b) shows time series graph of trade balances between England and Norway/Denmark 1786.

Since then the field of information visualization has evolved till today leading to the creation of dynamic visualization tools and techniques for visualizing the high-dimensional data. Currently, it is being used on a daily basis in one form or the other for example subway routes, or map of a city or a country, weather reports, and sports activities. Today there exists a tremendous amount of visualization tools and techniques to see the data structure in detail, find patterns, discover outliers, gain knowledge and make informed decisions. The section below describes the classification of these information visualization methods.



(a)

(b)

Figure 2.5 – Early visualizations of William Playfair a) exports and imports of Scotland to and from different parts for one year from Christmas in 1780 to Christmas in 1781, b) the time series graph of trade balances between England and Norway/Denmark 1786. (Image from Wikimedia Commons)

2.3.1 Information visualization classification schemes

There exist numerous classification schemes for the information visualization methods. This section describes the most relevant taxonomies proposed till date to classify different information visualization methods.

2.3.1.1 Classification scheme by Shneiderman

In order to categorize the information visualization techniques till the year 1996, Ben Shneiderman proposed a Task by (Data) Type Taxonomy [56]. This taxonomy mainly focuses on two aspects, namely data type and the task type. The data is categorized into seven data types: 1-dimensional, 2-dimensional, 3-dimensional, temporal, multidimensional, tree, and network data. The tasks are also classified into seven task types: overview, zoom, filter, details-on-demand, relate, history, and extract. This taxonomy is further summarized below.

- ***1-dimensional data***: This type of data consists of textual data or source code, and the most common tasks which can be performed on this type of data are related to finding certain sections having certain properties.

- ***2-dimensional data***: This data type mostly consists of geographical displays or building plans. The tasks which can be performed are finding adjacent items, or finding paths between different items.

- ***3-dimensional data***: This type consists of various physical objects having volume. For example, buildings, or molecules. Basic tasks include determination of inside and outside relationships or finding of patterns, outliers in the data.

- ***Temporal data***: It includes timelines from the financial or medical fields or the time plan of the software project. The task which is normally performed on this type of data is finding of events during some period.

- ***Multidimensional data***: Relational and statistical databases are considered as the multidimensional data. Finding outliers, patterns, clusters, correlations, and errors are most common tasks which can be performed on the multidimensional data.

- ***Tree data***: It includes hierarchical data, where each node or item in the tree is linked to its parent except the root node. Tasks normally involve finding structural properties.

- ***Network data***: It involves the data which cannot be represented hierarchically. Each data item can be related to an arbitrary number of data items. It involves finding the shortest or the longest path between the data items as one of the user tasks.

Further description of Task by (Data) Type Taxonomy is followed by seven task types as described below:

- ***Overview***: helps to gain an overview of the entire dataset.
- ***Zoom***: helps to zoom in the data of interest.
- ***Filter***: involves filtering out uninteresting data items from the dataset.
- ***Details-on-demand***: involves selecting a data item and getting details when needed.
- ***Relate***: enables to view relationships among various data items.
- ***History***: allows to keep history of actions to perform undo and redo.
- ***Extract***: allows extraction of sub-collections and query parameters.

This classification scheme did not consider various interaction techniques which can be performed on the above-mentioned data types. Moreover, the focus of this classification was not to classify different software frameworks or methods which are used in the field of information visualization. Therefore, Keim presented a new information visualization classification scheme in the year 2002 as described in the next section.

2.3.1.2 Classification scheme by Keim

The classification scheme presented by Keim [57] is based on three main factors namely, data type to be visualized, the visualization technique, and the interaction and distortion technique as shown in Figure 2.6 and described below.

The first category, **data to be visualized** is extended from classification scheme given by Schneiderman [56]. It includes 1-dimensional data, 2-dimensional data, multidimensional data, hierarchies and graphs as described in the previous section. Additionally, this classification consists of text/web which includes news articles, documents, or web documents and software, algorithms which includes various debugging operations.

- **1-dimensional data**: This type of data usually contains one dimension, for example, temporal data, time series of stock prices or time series of new data.
- **2-dimensional data**: It has two unique dimensions, for example, geographical data containing two dimensions: latitude and longitude. X-Y plots are most commonly used to visualize two-dimensional data.

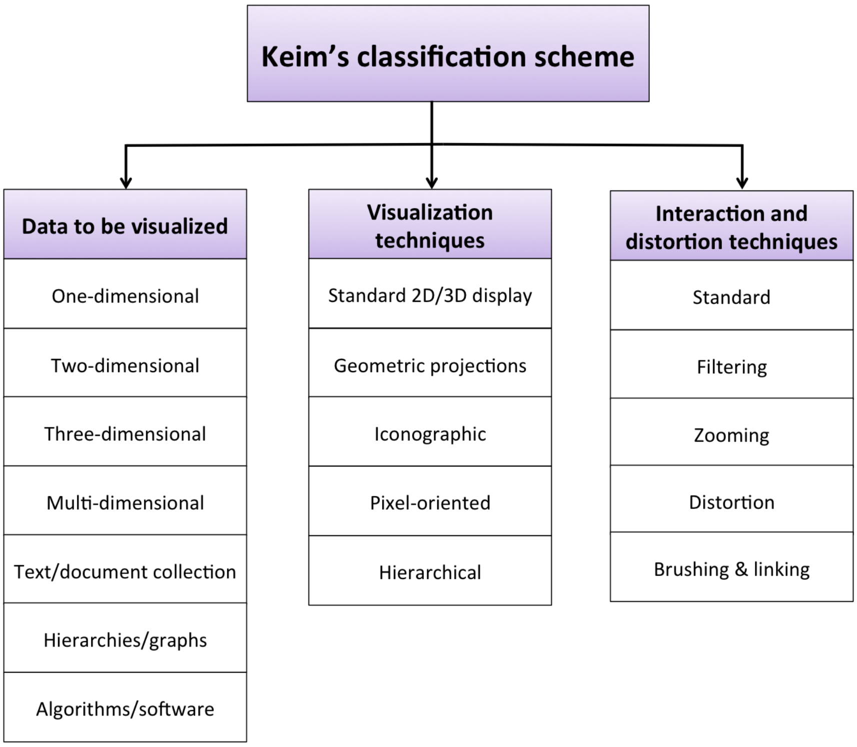


Figure 2.6 – Keim's [57] classification scheme. (Image from [57])

- **Multidimensional data**: The dataset containing more than three attributes. A typical example of multidimensional data are the relational databases. Such databases normally contain tens to hundreds of columns. Parallel coordinate plot and scatter plot matrix allows visualization of multidimensional data.

- **Text and hypertext**: This type of data cannot be described regarding dimensionality. It consists of news articles, reports, textual documents, or web pages.

- **Hierarchies and graphs**: In this type, data records share some relationship with the other records. Graphs are most commonly used to visualize such relationships. Each data record is a node, and the relationship between these nodes is described by links.

- **Software and algorithms**: This type includes written software code. Keim [57] stated that the goal of this kind of visualization is to support software development by helping developers understand the algorithms.

The second category, which is **visualization techniques** can be classified into following classes:

- **Standard 2D/3D displays**: This display technique visualizes 1-dimensional, 2-dimensional and 3-dimensional data. Examples include bar charts, line charts, x-y plots, 3D bar chart, etc.

- **Geometric projections**: The main goal of this display is to find an interesting pattern in the data. Typical examples of such displays include a scatter-plot matrix, parallel coordinates, star coordinates.

- **Iconographic displays**: This display maps the attribute values of multidimensional data to icons which may be stick figures icons, Chernoff faces or shape coding.

- **Pixel-oriented displays**: Each value of the dimension is mapped onto a colored pixel. This display can show about 1,000,000 data values as pixel gets one data value. Example of dense pixel display includes circle segments.

- **Hierarchical displays**: This display represents data in a hierarchical form where the child node shares a relationship with the parent node except the root. For example treemap, hyperbolic tree.

The third category is the **interaction and distortion techniques**. Here, interaction techniques allow the users to interact with the information visualization techniques and distortion techniques allow the users to focus on details of the information visualization without losing out the overview. The interaction and the distortion techniques is further classified as follows:

- **Interactive Filtering**: This type of technique filters out the most interesting part of the visualization. For example, Magic Lenses which has a magnifying glass. The data which comes under the magnifying glass is filtered out and shown in a different view.

- **Interactive Zooming**: Zooming not only makes the items larger but it can also show details very well.

- **Interactive Distortion**: This type of interaction technique allows to explore small parts of the data without losing the overview of the data.

- **Interactive Linking and Brushing**: Brushing and linking helps to combine two or more visualization techniques. Interactive changes made in one visualization are automatically reflected in the other visualization.

In the classification scheme mentioned above, the author did not distinguish between the basic visualization techniques and the systems or frameworks that implement those techniques. Therefore, Liu in 2014 presented a survey on state-of-the-art InfoVis techniques which is described in the next section.

2.3.1.3 Classification scheme by Liu

The information visualization survey presented by Liu [58] focuses on empirical methodologies, interactions, frameworks, and applications in the domain of information visualization. Liu [58] in his survey stated that real-world applications are the driving force behind information visualization research. In this context, numerous models, techniques, and methods are being proposed by visualization researchers worldwide. Therefore, he classified the recent information visualization research into four categories which are described as follows.

The first category, **empirical methodologies** includes visualization models, theories, and evaluation studies. Here, visualization models and theories provide a strong theoretical base for different applications and their evaluation. Most of the current visualization techniques employ usability studies and controlled experiments to investigate how the real-world users carry out a task and interact with the visualization technique.

Liu [58] classified **interaction category** into WIMP (windows, icons, mouse, pointer) interactions and post-WIMP interactions. WIMP interactions mainly focus on studying how the real-world users interact with the mouse and the keyboard. Post-WIMP includes how the users make use of touch interactions to interact with the devices which are beyond the paradigm of WIMP interactions.

The main aim of the third category, **frameworks**, is to design either a generic visualization framework which can be used for a large number of applications or a specific set of applications.

The fourth category consists of **real-world applications**: graph visualization, text visualization, map visualization, and multivariate data visualization. For example graph visualization can be used to explore the relationships between people or some entities. To visualize the text or a large number of document collections, the semantic meaning in the content is most concentrated, and spatial distributions are used to represent the geographic data.

2.3.2 Visual exploration

Visual exploration aims to provide an overview of the data and allows the users to browse through multiple levels of the visualization interactively. The user starts to explore expected information in the data and discovers unexpected information. Visual exploration is usually performed using the information visualization mantra: Overview first, zoom and filter, and then details-on-demand [56]. To execute this mantra, interactions on various visualizations play a significant role in the process of knowledge discovery and decision making. In theory, there exist two interaction paradigms for visual exploration which are described below [59].

- The first one, overview+detail-on-demand is the simultaneous display of the overview and the detail in the display space. It helps users to get an overview of the whole data set. The user then identifies an unusual pattern in the overview to further explore and drills down to access the detail of the particular pattern. This paradigm can also be extended to multiple levels where each level shows a different level of detail. The overview+detail-on-demand paradigm is widely being used in various everyday applications, such as Adobe Acrobat Reader, Microsoft Powerpoint, and Microsoft Word for showing the overview of the documents [60]. These applications support *thumbnail* document overview which allows users to get the overview of the whole document and access the detailed view when needed. Google Maps is another example where an overview is provided in an interactive rectangular region that corresponds to the area in the overview [61]. This paradigm has also been applied to various visualization systems, such as KronoMiner Interface [62], which is a visualization system which combines multiple coordinated components for exploring the time-series datasets as a whole and in detail.

- The second one, focus+context, distorts the overview to focus on particular parts of the visualization. In focus+context view, the user can access the details of the data without losing the overview. This paradigm allows users to comprehend and manipulate data in the large display spaces where all the parts are visible concurrently. Most commonly used focus+context interfaces is the fisheye view. In the past years many visualization systems showing focus+context paradigm have been developed, such as Document lens [63] uses continuous functions to distort the document regions. Here, the focused content of the document is demonstrated in the center, and this focused content is surrounded by the other parts of the document that provides an overview of the complete document. TableLens [64] allows providing the overview of the large datasets. It displays all the data values in the rows and columns as small bars. The fisheye effects allow to expand the rows and columns and explore the data values in detail.

2.3.3 The eight visual variables

The understanding of basic graphic primitives can be used to create a large number of information visualization systems. These graphic primitives are termed as elementary marks. Bertin [65] first presented seven visual variables (see Figure 2.7) for encoding the information with elementary marks, such as lines, areas, surfaces, points, and volumes. These *seven visual variables* are: position, size, shape, value, color, orientation, and texture. These seven variables were then extended to include motion [66]. Therefore, in total there exist eight visual variables. Each of these variables can be used in different ways to create information visualization design. These are described in following sections:

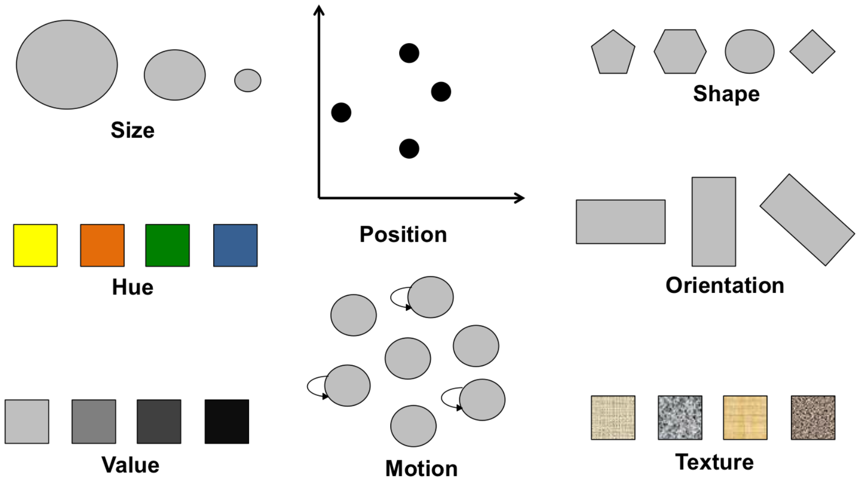


Figure 2.7 – Bertin's visual variables.

- **Position**: This is the first visual variable, and it represents the spatial arrangement or placement of various graphics primitives in a display space.
- **Size**: This variable is used to convey quantitative information, such as length, area, or volume.
- **Shape**: The shape represents basic graphic primitives, such as point, lines, or area.
- **Color**: Color is the visual phenomenon perceived especially by a human. It is defined by hue and saturation. Here, hue is the dominant wavelength of the visual spectrum and saturation defines the intensity of the hue relative to gray [67].
- **Value**: Value defines how light or dark a color is displayed.
- **Orientation**: This variable defines the direction of an elementary mark or how much a mark is rotated.

- **Texture**: This is the aggregation of other visual variables, such as elementary marks, color, and orientation.

- **Motion**: This can be associated with any of the visual variables, and it is used to convey information which changes over time.

2.4 Text and document visualization

As the primary goal of this thesis is related to the layout analysis and visual exploration of the document image, the text/hypertext and multidimensional data from the Keim's classification scheme and text visualization from Liu's classification are considered most interesting. Text or document visualization is gaining in importance due to the availability of online textual documents which exist in the form of news data, journals, books, historical documents or corpus. Currently, text visualization techniques either focus on the visualization of raw textual data or the results of some text mining algorithms.

An enormous amount of research has been done in the field of text and document visualization to support various analysis tasks. For instance, Kucher et al. [68] presented a visual survey for text visualization techniques. They proposed an interactive web browser and constructed a taxonomy of text visualization techniques as shown in Figure 2.8.

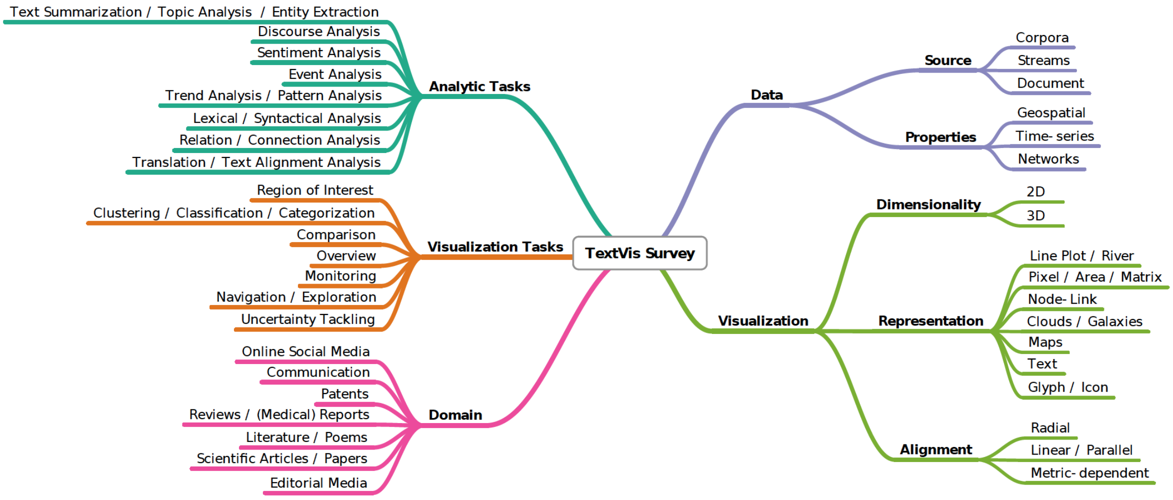


Figure 2.8 – Taxonomy of text visualization techniques presented by Kucher et al. [68], © 2015 IEEE

.

This taxonomy consists of following categories:

- **Analytic tasks**: This category consists of various high-level analysis tasks, such as text summarization, discourse analysis, sentiment analysis, and, trend/pattern analysis.
- **Visualization tasks**: This category contains low-level visualization and interaction tasks, such as determination of the region of interest, clustering, comparison, providing an overview, monitoring the changes in the data, exploring through the data set, and detecting or visualizing uncertainty.
- **Domain**: This category describes the domain for which text visualization techniques are developed. It includes online social media, medical reports, literature, which includes historical and documentary texts, scientific research papers, or newspapers.
- **Data**: This includes various data sources, such as a document, corpora, streams, and data properties, such as time-series, geospatial, and networks.
- **Visualization**: This category classifies the information visualization techniques according to the visualization approach, i.e., 2-dimensional/3-dimensional techniques. The representation consists of line plot, area diagrams, node-link diagrams, maps, text, icons. Alignment or layout includes radial, parallel, or metric-dependent.

The aforementioned taxonomy describes generic text visualization techniques. Some of the specific techniques based on the above taxonomy are described below:

Collins et al. [69] proposed DocuBurst, as shown in Figure 2.9. It breaks down a document into a tree-like structure with structures like sections, paragraphs, and sentences.

It visualizes the semantic content of the text document and the word frequency of the lexical databases created by human using SunBurst visualization. Interaction techniques like geometric and semantic zoom focus on individual words and linked access to the source text are provided.

The Bohemian Bookshelf [70] is used to facilitate an open-ended exploration of digital library collections. It consists of five interlinked visualizations where each of the visualizations provides a different perspective on the books collection. Bohemian bookshelf visualization shows the aesthetic and tangible qualities of books, such as the color of the book cover and page count, the temporal aspects of the books, such as publication year and content data, such as keywords and books author.

Literature fingerprinting [71] also summarizes the document content using a heatmap visualization where each cell of the heatmap represents a text block, and the color of

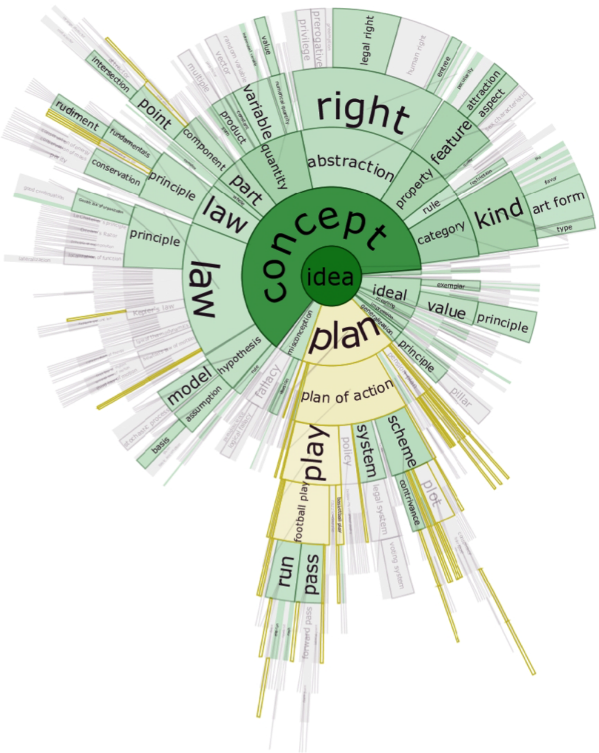


Figure 2.9 – DocuBurst of a science textbook rooted at idea (Image from [69] © 2009 by John Wiley Sons, Inc. Reprinted by permission of John Wiley & Sons, Inc.).

the text block shows the feature values per text as shown in Figure 2.10. The structural information of the document, such as average word length, average sentence length, and average number of syllables per word are visualized on different levels of resolution.

Semanticons [72] is a visualization system to represent text files by the graphical user interface icons. The icons serve as pictographic representations of the file contents. The current desktop interfaces have natural visual representation of the file, such as thumbnails used for the image files. This leads to identical icons that cannot be distinguished physically (visually distinguishable) or perceptually (viewer's understanding of semantics of file). The graphical user interface icons are automatically generated in this work to reveal the semantic content of the file.

The semantic content of the text file is extracted by its name, location, and content represented in an image. These resulting images are then further segmented by removing the unimportant regions. The resulting segmented image is then represented with the help of well-known graphical user interface icons.

Jigsaw [73] is an information visualization tool, which is designed to explore large

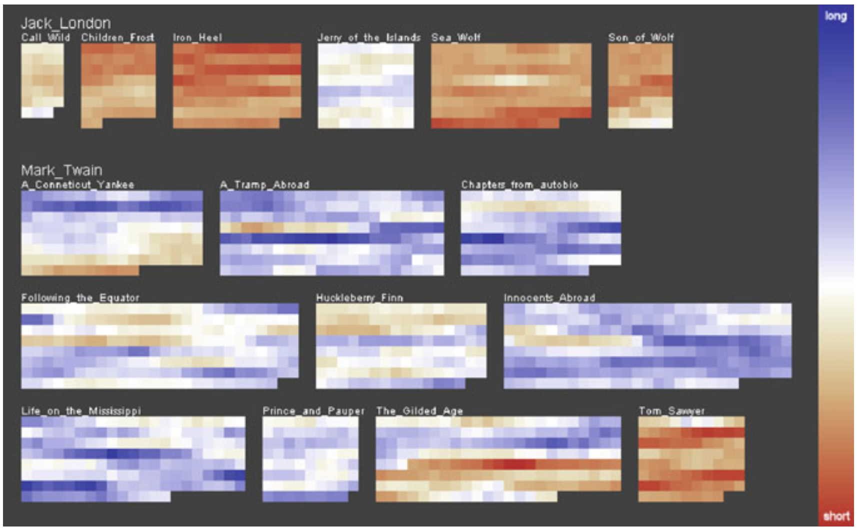


Figure 2.10 – Literature fingerprinting technique: here, each pixel represents a text block, and the pixels are grouped into books. Color is mapped to the feature value in this case showing the feature "average sentence length" of books written by Jack London and Mark Twain (Image from [71], © 2007 IEEE).

document collections and find hidden stories in the large document collections. It consists of multiple visualizations that highlight the relationships between the entities present inside the document collections.

Strobelt et al. [74] proposed a visualization system which converts a document into cards which look similar to top trumps game as shown in Figure 2.11. These cards summarize the whole document based on keywords and figures which are extracted from the document. The main keywords are extracted using the text mining approach which relies on the extraction of complete document structure. The picture regions and the captions present in the document are extracted using the graphical heuristic.

Data Mountain [75] is a visualization system which employs focus+context view. Here, the focused document are shown in front in a larger size, as compared to the unfocused ones which are shown in a smaller size at the back. This visualization system allows users to dynamically switch between the document by simply clicking on them and by focusing on a particular document.

Diggersdiaries [76], is a web interface for exploring historical, textual document collections. Each part of the web interface shows the letter and diary pages. These pages are represented as color-coded squares according to various data analysis topics, such as Personal (Christmas, Education, Food), War (Air, Health), Military Life, Traveling. The exploration is carried out by providing the overview of the data by-pages, by-diaries, and by-date.

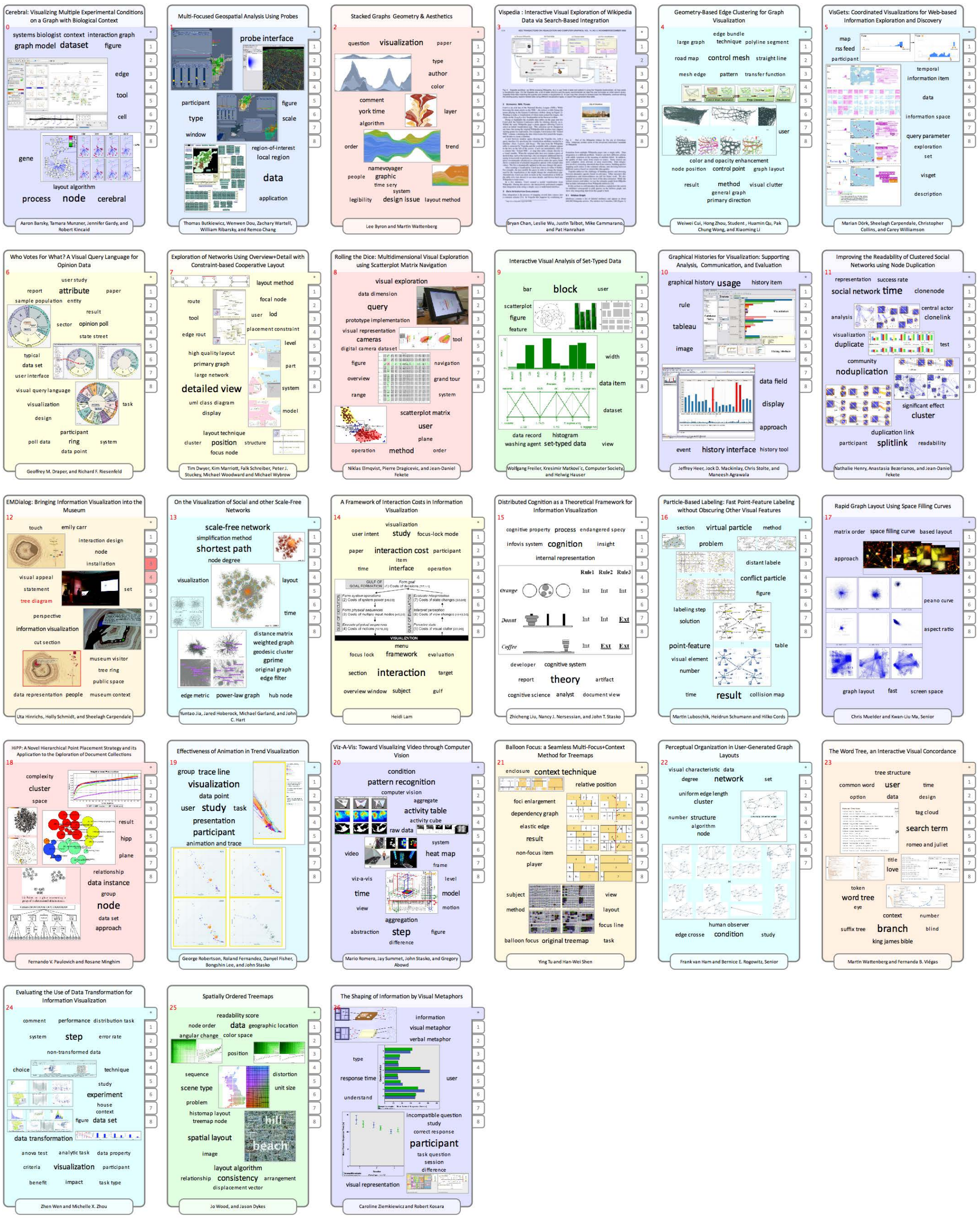

Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context
Multi-Focused Geospatial Analysis Using Probes
Stacked Graphs Geometry & Aesthetics
Vispedia: Interactive Visual Exploration of Wikipedia Data via Search-Based Integration
Geometry-Based Edge Clustering for Graph Visualization
VisGets: Coordinated Visualizations for Web-based Information Exploration and Discovery
Who Votes For What? A Visual Query Language for Opinion Data
Exploration of Networks Using Overview+Detail with Constraint-based Cooperative Layout
Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation
Interactive Visual Analysis of Set-Typed Data
Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation
Improving the Readability of Clustered Social Networks using Node Duplication
EMDialog: Bringing Information Visualization into the Museum
On the Visualization of Social and other Scale-Free Networks
A Framework of Interaction Costs in Information Visualization
Distributed Cognition as a Theoretical Framework for Information Visualization
Particle-Based Labeling: Fast Point-Feature Labeling without Obscuring Other Visual Features
Rapid Graph Layout Using Space Filling Curves
HiPP: A Novel Hierarchical Point Placement Strategy and its Application to the Exploration of Document Collections
Effectiveness of Animation in Trend Visualization
Viz-A-Vis: Toward Visualizing Video through Computer Vision
Balloon Focus: a Seamless Multi-Focus+Context Method for Treemaps
Perceptual Organization in User-Generated Graph Layouts
The Word Tree, an Interactive Visual Concordance
Evaluating the Use of Data Transformation for Information Visualization
Spatially Ordered Treemaps
The Shaping of Information by Visual Metaphors

Figure 2.11 – IEEE InfoVis 2008 proceedings corpus, represented by a matrix of document cards. The frequency of term is shown on right side of the document card (the more red, the higher the frequency). (Image from [74], © 2009 IEEE

.

Regarding visualization techniques which have been proposed for historical documents, Berzak et al. [77] proposed a visualization method for the document collections.

This visualization is a graphical representation which presents similarities between the documents according to the overlap of historically significant information, such as named entities or general vocabulary. Similarities between the documents are shown on the edges of the graph which reveals links between various documents. Their visualization method consists of a graphical user interface through which querying the database and visualizing the results are possible.

Jänicke and Wrisley [78] used a visualization and computational approach for identifying variance to allow the analysis of various medieval poetic using the transriptions of how they are found in manuscripts. They introduced a meso-level visualization, for representing the aligned text used for providing a comparitive study on the screen.

All of the visualizations as mentioned above either focus on the document exploration via summarization or they concentrate on visualizing the semantic content of the text to retrieve the required information. However, a document as a whole is just more than a text, especially handwritten historical documents with overlapping and irregular physical layouts. The humanities researchers studying the physical layout of the historical documents want to analyze the layout structure of each historical document in detail to explore the relationships between many document collections. To the best of author's knowledge, the visualization techniques described above do not help the humanities researchers to explore physical layout structure of the documents in detail.

Therefore, this thesis proposes a visual exploration method, which can help the humanities researchers to explore a large number of handwritten historical document images, their physical layout structure, and their corresponding layout features.

2.5 What is missing and trends?

Generally speaking, physical layout analysis of the handwritten historical document images has been an ongoing concern. In the recent past, many research projects have been started to solve the challenges associated with the historical documents which are also described in this chapter. However, the methods proposed in current research projects are still limited to solve particular kind of challenges associated with the historical documents. There exist tools like Transkribus [79] which can be applied to a variety of documents, but then the users have to compromise the accuracy and correct the results for each document manually which affects the reproducibility and the deterministic nature of the results. Also, the results generated by such tools are produced in the proprietary data format which cannot be read by other available tools. Such kind of tools focus only on identifying the physical regions enclosed in the document image but they do not extract the layout features of these identified physical regions. Therefore, the precise information describing the identified physical regions is still missing.

In order to fill this gap, a scientific contribution in the form of a generic approach which can be developed and applied to a wide variety of documents, identifies the physical regions and extract their corresponding layout features in a reproducible and deterministic manner is of utmost importance. Table 2.1 summarizes the research work related to layout analysis of historical document images and also highlights the generic and automated approach for layout analysis proposed in this thesis which will be described in the next chapters.

Additionally, approaches to further cope with the complexity of the data in order to extract relevant information is still missing. Currently, visualization researchers are designing a plethora of information visualization techniques to explore the data interactively through visual interfaces. However, the knowledge on how to choose an appropriate visualization technique which could be applied to a particular kind of application is of particular concern. Currently, various information visualization techniques and tools are available, but it is hard for users to decide which technique or tool is best suitable for their data when the analysis tasks are already ill-defined. For instance, with respect to the exploration of the physical regions and their layout features present in handwritten documents, the current text/document visualization techniques focus only on the semantic content of the text but could not be directly applied to explore the physical regions and the features of the documents. This leads to the question of choosing a visualization technique which could be used to examine the physical regions and the features present in documents.

Therefore, to fill this gap, a provision of generic design strategy is advantageous for the visualization researchers as well as for the users to identify the best possible technique and tools suitable for their kind of data and similarly for the handwritten historical documents in case of this thesis.

Table 2.1 – Summary of research work related to layout analysis of historical document images and automatic approach presented in this chapter.

Ref.	Research project	Goal	Dataset	Research work	Strategy	Data storage format	Visual exploration
[39]	**DEBORA**	Retro-conversion of historical documents	Renaissance books from 16^{th} century	Indexing, annotation and compression of document pages	Bottom-up (connected component analysis)	-	-
[42]	**AGORA**	Management of book contents and description	85 historical printed documents	Web-portal to provide accessibility, indexing of documents, enable users to search, retrieve specific documents	Bottom-up (connected component analysis)	-	-
[44]	**DMOS**	Recognition of 2D structures	88,745 military forms from French archives	Extraction of specific 2D structures from military forms	EPF (Enhanced Position Formalism), an parser which introduces context in segmentation	-	-

continued ...

. . . continued

Ref.	Research project	Goal	Dataset	Research work	Strategy	Data storage format	Visual exploration
[45]	**STATE**	Transcription of Spanish documents	Ancient Spanish documents	Complete assisted transcription system	Modular interactive application for image restoration, layout analysis and transcription	-	-
[53]	**Inside the Bible**	Image retrieval from illuminated documents	Renaissance Bible of Borso d'este	System for automatic segmentation, annotation and image retrieval	Bottom-up (connected component analysis)	-	Sammon Mapping
[2]	**Transkribus**	Automated recognition, transcription and searching of historical documents	Handwritten historical documents	Handwritten Text Recognition (HTR) system	Machine learning approach	Modified PAGE schema	-

continued . . .

[2] https://transkribus.eu/Transkribus/

…continued

Ref.	Research project	Goal	Dataset	Research work	Strategy	Data storage format	Visual exploration
[48]	**HisDoc, HisDoc 2.0**	Analysis, recognition, indexation, and retrieval of the historical documents	IAM-HistDB (74 documents from Parzival, 60 documents from Saint Gall and 20 documents from George Washington)	Modules for layout analysis, handwritten recognition and information retrieval	machine learning approach	Modified PAGE schema	-
[15]	**Chandna et. al**	Generic and extensible automated approach for layout analysis	Generic: 150,000 historical documents, 70 printed documents, 74 documents from Parzival and 60 documents from Saint Gall	Modules for preprocessing, region segmentation, and feature extraction	Machine learning approach	PAGE 2017 schema, CSV format	Multiple coordinated visual exploration

Chapter 3

Automatic layout analysis of handwritten historical documents

The aim of document layout analysis is to automatically identify and categorize different physical regions enclosed in document images. Moreover, the precise information about the physical regions helps to characterize the physical structure of the handwritten historical documents. The general approach is illustrated in Figure 3.1 and each part of the layout analysis approach is described in the following sections.

3.1 Dataset

As a starting point, this thesis uses 386 handwritten historical documents written between the eighth and the eighteenth century. These historical documents can be currently found in the City Library of Trier, the Diocese's Archive of Trier, and at the St. Matthias' Abbey itself [80]. It consists of approximately 150,000 handwritten historical document pages. Each of the document pages was digitized individually and enriched with bibliographic information, such as century of production, material, and binding format within the scope of the project "Virtual Scriptorium St. Matthias."

The primary goal of this digitization project was the digital reconstruction of the library to support further browsing and exploration. Each of the resultant digitized manuscript pages is in Tagged Image File Format (TIFF or TIF) format, which varies in size (5 MB – 150 MB) and has a high-resolution of 300 – 400 dots per inch (DPI). Additional high-resolution derivatives are also available in JPEG and PDF format for the web representation. The web representations of digitized medieval manuscripts and their associated bibliographic information can be accessed from the project homepage[1]

[1] http://stmatthias.uni-trier.de

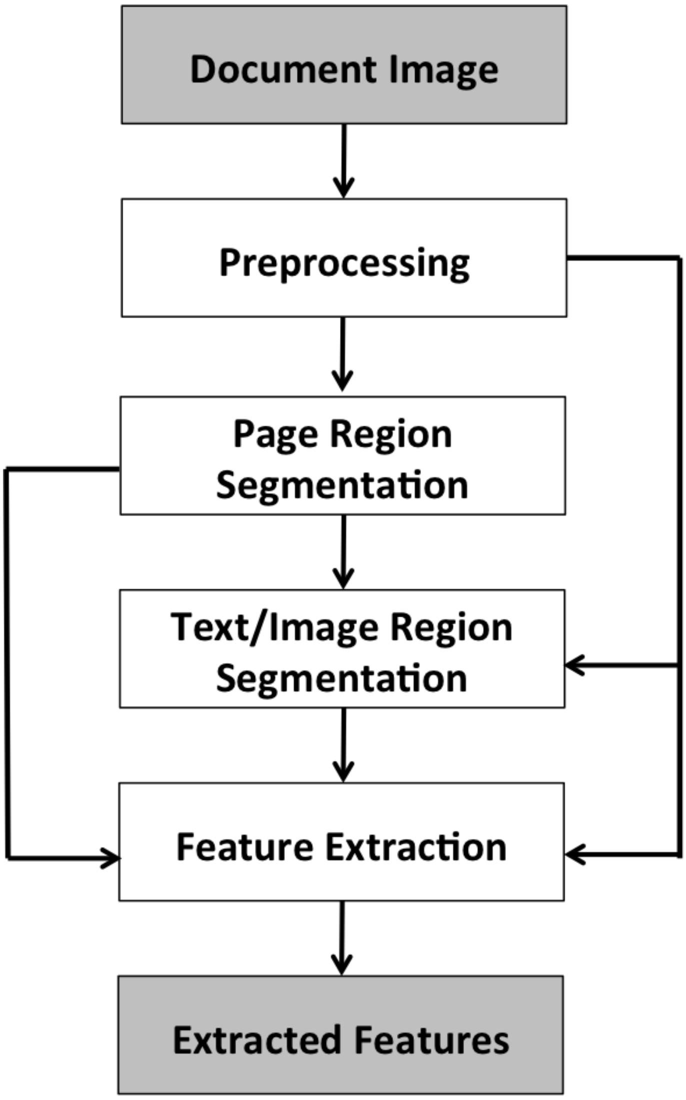
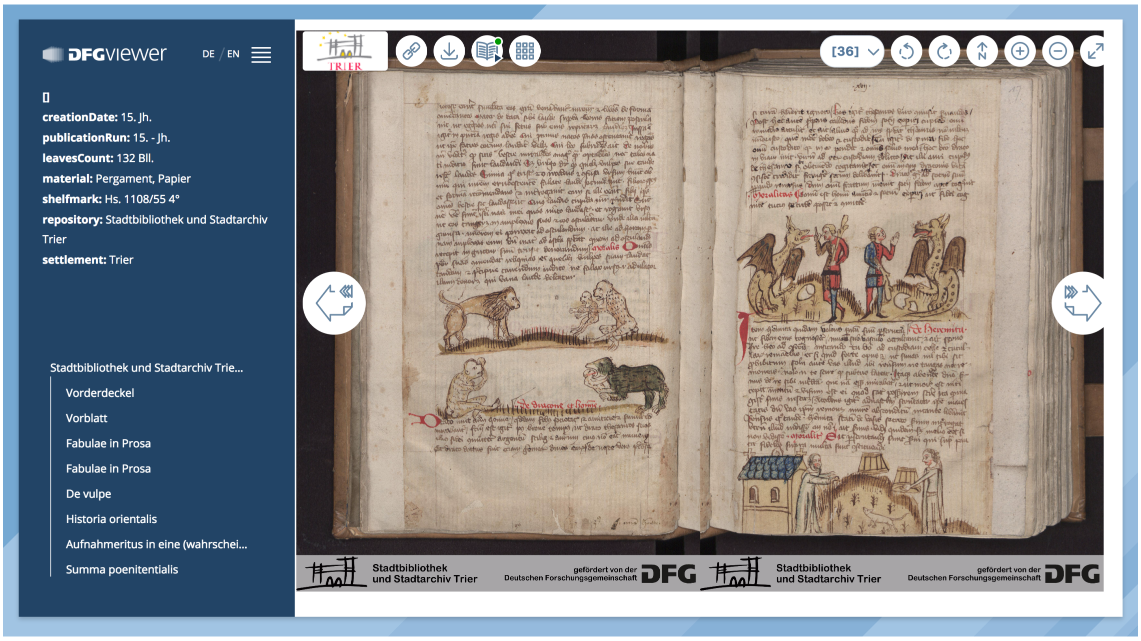


Figure 3.1 – General approach for identifying different physical regions enclosed in the document image and extracting their corresponding layout features.

using a user interface called DFG-Viewer, as shown in Figure 3.2. This homepage allows researchers to browse, zoom, and download these documents. The bibliographic information is presented in an XML format, which is structured according to the Metadata Encoding Transmission Standard (METS)[2] with a Text Encoding Initiative (TEI)[3] header because TEI is widely used in the digital humanities community to store information about textual data.

[2] https://www.loc.gov/standards/mets/METSOverview.v2.html

[3] http://www.tei-c.org/index.xml


DFGviewer
DE / EN
Stadtbibliothek und Stadtarchiv Trie...
Vorderdeckel
Vorblatt
Fabulae in Prosa
Fabulae in Prosa
De vulpe
Historia orientalis
Aufnahmeritus in eine (wahrschei...
Summa poenitentialis
TRIER
[36]
Stadtbibliothek und Stadtarchiv Trier
gefördert von der Deutschen Forschungsgemeinschaft DFG
Stadtbibliothek und Stadtarchiv Trier
gefördert von der Deutschen Forschungsgemeinschaft DFG

creationDate: 15. Jh.
publicationRun: 15. - Jh.
leavesCount: 132 Bll.
material: Pergament, Papier
shelfmark: Hs. 1108/55 4°
repository: Stadtbibliothek und Stadtarchiv Trier
settlement: Trier

Figure 3.2 – Digitized medieval manuscript T Hs 1108/55 4° 36v and 37r, from St. Matthias[1] (City Library of Trier, Germany) in DFG-Viewer.

3.1.1 Dataset handling

A research data repository was built using the software services provided by the research data repository software, KIT Data Manager[4]. It provides a generic and flexible architecture for the management of large-scale data as compared to other data management solutions, such as FedoraCommons[5], DSpace[6], which are specific to individual communities. The architecture of KIT Data Manager [81] is shown in Figure 3.3. It is organized into several layers, where each layer offers different services. It starts with a basic services and resources layer, then a high-level services layer, and an access layer.

- ***Basic services and resource layer***
  This layer includes standard services that are required to store, manage, process, and archive the data and its associated metadata.

- ***High-level services layer***
  This layer is the main building block of the whole architecture. All the core functionalities are provided by this layer, including data management, metadata management, search, staging, data processing, and lifecycle management.

[4] http://datamanager.kit.edu/index.php/kit-data-manager
[5] http://fedorarepository.org/
[6] http://www.dspace.org/

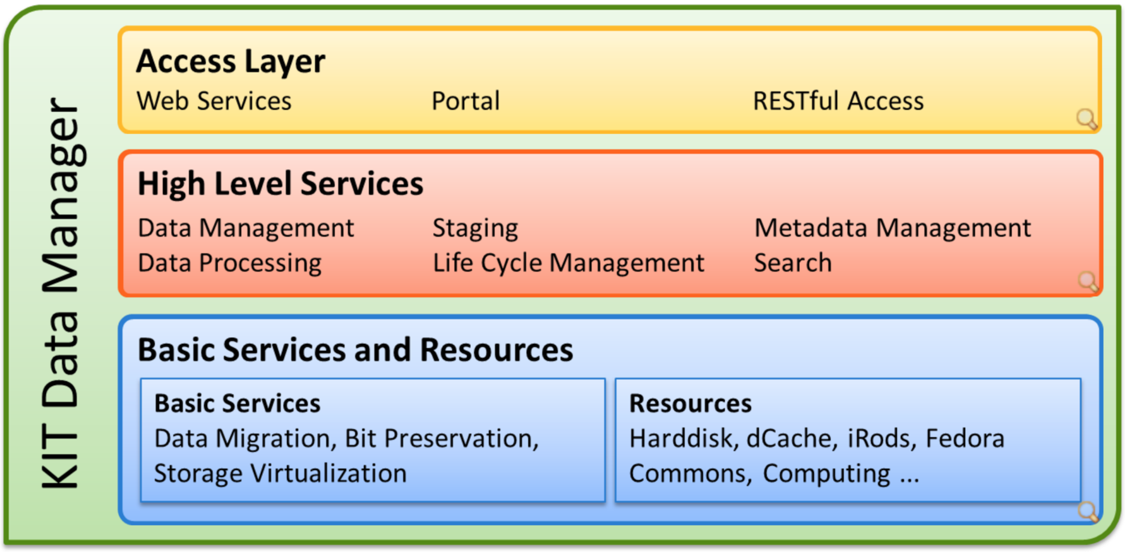


Figure 3.3 – Architecture of the research data repository KIT Data Manager.

- *Access layer*
  Each service provided by the high-level service layer can be accessed by this access layer via a set of RESTful Web Service or some community-specific web portals.

The minimum requirement for ingesting a rich and vast dataset is to provide metadata corresponding to the base metadata model of KIT Data Manager. The base metadata model of KIT Data Manager relies on the Core Scientific Metadata Model (CSMD) [82]. It creates a hierarchical structure consisting of three main elements: a) study, b) investigation, and c) digital objects. A study consists of investigations, and each investigation contains at least one digital object.

- *Study*
  The study represents the root element of the base metadata model. It describes the main topic of this research. For example, as this research work is primarily focused on the handwritten historical documents and the experimental dataset, in this case, was St. Matthias dataset. Therefore, the study's topic was set to "St. Matthias".

- *Investigation*
  The investigation describes the subject under consideration. Multiple investigations can be created for a given study. For example, in this case, an investigation was used to represent one manuscript or one document.

- *Digital Object*
  The digital object, in this case, is the representation of a single historical document

page including its related metadata. Each digital object has a unique identifier associated with it. Multiple digital objects were associated with a given investigation. In this case, there were three types of digital objects in one investigation. The first type of digital object corresponds to one manuscript page. The second type of digital object corresponds to the bibliographic information associated with the document images, and the third type collects all the representations of one manuscript, i.e., PDF, JPEGs.

According to the structure mentioned above, approximately 150,000 historical documents, 1,000,000 in other file formats, and 386 XML files with the bibliographic information were ingested to a research data repository. The remaining 20,000 historical documents could not be ingested because of incomplete datasets and the missing information in the XML files.

3.2 Layout analysis

This section describes the complete process of identifying the physical regions of a document image and extracting their corresponding layout features with the help of various image processing methods. Here, the layout features are any measurable property of a layout region enclosed in a document image – a number of text regions in the document image, bounding box measurements of the text regions, or color features of text regions. The main aim of the layout analysis is to collect multidimensional information, which helps to characterize the physical structure of the handwritten historical document images.

3.2.1 Preprocessing

Preprocessing involves the processing steps that are needed to make the acquired document image more suitable for the automatic document layout analysis. The aim is to reduce the dependency of the document images on data acquisition hardware, i.e., scanners or digital cameras caused when acquiring the images. It consists of methods like color calibration, spatial conversion, and scaling. These methods are described in following subsections.

3.2.1.1 Color calibration

At first, a median filter was applied as a precautionary step to remove the physical noise occurred due to analog-to-digital conversion by smoothing while preserving the discontinuities. This filter is applied because of its simplicity but it also changes brightness of the image. Better but costlier options than median filter would be bilateral filters.

Median filter helps to reduce the noise while preserving the edges of the text and image regions enclosed in the document image [83]. Here, the median filter works by sliding through whole image pixel by pixel and replacing each pixel value with the median value of the neighboring pixels. The arrangement of neighbors that slide pixel by pixel over the entire image is called the "kernel" or "window." In the case of document images, a simple 3-by-3 kernel was used.

Color calibration is the next essential step, which was performed after applying the median filter since the historical documents were scanned by different scanners having different representations of the same color (see Figure 3.4).

This made the digitized historical document images heavily dependent on the data acquisition hardware, i.e., scanner hardware. Color calibration is a process that helps to eliminate this dependency by transforming different color representations of the digitized images into a color representation that is similar to each other. The standard calibration methodology is to determine the transformation model between the scanned values and

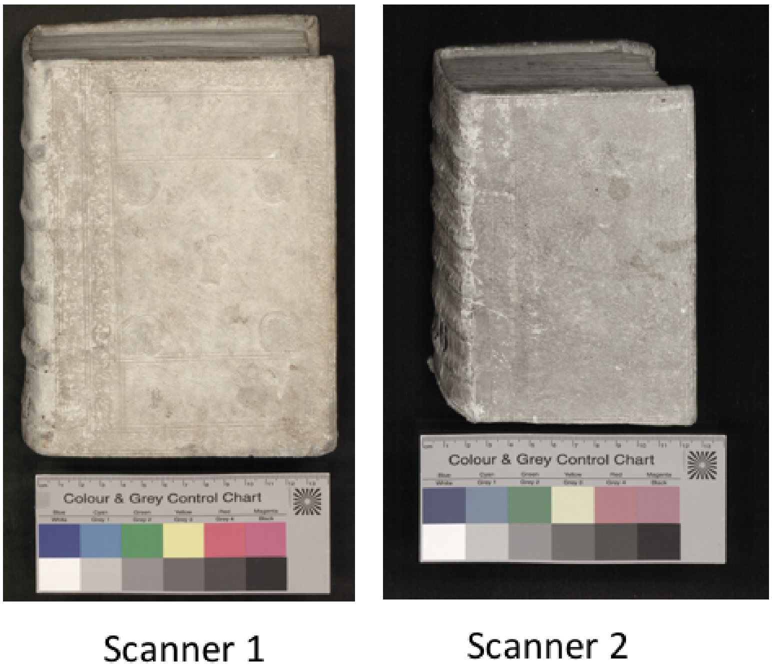


Figure 3.4 – Historical documents scanned by different scanners having different representations of the same color. For example, yellow color patch of the image digitized with scanner 1 is different to the yellow color patch of the image digitized with scanner 2.

the original values of the color chart. Therefore, to make variability of document images more straightforward and to have the same starting point for further layout analysis processes, it is assumed that there exists a linear RGB space. Therefore, a standard linear RGB (red, green, blue) transformation model was used for color calibration where gamma has been accounted for. The transformations inbetween can be described by affine mappings. The coefficients were determined based on the RGB color values of each color patch located on the color chart. The RGB values of each color patch of the original color chart are called target color values, and the R′G′B′ color values of each color patch scanned with different scanners are called input color values.

$$\begin{aligned} R_i &= a_{11}R_i' + a_{12}G_i' + a_{13}B_i' + a_{14} \\ G_i &= a_{21}R_i' + a_{22}G_i' + a_{23}B_i' + a_{24} \\ B_i &= a_{31}R_i' + a_{32}G_i' + a_{33}B_i' + a_{34} \end{aligned} \tag{3.1}$$

In this case "B.I.G. Color and Grey Control chart" was used for calibrating the images of St. Matthias as shown in Figure 3.5. This color chart has 12 color fields of size 2*2 centimeters. Therefore, a set of 36 equations was formed, and the required transformation coefficients were determined with the least square method. These transformation coefficients were then applied on the document image to get a resultant color-calibrated image as shown in Figure 3.6. Moreover, Figure 3.7 shows that the document images digitized with two different scanners resulted in RGB values which were similar to each other.

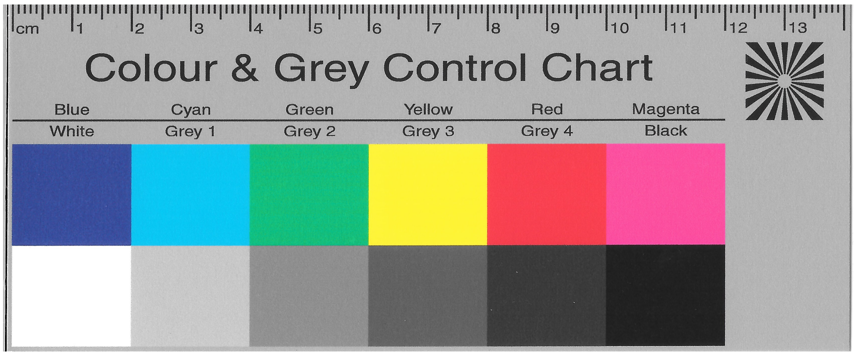


Figure 3.5 – B.I.G. color and grey control chart digitized with the St. Matthias document images.

The target and input RGB color values were stored in an XML format using proprietary XML schema. This was done in order to make the process of color calibration generic so that it can be utilized if other digital libraries use different color charts during digitization. Only the requirement of providing input and target color values using the proprietary XML schema needs to be fulfilled for the application of color calibration process.

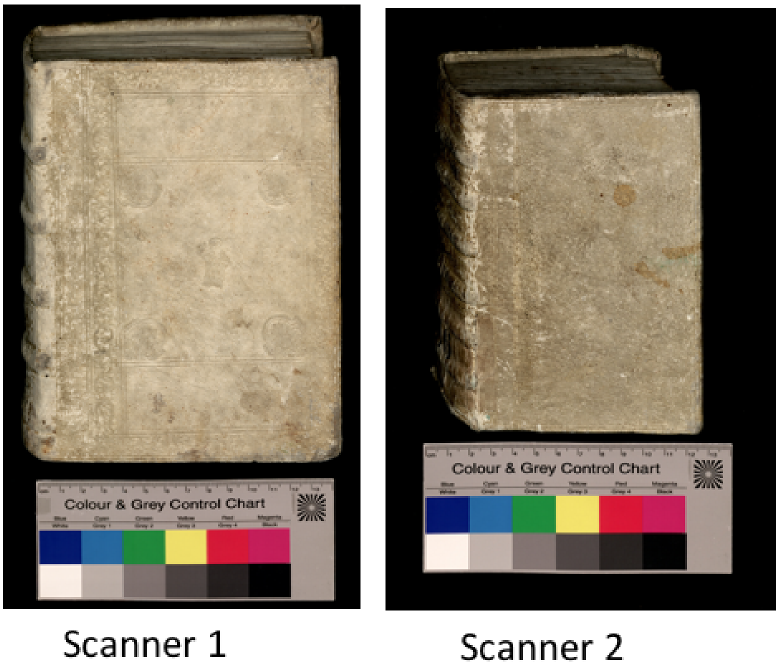
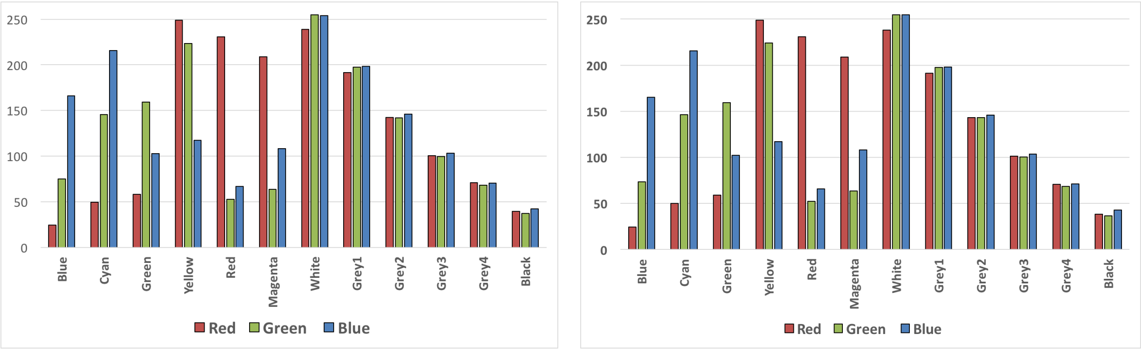


Figure 3.6 – Color calibrated document images having similar color representations.



(a) Scanner A

(b) Scanner B

Figure 3.7 – The document images digitized with two different scanners A and scanner B resulted in RGB values that are similar to each other after color calibration.

3.2.1.2 Spatial conversion

Spatial conversion helps in co-relating the pixels of an acquired image to the real-world units [84]. This method helps to make accurate measurements in real-world units, such as centimeters, millimeters, and inches. In this step, the original resolution of the image was used and transformed into centimeters. If the original resolution was not present with the image, it was provided externally, or it was assumed that the original resolution of the image was 300 DPI being a high-resolution image.

3.2.1.3 Scaling

The last step of preprocessing was to resize or scale the high-resolution images to a smaller resolution. The high-resolution handwritten historical document images were scaled down with the resizing factor of 0.2 to decrease the computation time of further processing steps. The bicubic interpolation method was used to downsample the image as it uses 4×4 samples to interpolate the pixels which result in a smoother image and with very negligible artifacts as compared to Nearest Neighbor and Bilinear interpolation algorithms which uses 2×2 neighborhood [85].

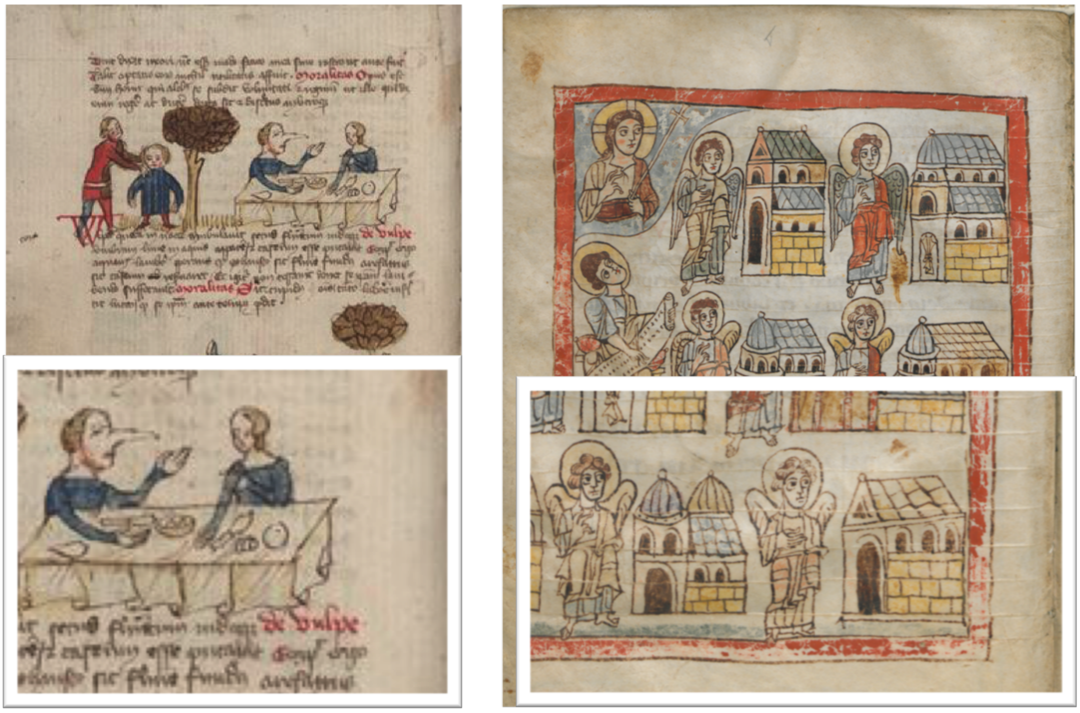


Figure 3.8 – Historical handwritten document images containing image regions with the same background color as document image itself.

3.3 Region segmentation

Region segmentation refers to the process that divides the whole document image into the constituent physical regions enclosed in the document and the background. Different constituent physical regions enclosed in the document image, include the page regions, text regions, and picture regions. It is one of the most challenging tasks in document layout analysis. The complexity of this process varies with factors, such as the non-uniform intensity of the background and intensity variations within the foreground regions. The background variations make it difficult to differentiate between the background and foreground as shown in Figure 3.8.

This section describes the segmentation of various physical regions enclosed in document images, which are page regions, text regions, picture regions, and other regions, such as red-colored regions. Here, each physical region is a circumscribed rectangle, or bounding box around the physical regions enclosed in the document images, such as page regions, text regions, or picture regions.

3.3.1 Page region segmentation

Page region segmentation refers to the process of detecting the border of the document image. The resulting circumscribed rectangle around the detected border is called the page region of the document image.

As these document images were scanned individually with the high-resolution scanners, following assumptions were taken into account.

- document images are nearly rectangular;
- document images are parallel to axes;
- document images are bright on dark border;
- the written ink on the document images is darker than the document image background;
- the written ink is sparse.

According to these assumptions, the projection profile method was used to identify the page region[7]. This method projects the foreground pixels on horizontal projection (top to bottom) and vertical projection (left to right) simultaneously [86] as shown in Figure 3.9. In order to identify the page region, the document image was first partitioned using a thresholding method. Here, thresholding is an image segmentation method that

[7] if these assumptions are not true for the document images, then the document images should be modified to meet the assumptions before applying the page region segmentation algorithm.

separates the foreground and background objects using a thresholding value. In this case, the document image was binarized using the Shanbhag thresholding method [87], which is an extension of the Kapur [88] thresholding method. It considers the image as the composition of two fuzzy sets which corresponds to the two classes with membership coefficient. This membership coefficient is to define how strongly a gray value is mapped between 0 (background) and 1 (foreground). The Shanbhag thresholding method also

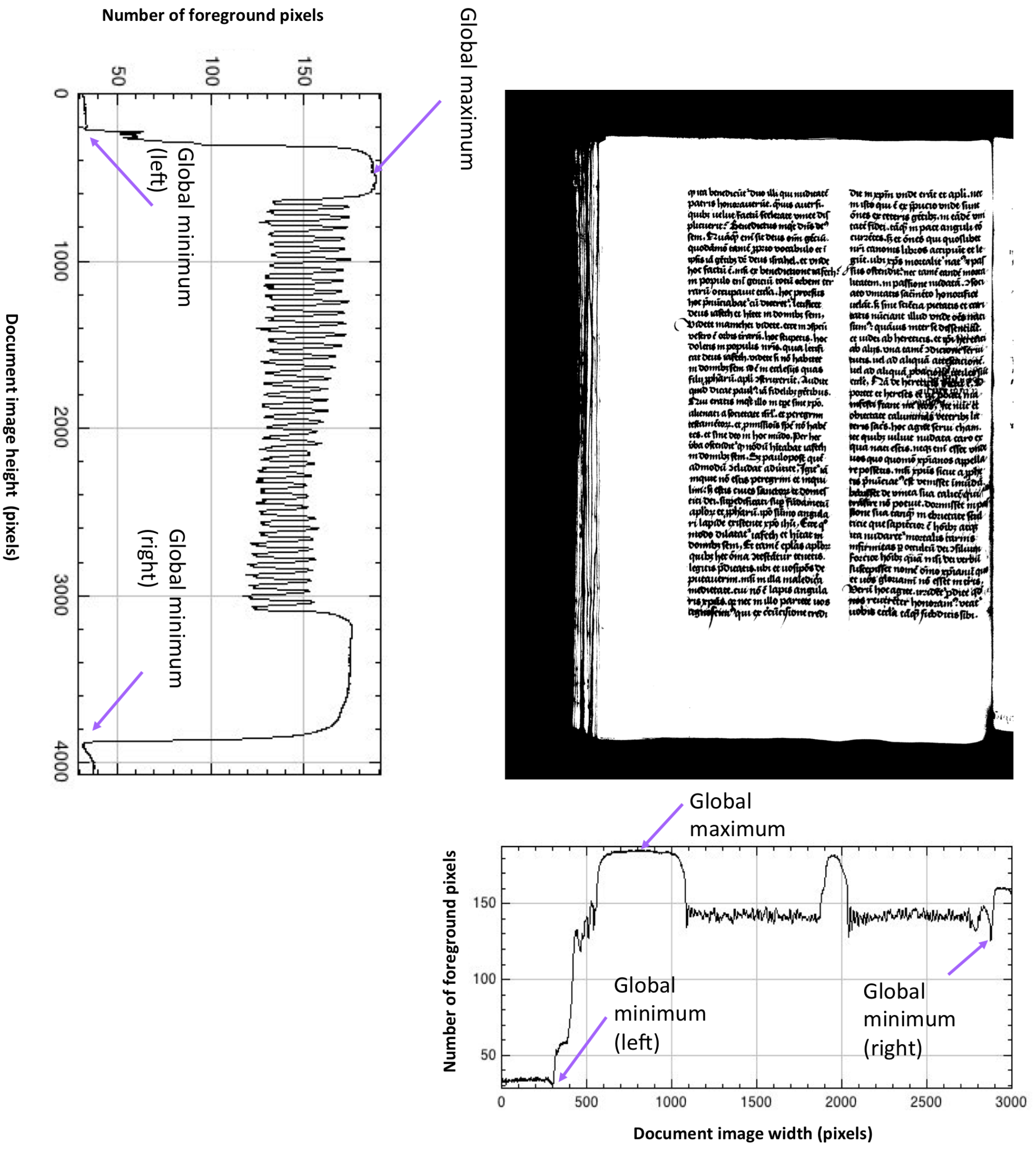


Figure 3.9 – Horizontal and vertical projections of the document image.

considered semantic content of the image for thresholding. As a result of applying Shanbhag thresholding method, a binary image was obtained; foreground pixels were set to a value of 1, and background pixels were set to a value of 0. This is because of the assumption that document images are bright on the dark background.

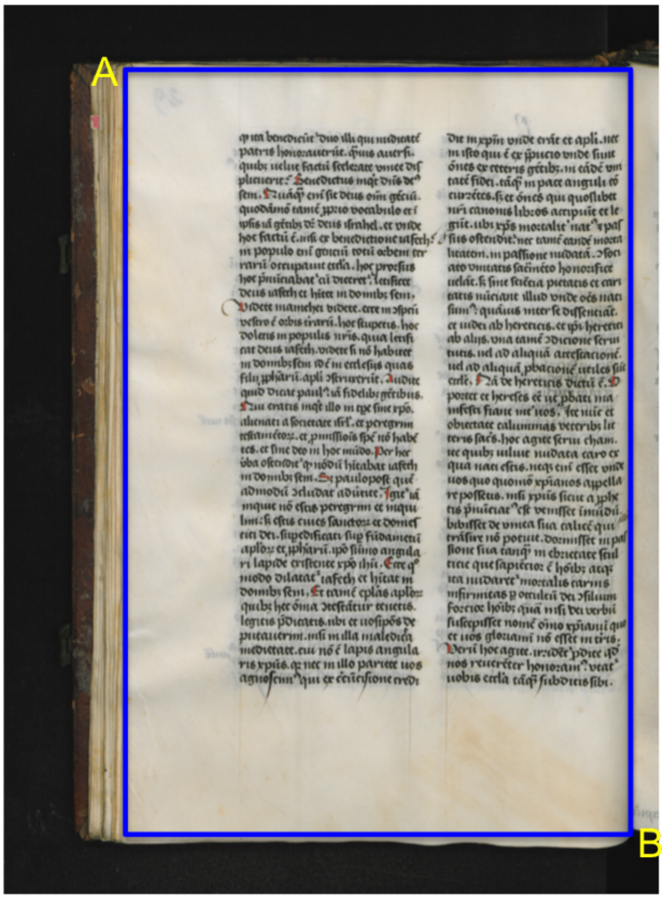


Figure 3.10 – Bounding rectangle representing the page region and the two boundary points, i.e. A represents the upper-left corner coordinate and B represents the lower-right corner coordinate.

Then, these foreground pixels were projected to create a vertical projection and a horizontal projection. After that, the global maximum was determined as shown in Figure 3.9 from the horizontal projection and the vertical projection and here it was assumed that this global maximum value of each projection lies within the document to be segmented.

In case of horizontal projection, the projected values were spanned to the left and right of the global maximum value along the *x-axis* to find the global minimum in both left and right directions.

This spanning was performed until the atomicity of the page region was met. In this case, the atomicity of the page region was defined by *model width* and *model height*, which was measured semi-automatically by outlining the page border of any arbitrary document.

As a result, the y values of the upper-left corner coordinates and lower-right corner coordinates were computed. The difference between these two y values represents the height of the page region enclosed in the document image. Similarly, by using the vertical projection, x values of the upper-left corner coordinates and lower-right corner coordinates were computed. The difference between these two x values represents the width of the page region enclosed in the document image. The bounding box representing the page region is shown in Figure 3.10.

3.3.2 Text region segmentation

Text region segmentation refers to the process of detecting written spaces enclosed in the document images. In this thesis, the bounding box circumscribed around the handwritten spaces is considered as text region of the document image as shown in Figure 3.11.

Machine learning

A machine learning process enables computers to learn the data without the help of any explicit programming instructions. Various machine learning algorithms can be categorized into:

- Supervised learning algorithms: These algorithms learn from the labeled training data and to generate a predictive function. This predictive function is then further used to perform the task of classification and decide the class label for unknown instances [89].
- Unsupervised learning algorithms: These algorithms learn from the unlabeled training data by clustering the data into classes by similarity and by reducing the dimensionality of the data while maintaining the structure.

This thesis will mainly focus on the supervised machine learning algorithms as they are most commonly used and can also be trained from the input provided by the users to build a classifier. Classifiers are the machine learning algorithms that perform the task of classification by assigning class labels for the new or unknown data by the training set for which the class labels were known in beforehand.

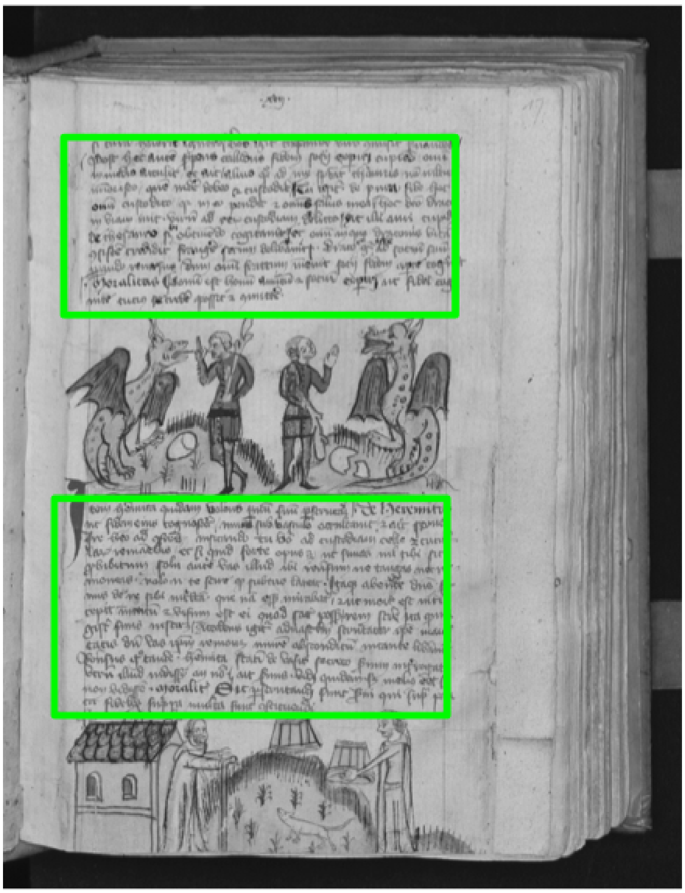


Figure 3.11 – Bounding rectangle representing the text regions.

Platforms for machine learning

There exists many platforms that provide various machine learning and image processing tools, such as MATLAB, the Trainable Weka Segmentation (TWS) [90], the Konstanz Information Miner (KNIME) [91], RapidMiner [92], and Vision with Genetic Algorithms (VIGRA) [93]. Here, KNIME and RapidMiner are the open source systems for the data mining but comprise a minimal set of image processing algorithms. On the other hand, tools, such as VIGRA offer various algorithms for mining and image processing, but this tool does not contain interfaces for the visualization of results. TWS provides a user-friendly interface for segmentation of images. It is included in the image processing toolkit Fiji (Fiji Is Just ImageJ) [94] as a plugin. It uses a machine learning process provided by Waikato Environment for Knowledge Analysis (WEKA) [95]. For this thesis, a machine learning approach provided by TWS was chosen for the segmentation of text regions because of the user-friendly convenience provided by TWS to learn the patterns in the image data without following the static programming instructions. One can interactively provide the training sample and obtain the results on the loaded image, allowing users to judge the segmentation results better.

Also, TWS provides a set of methods by which users can extract statistical properties of an image, called image training features. It provides the users complete freedom to select and tune these image training features. It consists of a variety of image training features to perform segmentation, such as Laplacian filter, Gaussian, and Sobel filters. The image training information includes image training features and the labels assigned to each class, and this information is stored in an ASCII text file called the Attribute-Relation File Format (ARFF) file[8]. It can be quickly loaded and manipulated with a linked Weka software package. The ARFF files require the declarations of @RELATION, @ATTRIBUTE, and @DATA. Here,@RELATION associates a name with the dataset, @ATTRIBUTE associates the name and datatype of the attribute (numeric, nominal, string, or date), and @DATA is the single line which denotes the start of the data segment.

The training features provided by TWS are grouped into following categories [90]:

- ***Edge detectors***
  These training features are used to detect the boundaries of the regions or edges in an image and include edge detectors, such as Laplacian filter, Sobel filter, and the Hessian matrix.
- ***Texture filters***
  These training features are used to extract the texture information from an image: minimum, maximum, mean, variance, entropy, and structure tensor.
- ***Noise reduction filters***
  These are noise-reducing features, such as Gaussian blur, as well as median, bilateral, Kuwahara, and Lipschitz filters.
- ***Membrane detectors***
  These detectors are membrane-like structures of a certain size and thickness.

Using the TWS plugin, a set of pixels belonging to the region of interest can be labeled. Then, these labeled pixels can be represented in an image training feature space and can be used as a training set. After that, a classifier can be trained with this training set to learn image training features associated with the labeled text regions. This classifier can then be applied to images to segment the regions of interest from a variety images. The detailed description of how this process was applied to the handwritten historical document images is described below.

[8] `https://www.cs.waikato.ac.nz/ml/weka/arff.html`

Training set

A stack of two-dimensional preprocessed gray-scale historical handwritten document images was used for training because they reduce the overall complexity of the images. Moreover, color information was not required to identify the text regions enclosed in the document images.

After that, the freehand selection tool provided by Fiji was used to label the pixels and collect the instances. For example, the instances belonging to the background of the document images were added to class 1 (background), and instances belonging to text region were added to class 2 (text region), provided in the graphical user interface of the TWS plugin as shown in Figure 3.12. As it was not known beforehand which of the image training features were relevant, all the image training features as described above were selected, and a training set with 120 training images was created and stored in an ARFF file. It stores the feature vectors derived from the pixels belonging to each instance into an ARFF file at a location chosen by the user.

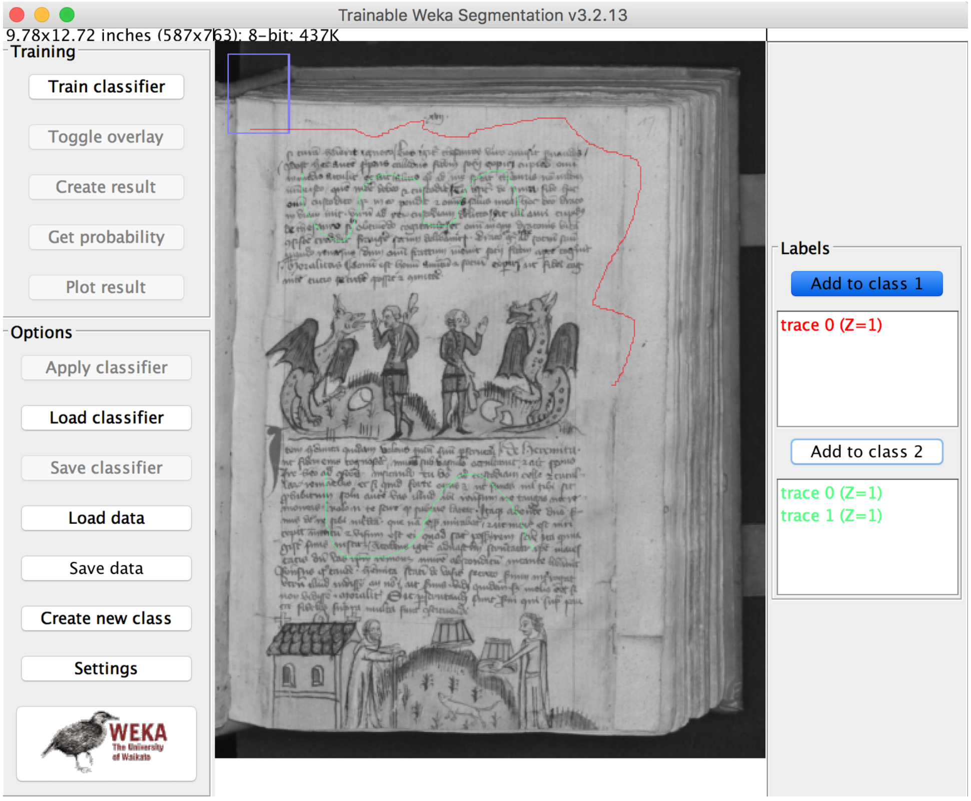


Figure 3.12 – Graphical user interface of the TWS plugin and the instances belonging to background and text regions added to class 1 and class 2 respectively.

Image training features selection methods

The feature selection method provided by the WEKA software package was used to remove irrelevant image training features from the training set (which does not influence the output) and capture only the relevant training features. The process of feature selection helps to reduce the computational time and increase the efficiency of the machine learning algorithms. This process consists of two selection methods: the attribute evaluator method and the search method.

- **Attribute evaluator method**: This method evaluates the set of image training features (attributes in Weka's terminology) and ranks them according to specific criteria. It includes correlation, information gain, and learner-based feature selection methods.

  - The **correlation-based evaluator** [96] calculates the relevant image training features by determining the correlation between the training feature and the class (output variable) itself. Here, correlation is defined by a Pearson's correlation coefficient. This coefficient measures the linear correlation between two variables and where the range of values is between +1 and -1. A value greater than 0 indicates a positive correlation, 0 means no correlation, and a value less than 0 indicates negative correlation.
  - The **information gain-based evaluator** calculates the entropy for each attribute/feature of the class. The results vary from 0 to 1. Here 0 means no information and 1 means maximum information.

    The entropy, H(A) can be defined as follows:

    $$H(X) = -\sum_{k=1}^{n}(P_k * log_2(P_k)) \tag{3.2}$$

    where P_k is the probability of the class k and n is the number of outcomes. Furthermore, Information Gain (IG) can be calculated as follows:

    $$\text{Information gain (A, B)} = H(A) - H(A|B) \tag{3.3}$$

    where A represents the class and B represents the attribute. A big weakness of information gain is that it gets biased in favor of the features with more data values even if they are not giving any useful information [97].
  - The **gain ratio** attribute evaluator is introduced to compensate for the bias of

information gain. It is defined as follows:

$$\text{Gain ratio} = \frac{\text{Information gain}}{H(A)} \tag{3.4}$$

where A represents the class. The information gain is normalized by dividing it by entropy. As a result of normalization, the values of gain ratio lie between 0 and 1. A value of 0 indicates that there is no relation between A and B and value of 1 indicates that the knowledge of A predicts B.

  - **Learner-based selection** uses a generic machine learning algorithm, such as a decision tree algorithm. Different subsets of features are selected, and the performance of this decision tree algorithm is evaluated. The subset that gives the best performance is used as the selected subset of features or attributes.

- **Search method**: This method searches the subset of all possible features to determine the best list of features. It includes exhaustive search, best first, and greedy stepwise search methods.

  - Exhaustive search or brute-force methods test all the combinations of features.
  - Best-first search methods search for the best possible set of features.
  - Greedy stepwise uses a greedy forward (additive) or greedy backward (subtractive) approach to search through features.

Evaluation of image training features selection methods

Since there exist many feature selection methods to choose from, a performance analysis was conducted to choose an attribute evaluator and the search method. The section below describes the performance analysis of various evaluator and search methods.

A number of attribute evaluation and search methods were used to create different views of the data. Each of these views was then further evaluated by building a classifier model. This classifier model was built using a generic decision tree algorithm (J48) and using a k-fold cross-validation method.

Here, the k-fold cross validation method is used to test the proficiency of a machine learning algorithm on unseen data. It systematically creates and evaluates multiple classifier models that are built on the multiple subsets of the dataset. The most commonly used value of k is 10 (10 folds). Here, 10-fold cross validation splits the dataset into 10 partitions. Each time the machine learning algorithm is run, 90% of the partitions are trained, and 10% of the partitions is used as the test set. After that, the average performance of all 10 models was calculated regarding classification accuracy, the time taken to build the model, and the root mean square error. Classification accuracy can be defined as follows:

$$\text{Classification accuracy} = \frac{\text{Total number of correct predictions}}{\text{Total number of predictions}} \tag{3.5}$$

The root mean square error (RMSE) is quadratic scoring rule, which measures the magnitude of error. It is calculated by the difference between the predicted values and the values actually observed. This helps to indicate the degree to which a given prediction may be wrong. It is the square root of the mean of squared differences between predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k - y'_k)^2} \tag{3.6}$$

where y_k are the predicted values, i.e., the probabilties of the class 1 and class 2 and y'_k are the actual values of the class 1 and class 2 i.e. 0 and 1.

Table 3.1 – Performance analysis of attribute evaluators and search methods

Attribute evaluator	**Search method**	**Classification accuracy (%)**	**Time taken (sec)**	**Root mean square error**
CfsSubsetEval	Best-first search	98.82	46.1	0.106
CfsSubsetEval	Greedy stepwise	98.77	51.88	0.1035
Infomation gain	Ranker	97.81	63.02	0.137
Gain ratio	Ranker	97.95	44.69	0.133

As a result of the performance analysis, it was found that *CfsSubsetEval*, which is a correlation-based evaluator, and the *best-first search* method outperformed other views, with a classification accuracy of 98.82 % and the RMSE of 0.106. This view generated a set of 30 training image features. This set included features, such as Hessian, membrane projections, maximum, Laplacian, entropy, difference of gaussians, variance, and derivatives. The labeled pixels were mapped to the feature space of this training set.

Machine learning algorithm selection

The environment of Weka also supports a variety of machine learning algorithms. But, it was also not known which of the machine learning algorithms will perform better in which conditions. Therefore, an evaluation of the suite of machine learning algorithms was done by selecting top five machine learning algorithms provided in the Weka for the initial comparison. Therefore, the 30 training image features and features vectors selected in the previous step were given as input to each of the machine learning algorithm.

- **Naive Bayesian (NaiveBayes)**: The algorithm employs the Bayes formula to perform a classification task. It provides a way to calculate the posterior probability of each class, and then output the class with the highest probability. More formally, the Naive Bayesian classifier can be defined as follows:

  $$P(M \mid N) = \frac{P(N \mid M)\, P(M)}{P(N)} \tag{3.7}$$

  - $P(M \mid N)$ is the posterior probability of a class given the attribute values
  - $P(M)$ is the prior probability of a class
  - $P(N \mid M)$ is the likelihood which is the probability of the attribute given a class
  - $P(N)$ is the prior probability of the attribute

- **Classification and Regression Tree (REPTree)**: This algorithm was introduced by Leo Breiman [98] that builds a binary decision tree that is first constructed by splitting a node continuously. Each node represents an input variable (i) and splits on the point where it is assumed that the variable is numeric. The leaf nodes of the binary decision tree consist of the target variable (t) that is used to make further predictions. The most commonly used method to stop the recursive splitting process is to utilize the minimum number of the training instances. When the total number of instances is less than this minimum number, the splitting procedure is stopped, and the leaf node is considered as a final leaf node. To further increase the performance, the resulting decision tree can be pruned. The most commonly used procedure to prune the trees is by using the reduced-error tree method. In this method, each leaf node of the decision tree is evaluated by calculating the effect of removing it and if the overall cost function is dropped then the leaf node is removed [99].

- **k-Nearest Neighbors (IBk)**: This algorithm is used to make predictions in the data using the entire training dataset. Here, the predictions on the unknown or new dataset are made by exploring the entire dataset to find the k nearest neighbors (similar instances) and then the output variable is summarized based on the resulting k-nearest neighbors.

  There are many measures to find k most similar instances in the unknown dataset, such as Euclidean distance, Hamming distance, and Manhattan distance. However, the most popular measure is the Euclidean distance that is defined [100] formally as follows:

  $$EuclideanDistance(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{3.8}$$

Euclidean distance is calculated as the square root of the sum of the squared differences between a point a and point b across all input attributes $\mathbf{i}$.

- **Logistic Regression (Logistic)**: It is a binary classification algorithm [101]. It is based on the logistic (or sigmoid) function that takes real-valued numbers and transforms it into values that lies between 0 and 1. This logistic function can be defined as follows:

$$\text{Transformed value} = 1/(1 + e^{\text{-}value}) \tag{3.9}$$

  where e is the constant Euler's number (base of the natural logarithms), and *value* is the actual value which needs to be transformed.

  The logistic regression model utilizes these transformed values to make a prediction of the probability of the default class (class 1) to which the value belongs. If the probability is $\geqslant 0.5$, the output for the prediction is in favor of class 1, and if the probability is < 0.5 the prediction is in favor of class 2.

- **Support Vector Machines (SMO)**: This algorithm works by finding a line that separates the data into two groups. The data instances that are close to the line and can best separate the classes are considered. These selected data instances are called support vectors. Sometimes, a margin is also added to relax this constraint [102].

- **Random Forest (Random Forest)**: Random forest is a machine learning algorithm used for classifying large amounts of data. It is based on an ensemble learning technique which means that many decision trees are created, and the response of each decision tree is taken into consideration. The final response is determined by the evaluating the responses of the individual decision trees. Each decision tree is grown according to the CART algorithm [103]. However, it considers only a small subset of randomly selected splits and the best split is chosen from this subset. An advantage of using such a machine learning approach is that the classifier can be tuned to various types of images and also would not overfit the model [104].

To compare the algorithms mentioned above, the training set and 10-fold cross-validation were used. Then, the classification accuracy rate for each of the algorithms was calculated. The results of the algorithms were compared to one result of the baseline algorithm, called the ZeroR algorithm. This algorithm mainly works by focusing on the target values specified in the class, and it has no prediction-making power. It only predicts the majority class and is most commonly used for determining a baseline performance and acts as a benchmark for other classification methods. Thus, any machine learning algorithm has to achieve classification accuracy greater than that of ZeroR to prove its accuracy. The number of correctly and incorrect predictions are described in a confusion matrix which is shown as follows:

		Actual Class	
		Class 1	**Class 2**
Predicted Class	**Class 1**	True Positive	False Positive
	Class 2	False Negative	True Negative

where True positives (TP) is the one that detects the class, when the class is actually present, true negative (TN) is the one that does not detect the class when the class is absent, false positive (FP) is the one that detects the class when the class is absent. False negative (FN) is the one that does not detect the class when a class is present.

Precision is the ratio of the true positives to the total positives predicted.

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \tag{3.10}$$

Recall is the ratio of true positives to the actual positives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \tag{3.11}$$

Table 3.2 shows the classification accuracy rates, true positive rates, false positive rates, precision and recall for the algorithms mentioned above. As a result of the comparison of classification accuracy rates, true positive rates, false positive rates, precision and recall, it was clear that the random forest machine learning algorithm outperformed the rest of the algorithms. Thus, this algorithm was chosen to train the training set. However, it was not known how many trees were sufficient to create a random forest. Therefore, to get the maximum amount of trees, a parameter tuning was needed. The following section describes the parameter tuning of the *random forest* machine learning algorithm.

Parameter tuning of the machine learning algorithm

Here, the parameter tuning of the random forest machine learning algorithm was investigated to perform the process of algorithm tuning or algorithm hyperparameter optimization. The process of parameter tuning is an empirical process of trial and error. In this case, the parameter that needs to be configured is called number of trees *(NumTrees)*.

Generally, the greater the number of trees in the random forest, the higher the accuracy of the results will be. In this case, parameter tuning indicates the maximum number of trees required to create an optimized classifier. The *Experiment Environment* provided by *Weka* allows to try various parameter of an algorithm and analyze their

Table 3.2 – Performance comparison of the machine learning algorithms

Machine learning algorithm	Classification accuracy (%)	Root mean square error	TP rate	FP rate	Precision	Recall
ZeroR	65.2307	0.4762	0.652	0.652	0.426	0.652
Naive bayes	84.6749	0.3772	0.847	0.184	0.847	0.847
REPTree	93.6553	0.226	0.937	0.080	0.936	0.937
IBk	91.6461	0.2889	0.916	0.095	0.917	0.916
Logistic	90.9002	0.2611	0.909	0.119	0.908	0.909
SMO	91.1762	0.297	0.912	0.116	0.911	0.912
Random forest	95.738	0.1829	0.957	0.055	0.957	0.957

results. *NumTrees* was arbitrarily chosen to be 20, 50, 100, and 200, and in this case, to get the best estimate of a number of trees. Here, a random forest with 100 trees was considered as the base configuration to perform the comparison. Table 3.3 shows the classification accuracy results.

Table 3.3 – Parameter tuning of the random forest machine learning algorithm

Number of Trees	Percentage correct (%)
20	95.21
50	95.40
100	95.47
200	95.49

The result of parameter tuning showed no significant difference between the base configuration and the other three configurations. When *NumTrees* was 50 or *NumTrees* was 200, it did not yield better results than the base configuration. Therefore, *NumTrees* = 50 was considered for the random forest machine learning algorithm.

Segmentation of text regions by pixel classification

In general, TWS transforms the segmentation into pixel classification where each pixel is classified as belonging to the particular class. Segmentation of text regions was done by pixel classification where a set of pixels that were labeled were represented in the feature space and were used as the training set for the selected classifier.

Therefore, a random forest classifier was trained with a selected list of image training features and number of trees, which was called the *text classifier model*. This classifier was then applied to the other handwritten historical document images. Figure 3.13 shows some of the results of the text regions segmentation which was done through pixel classification. After that, the bounding boxes around the classified pixels were measured to obtain the text regions.

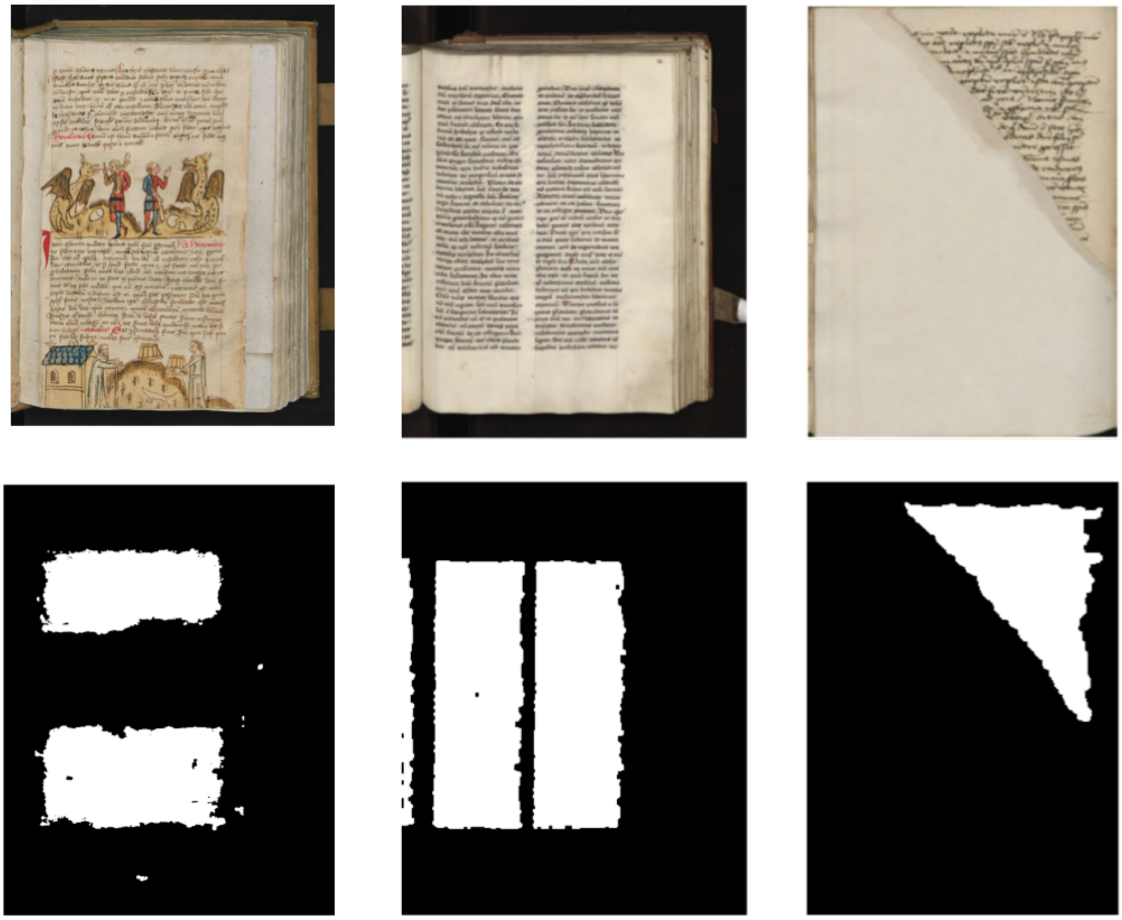


Figure 3.13 – Segmentation results for the text regions performed by pixel classification.

3.3.3 Picture region segmentation

Picture region segmentation refers to the process of detecting picture spaces enclosed in handwritten historical document images. In this thesis, the bounding box around the pictorial spaces was considered as a picture region of the document image, as shown in Figure 3.14. The segmentation of picture regions was also completed through the pixel classification process described above. The image training feature set, machine learning algorithm, and other parameter configurations, which were used for segmentation of text regions, were also used for the segmentation of picture regions. At first, the freehand selection tool of Fiji was used to label the pixels and collect the instances. For example, the instances belonging to the background of the document images, written spaces were added to the class 1 (background) using TWS, and the instances belonging to the picture regions were added to class 2 (picture region).

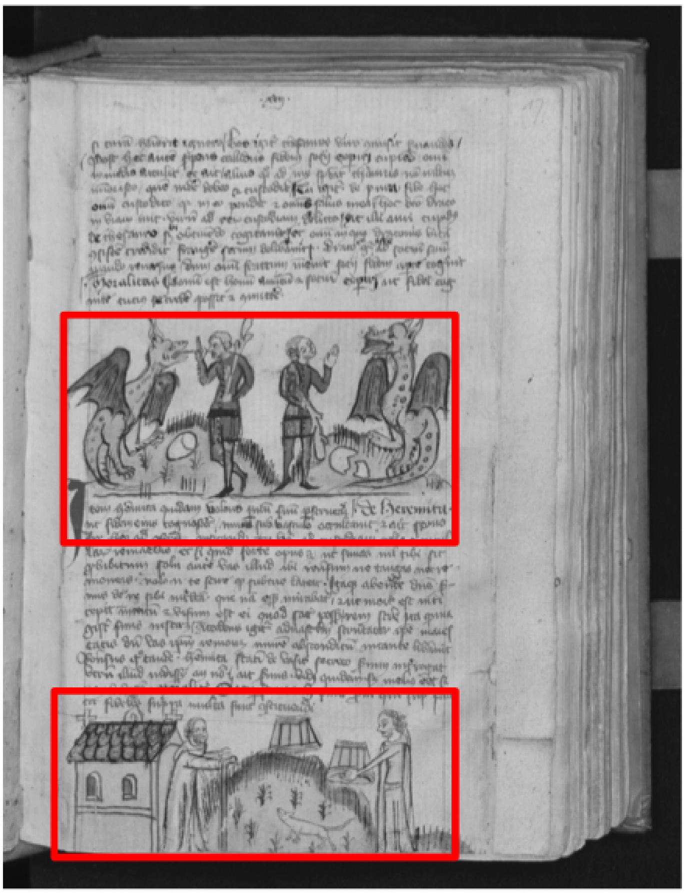


Figure 3.14 – Bounding rectangle representing the picture regions.

These instances were mapped to the same set of image training features as created for text regions. Then, a new classifier model was created with the random forest machine learning algorithm with 50 random trees. This classifier is called *picture classifier model*, which was then applied to historical document images to segment picture regions enclosed in the document images. Figure 3.15 shows the results of the picture region segmentation, which was done by pixel classification.

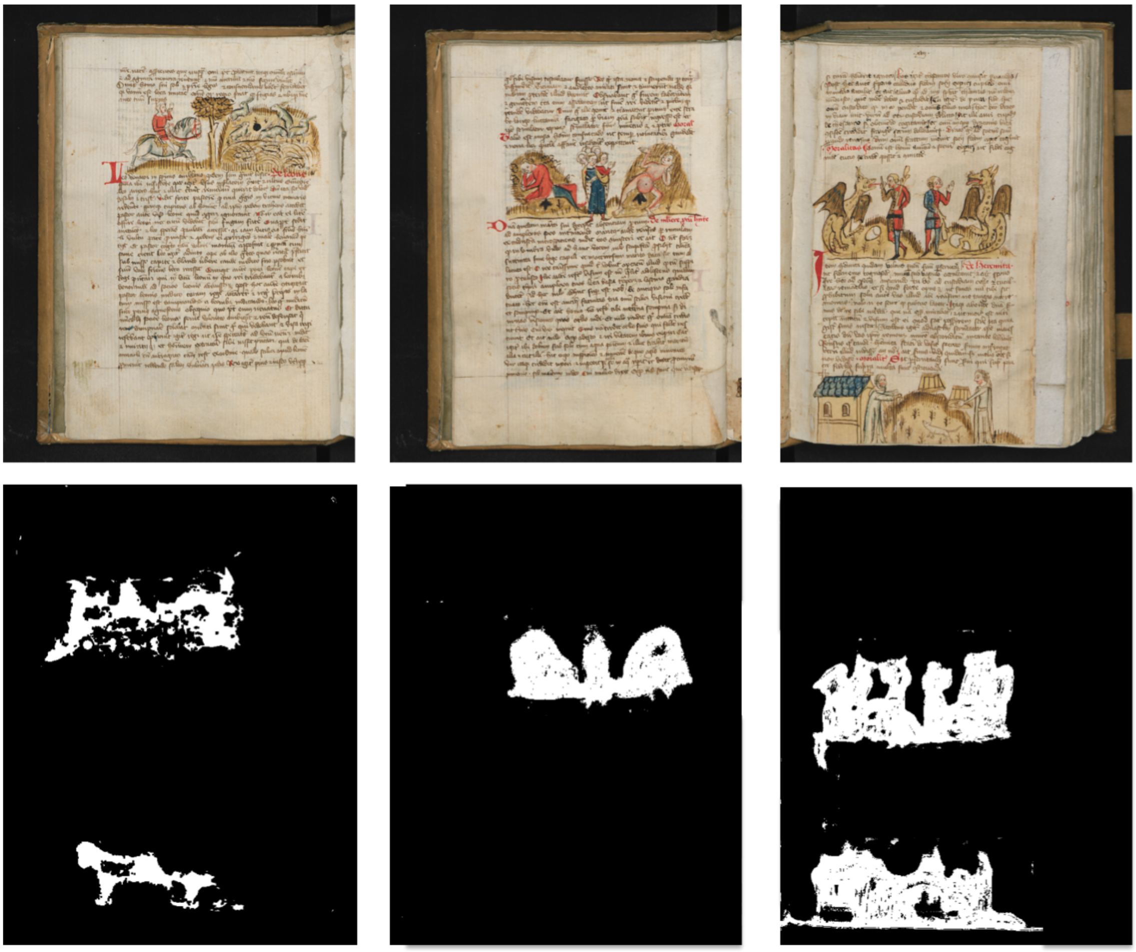


Figure 3.15 – Segmentation results for the picture regions performed by pixel classification.

3.3.4 Red-colored region segmentation

Handwritten historical document images often contain regions written with red-colored ink. It is important for the humanities researchers to identify these regions so that knowledge about the usage of the color red in different centuries could be known. RGB (red, green, and blue) is a widely used color model, but the HSB (hue, saturation, and brightness) color model is preferred in this case because hue is invariant under different lightning conditions; moreover, the segmentation is performed only on one dimension, namely hue [105].

Therefore, in order to identify regions with red-colored ink, the HSB color model was used. The detailed description of each of the component of the HSB color model is given below and shown in Figure 3.16.

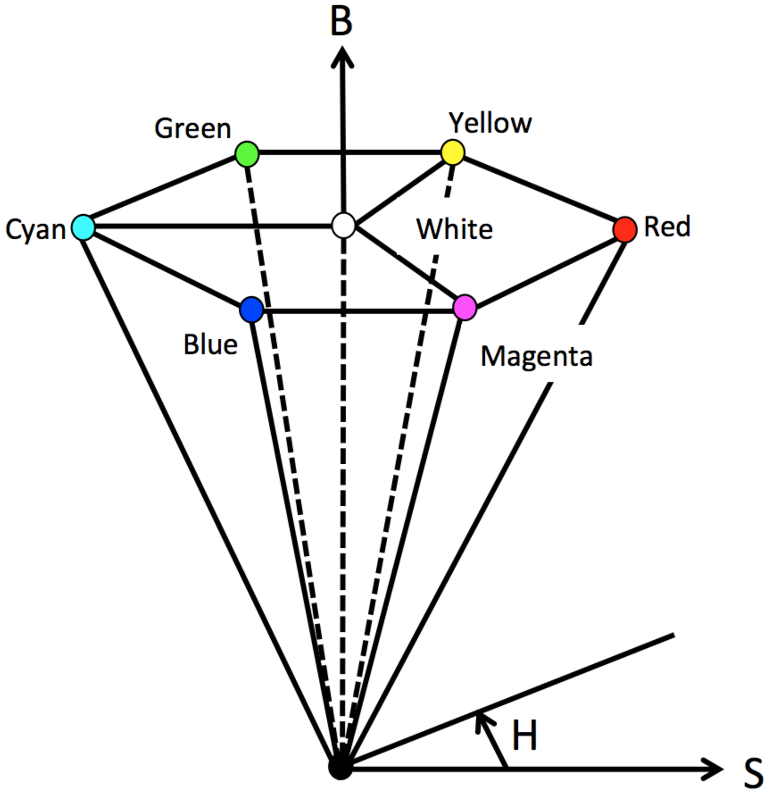


Figure 3.16 – HSB color model.

- **Hue**: Hue represents the actual color. It is usually measured in angular degrees counter-clockwise around a cone. It starts and ends at a red color having value of 0° or 360°.

- **Saturation**: Saturation defines the intensity of the color. It is usually measured in percentage from 0 % to 100 %.
- **Brightness**: Brightness is also measured in percentage from 0%, which defines no brightness, to 100%, which defines full brightness.

To identify the red-colored regions, at first, the HSB components of were computed. Then the range of thresholds defining the red color was chosen based on the hue and saturation component in the image using the image processing toolkit Fiji. The lower and upper thresholds for the hue component was selected as 10 and 241, respectively whereas, the lower threshold for the saturation component was chosen as 132. The regions in the document images that had the hue value in the range of 10 to 241 and a saturation value greater than 132 were segmented as the red-colored regions. Figure 3.17 shows results of the red-colored segmentation of the handwritten historical document image.

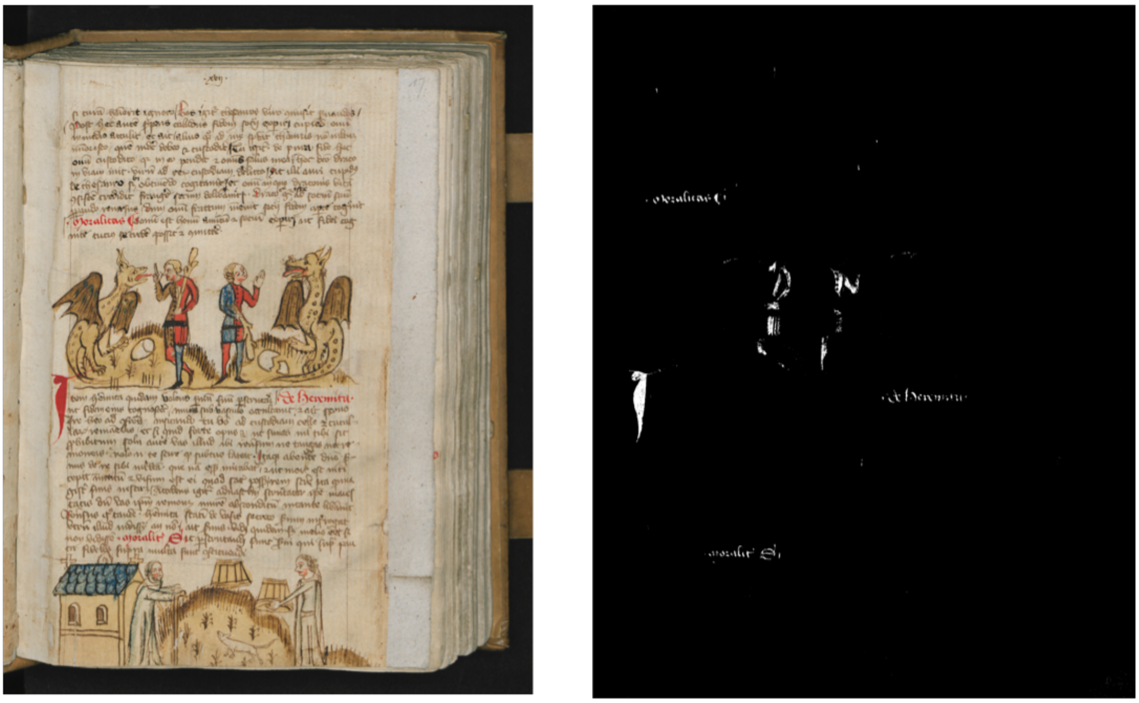


Figure 3.17 – Binary image where red-colored regions are represented by white pixels.

3.4 Feature extraction

In this feature extraction step, relevant quantitative values describing the physical regions identified by the region segmentation were extracted. These quantitative values included the size features of the physical regions, area measurements, color features, and image-specific features, such as the number of text regions and the number of picture regions enclosed in the historical document images. Other derived features, such as the text region to left margin region ratio and the text region to right margin region ratio were also extracted. The extraction of each of these features is described below.

- **Size features**: Size features were extracted by counting the number of pixels corresponding to a layout region. For example, in the binarized text segmented image, all the white pixels representing the text regions were counted to get the size features. These features included bounding box measurements (i.e., upper left corner coordinates, width, and height of the region, and area measurements).

- **Color features**: Color features were extracted so that knowledge about usage of different colors in historical times could be known. At first, hue, saturation, and brightness channels were extracted for each of the layout region (i.e., page regions, text regions, and picture regions). Afterwards, each region was partitioned into foreground and background using a thresholding method on the brightness channel. For example, the foreground for text region included only written spaces, foreground for picture regions included only pictorial spaces, and foreground for page regions included entire page region. Then, for each of the region, the foreground was set to dark and background was set to bright. At last, image statistics, such as mean, standard deviation, median, and skewness for the foreground and the background regions for the hue, saturation, and brightness channels were calculated.

- **Image-specific features**: Here, image-specific features, such as the number of text regions, picture regions, and red-colored regions enclosed in the document images were calculated.

- **Derived features**: Derived features included the ratio of the text region area to the left margin area and the ratio of the text region area to the right margin area. As a first step, the minimum bounding rectangle enclosing the text regions was calculated. Then, the left margin area and the right margin area were calculated based on the areas of the page region and the minimum bounding rectangle. Finally, the ratio of the text region area to the left margin area was calculated, and similarly, the ratio of the text region area to the right margin area was calculated.

3.5 Data storage

Each digital object representing a handwritten document image present in the research data repository (see Section 3.1.1) was accessed and processed. This process resulted in approximately 162,036,280 objectified and reproducible layout features (multidimensional information) for 150,000 handwritten historical document images, where each physical region had approximately 70 features. The detailed distribution of the number of page regions, text regions, picture regions, red-colored regions, and their layout features is shown in Figure 3.18 and Figure 3.19 respectively.

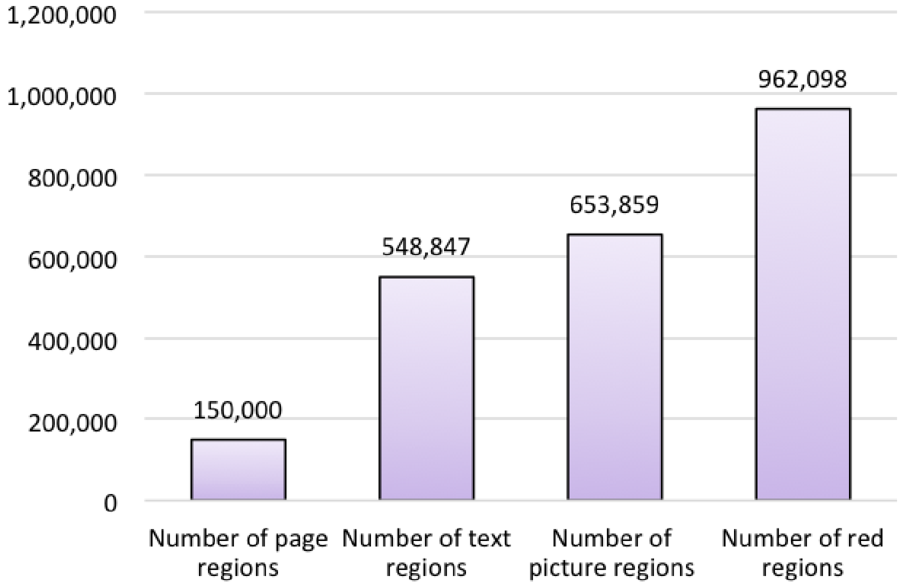


Figure 3.18 – Distribution of number page regions, text regions, picture regions, and red-colored regions identified.

These results were stored in an XML file according to the PAGE (Page Analysis and Ground-truth Elements) 2017 schema[9], which is an established format for representing individual stages of document image analysis from document image enhancements to layout analysis to OCR [106].

It records the information about image borders and the different physical regions enclosed in the document images. The PAGE XML file starts with the *metadata* attribute, which provides the details about the creator, the creation date, and the last modified date. After the *metadata* attribute, the PAGE XML file contains the *page* attribute. This attribute includes the details of identified physical regions, and their corresponding extracted

[9] http://www.primaresearch.org/schema/PAGE/gts/pagecontent/2017-07-15/pagecontent.xsd

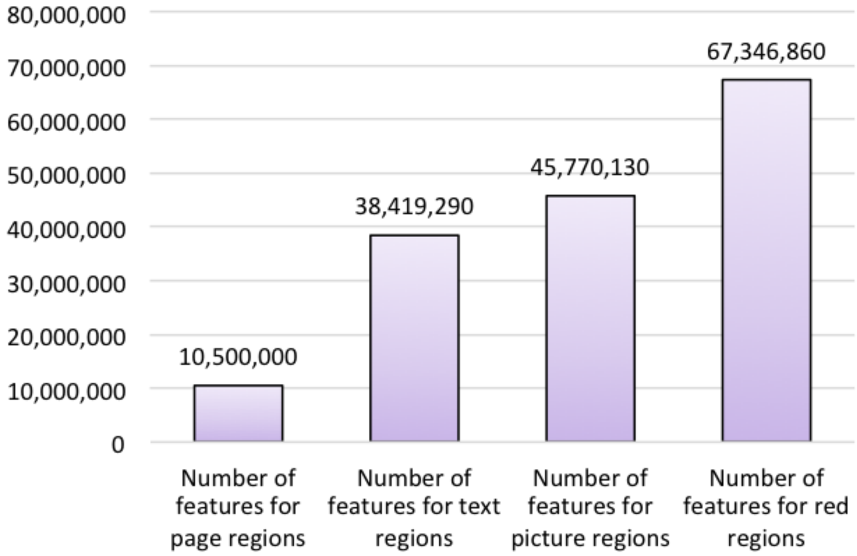


Figure 3.19 – Distribution of number of layout features for page regions, text regions, picture regions, and red-colored regions.

layout features. These extracted features, such as color features and size features, were recorded in the user-defined part of the XML file. All the physical regions were identified with an ID that is unique in the whole XML file. The ID was stored in an alphanumeric form, such as r1, r2,....r*n* where *n* represents the total number of the physical regions identified in the document images. The structure of PAGE XML is shown below:

Listing 3.1 – Structure of PAGE XML for representing layout analysis results

```
<?xml version="1.0" encoding="UTF–8" standalone="no"?>
<PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent
    /2017–07–15" xmlns:xsi="http://www.w3.org/2001/XMLSchema–
    instance" xsi:schemaLocation="http://schema.primaresearch.org/
    PAGE/gts/pagecontent/2017–07–15 http://schema.primaresearch.org/
    PAGE/gts/pagecontent/2017–07–15/pagecontent.xsd">
    <Metadata>
        <!–– Various attributes regarding the PAGE XML ––>
        <Creator>LayoutAnalysis–Version 1.3</Creator>
        <Created>2017–11–13T10:16:45Z</Created>
        <LastChange>2017–11–13T10:16:45Z</LastChange>
        <Comments>PageXml created according to 2017 schema</
            Comments>
    </Metadata>
    <Page custom="MeasurementUnit:Pixel" imageFilename="S Hs 47
        _00052_T_0.tif" imageHeight="4071" imageWidth="3002">
```

```
         <!— Page region coordinates —>
        <Border>
            <Coords points="3000,305 3000,3860 425,3860 425,305"/>
        </Border>
         <!— Features for page region —>
        <UserDefined>
            <UserAttribute description="MeasurementUnit:cm" name="
                Height" type="xsd:string" value="30.099"/>
            <UserAttribute description="MeasurementUnit:cm" name="
                Width" type="xsd:string" value="21.801666"/>
             ...
        </UserDefined>
        <TextRegion id="r1">
            <!— Text region coordinates —>
            <Coords points="2865,590 2865,3130 2000,3130 2000,590"/
                >
              <!— Features for text region —>
            <UserDefined>
                <UserAttribute name="NumberHorizontalLines" type="
                    xsd:string" value="39"/>
                 ...
            </UserDefined>
            <TextEquiv>
                <Unicode/>
            </TextEquiv>
        </TextRegion>
        <ImageRegion id="r2">
            <!— Image region coordinates —>
            <Coords points="1970,1330 1970,2120 270,2120 270,1330"/
                >
            <!— Features for image region —>
            <UserDefined>
                <UserAttribute description="MeasurementUnit:cm"
                    name="Width" type="xsd:string" value="14.3933325
                    "/>
                 ...
            </UserDefined>
        </ImageRegion>
    </Page>
</PcGts>
```

The quantitative results were also stored in a comma-separated format in a CSV file for a multiple-coordinated visual exploration approach (see Chapter 4). In the CSV file, the first row represented the names of the extracted layout features. From the second row onwards, the measurement values of the extracted layout features were shown.

3.6 Evaluation

After the development of the feature extraction workflow, the automated approach for the layout analysis was needed to be evaluated before applying it to other larger number of handwritten historical document images. Therefore, in order to evaluate the automated approach for the layout analysis, at first, a set of document images were arbitrarily selected, and then the ground truth of these images was identified. Afterwards, the ground truth was compared with the results obtained from the state-of-the-art tools and the automatic layout approach proposed in this thesis.

3.6.1 Dataset selection

All the experiments in this section were carried out using the handwritten historical documents digitized within the scope of the project "Virtual Scriptorium St. Matthias". For the first evaluation, different handwritten historical document images were arbitrarily chosen as an experimental dataset from the St. Matthias database. The physical characteristics of these images mainly included overlapping physical regions, irregular spacing between words, margins notes, and varying text column widths, and so forth. Some examples of the selected set of the document images are shown in Figure 3.20.

The handwritten historical document images were composed of:

- 30 document images with black borders.
- 14 document images containing text in single column where text regions overlap with text regions in margins or picture regions.
- 16 document images containing text in two columns where text regions overlap with text regions in margins or picture regions.

3.6.2 Ground truth evaluation

Measurements of different physical regions were manually acquired by some computer scientists and domain experts using a ruler and semi-automatically by drawing a rectangular pattern around a layout region using an image processing software (Fiji). However, the bounding box measurements usually vary depending on the observer. If the two experts mark the same layout region within an image, the results may vary. This is called inter-observer variability. The same holds, if one expert marks the layout region several times, which is called intra-observer variability in this case.

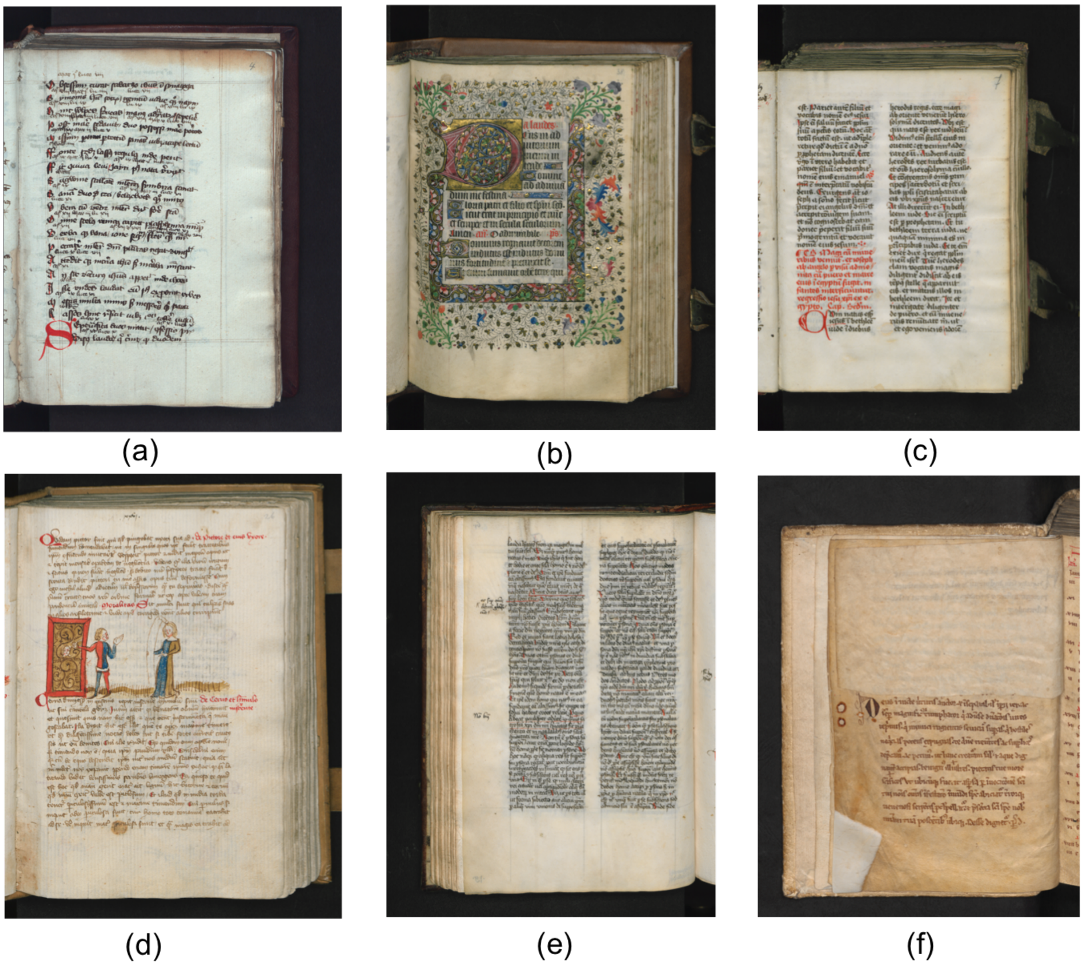


Figure 3.20 – Document images with (a) black borders (b) text in the single column (c) and (d) text with two or more columns (e) and (f) with overlapping picture regions and text regions.

For this experiment, computer scientists and domain experts manually and semi-automatically measured the size features, such as width and height of the page, as well as the text and the picture regions of the datasets described above. Afterwards, the mean values of these measurements were taken as ground truth. Then, various state-of-the-art tools, such as *Transkribus* and *Aletheia* were used to acquire the size features, such as width and height of the page and text regions. And, lastly, the ground measurements, size features from the state-of-the-art tools, and the automatic layout analysis approach presented in this thesis were compared.

3.6.2.1 Thirty handwritten historical documents digitized with black borders

In this case, the domain experts measured the page region height and the page region width of 30 handwritten historical documents digitized with black borders. Figure 3.21 and Figure 3.22 show the line chart displaying ground truth values of page region height and page region width, as well as, the values acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis. The values measured by the presented approach for the page region height lie within the range of the ground truth as compared to that of Transkribus and Aletheia (Figure 3.21).

The layout analysis approach presented by Transkribus was not able to detect the black borders. Therefore, the original document height was considered as the page region height. Aletheia was able to detect the black borders of the document, but it included the hardcover of the book which was also digitized with the document.

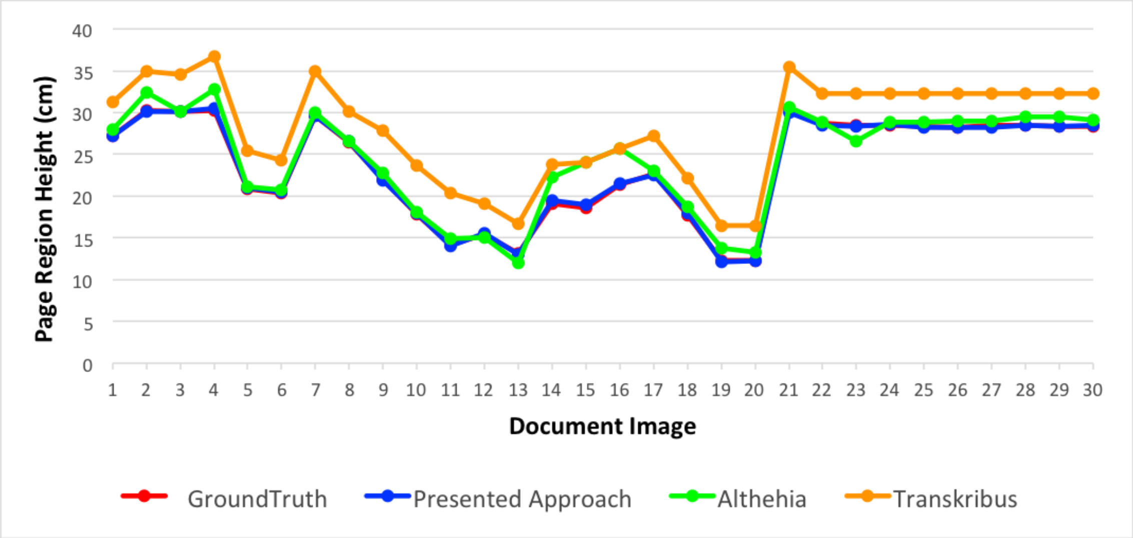


Figure 3.21 – Ground truth comparison of page region height values for thirty handwritten historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

Also, the values measured by the presented approach for the page width lie within the ground truth range in most of the cases (Figure 3.22). However, this is not true for all the cases because, for some of the document images, the page region segmentation algorithm found the global minimum somewhere inside the page region.

As a result, a part of the adjacent document page that was digitized together with the main document page was also included in the overall page width, which was measured automatically. However, still the results generated by presented approach are comparable to the results generated by Transkribus and Aletheia.

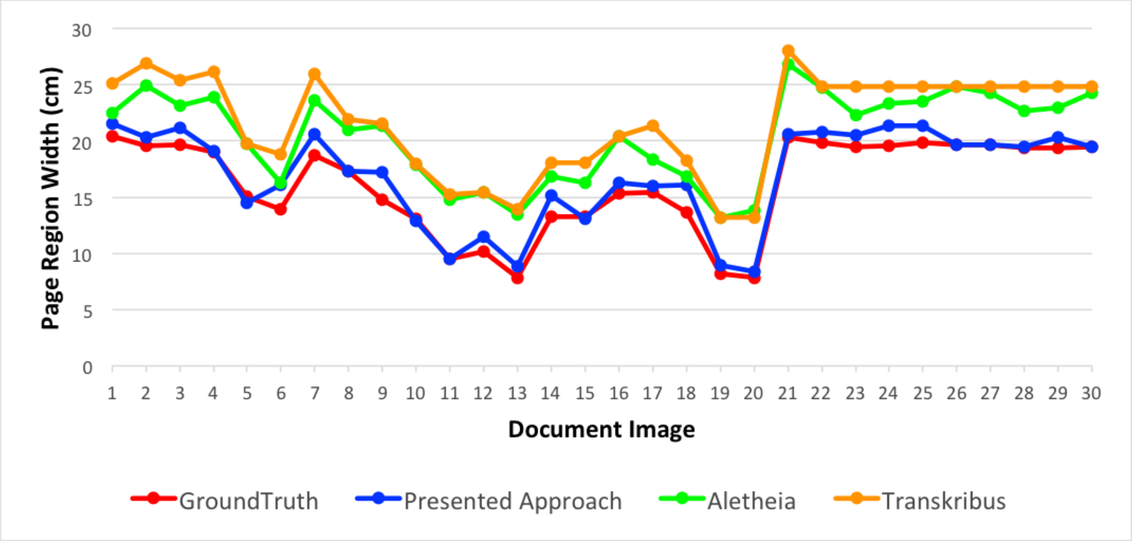


Figure 3.22 – Ground truth comparison of page region width values for thirty handwritten historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

3.6.2.2 Fourteen document images containing text regions in a single column

In this case, the domain experts measured the text height and the text width of fourteen handwritten historical documents containing text regions in a single column. Figure 3.23 shows the line chart displaying the ground truth values, as well as, the values acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis for text region height in a single column. Figure 3.24 shows the line chart displaying the ground truth values, as well as, the values acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis for text region width in a single column.

The values measured by the presented approach for the text height lie within the range of the ground truth for the majority of the cases (see Figure 3.23). However, in some cases, there exists a difference between the ground truth values and the presented approach values. The reason for this behavior is the presence of marginal text present on the top and bottom margin of the document images, which was closely connected to the main text and therefore was segmented with the main text.

There exists an offset for the Transkribus and Aletheia because these tools in some cases skipped detecting the text regions and in some cases detected the text regions outside of the page region. Figure 3.24 shows the text region width values; here it is seen that for one of the cases out of 10 document images the text width lies far outside the ground truth range. The reason for this behavior was the presence of highly illuminated regions in the document image which were also detected as text regions.

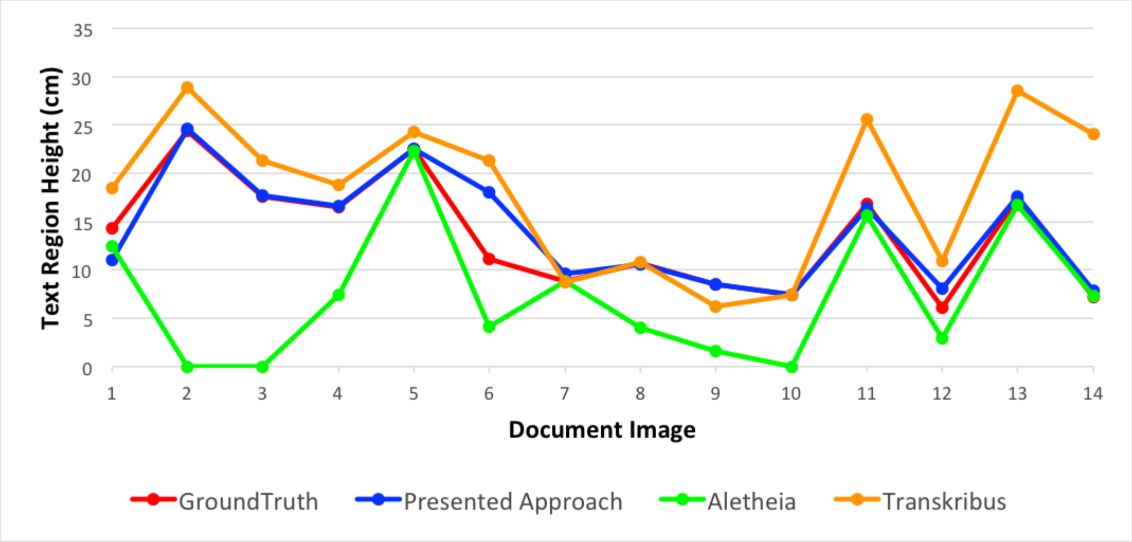


Figure 3.23 – Ground truth comparison of the text region height in a single column layout for fourteen handwritten historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

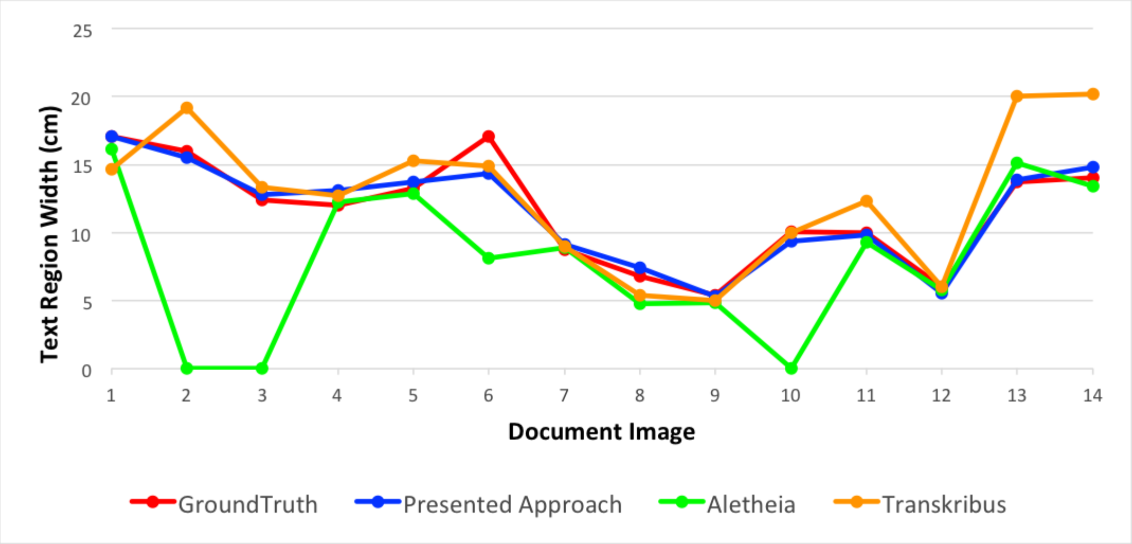


Figure 3.24 – Ground truth comparison of the text region width in a single column layout for fourteen handwritten historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

3.6.2.3 Sixteen document images containing text regions in two columns

In this case, the domain experts measured the text region height and the text region width of sixteen handwritten historical documents containing text regions in two columns. Figure 3.25 and Figure 3.26 show the line chart displaying ground truth values of text region height in two column layout, as well as, the values acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

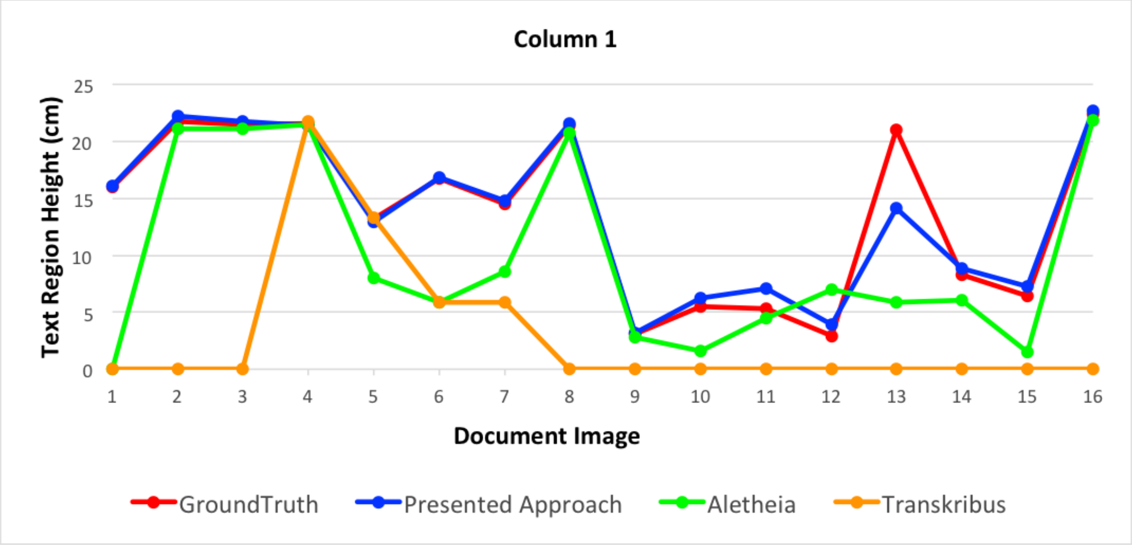


Figure 3.25 – Ground truth comparison of the text region height in two column layout for sixteen historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

In Figure 3.25 it is seen that the automatically measured values for the text region height for both the text columns present in the document image lie within the range of the ground truth for the majority of the cases. While in Figure 3.26 it can be observed that, for some of the cases, the automatically measured values for the text region height for both the columns have a difference with the ground truth values. These cases represent the text width where two columns were identified as a single column due to the presence of connecting text between these two columns and thus giving them an appearance of a single column. Aletheia was able to detect two columns in some of the cases as compared to Transkribus which couldn’t detect written spaces in two column layout. This behavior was also similar for text region width in two columns (Figure 3.27 and Figure 3.28).

3.6.3 Visual comparison with state-of-the-art tools

The identified regions of the selected dataset (see Section 3.6.1) were also compared visually to state-of-the-art tools, such as Transkribus and Aletheia. Figure 3.30 and Figure

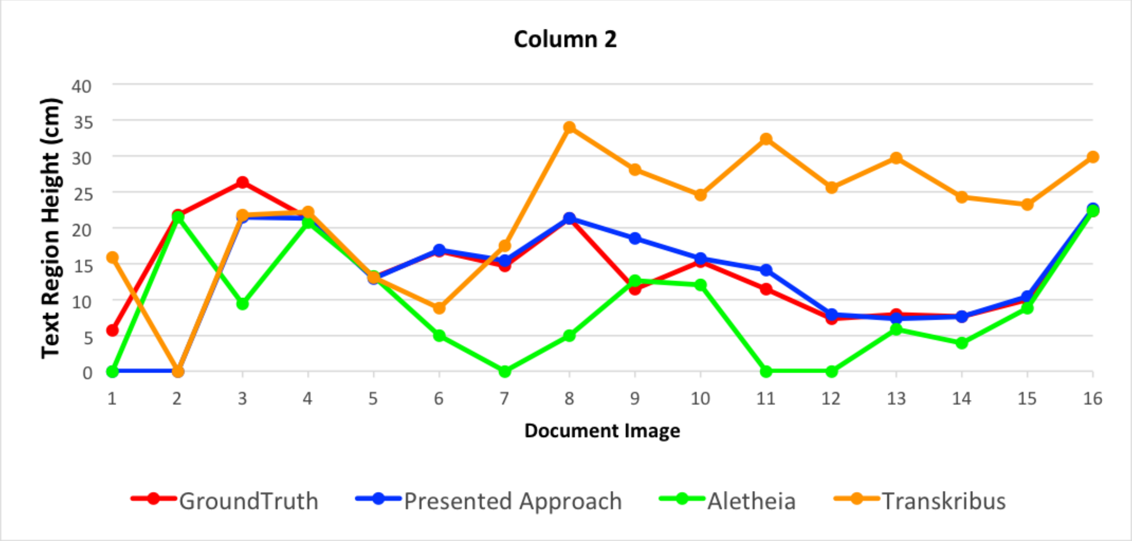


Figure 3.26 – Ground truth comparison of the text region height in two column layout for sixteen historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

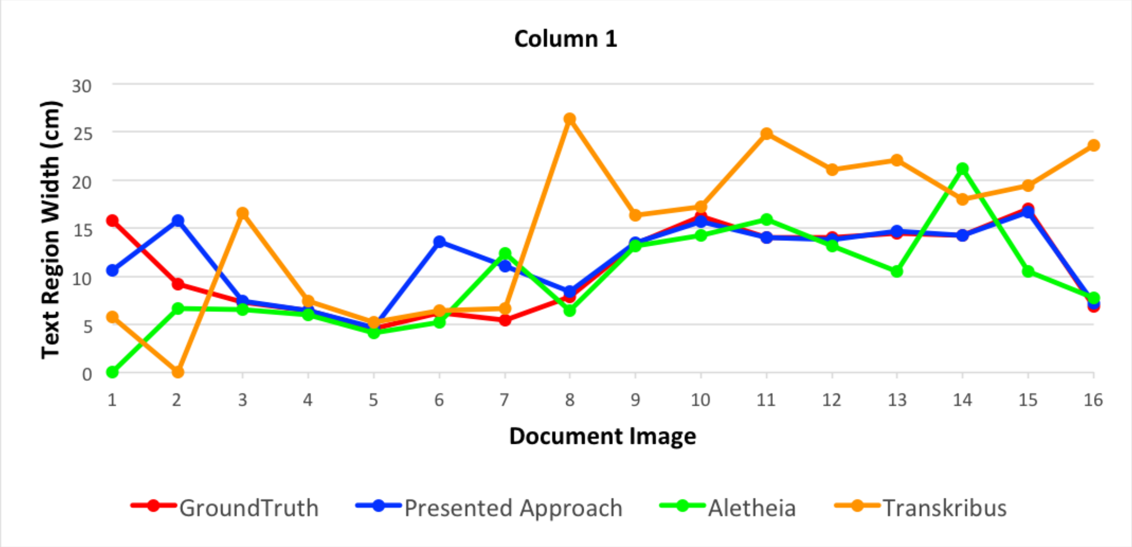


Figure 3.27 – Ground truth comparison of the text region width in two column layout for sixteen historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

3.29 show the visualization of the identified page regions and text regions. As Figure 3.30 and Figure 3.29 illustrate, Transkribus cannot identify the two different columns in the handwritten document. Moreover, the parts outside the page region are also identified as text regions as compared to the approach presented in this thesis, where even smaller text

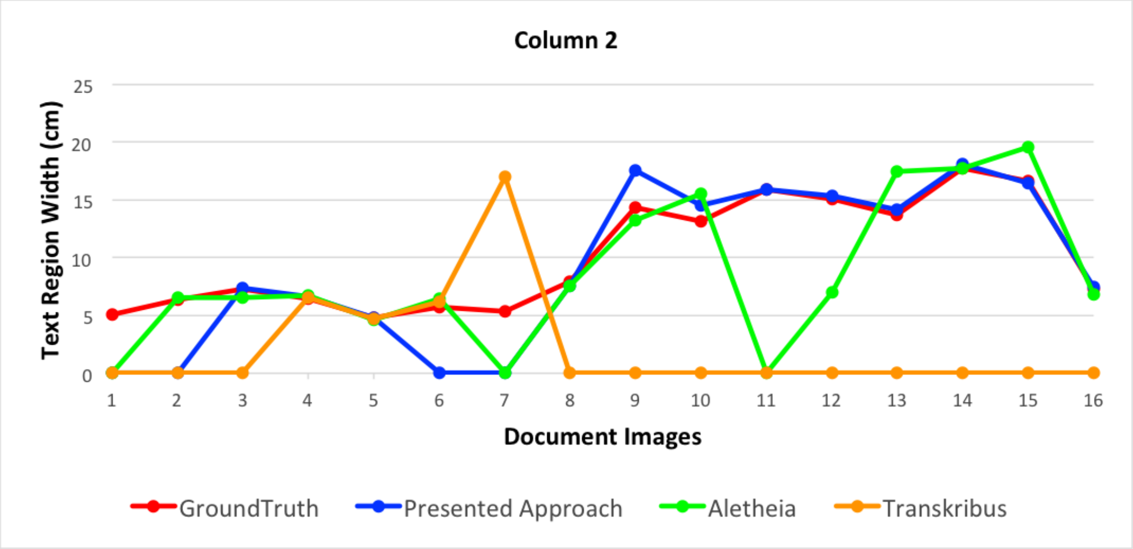


Figure 3.28 – Ground truth comparison of the text region width in two column layout for sixteen historical document images acquired from Transkribus, Aletheia, and the layout analysis approach presented in this thesis.

regions located on the margins of the text are identified. Similarly, it is seen that smaller text regions are completely missed by *Transkribus*. Furthermore, it is seen that nearly full document page is identified as the text region. This shows that the layout analysis approach presented in this thesis yield much better results than the state-of-the-art tool.

3.6.4 Adaptability of the automated approach for layout analysis to other datasets

One of the main goals of this research was to design and develop a generic layout analysis method so that it can be applied to a wide range of document images. Therefore, in order to evaluate the adaptability of an automated approach for layout analysis to other datasets, a set of heterogeneous document images were selected from other publicly available datasets. Some examples of these document images are shown in Figure 3.31. These datasets are composed of:

- 69 document images from a database of modern printed Spanish magazines[10].
  - Type: printed
  - Material: paper
  - Century: early modern period, i.e., 19th century

[10] http://www.revistas-culturales.de/

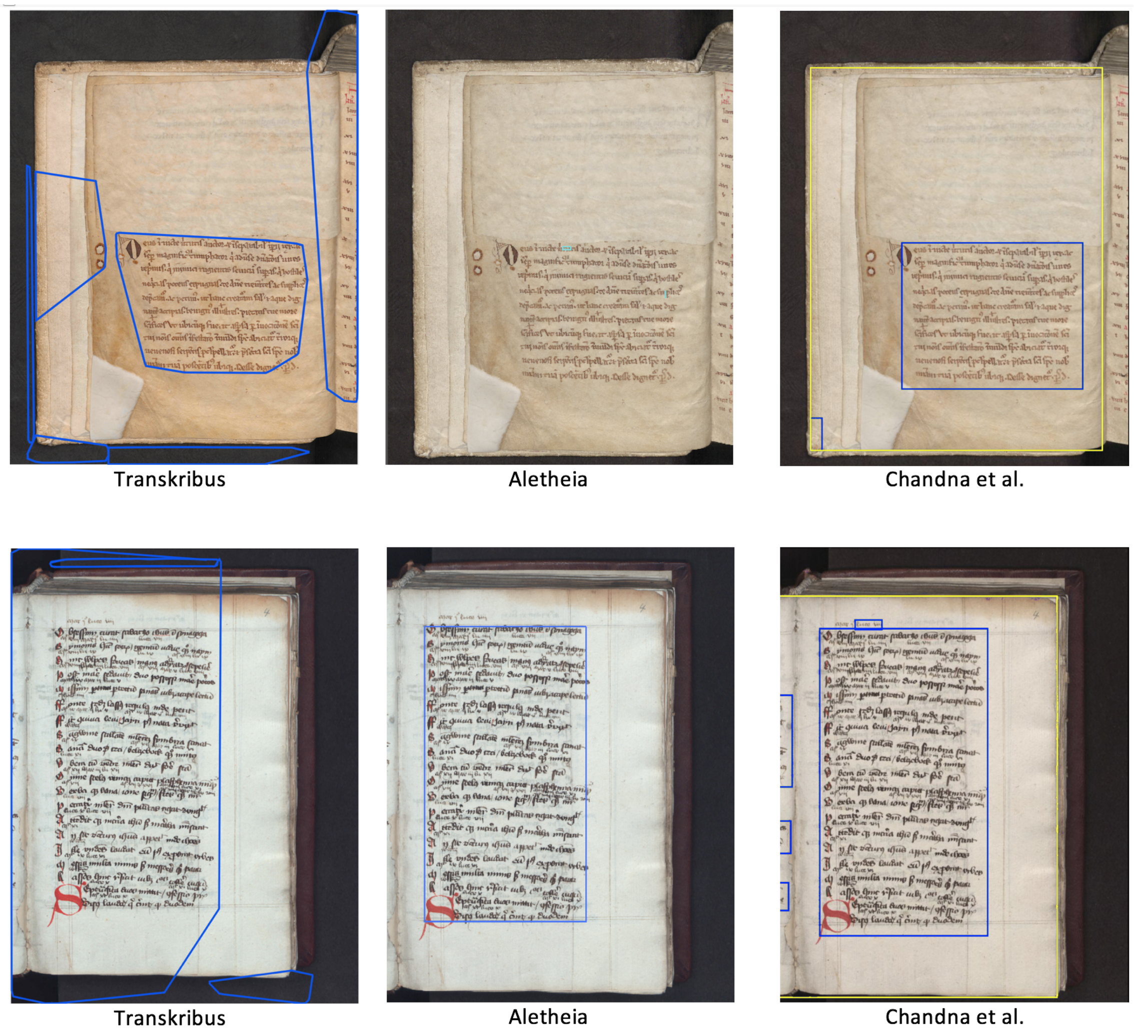
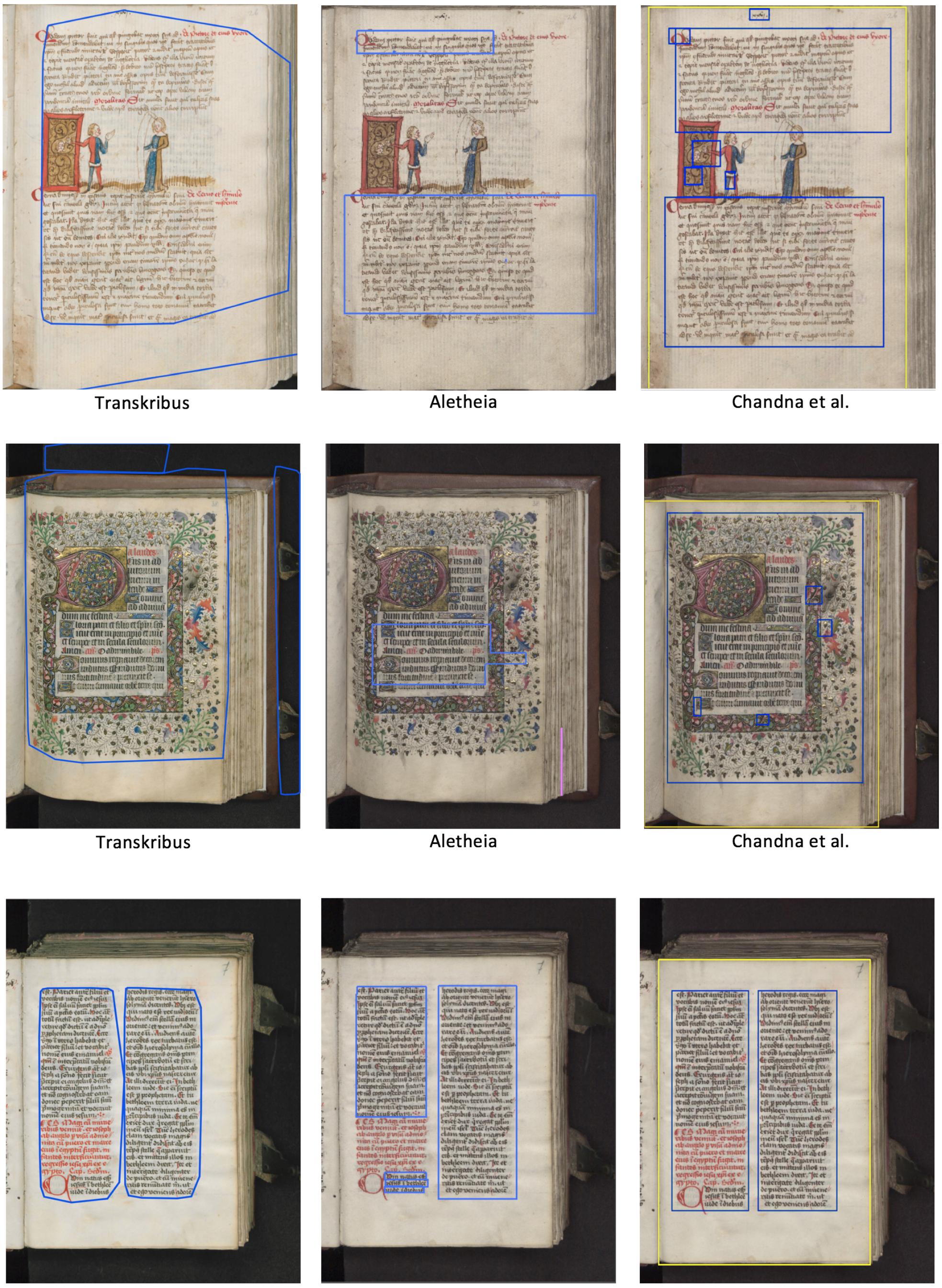


Figure 3.29 – Comparison of layout analysis approach presented by *Transkribus*, *Aletheia*, and the automatic layout approach for identification of text region presented in this thesis.

  - Spatial resolution: 300 DPI
  - Layout: multiple column layout

- 47 document images from the Parzival database[11].
  - Type: handwritten
  - Material: parchment

[11] http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/parzival-database



Transkribus

Aletheia

Chandna et al.

Transkribus

Aletheia

Chandna et al.

Transkribus

Aletheia

Chandna et al.

Figure 3.30 – Comparison of layout analysis approach presented by *Transkribus*, *Aletheia*, and the automatic layout approach for identification of text region presented in this thesis.

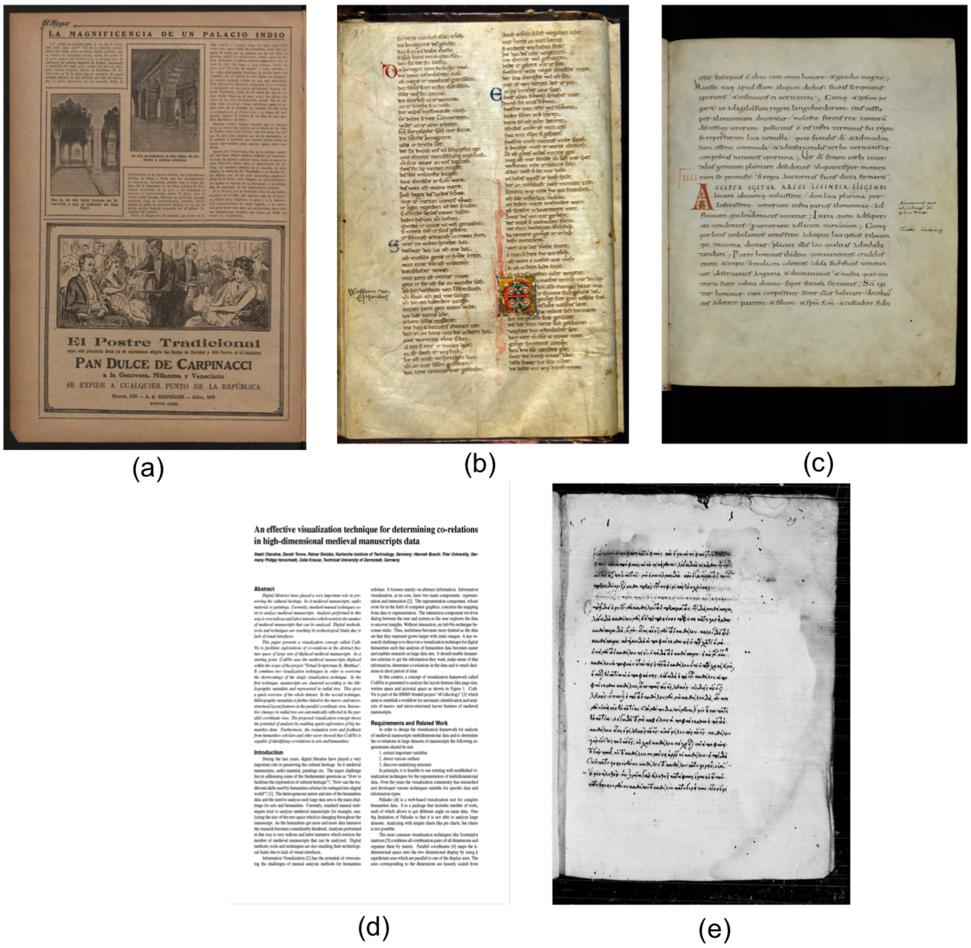


Figure 3.31 – Examples of other publicly available datasets: (a) printed Spanish magazines, (b) handwritten documents from the Parzival database, (c) handwritten documents from the Saint Gall database, (d) PDF, and (e) Aristoteles documents.

  - Century: 13th century
  - Spatial resolution: 300 DPI
  - Layout: two column layout, text in margins, overlapping regions

- 60 document images from the Saint Gall database[12].

  - Type: handwritten
  - Material: parchment

[12] http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/saint-gall-database

  - Century: 9th century
  - Spatial resolution: 300 DPI
  - Layout: single column layout, text in margins, overlapping regions

- 10 Portable Document Format (PDF) documents published as research papers during the course of this thesis [13].
  - Type: computer typed
  - Spatial resolution: 200 DPI
  - Layout: rectangular layout, one-two columns, text in headers and footers

- 80 digitized images of Aristoteles documents
  - Type: handwritten, microfilms
  - Material: paper, parchment
  - Century: 9th - 16th century
  - Spatial resolution: 300 - 400 DPI
  - Layout: single columns, overlapping regions

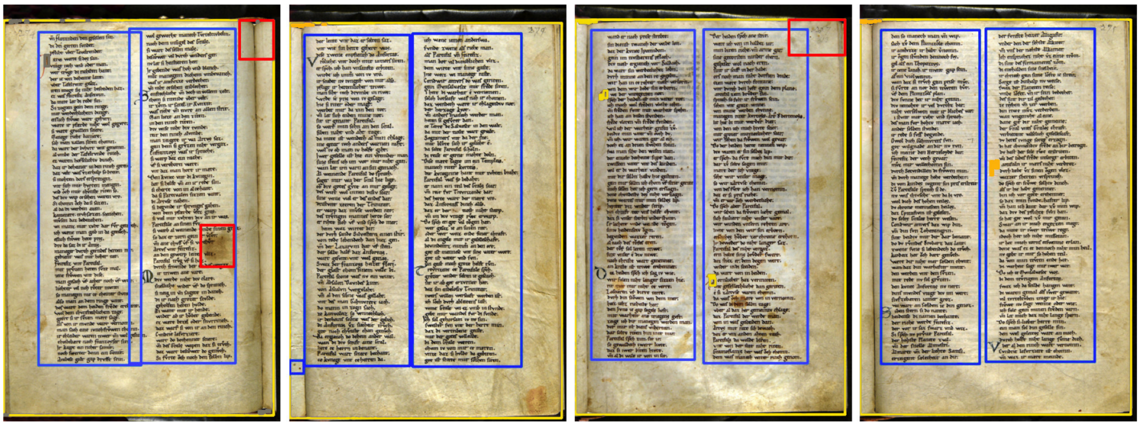


Figure 3.32 – Identified physical regions of a subset of the all document images from the Parzival database.

[13] //www.spiedigitallibrary.org/conference-proceedings-of-spie/9402/940206/Software-workflow-for-the-automatic-tagging-of-medieval-manuscript-images/10.1117/12.2076124.full?SSO=1,http://www.ingentaconnect.com/content/ist/ei/2016/00002016/00000001/art00013

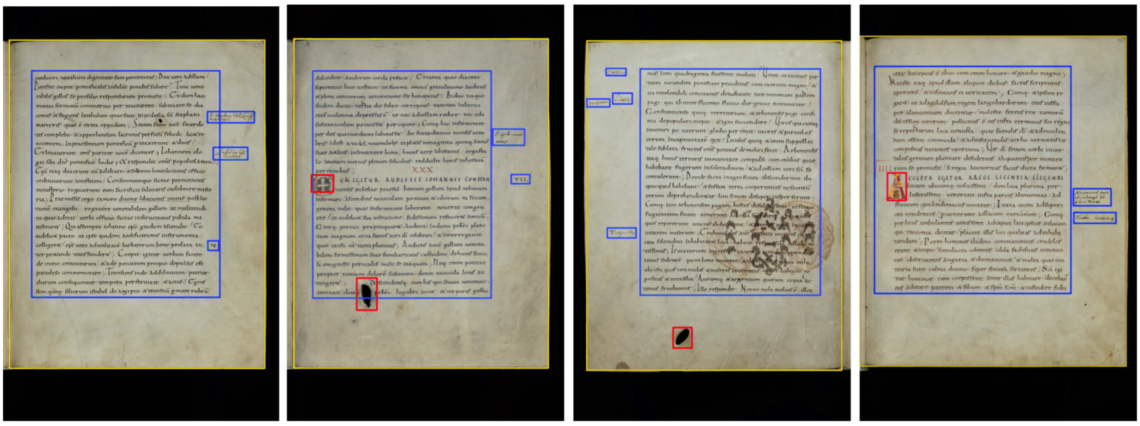


Figure 3.33 – Identified physical regions of a subset of the all document images from the Saint Gall database.

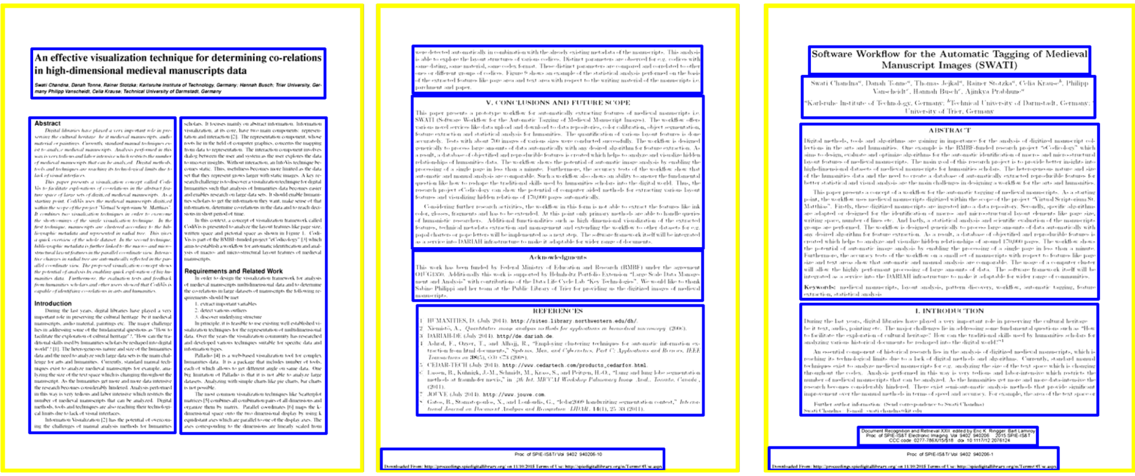


Figure 3.34 – Identified physical regions of a subset of the all document images from PDFs.

The classifier model which was trained for identification of text regions for the St. Matthias was directly applied on the Parzival, Saint Gall, and PDF database. Figure 3.32, Figure 3.33, and Figure 3.34 show results of the application of the layout analysis approach on the Parzival database, the Saint Gall, and PDF respectively. It is seen that the two columns, the overlapping text regions, and overlapping text regions enclosed in margins are clearly identified.

However, as Spanish magazines contained text in the picture regions and the Aristoteles documents were digitized microfilms, a new classifier was trained with new instances and same image training features, machine learning algorithm and number of trees used for St. Matthias database (see Section 3.3.2). Figure 3.35 and Figure 3.36 show the results

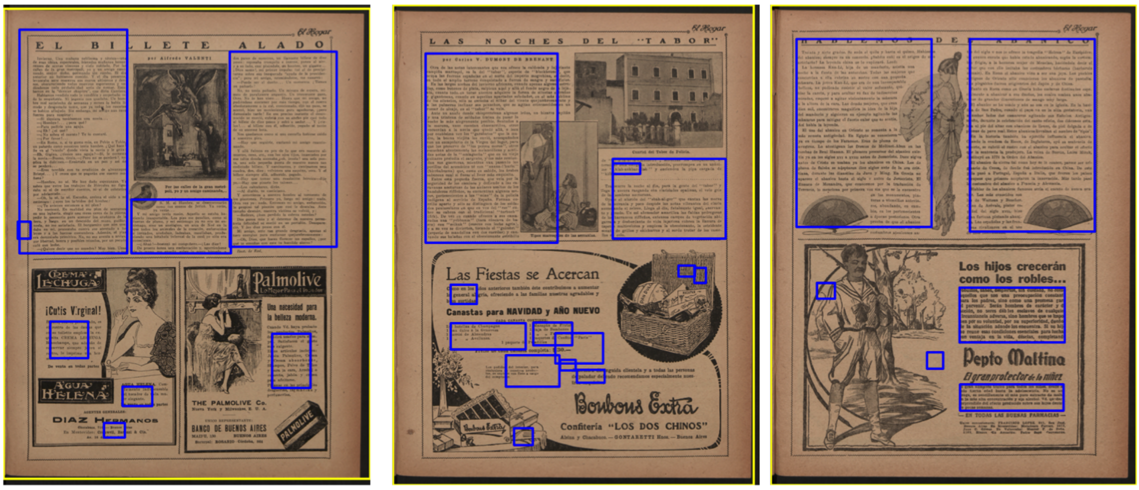


Figure 3.35 – Identified physical regions of a subset of the all document images from the printed Spanish magazines.

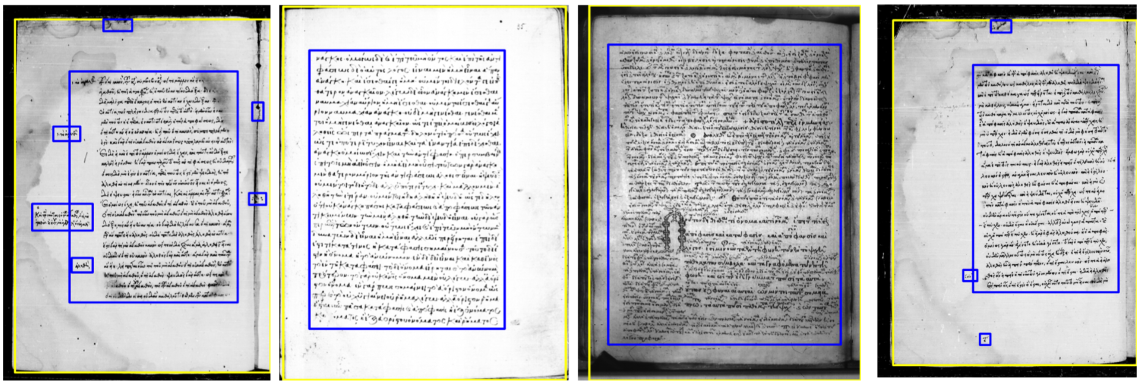


Figure 3.36 – Identified physical regions of a subset of the all document images for the Aristoteles documents.

of the application of layout analysis approach on Spanish magazines.

3.7 Summary

This chapter describes a generic, automated approach to identify the physical regions and extract their corresponding layout features. The automatic approach is organized according to the processing steps, which must be performed. After providing the details of the "Virtual Scriptorium" dataset, the chapter includes an explanation of the preprocessing of the dataset. Preprocessing involves the processing steps necessary to make the acquired document images more suitable for the layout analysis. Preprocessing utilizes methods like noise filtering, color calibration, spatial calibration, and scaling.

After preprocessing, the segmentation of different regions enclosed in the document image was performed. This starts with the segmentation of page regions followed by the segmentation of text and picture regions, which is done by pixel classification. This section further describes the extraction of features or attributes describing these physical regions, namely color features, size features, or count features. As a result, a multidimensional database of objectified and reproducible features was created and stored. This approach was tested on the "St. Matthias" dataset and was also applied to other datasets to prove its generality. The main aim of this part of the research was to deal with the large variety of document images and to analyze their layout rather than providing a high accuracy rate for document layout analysis. The automatic approach of identifying physical regions and extracting their corresponding layout features without any user intervention was expected to be more significant for the scholars studying the layouts of historical document images as compared to the manual approach.

Chapter 4

Visualization interface design strategy

The application of an automated approach for layout analysis on the St. Matthias data resulted in a database of multidimensional (more than three dimensions) extracted layout features. Exploration of these multidimensional layout features in order to define relationships between many documents gets challenging where it is not known beforehand what needs to be searched or determined. Here, a visual exploration approach may help to explore this data interactively with the help of information visualization techniques. But there exist a plethora of information visualization techniques to choose from, and thus it is also challenging to decide which information visualization technique is better suited for such an application.

Therefore, this chapter describes a generic design strategy to design and develop a visualization interface for a particular kind of application. This generic design strategy is then applied on St. Matthias dataset to design a visualization interface for the multidimensional layout features.

4.1 Methodology

A generic design strategy for designing a visualization interface consists of seven steps (see Figure 4.1). Each of these steps is described as follows.

Identify problem

The first step involves interacting with the domain experts and identifying their research questions, which they would like to answer with the help of the visualization interface. Moreover, it would be advantageous to understand that which methods do they use to solve their research questions and which are the limitations of those methods.

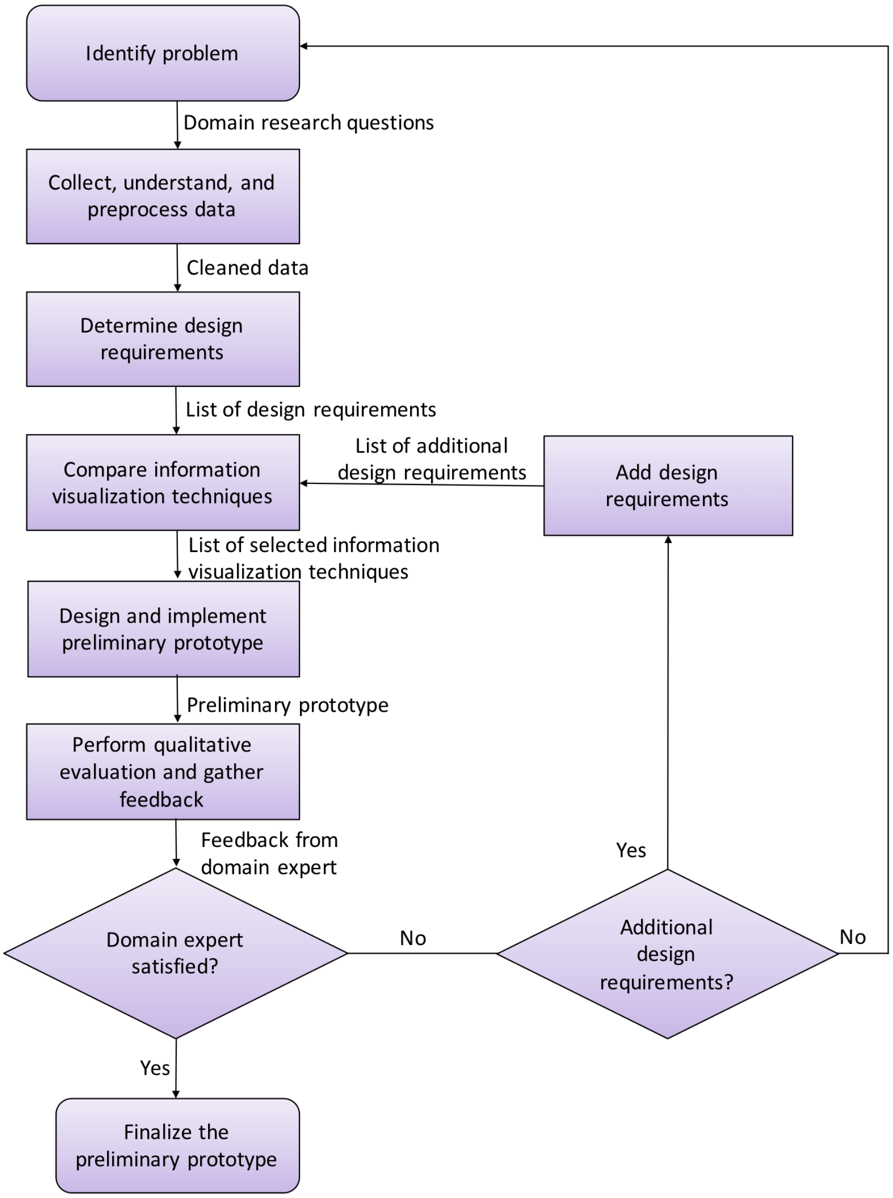

Identify problem
Domain research questions
Collect, understand, and preprocess data
Cleaned data
Determine design requirements
List of design requirements
Compare information visualization techniques
List of additional design requirements
Add design requirements
List of selected information visualization techniques
Design and implement preliminary prototype
Preliminary prototype
Perform qualitative evaluation and gather feedback
Feedback from domain expert
Yes
Domain expert satisfied?
No
Additional design requirements?
No
Yes
Finalize the preliminary prototype

Figure 4.1 – Design strategy to design and develop a visualization interface.

By identifying their research questions and knowing their current methods and their limitations, a better understanding of their domain can be obtained which will help in refining the design requirements at a later stage.

Understand and preprocess data

The main objective of this step is to understand and preprocess the data. In this step, mainly the following aspects are considered:

- **Understanding of the origin of the data:** It involves an understanding of how the data is generated. For example, is it created as a result of digitization of analog material, manually by domain experts, or as a result of execution of a workflow or an algorithm?
- **Measurement units associated with the data:** It involves an understanding of the units which are associated to describe the data. For example, whether the data related to temperature is in Celsius or Fahrenheit?
- **Format and type of the data:** It involves an understanding of the format in which data is stored and type in which data is present (image, data table). For example, is it text documentation stored in XML, PDF, or Plain text, or is it an image stored in TIFF or JPEG?
- **Relationships which exist between data and its corresponding information:** It involves an understanding of how changes in one type of the data can influence the other data or whether there are any inter-dependencies within the various sets of data.

Preprocessing data involves an important step of visualization pipeline, i.e., filtering [11]. This involves identifying the irrelevant, missing or incomplete data which needs to be removed or refined by completely discarding the values or by placing a sentinel value. It also requires removal of the inconsistencies and redundancies present in the data and making it consistent to make it compatible with the visualization interface. As a result of this step, a cleaned data is generated.

Determine the design requirements

This involves understanding the needs of the domain experts in detail as described in the visualization reference model given by Robertson and DeFerrari [107]. For instance,

- the kind of interface required by the user;
- the level of data details required by the user, i.e., whether a detailed or an abstract view of the data is needed;
- the kind of analysis task needed to be performed by the visualization interface, such as finding correlations, detecting outliers or determining the range of values.

As a result of this step, a list of design requirements is generated. For instance, a static visualization with no interactions and just some animations where the visualization should be understood without requiring the set of additional actions, or a dynamic visualization where multiple visualization techniques interact with each other in a dashboard. Another design requirement could be where the domain experts prefer just to view the overview of the complete dataset.

Compare information visualization techniques and tools

Chapter 2 describes various taxonomies of information visualization techniques which exist in the field of information visualization, i.e., taxonomy by Shneiderman [56], Keim [57], and Liu [58]. Moreover, there exist various tools which provide the possibility to visualize the data. For example, the ManyEyes [108] tool visualizes data using the bar chart, bubble chart, scatter plot, treemaps, etc. Another example of such a tool is Tableau [109], which also consists of basic visualization techniques such bar chart or line chart. The list of possible information visualization techniques, tools, and systems is described below:

1. **Standard 1D-3D techniques**: These techniques visualizes 1-dimensional, 2-dimensional, and 3-dimensional data. It includes information visualization techniques, such as bar chart, line chart, and scatter plot [110].

   (a) **Bar chart:** It is mainly used to visualize discrete, numerical data across different categories. One of the axes of the bar chart shows the discrete data value and other axis shows the specific categories. They are mainly used for direct comparison of the data.

   (b) **Line chart:** It is used to visualize the quantitative data values over a period of time. It is mainly used to show how the data has evolved over the period of time. They are most commonly used to discover trends in the data and help make decisions. The line charts are drawn by first plotting the data points on the Cartesian coordinate and then by connecting a line segment between all of these points. Normally, the y-axis represents the quantitative value and x-axis represents the time scale. The direction of line segments indicates whether the values have increased or decreased over the period of time.

   (c) **Scatter plot:** It uses X-axis and Y-axis in the Cartesian coordinate to visualize the series of data points. These data points represent its X and Y values which show a relationship or a correlation between each other. Various type of correlation which can be spotted with the scatter plot are a positive correlation (both the values increases together), negative correlation (one values increases and the other value decreases), or none (no correlation at all). The data points which are far outside the cluster are called outliers.

2. **Iconographic techniques**: These techniques map the multidimensional data to icons. Various iconographic visualization techniques which could be mapped to multidimensional layout features of the historical documents are stick figures, and Chernoff faces.

   (a) **Stick figures:** This technique maps multidimensional data to five interconnected line segments. One segment out of these five segment is the body, and other four are called as limbs. The two dimensions are mapped on the two axes, and the remaining dimensions are mapped to the angle or lengths of the limbs [111].

   (b) **Chernoff faces:** This technique is used to visualize the multidimensional data (maximum 18 dimensions) using the properties of the human face. The individual property of the human face, such as eyes, ears, or nose presents the value of each dimension using shape, size, or orientation [112].

3. **Geometric projection techniques**: This type of visualizing techniques projects multidimensional data in a two-dimensional plane.

   (a) **Parallel coordinate plot:** This visualization technique is used to visualize high-dimensional data where individual data elements are plotted across multiple attributes. Each attribute is given a vertical axis placed parallel to each other. Each of the vertical axes is given its scale as each variable deals with a different unit of measurement. All the data values are plotted as a series of line segments connected across each of the attribute [113].

   (b) **Scatterplot matrix:** The scatterplot matrix visualizes a set of pairwise scatter plots in a matrix form. It is mainly used to analyze the linear correlation between variables [114].

   (c) **Star coordinates:** This visualization technique arranges coordinate axis on the circle in a two-dimensional domain. Each of the coordinates axes is equidistant from each other with equal angles and originating at a center [115].

4. **Pixel-oriented techniques**: These type of visualization technique maps each value of the dimension on to a single colored pixel. It allows visualization of the largest amount of data. Various visualization techniques which could be applied to historical documents are as follows.

   (a) **Pixel bar chart:** It is derived from a standard bar chart where each data value is mapped on to single colored pixel and plotted on the bar chart. One needs to specify dividing, ordering and coloring dimensions to design a pixel bar chart [116].

   (b) **Circle segment:** This technique visualizes the whole dataset in a circle. The circle is divided into segments where one segment corresponds to one dimension of the data. Within these segments, each value of the dimension is mapped on a single colored pixel [117].

5. **Hierarchies and graphs**: These type of visualization technique is mainly used to visualize hierarchical or tree-structured data.

   (a) **Treemap:** This visualization technique visualizes a large amount of hierarchical data using rectangles which are nested inside other rectangles. Each category of the data table is assigned a rectangle, and each subcategory is assigned another rectangle. Here, the division and ordering of rectangles are completely dependent on the tiling algorithm, such as "squarified algorithm". This algorithm tries to keep each rectangle as square as possible [118].

   (b) **SunBurst:** This type of visualization technique visualizes hierarchical data through a series of concentric circles, that are sliced for each category of the node. Each concentric circle refers to the level of the hierarchy. The circle at the center corresponds to the root node of the tree and hierarchies move outward from the center. Circles are sliced and divided based on the hierarchical relationship to the parent [110].

   (c) **Radial tree:** It maps the hierarchical or tree structure of the data in the polar coordinates [119].

   (d) **Dimensional stacking:** It visualizes multidimensional data in a two-dimensional plane by embedding one dimension into another. It starts with the discretization of the ranges of each dimension. After that, each dimension is assigned an orientation and order. This procedure is repeated until all the dimensions are embedded.

Other visualization techniques from the Liu's taxonomy includes:

1. **Table lens:** This information visualization technique is mainly used to explore the high-dimensional tabular data. All the data values are represented in the form of columns and rows as represented in a spreadsheet. Each cell of the data table is filled with small colored horizontal bars, and each column represents a specific indicator. The main advantage of table lens visualization technique is that

it can focus on the interesting part and remove the unwanted data values using focus+context interaction technique [64].

2. **Heatmap:** It is used to visualize multivariate data by placing the variables in the row and column of the data matrix. Each cell of the data matrix is then assigned a color to reveal patterns in the data or to identify similar variables. They are good for visualizing patters, and for showing which variables are similar to each other. All the rows of the heatmap are one category, and all the columns are another category. Each of the individual row and column is further divided into subcategories to match each other in the matrix. The data within each cell of the heatmap is based on the relationship of the variables connected by row and column [110].

3. **Andrew Curve:** These are useful for visualizing multivariate data which cannot be easily separated in the tabular representation. The curves which are similar overlap with each other while other dissimilar observations are plotted in a different group [120].

4. **TileBars:** This visualization technique provides an iconic representation of the content of the document concerning various query terms. It allows users to make decisions based on the distribution of query terms in the documents. The main purpose is to indicate relative length of the document, frequency, and distribution of terms sets in the document [121].

Various visualization systems listed by Liu [58] to help visualization researchers create visualization interfaces are as follows.

1. Improvise [122]: This visualization system allows users to create multiple coordinated views of the relational data interactively.

2. the InfoVis Toolkit [123]: It is a Java-based library with generic data structures and visualization algorithms to help the visualization designers.

3. Prefuse [124]: It is most widely visualization system providing a library of various layout algorithms and interaction and animation techniques.

4. Protovis [125] : This visualization system employs JavaScript and Scalable Vector Graphics (SVG). It overcomes the problem of the traditional systems by providing the possibility to create interactive web-based visualizations.

5. Data-Driven Documents (D3) [126]: Protovis was further extended to build D3 also to create interactive visualizations on the web. It supports direct manipulation of the web elements by binding data to these web elements.

Various existing visualization tools are listed below:

1. ManyEyes [108]: It provides the possibility of using various information visualization techniques, such as bar chart, scatterplot, line graph, treemaps etc.
2. Xmdvtool [127]: It includes parallel coordinate plot, scatterplot, and dimensional stacking.
3. Mondrian [128]: It includes bar charts, parallel coordinate plot, box plots.
4. Tableau [109]: The trial version of this tool contains only few basic visualizations, such as line charts, bar charts.

It is essential to decide which of the aforementioned visualization technique or tool is well-suited for a particular kind of application. The improper use of any of the visualization techniques can lead to false conclusions or analysis [67]. Therefore, a list of possible state-of-the-art information visualization tools and techniques is prepared. Then a theoretical assessment can guide the visualization designers to search for the most suitable visualization technique or tool for a particular kind of application.

This comparison is motivated by the previous research of Dias et al. [129] where the authors did a study to evaluate various information visualization techniques and made a list of visualization techniques which may be considered as best for an individual application. However, this research study mainly focused on comparing the visualization techniques only according to Keim's taxonomy and also missed to compare various visualization tools to visualize the data. The study even did not consider to identify the problems and the design requirements of domain experts.

Comparison based on type of visualization and analysis tasks Each of the visualization technique needs to be compared based on the visualization and analysis tasks such as:

- **overview** provides a general view of the complete dataset at once;
- **clustering** groups data values which are similar to each other;
- **correlation** shows the relationship between two variables;
- **outliers** are the data values which lie outside the range of other data values in the dataset;
- **trends** is a regular pattern which is observed in the dataset;
- **comparison** is an estimate of determining similarities and dissimilarities between the data values;

- **deviation** shows the amount by which any particular measurement of a dataset deviates from a fixed value, such as the mean of the dataset.

Comparison of the visualization techniques based on tasks, such as overview, correlation, clustering, and outliers are followed from the previous research of Dias et al. [129] whereas comparison based on deviation, trends, comparison are added as a part of this thesis.

Comparison based on type of data Various visualization techniques differ from each other with respect to the type of data they can visualize, i.e., qualitative data and quantitative data. This comparison is also followed from the previous research of Dias et al. [129]

- **Quantitative data**: This type of data is mainly associated with quantities. It includes numbers or information which can be measured objectively, such as height or width. It can be further categorized into continuous and discrete.
  - discrete data - it takes only the integer values, such as 2, 4, 8;
  - continuous data - it represents the real values, such as 2.5, -1.2;
- **Qualitative data**: This type of data cannot be measured. It includes subjective descriptors about an entity, such as color, taste, texture, etc.
  - binary data - it works on two possible states, i.e., true/false, right/wrong, accept/reject;
  - nominal data - it represents data items which do not have any implicit order;
  - ordinal data - it represents data items which do have an implicit rank or order.

Comparison based on the number of records and dimensions Visualization techniques also differ from each other with respect to the number of records or dimensions that they can visualize. Some visualization techniques can represent a large number of records and dimensions as compared to others. Thus, such a comparison is important while choosing the information visualizations techniques depending upon the magnitude of the data.

Comparison based on visual composition methods Javed et al. [130] stated that by combining different visualization techniques, strengths and weaknesses of individual visualization techniques could be balanced. They identified the design space of composite visualization views (CVVs) that allows combining two or more visualization techniques. In their paper, they contributed four CVV design patterns for creating a composite visualization which are shown in Figure 4.2 and are described below:

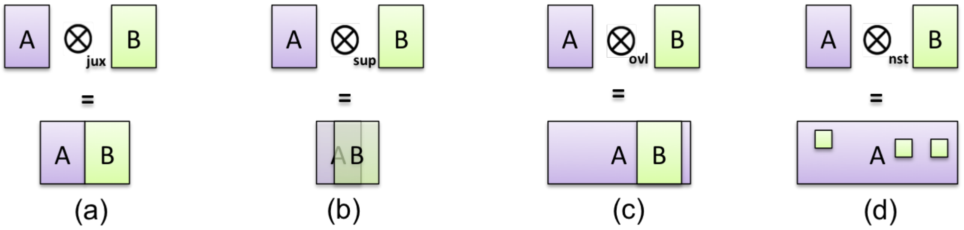


Figure 4.2 – Four CVV design patterns for creating a composite visualization (a) juxtaposition, (b) superimposition, (c) overloading, (d) nesting [130], © 2012 IEEE.

1. **Juxtaposition**: This design pattern is created by juxtaposing two or more visualization techniques side by side as shown in Figure 4.2a. In literature, there exists a large number of visualization frameworks which use juxtaposition to create a composite visualization [131], [132]. For example, ComVis [133], is a multidimensional visualization framework for complex meteorological datasets. It consists of eight different visualizations juxtaposed to each other and connected to a data table using interaction technique called as brushing and linking.

2. **Superimposition**: In this design pattern, two or more visual spaces are overlaid on top of each other as shown in Figure 4.2b. Most often, transparency is used to see the spatial linking present in the views. Javed stated that [130] "*spatial linking in the superimposed views allows for easy comparison across different datasets because the user does not have to split their attention between different parts of the visual space*". For example, GeoSpace [134] is a visualization where crime data is superimposed on the geographical map of the Cambridge.

3. **Overloading**: Overloaded views are created when one visualization technique is rendered inside another visualization technique as shown in Figure 4.2c. As in superimposition, this design pattern does not share one-to-one spatial linking. For example, Scatter Plots in Parallel Coordinates (SPPC) [135] is a visualization system which allows to overload two-dimensional scatter plot on the parallel coordinate plot. The space between the axis of the parallel coordinate plot is overloaded by a scatterplot.

4. **Nesting**: Nested views are created like overloaded views, but in this case, one or more visualizations are rendered inside another visualization as shown in Figure 4.2d. One famous example of a nested view is NodeTrix [136]. This visualization tool was created to visualize large social networks; it renders various adjacency matrices inside a node-link diagram. This visualization was used to find dense and sparse connections between the cliques in the node-link representation.

Comparison of visualization tools Existing visualization tools, such as R, Tableau, ManyEyes and systems, such as the InfoVis toolkit, D3, Protovis, also differ from each other with respect to the visualization techniques that they use to visualize the data and also their effectiveness to produce a desired result. It is important to collect a list of various visualization tools and systems and compare them to determine which are suitable for a particular kind of application. Thus instead of re-inventing the tools or writing the systems from scratch, the already available tools and systems can be compared to assess their suitability for answering the research questions of the domain experts.

Design and implement preliminary prototype

After the comparison of various information visualization tools and techniques, a preliminary prototype of the visualization interface can be designed in order to validate the requirements of the domain experts. The prototype will also the help to identify the usability issues faced by the domain experts and will also help to acquire additional requirements from the domain experts.

Perform qualitative evaluation and gather feedback

This involves performing a qualitative evaluation of the preliminary prototype. The qualitative evaluation is preferred over a quantitative evaluation because the quantitative user evaluation requires typically large sample sizes to make any statistically significant statements [137] which may not be possible in all the domains.

The qualitative evaluation involves preparing a set of questions based on the design requirements of the domain experts. Afterwards, the domain experts are asked to solve those questions using the prototype in order to evaluate their performance with regards to the usage of the prototype. It also involves identifying any usability issues or design problems that the domain experts may face while answering those questions. This can be considered as the feedback from the domain experts and taken into account for future improvements of the preliminary prototype. If the domain experts are satisfied, the preliminary prototype can be finalized to solve the domain experts research questions. However, if the domain experts are not satisfied, the preliminary prototype must be refined by gathering additional design requirements or checking the identified problem or prepossessed data.

4.2 Application of the design strategy on handwritten historical documents data

The design strategy described in the previous section is applied to the handwritten historical documents of St. Matthias dataset. The detailed explanation of this application is described below.

Identify problem

This step aimed to identify the research questions and challenges faced by domain experts in order to analyze their data. Through a series of interviews and regular meetings, a set of research questions were formulated:

- Determine the geometric proportions which medieval artisans used to follow to write the historical documents?
- At what times have historical document been put together on two different materials, i.e., paper and parchment?
- Is the number of blank pages changing with respect to the writing material of the historical documents, e.g., is the number of blank pages in documents made of paper rising in proportion to the number of pages?
- Are there any differences between manuscripts made of parchment and manuscripts made of paper with respect to the page size and page height?
- In which centuries have several texts with different contents been assembled into one document?

With the lack of efficient visualization interfaces to address their research questions, domain experts struggle to answer such questions effectively. Currently, they are using static visualization techniques, such as a bar chart or a scatter plot to analyze the data.

Understand and preprocess data

In case of the St. Matthias dataset, it was known that handwritten historical document images are mainly located in Trier, Germany.

- **Understanding of the origin of the data:** The first category of the data includes the bibliographical information, such as the century of the production, material, binding format, and a number of leaves which is structured according to the

METS with the TEI header. The second category includes all the layout features extracted from the automatic layout analysis method. This includes the bounding box measurements of the physical regions and the corresponding layout features (see Section 3.4). The bounding box measurements are in centimeters (cm) whereas the units of various color features, such as hue is represented in degrees (°) and ranges from 0° to 360°. The saturation and brightness are represented in percentage (%) and ranges from 0% to 100%.

- **Measurement units associated with the data:** The century of production has a unit "AD" associated with it. The binding format has degree (°) symbol associated with it. This degree symbol has a special meaning, i.e., the number of leaves after a writing material is folded. For example, 2° means two leaves after folding a writing material one time.

- **Format and type of the data:** The data is stored in XML and CSV format whereas, the handwritten historical documents are stored in TIFF/JPEG format.

- **Relationships which exist between data and its corresponding information:** On one hand, bibliographic information, such as the century, library, writing material, and the binding format seemed relevant to draw general conclusions. On the other hand, bibliographic information, such as, library location, library name, writers, and binding of the document covers was considered irrelevant as this information was described in long text sentences with many missing values in between.

Determine design requirements

This involves understanding the needs of domain experts in detail. After various regular weekly meetings and various face-to-face meetings with domain experts, a lot about physical layout analysis for the historical documents was learned. In response to various research questions mentioned above, following design requirements were formulated for designing a visualization interface for visualizing multidimensional layout features of handwritten historical document images derived from the automatic layout analysis approach.

- provide an overview of the complete dataset to help achieve an overall picture of the data;

- provide an overview of the complete physical structure of the document images and also allow the analysis of the layout features of the physical regions of the documents in detail;

- support different type of interactions, such as selecting, dragging, clicking, and mouse-over to gives domain experts much-required freedom. Interaction techniques make the visualization design easy to learn and can also be used to represent different subsets of the information;
- enable domain experts to discover errors of automatic layout analysis, or to perform the quality check of the automatic layout process interactively;
- facilitate the comparison of layout features between many documents;
- support determination of correlations, detection of outliers, or extraction of important variables.

Compare information visualization techniques, tools, and systems

Comparison based on type of visualization and analysis tasks

Standard 1D-3D techniques are used for providing a distribution of data values. For example bar chart makes use of rectangular bars to give the distribution of values for a category of data. It is capable of comparing two or more classes of data. As the number of data values in the standard 1D-3D visualization techniques, such as line chart, bar chart, or scatter plot increases, gaining an overview and determining clusters in the data gets challenging.

Iconographic techniques as mentioned above make use of icons to display multidimensional data. Techniques like Chernoff faces providing a method of visualizing multidimensional data using human cartoon faces. Here, identification of similar attributes and outliers is possible as compared to getting an overview of the complete dataset with the bird's eye view. Lee [138] performed an empirical evaluation of Chernoff faces and star glyphs with 32 participants and stated that these visualization techniques need long response time to come to a conclusion and low confidence is reported in the participants while using iconographic techniques. Stick figures also give a general overview of the data which might be difficult to interpret. Moreover, these techniques require large screen space as compared to other visualization techniques.

Geometric projections techniques visualization techniques use two-dimensional space to visualize multidimensional data to determine interesting correlation in the dataset. It includes visualization techniques, such as the parallel coordinate plot, the scatterplot matrix, and the star coordinates. Most of these techniques are capable of identifying outliers, correlations, and trends. But these techniques also have their limitations; for example, scatterplot matrices do not work well with discrete values because the measurement with decimal places are not accurate enough when they are rounded off [139]. Moreover, as the number of dimensions increases the problems of visual cluttering and overplotting occur in parallel coordinate plot and scatterplot matrices [140]. This issue of visual clutter and overplotting makes it challenging to identify the clusters in the dataset [141].

Pixel-oriented visualization techniques use each pixel of the complete display to visualize a data value. There exist two types of pixel-oriented visualization techniques (a) query-independent visualization techniques (b) query-dependent visualization technique. Here query-independent techniques visualize the complete or a portion of data set while query-dependent techniques visualize data in the context of a query [142]. The visualization techniques falling under this category are capable of identifying clusters and correlations in huge datasets, but these methods are not suitable for the identification of outliers or detection of uncertainty.

Hierarchical techniques are mainly used for tree-structured data. Methods like treemap, node-link diagram, and sunburst are useful in providing an overview of the complete dataset on the one hand but are not capable of identifying outliers quickly on the other side. Visualization techniques belonging to this category are also capable of determining clusters in the dataset. Table 4.1 summarizes the comparison of visualization techniques with respect to various visualization and analysis tasks.

Table 4.1 – Comparison of visualization techniques concerning visualization and analysis tasks

Category	**Visualization technique**	**Visualization and analysis tasks**						
		overview	**clustering**	**distribution**	**correlation**	**outliers**	**patterns**	**comparison**
1D to 3D	Bar chart	+		+		+	+	+
	Line chart		+		+	+	+	+
	Scatter plot		+	+	+	+	+	+
Iconographic	Stick figures	+	+			+	+	
	Chernoff faces	+	+	+		+	+	+
Geometric projection techniques	Parallel coordinate plot	+	+	+	+	+	+	+
	Scatterplot matrix		+	+	+	+	+	+
	Star coordinates				+	+	+	
Pixel-oriented	Pixel bar chart			+	+		+	
	Circle segment	+		+	+		+	
Hierarchical	Treemap	+	+	+		+	+	
	Sunburst	+	+	+			+	+
	Radial tree	+	+	+		+	+	+
	Dimensional stacking	+	+	+		+	+	+
Misc.	Table lens			+	+	+	+	
	Heatmap	+	+				+	
	Andrew curve	+	+	+		+	+	+
	Tilebars	+		+			+	

Comparison based on type of data

Standard 1D-3D techniques are mainly used to visualize the quantitative type of data. For example, line chart or scatter plot are mainly used to represent quantitative data as compared to bar chart which can be used to represent qualitative as well as quantitative data type [143].

Iconographic techniques, such as stick figures and Chernoff faces are used to represent quantitative data type [142]. The shape and features of the icons used by techniques are highly dependent on the quantitative value of an attribute.

Geometric projections techniques can represent qualitative as well quantitative data type. For example, parallel coordinate and star coordinate plot can be used to represent both data types as compared to scatter plot which is more suitable for the quantitative data type.

Pixel oriented visualization techniques, such as circle segments and pixel bar charts are also used for representing quantitative data type [142].

Hierarchical techniques as mentioned before are used to represent hierarchical relationships. The techniques, such as the radial tree or tree map can both be used to represent qualitative as well as the quantitative data type.

Table 4.2 summarizes the comparison of visualization techniques with respect to various type of data.

Table 4.2 – Comparison of visualization techniques concerning various type of data

	Visualization technique	**Type of data**				
		Qualitative			**Quantitative**	
		binary	**nominal**	**ordinal**	**discrete**	**continuous**
1D - 3D	Bar chart	+	+	+	+	+
	Line chart				+	+
	Scatter plot					+
Iconographic	Stick figures				+	+
	Chernoff faces				+	+
Geometric projections	Parallel coordinate plot	+	+	+	+	+
	Scatterplot matrix					+
	Star coordinates	+	+	+	+	+
Pixel-oriented	Pixel bar chart				+	+
	Circle segment				+	+
Hierarchical	Treemap	+	+	+	+	+
	Sunburst	+	+	+	+	+
	Radial tree	+	+	+	+	+
	Dimensional stacking	+	+	+	+	+
Misc.	Table lens	+	+	+	+	+
	Heatmap	+	+	+	+	+
	Andrew curve				+	+
	Tilebars				+	+

Comparison based on number of records and dimensions

Visualization techniques also differ from each other with respect to a number of records or dimensions that they can visualize. Some visualization techniques can represent a large number of records and dimensions as compared to others. For example, the parallel coordinate plot can be used to represent a large number of records as well as a large number of dimensions as compared to standard one-dimensional display, such as bar chart. Additionally, iconographic visualization techniques, such as stick figures or Chernoff faces can handle a large number of records and dimensions as compared to line charts or scatter plot, but these techniques are limited to the screen space occupied by the icons and the stick figures. Geometric projection techniques can be used to represent multidimensional data, but pixel-oriented visualization techniques outperform when it comes to visualizing very large number of data.

Therefore, an exact number of dimensions and number of records for each of the visualization techniques is difficult to define because most of the visualization techniques are subject to individual perception skills and also dependent on the size of the display screen. In 2001, Keim [142] performed a comparison of various visualization techniques and stated that a parallel coordinate plot could represent approximately 1000 data records.

However, various interaction techniques, such as selecting, zooming, distorting, and filtering can be applied to each of the visualization technique to overcome the limitation of scalability and dimensionality in large datasets. Keim [142] and Shimabukuru [144] stated that with the help of interaction techniques and selection of a subset of a large amount of data the problem of representing a large number of records and high-dimensions can be solved.

Comparison of visualization tools and systems

ManyEyes visualizes 1D-3D, hierarchical, network, and text data using visualization techniques, such as tag cloud, treemaps, line charts, stack graphs, bar chart, and bubble chart. Xmdvtool visualizes only multidimensional data using parallel coordinate plot, scatterplot matrix, star glyphs, and dimensional stacking. Mondrian also visualizes the multidimensional data using bar charts, histograms, and parallel coordinate plot. Here, ManyEyes provides only dimension manipulations as one of the interaction techniques. Various other interaction techniques, such as brushing and linking or focus and context are missing. Xmdvtool and Mondrian tool provides the possibility of creating only one level of multiple coordinated views. Moreover, getting the details of each of the data value could not be obtained.

Comparison based on visual composition methods

Javed et al. [130] also defined various advantages and disadvantages of each of the visual composition technique as described above. These advantages and disadvantages are summarized in the Table 4.3.

Table 4.3 – Advantages and disadvantages of visual composition techniques

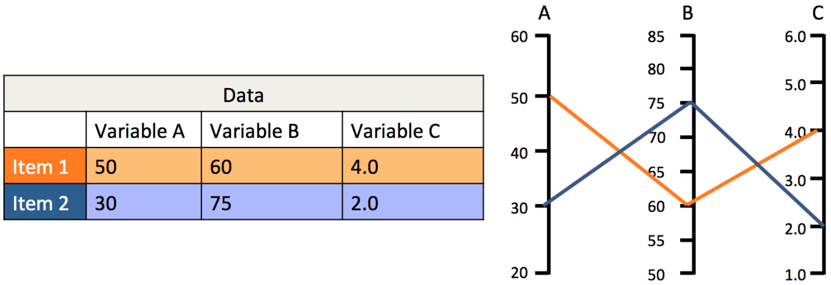
Composite visualization Views	**Advantages**	**Disadvantages**
Juxtaposition	Independent, no interference	Implicit visual linking is difficult
Superimpostion	one-to-one spatial linking, direct comparison	Occlusion, high visual clutter
Overloading	no one-to-one spatial linking	Visual clutter
Nesting	compact representation	High visual dependencies, limited space

Comparison results

Based on the comparison table 4.1 and 4.2, it is seen that standard visualization techniques, such as the bar chart, the line chart, and the scatter plot are mostly suitable for determining correlations and patterns in the data. But, they are not good enough to provide an overview of the entire dataset and are also not adaptable to qualitative data type except for the bar charts. Whereas, iconographic visualization techniques, such as stick figure and Chernoff faces can be used to get an overview of the entire dataset, they are not suitable for determining relationships between two or more variables or for visualizing the qualitative data type. Moreover, geometric projection techniques, such as, the parallel coordinate plot and the scatterplot matrix are suitable for viewing the clusters, finding the groups, correlation, and outliers. They also allow to compare multidimensional datasets directly. However, the scatterplot matrix is not suitable for the qualitative data type. The pixel-oriented techniques, such as pixel bar chart, and circle segment are not ideal for the qualitative data type. The pixel-oriented methods are only useful for finding correlations or distribution. Hierarchical techniques, such as radial tree perform well in case of qualitative datatype or getting an overview, finding clusters, distributions, or outliers in the data as compared to a treemap or a heatmap.

Thus, geometric projection techniques, such as the parallel coordinate plot [113] and the radial tree are considered for visualizing multidimensional extracted layout features of handwritten historical documents. As described briefly in Section 4.1, the parallel coordinate plot is mainly used for comparing many dimensions and seeing correlations between the data. In the parallel coordinate plot, each attribute of the data is given a vertical axis, which is placed parallel to each other. Here, each vertical axis of the parallel coordinate is assigned a different scale because each attribute has a different measurement unit. The data values are connected through a series of line segments as shown in Figure 4.3.

The main limitation of the parallel coordinate plot is that as the number of dimensions or number of records in the dataset increases visual cluttering also increases. But, as mentioned above, various interaction techniques such as, brushing and linking or focus and context can help to overcome this limitation where these interaction techniques assist in selecting the subset of the entire dataset.



Data			
	Variable A	Variable B	Variable C
Item 1	50	60	4.0
Item 2	30	75	2.0

Figure 4.3 – Data values present in the data table are represent as data points on the parallel coordinate plot. These data points are connected through a series of line segments in parallel coordinate plot. Such a plot is mainly used to detect correlations or patterns in the dataset.

The radial tree [145] is a node-link diagram and is mainly used to represent hierarchical relationships in a tree structure as shown in Figure 4.4. It consists of a node called as root node situated at the center and has no superior or parent. There are other nodes called as child nodes which are linked together through a line segment called as the link. These links represent the relationships between various connections. And lastly, there exist leaf nodes which do not have any children or child nodes. The child and leaf nodes are placed in the concentric circles around the root node in the level of their hierarchy. For example, immediate child nodes of the root node are arranged on the smallest inner ring of the radial tree and so on.

Based on the comparison table 4.3, it was difficult to decide that which of the composition methods would be suitable for multidimensional datasets of handwritten historical documents. Therefore, considering the limitations of visual composition methods, such as occlusion, high visual dependencies, or the complexity of the underlying dataset. Also, providing the possibility to perform the direct comparison of the data only juxtaposing different visualizations was considered as a valid option to design a first prototype of the visualization interface for historical documents.

As far as visualization systems are concerned, Data-Driven-Documents (D3) seemed to be a better option than using other visualization systems because oftheir ability to provide support for constructing visualization systems on the web.

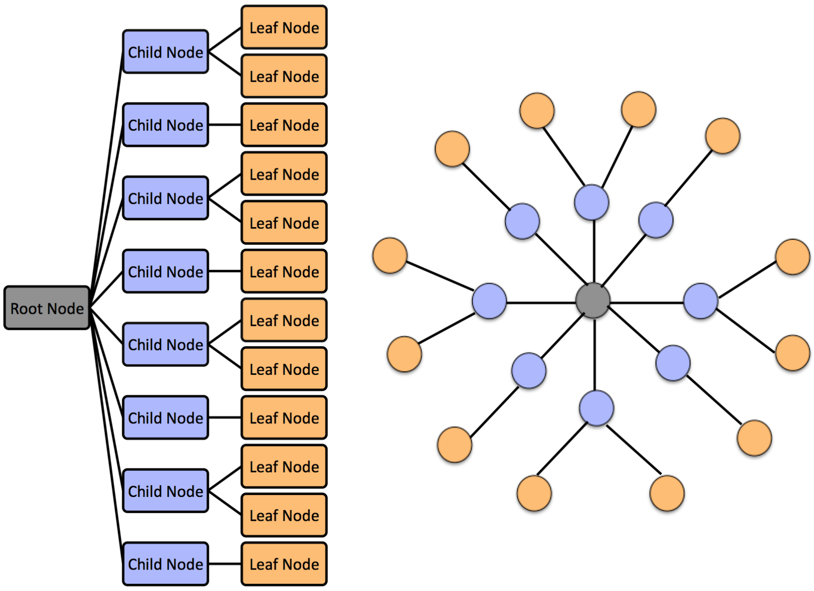


Figure 4.4 – Hierarchical relationship in a tree structure represented by the radial tree.

Design and develop preliminary prototype

A preliminary prototype called CodiViz-I using D3 was designed. In this prototype, a parallel coordinate plot was juxtaposed with the radial tree. This design is explained in detail in chapter 5.1.

Perform qualitative evaluation and gather feedback

In order to validate and assess how well the domain experts can use the CodiViz-I visualization interface to explore and discover interesting findings, and find any usability issues, a qualitative evaluation (laboratory study) was conducted. The questions were designed on the basis of research questions of the domain experts and various visualization and analysis tasks. After that, feedback from the domain experts was gathered for the prototype refinement. This is also explained in detail in chapter 5.1. Based on the feedback a revised system was designed and developed called CodiViz-II. This is explained in detail in chapter 5.2.

4.3 Summary

This chapter describes a generic design strategy for designing and choosing the appropriate information visualization techniques for a particular kind of application. It also describes the application of this design strategy to the real-world digital humanities datasets and finds out the best possible techniques to visualize the multidimensional layout features generated from automatic layout analysis method. The inadequate use of visualization techniques can result in incorrect exploration results where already analysis problems are ill-defined. Therefore, the problem of the domain experts should be defined clearly as the first step to design a visualization interface. Various visualization techniques can be compared based on the type of the data, number of dimensions, number of records and type of visualization and analysis tasks. This chapter also evaluates existing composition theories to merge two or more visualization techniques into one. As a result of this evaluation, it is found that the parallel coordinate plot and radial tree are the best possible techniques to visualize multidimensional layout features of the historical handwritten documents whereas juxtaposition is considered as the best option to merge two or more visualization techniques to create a composite visualization.

Chapter 5

Interactive visual exploration of handwritten historical documents

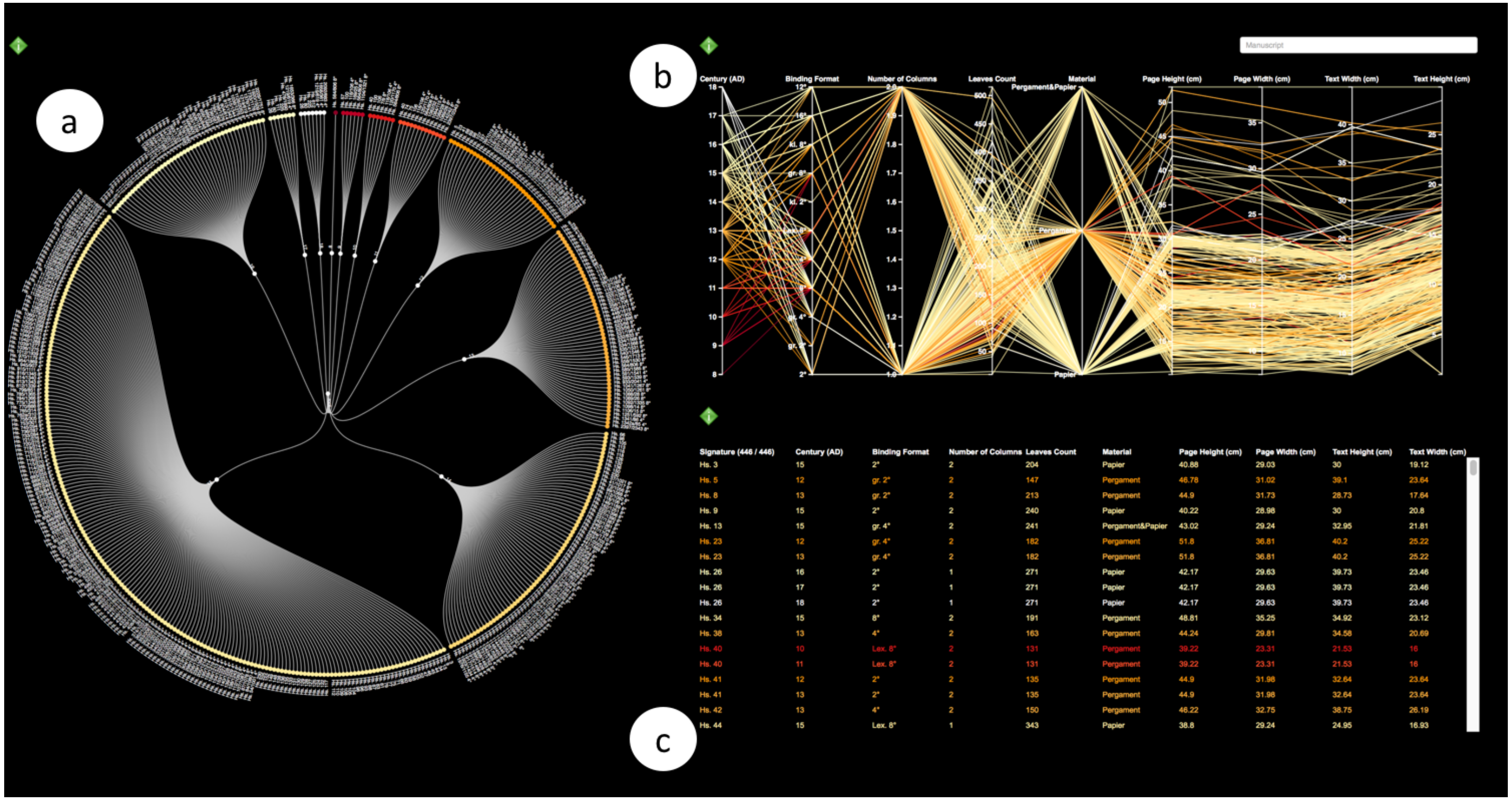
This chapter presents two interactive information visualization designs with a primary focus on the visual exploration of multidimensional layout features of handwritten historical document images extracted from the automatic layout approach described in Chapter 3. In recent years, the visualization research community has researched and developed many visual representations for the text and document data (see Section 2.4). These visualization techniques have enabled the analysis of many documents available online in the form of news data, journals, books, and historical documents. Most of these visualization techniques focus on the exploration of the semantic content of the document collection via summarization. However, there has been comparatively less research done on visualization of the physical layout structure of the historical documents, primarily handwritten historical documents with irregular and overlapping physical layouts. Chapter 4 explains this aspect in detail by describing a design strategy in order to design a visualization interface and find the best possible visualization techniques to support exploration of the physical layout of the documents by detection of outliers, extraction of essential variables, determination of correlations, etc. As a result of the application of this strategy on documents data, it is indicated that the parallel coordinate plot and the radial tree are some of the best possible techniques to explore the physical layout of the documents. Whereas, the superimposition and the juxtaposition can be used to merge two or more visualization techniques to create a composite visualization for text and document data.

5.1 CodiViz-I visualization design interface

CodiViz-I is an interactive visualization interface designed for the visual exploration of multidimensional and documents data. The objective of designing such a visualization interface is to enable domain experts to explore the physical layout structure of the documents with irregular and overlapping physical layout. As shown in Figure 5.1, CodiViz-I mainly combines two coordinated visualization techniques (i.e., the radial tree and the parallel coordinate plot). In this section, various aspects of this visualization design are described, including various interactions that visual exploration supports by getting an overview of the data or by drilling down the details. The combination of the radial tree and parallel coordinate plot is described in the following sections.

5.1.1 Overview: Exploring the whole dataset

The overview of the documents is provided by the bibliographic features displayed in the radial tree. The radial tree is a node-link tree which is positioned in polar coordinates [145]. This visualization technique can directly see the clusters in the hierarchical levels of the dataset. Therefore, from the St. Matthias dataset, bibliographic features, such as the century, the material, and the binding format, are used to create three different radial trees. On the first level of the hierarchy, the nodes in a radial tree represent values belonging to a particular bibliographic feature. For example, Figure 5.2 shows a radial tree where each node on the first level of hierarchy shows a different century. On the second level of the hierarchy, signatures that are identifiers (IDs) of the documents belonging to each bibliographic feature are displayed. For instance, the second level of hierarchy in Figure 5.1 shows the signatures of the documents pertaining to different centuries. Furthermore, the nodes at the second level of the hierarchy are assigned different colors with varying brightness. This color scheme helps to distinguish various clusters of data values and also increase the appeal of the visualization.



Signature (446 / 446)	Century (AD)	Binding Format	Number of Columns	Leaves Count	Material	Page Height (cm)	Page Width (cm)	Text Height (cm)	Text Width (cm)
Hs. 3	15	2°	2	204	Papier	40.88	29.03	30	19.12
Hs. 5	12	gr. 2°	2	147	Pergament	46.78	31.02	39.1	23.64
Hs. 8	13	gr. 2°	2	213	Pergament	44.9	31.73	28.73	17.64
Hs. 9	15	2°	2	240	Papier	40.22	28.98	30	20.8
Hs. 13	15	gr. 4°	2	241	Pergament&Papier	43.02	29.24	32.95	21.81
Hs. 23	12	gr. 4°	2	182	Pergament	51.8	36.81	40.2	25.22
Hs. 23	13	gr. 4°	2	182	Pergament	51.8	36.81	40.2	25.22
Hs. 26	16	2°	1	271	Papier	42.17	29.63	39.73	23.46
Hs. 26	17	2°	1	271	Papier	42.17	29.63	39.73	23.46
Hs. 26	18	2°	1	271	Papier	42.17	29.63	39.73	23.46
Hs. 34	15	8°	2	191	Pergament	48.81	35.25	34.92	23.12
Hs. 38	13	4°	2	163	Pergament	44.24	29.81	34.58	20.69
Hs. 40	10	Lex. 8°	2	131	Pergament	39.22	23.31	21.53	16
Hs. 40	11	Lex. 8°	2	131	Pergament	39.22	23.31	21.53	16
Hs. 41	12	2°	2	135	Pergament	44.9	31.98	32.64	23.64
Hs. 41	13	2°	2	135	Pergament	44.9	31.98	32.64	23.64
Hs. 42	13	4°	2	150	Pergament	46.22	32.75	38.75	26.19
Hs. 44	15	Lex. 8°	1	343	Papier	38.8	29.24	24.95	16.93

Figure 5.1 – Radial tree view (a) represents century of the historical documents, (b) corresponding layout features are represented in the parallel coordinate plot view on the right side, (c) table showing the numeric values of the layout features represented in the parallel coordinate plot in the row and the column format.

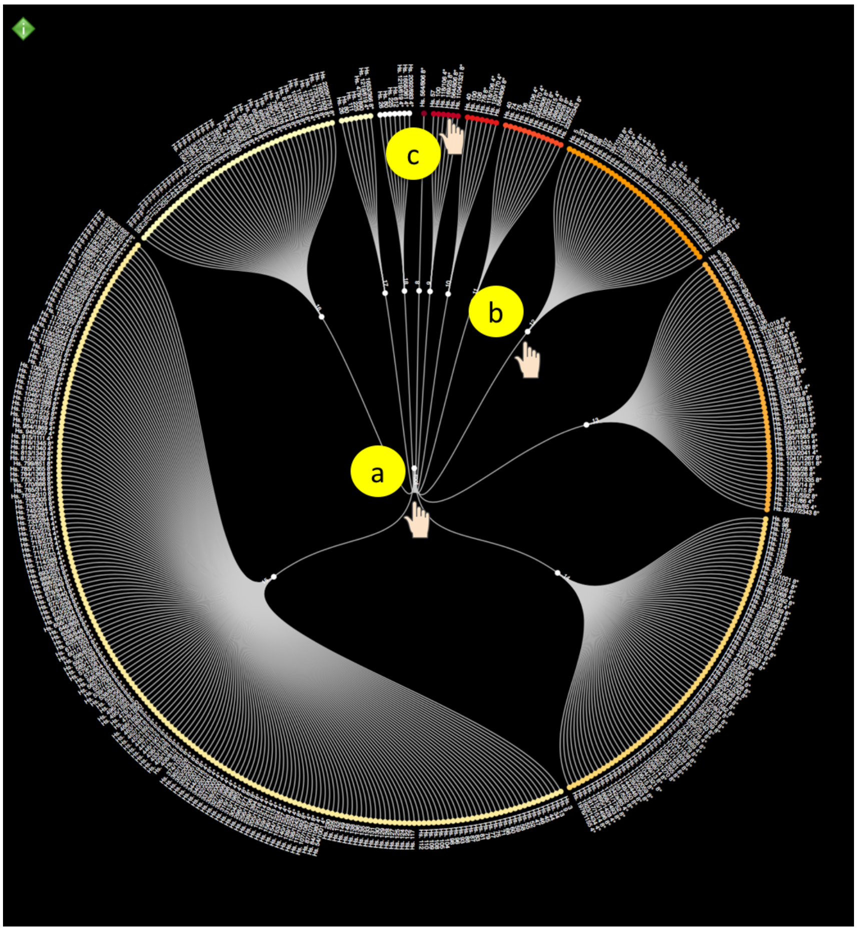


Figure 5.2 – Exploration of the bibliographical features of the St. Matthias dataset (a) root node of the radial tree (b) different centuries in which document were written (c) the signatures which are the identifiers of the documents given by library.

5.1.2 Drilling down: Further exploration

Various layout features, such as the size features of page regions and text regions, which are extracted from the automatic layout analysis approach, are represented using the parallel coordinate plot. This plot is juxtaposed with the radial tree described in the previous section. The vertical axes of the parallel coordinate plot are ordered in a way such that the first vertical axis in the parallel coordinate plot corresponds to the first level of the hierarchy of the radial tree, i.e., both represents the bibliographic feature, i.e., the century, in this case. The remaining vertical axes of the parallel coordinate plot shows other bibliographical features, such as binding format, number of leaves, and material. The parallel coordinate plot also shows the size features of the documents including document page width, document page height, the average height of the main text areas of the document page, and the average width of the main text areas of the document page. Here, each record of the size features of the document from the data table is represented by a connected line segment as shown in Figure 5.3. The same color is assigned to a particular bibliographic feature in the radial tree and the parallel coordinate plot so that the user could relate the corresponding data values across two visualizations. For example, the red color assigned to the 8th century in the radial tree is same as the red color assigned to the 8th century in the parallel coordinate plot (see Figure 5.1). There exists a data table which represents the values of the layout features in row and column format. The first row of this data table represents the names of the bibliographic and extracted layout features. From the second row onwards, the measurement values of the bibliographic features and the extracted layout features are shown.

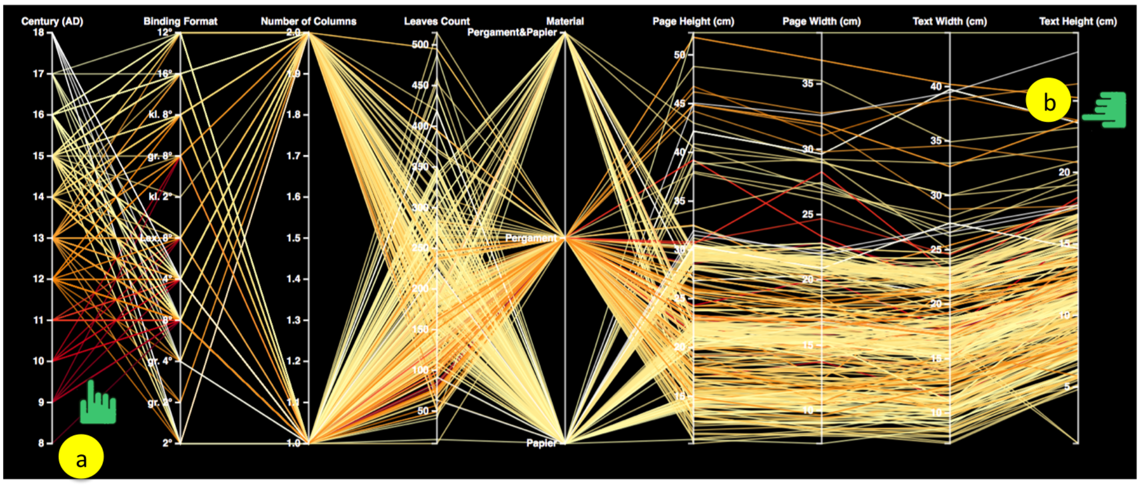


Figure 5.3 – Exploration of the layout features of the St. Matthias dataset (a) a line segment connecting data values of the layout features of the document (b) a vertical axis representing the layout features itself.

5.1.3 Interaction techniques

The radial tree and the parallel coordinate plot are linked together through the brushing and linking interaction technique. Two different kinds of brushes exist: a polar brush and a vertical brush. Here, the brush is referred to as an axis-aligned rectangle for making selections in the visualization technique [146].

- **Polar brush**: This brush allows the selection of a subset of nodes of a radial tree in polar coordinates. For example, in Figure 5.4, nodes at the first level or the second level of the hierarchy in the radial tree can be brushed to see the corresponding reflections in the parallel coordinate plot.

- **Vertical axis brush**: This brush allows the selection of the subset of the data records of the bounding box measurements represented on the vertical axis of the parallel coordinate plot.

The data table (see bottom right of Figure 5.4) is also linked to both the radial tree and the parallel coordinate plot. A mouse-over interaction technique is provided in this data table, where hovering over a single record in the data table allows automatic reflections in the radial tree and the corresponding line segments in the parallel coordinate plot.

5.1.4 Preliminary qualitative user study

In order to validate and assess how well the domain experts can use the CodiViz-I visualization interface to explore, and discover interesting findings and find any usability issues, a qualitative evaluation (laboratory study) was conducted. A qualitative user evaluation was preferred in this thesis as compared to quantitative evaluation mainly because of the sample size (number of users required to validate the visualization interface). A quantitative user evaluation requires high sample size to make any statistically significant statements about the design [137]. For instance, an estimate of 67 users was made in order to evaluate the CodiViz-I design by using the following equation for calculating the sample size (number of users):

$$\frac{(Z-score)^2 \times StdDev \times (1-StdDev)}{(\text{margin of error})^2} \tag{5.1}$$

Here, *Z-score* was assumed to be 1.645 at 90% confidence interval, *standard deviation (StdDev)* to be 0.5, and margin of error +/-10%.

This estimate for the number of users was not possible for the sake of scope of this thesis.

Participants and apparatus

For the qualitative evaluation of CodiViz-I, 20 participants of different age groups, and from different educational backgrounds, such as the humanities, medical sciences, physics, electrical engineering, and finances were recruited from the KIT and the research project network. These included students, doctoral researchers, and participants with several years of research experience. The experiment was conducted on MacOS and connected with a 27-inch 5k retina display (resolution 5120 x 2880). Participants used the computer mouse and keyboard during the entire user study.

User tasks and procedure

Based on the visualization and analysis tasks described in chapter 4 and various user evaluation tasks conducted in previous research studies [147], [148] a set of user tasks was formulated for evaluating CodiViz-I. These tasks were divided into two categories: a) the simple category, which refers to the primary user tasks, and b) the complex category, which refers to the secondary user tasks. Although classified into different categories, some tasks might involve some usability feedback or some open-ended questions. In the primary user tasks, at first, a set of two sample questions was solved with the participants to demonstrate them the usage of the visualization prototype and make them feel comfortable with various interaction techniques. After that, the prototype of CodiViz-I was given to each participant with various analytical questions to answer. In case of the secondary user tasks, the participants were asked to explore the data by using CodiViz-I themselves to find some interesting patterns.

Each of the users was called individually according to the availability of their time and as the first step of this user study, the features and various interaction techniques, such as brushing and linking of the CodiViz-I was explained to these participants in detail followed by a live demonstration (about 10 minutes). After the live demonstration, CodiViz-I was provided to the participants so that they could get familiar with the prototype of CodiViz-I by interacting with it (10 minutes). As a second step, primary and secondary user tasks were given to these participants to solve with the help of CodiViz-I. During the whole study, the users were encouraged to use "think aloud protocol". Also, they were allowed to ask furthers questions or for further explanations whenever necessary. The user study lasted for around 1 hour with each of the participants.

1. **Primary user tasks**
   - Which documents have the largest and smallest page size (width, height)?
   - Which documents have the maximum and minimum number of pages?
   - Which documents have largest and smallest text size (width, height)?
   - Which century has the maximum number of documents written?
   - Which documents have similar page height and page width?
2. **Secondary user tasks**
   - **Detect outliers and extract important information**: In this task, participants were asked to find any outliers or any other information that may seem important to them. As a result of this task, participants determined outliers for the document written in the 10th and the 11th centuries with varying page sizes.

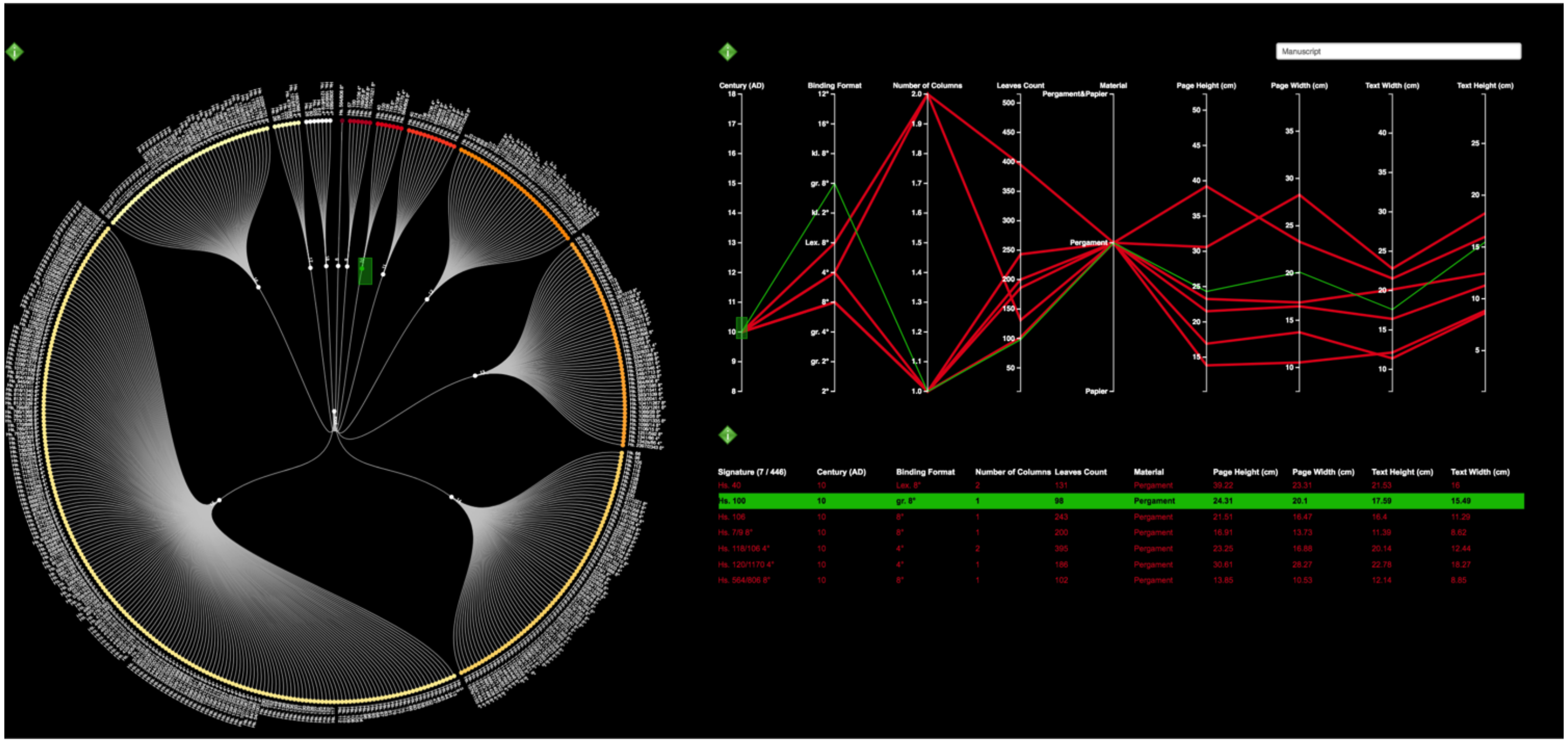
- **Determine documents sharing similar physical characteristics**: The participants were asked to provide an overview of documents sharing similar physical characteristics. In this case, participants interacted with the radial tree representing the centuries of the documents. By brushing the documents in the radial tree, similar documents, according to centuries, are reflected in the parallel coordinate plot. The participants were also asked to determine the relationships between various layout features shown in the parallel coordinate plot. They were asked to sort out the most influential of all the features considered. To determine the effect of various features, participants studied different clusters of the radial tree and linked the parallel coordinate plot. They learned that the manuscripts belonging to the 9th century have the largest page size. Thus, page area was the most influential feature of the documents belonging to the 9th century.

5.1.5 Experimental results

Observations

Participants were able to complete the primary tasks relatively quick, where each task took around 40-60 seconds. They mainly used brushing and linking between the parallel coordinate plot and the radial tree to explore the St. Matthias database. The participants appreciated the combination of two visualization techniques to get an overview of the whole data with the radial tree on the one hand and simultaneously get the details from the parallel coordinate plot on the other hand. Around 80 percent of the total participants solved the primary tasks correctly.

But, the secondary tasks took longer (10-15 minutes) because it involved open-ended questions and some feedback on the usability. Participants solved the first secondary task by exploring various radial tree generated according to selected bibliographic features and their corresponding information in the parallel coordinate plot. They were able to provide an overview with the single glance. They also solved the secondary user task 2 while solving user task 1. For finding the relationship between various layout features, they explored the clusters that covered the largest or the smallest feature space. Of the total participants, 60% to 80% of the participants were able to solve the secondary task; 20% to 40% of the inexperienced participants shared the opinion, that solving secondary tasks would require more time and understanding of the data.



Signature (7 / 446)	Century (AD)	Binding Format	Number of Columns	Leaves Count	Material	Page Height (cm)	Page Width (cm)	Text Height (cm)	Text Width (cm)
Hs. 40	10	Lex. 8°	2	131	Pergament	39.22	23.31	21.53	16
Hs. 100	10	gr. 8°	1	98	Pergament	24.31	20.1	17.59	15.49
Hs. 106	10	8°	1	243	Pergament	21.51	16.47	16.4	11.29
Hs. 7/9 8°	10	8°	1	200	Pergament	16.91	13.73	11.39	8.62
Hs. 118/106 4°	10	4°	2	395	Pergament	23.25	16.88	20.14	12.44
Hs. 120/1170 4°	10	4°	1	186	Pergament	30.61	28.27	22.78	18.27
Hs. 564/806 8°	10	8°	1	102	Pergament	13.85	10.53	12.14	8.85

Figure 5.4 – Application of polar brush in radial tree and automatic reflections in parallel coordinate plot and data table.

General feedback

Overall, participants liked the combination of the parallel coordinate plot and the radial tree for visual exploration. Most of the participants found the interactive approach to exploring data intuitive and fun to learn.

However, participants also complained that, with such visualization design, they were not able to relate back the numbers to the documents itself and see whether these numbers were telling the truth regarding the size features. Signatures (IDs) of the documents exist, but these IDs were not enough to get into the details of the document pages. Thus, the second iteration of CodiViz was designed called CodiViz-II. The design of this prototype is described in the following section.

5.2 CodiViz-II visualization design interface

CodiViz-II is also a multiple-coordinated visualization interface for the exploratory analysis of documents data and their multidimensional extracted layout features (see Chapter 3).

The design of CodiViz-II was mainly motivated by the preliminary evaluation results of CodiViz-I in the previous section. It creates an illusion of a virtual digital library of all the documents present in the database. It consists of three significant views: the document explorer view, and the document page explorer view, and single document page explorer view. The exploration is carried out at multiple levels with the fusion of various visualization techniques (see Figure 5.5). These levels are described below:

5.2.1 Document explorer view

Humanities scholars may have some domain-specific questions regarding their data, but they can be novices when it comes to the visual exploration of multidimensional datasets. The document explorer view is the fusion of parallel coordinate plot and document montage plot. This view is the first level of visual exploration. It helps humanities scholars to explore and determine correlations in the multidimensional feature space visually. In this level, the parallel coordinate plot and the document montage plot (the juxtaposition of cover pages of the document) are fused together (see Figure 5.7). This is called document explorer view which provides an overview of the entire dataset. It also allows the users to get access to the documents and the bibliographic features of the documents at a single glance. Users can select a subset of documents which they want to explore and see the details of the document on the second level of exploration.

The bibliographic features and the document page features (i.e., library, number of leaves, century, format, material, page model widths, and page model heights) are

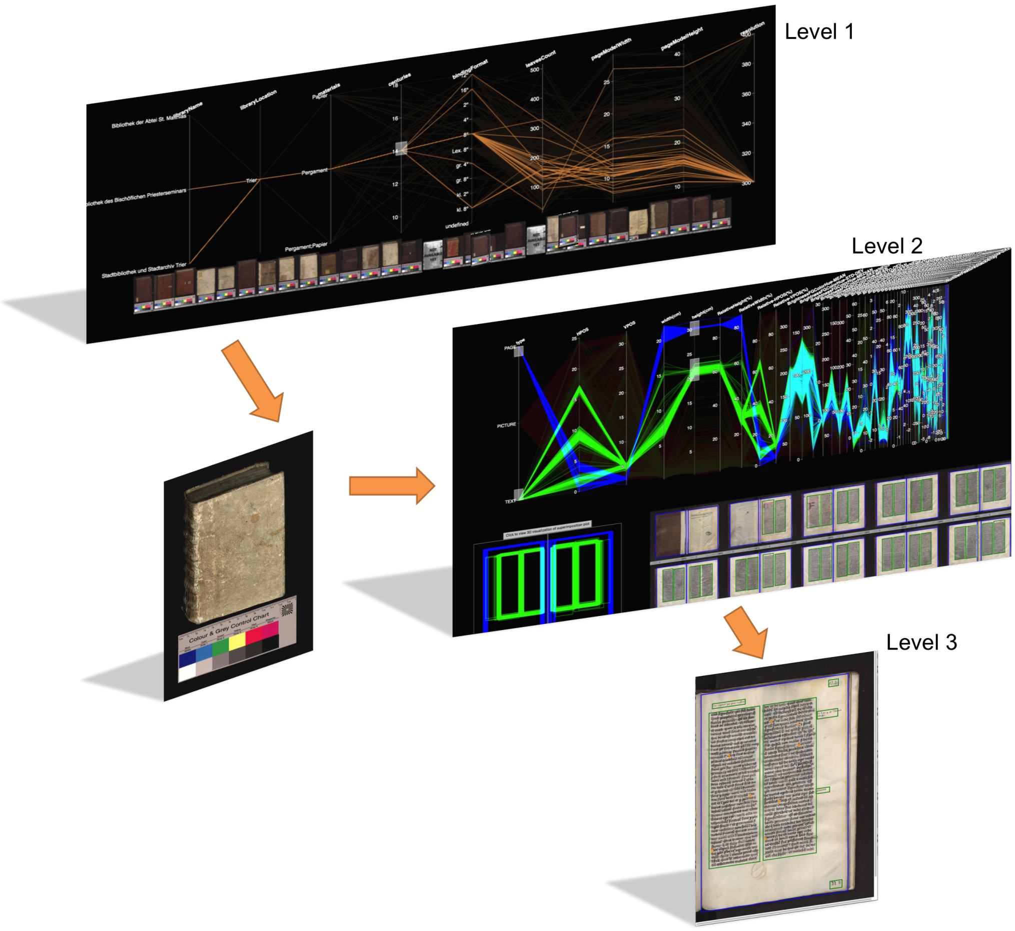


Figure 5.5 – CodiViz-II visualization interface (Level 1) Document explorer view, (Level 2) Document page explorer view, (Level 3) Single document page explorer view.

mapped onto the vertical axes of the parallel coordinate plot shown at the top of the screen display space.

Each record is represented along the line of the parallel coordinate plot. In this case, only a parallel coordinate plot is chosen to provide an overview of the complete dataset because the display space used for the parallel coordinate plot is much less than the radial tree. Moreover, one parallel coordinate plot can represent all the bibliographical features as compared to CodiViz-I where one radial tree was required to visualize one bibliographic feature. At the bottom of the screen display space, an interactive document montage plot is displayed. It shows the cover page of the documents juxtaposed to create one visual composite view. Such a document montage plot allows viewing an eclectic collection of documents in one single view. The documents are ordered according to the

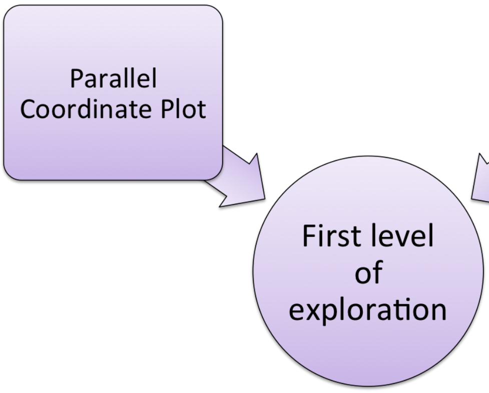


Figure 5.6 – Document explorer view, i.e., the first level of exploration is the fusion of parallel coordinate plot and document montage plot.

increasing order of their signatures where signatures are the identifiers (IDs) assigned to the document by the library. Although it is possible to scroll down with the help of computer mouse to view all the documents, this document montage plot is more useful if a subset of the document is selected from the parallel coordinate plot using brushing and linking as shown in Figure 5.8. Any of the vertical axes can be brushed to choose corresponding documents in the montage plot.

Other interaction techniques, such as axes reordering and mouse-over are also possible. In the axes reordering interaction technique, the vertical axes present in the parallel coordinate plot can be rearranged to discover some interesting patterns or correlations in the data. Mouse-over on any cover page of the document highlights the line segment representing the bibliographic feature of that particular document in the parallel coordinate plot and fades out all the other line segments as shown in Figure 5.9. This enables users to isolate the interesting sections of the parallel coordinate plot while filtering out others.

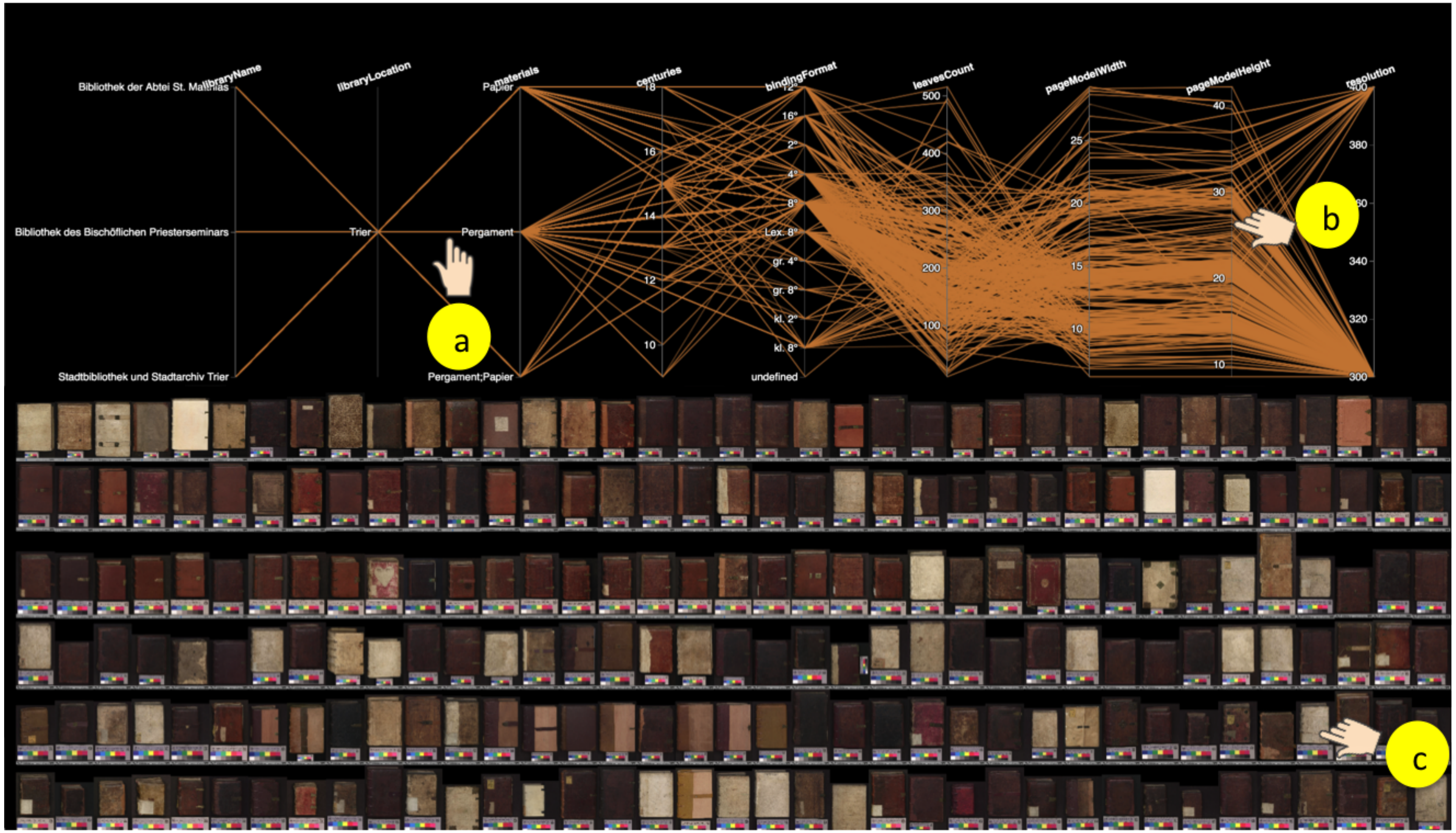


Figure 5.7 – Document explorer view: (a) a line segment connecting the data values of bibliographical features and page features of the document (b) a vertical axis representing the bibliographical features and page features itself, (c) cover pages of the document juxtaposed to each other.

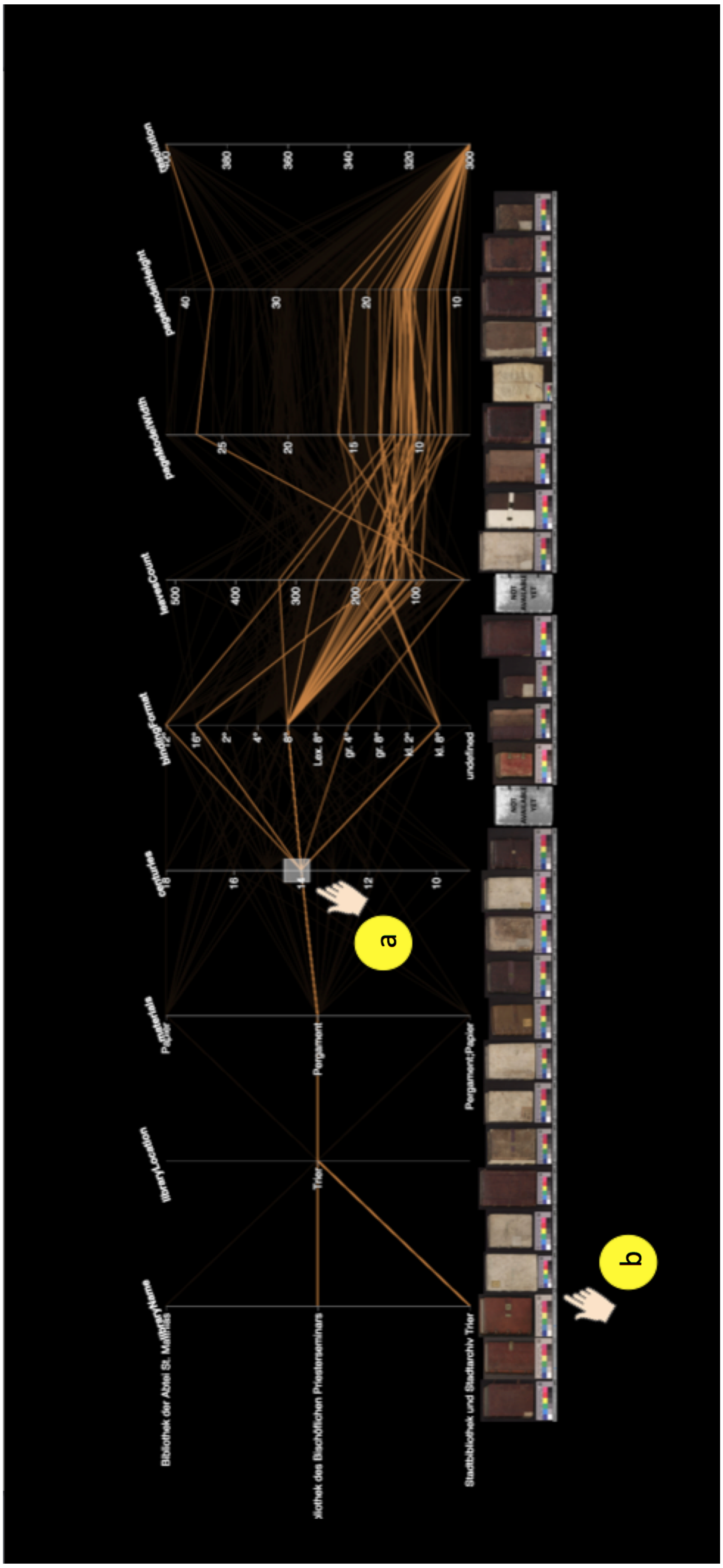


Figure 5.8 – Brushing and linking effect between a parallel coordinate plot and a document montage plot (a) a rectangular brush highlighting the documents written in 14th century (b) filtered cover pages of the documents written in 14th century.

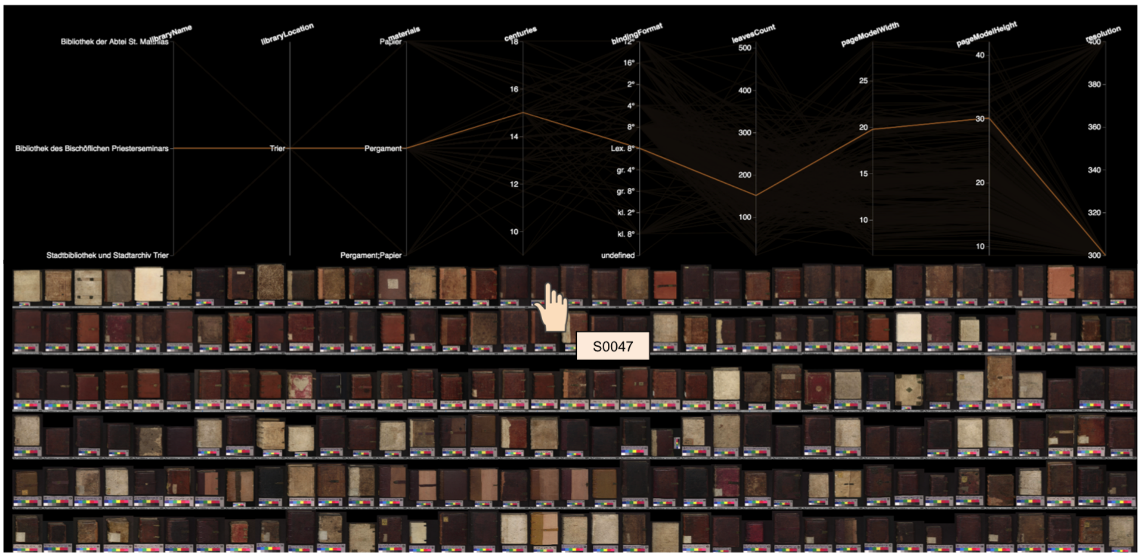


Figure 5.9 – Mouse-over effect between a parallel coordinate plot and a document montage plot.

5.2.2 Document page explorer view

From the preliminary evaluation of CodiViz-I, it was clear that there is a need to view the details of layout features of each document page individually. Thus, to explore the details of each document page, a detailed view called the document page explorer view was designed, which makes investigation or exploration reasonably simple to follow.

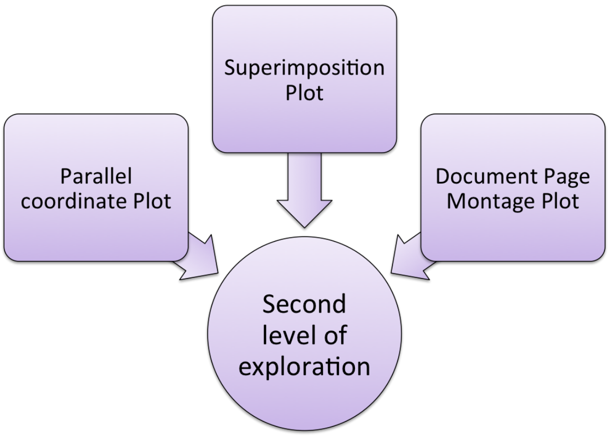


Figure 5.10 – Document page explorer view, i.e., the second level of the exploraion is the fusion of parallel coordinate plot, superimposition plot, and document page montage plot.

When a document cover page from the document montage plot present in the document explorer view is selected, the document page explorer is displayed as shown in Figure 5.11. This view is the second level of exploration which shows the fusion of three visualization techniques. These three visualization techniques are: a) parallel coordinate plot, b) superimposition plot (bounding boxes representing physical regions superimposed on top of each other), and c) document page montage plot (document images juxtaposed with each other and bounding boxes representing physical regions superimposed on document images themselves).

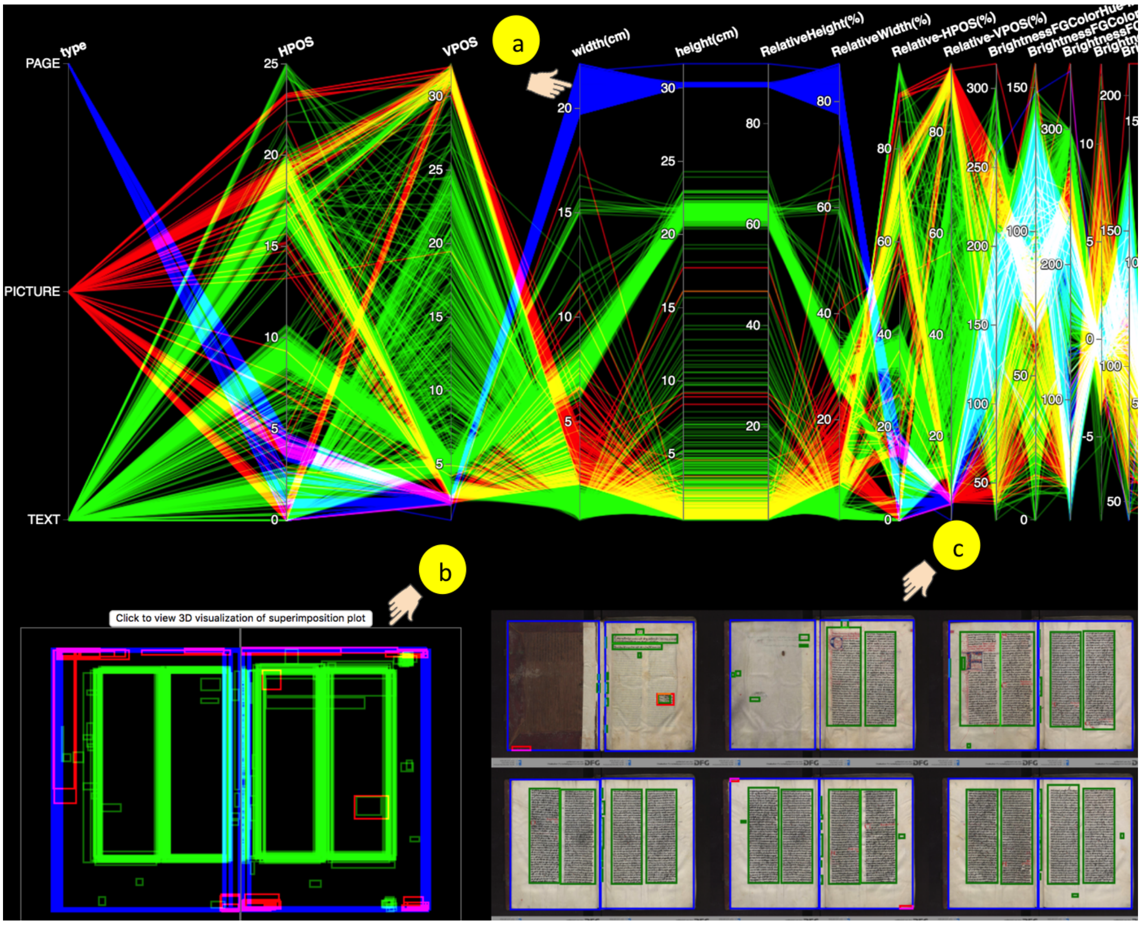


Figure 5.11 – Document page explorer view: It shows all the measurements of layout features of a single manuscript. (a) the parallel coordinate plot showing the data values of the layout features connected through a line segment. (b) the superimposition plot is shown where the size features of the page region, text region, and image region are superimposed to see the overall physical structure of hundreds of manuscript pages at a single glance. (c) the document page montage plot is shown where size feature are drawn on the respective document page.

5.2.2.1 Parallel coordinate plot

The parallel coordinate plot is displayed at the top of the document page explorer view (see Figure 5.11). All the automatically extracted layout features are visualized on this plot including the size features, the color features, the count features, and the other derived features. These features are represented along the vertical axes of the parallel coordinate plot. Each layout feature is given its own axis and is placed parallel to each other. Moreover, each axis has its own scale because each feature has its unit of measurement. For example, the width of the document is expressed in centimeters (cm). The data values are plotted as the series of line segments that are connected along the vertical axes.

The first vertical axis of the parallel coordinate plot always displays the type of physical region, namely the page region, the text region, and the image region. From the second axis onward, all the axes are ordered according to the class of layout features that are being displayed, such as size features, which include width, height, and the upper-left corner coordinates of a particular layout region. Another example of a class of layout features is the color features, which include the hue, saturation, and brightness of the foreground and background of the physical regions.

The main limitation of the parallel coordinate plot is that, as the number of line segments increases, it gets difficult to identify the individual lines resulting in a visual clutter [141]. To further optimize the parallel coordinate plot, the three primary colors, i.e., red, green, and blue are chosen to distinguish between the different physical regions of the documents. For instance, the blue-colored line segments represent the layout features of page region, the green-colored line segments represent the layout features of text regions and the red-colored line segments represent the layout features of picture regions.

Curve bundling and smoothing

Various state-of-the-art methods, such as curve bundling and smoothing algorithm [149], were applied to increase the visibility of the structure in the data across multiple axes and also to reveal the underlying structure within clusters. Here, the traditional polygon lines were replaced by the continuous piecewise Beźier curve [149]. These resulting curves are then bundled to increase the visual appeal of the clusters. These are called curve bundles because they provide the explicit visual representation of the data and also minimize the visual clutter. The properties of the piecewise cubic Beźier curve is described as follows:

- The piecewise cubic Beźier curve interpolates P_1, P_2,........P_N at the value axes.
- The piecewise cubic Beźier curve is C^1 continuous.

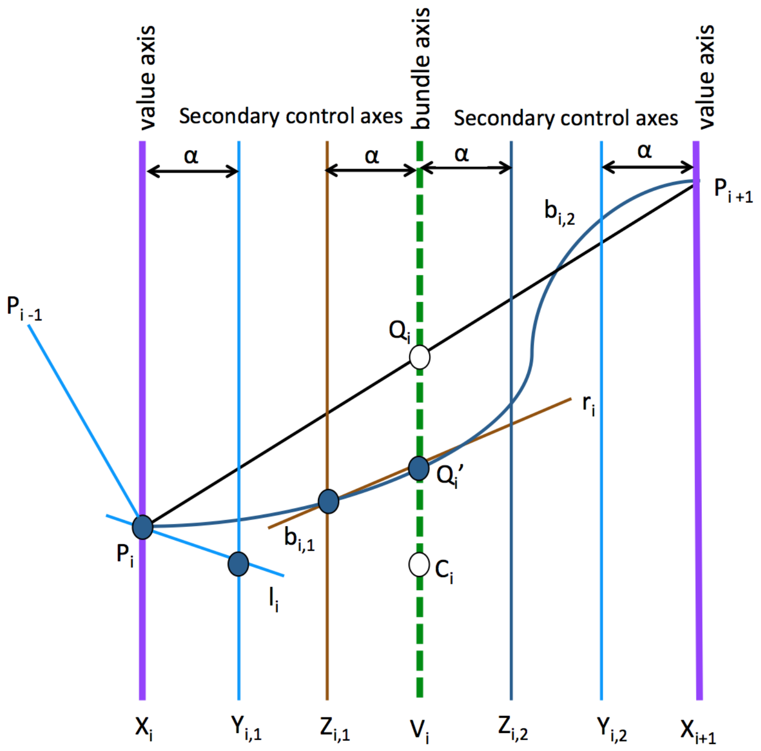


Figure 5.12 – Construction of piecewise Beźier curve between the two adjacent value axes X_i and X_{i+1}. A bundle axis V_i is inserted in the middle of X_i and X_{i+1}. The secondary control axes $Y_{i,1}$, $Z_{i,1}$, $Z_{i,2}$, $Y_{i,2}$ are also inserted. The adjacent Beźier curve shares the same tangent line and intersections between the tangent lines. The value, bundle, and secondary axes define the control points shown by blue points. The point C_i is the cluster centroid. This point pulls the constructed Beźier curve. As a result, the polyline of the parallel coordinate plot passes through Q_i'. (Image adapted from [149]

- Curves which are corresponding to the data point and belong to the same cluster are bundled together between the adjacent value axes of the parallel coordinate plot. This is done by adding a bundle axis in the middle of the value axes and by positioning the Beźier control points.

- Two parameters α and β are used to adjust the shape of the Beźier curve. Here, α is called smoothness scale, and it controls the extent the Beźier curve approximates the linear polyline of the parallel coordinate plot while the C^1 continuity is retained. β is called the bundling strength, and it indicates how tightly the curves in the cluster are bundled together. When $\alpha = \beta = 0$ the polygon parallel coordinates is the result.

Curve continuity and smoothness scale Let there be two adjacent value axes X_i and X_{i+1} connected with the two points P_i and P_{i+1} as shown in Figure 5.12. Let the intersection of the line segment P_i, P_{i+1} with the bundling axis V_i which is the midway of the X_i and X_{i+1} be Q_i. The line segment P_i, P_{i+1} is converted into two Beźier curve, i.e., $b_{i,1}$ between X_i and V_i and $b_{i,2}$ between V_i and X_{i+1}. The point Q_i is moved to Q'_i along V_i because of cluster centroid and this Q'_i serves as a common control point for two Beźier curves $b_{i,1}$ and $b_{i,2}$. Here, Heinrich et al. [149] used Holten and van Wijk's method [150] to construct the Beźier curves.

Curve bundling Let the set of polylines belong to a cluster C, and for each of the polyline, an intersection is recorded on the bundling axis V_i. Then, the centroid C_i of the intersection points corresponding to the polylines in the cluster C serve as the control point shared by Beźier curves. This point forces all the Beźier curves in the cluster to pass through this centroid C_i.

The bundling strength parameter β controls the extent to which Q_i will be pulled towards centroid C_i. A linear weighting scheme is used to calculate the new Q'_i which is given by

$$Q'_i = (1 - \beta)Q_i + \beta C_i, \tag{5.2}$$

where $0 \leq \beta \leq 1$. When the value of β = 0 the curve bundling is disabled.

In the current state of implementation, the user is given a possibility to adjust bundling strength and smoothing scale to detect the correlations in the bundled parallel coordinate plot. Figure 5.13 shows a bundled curve where α = 0.25 and β = 1.

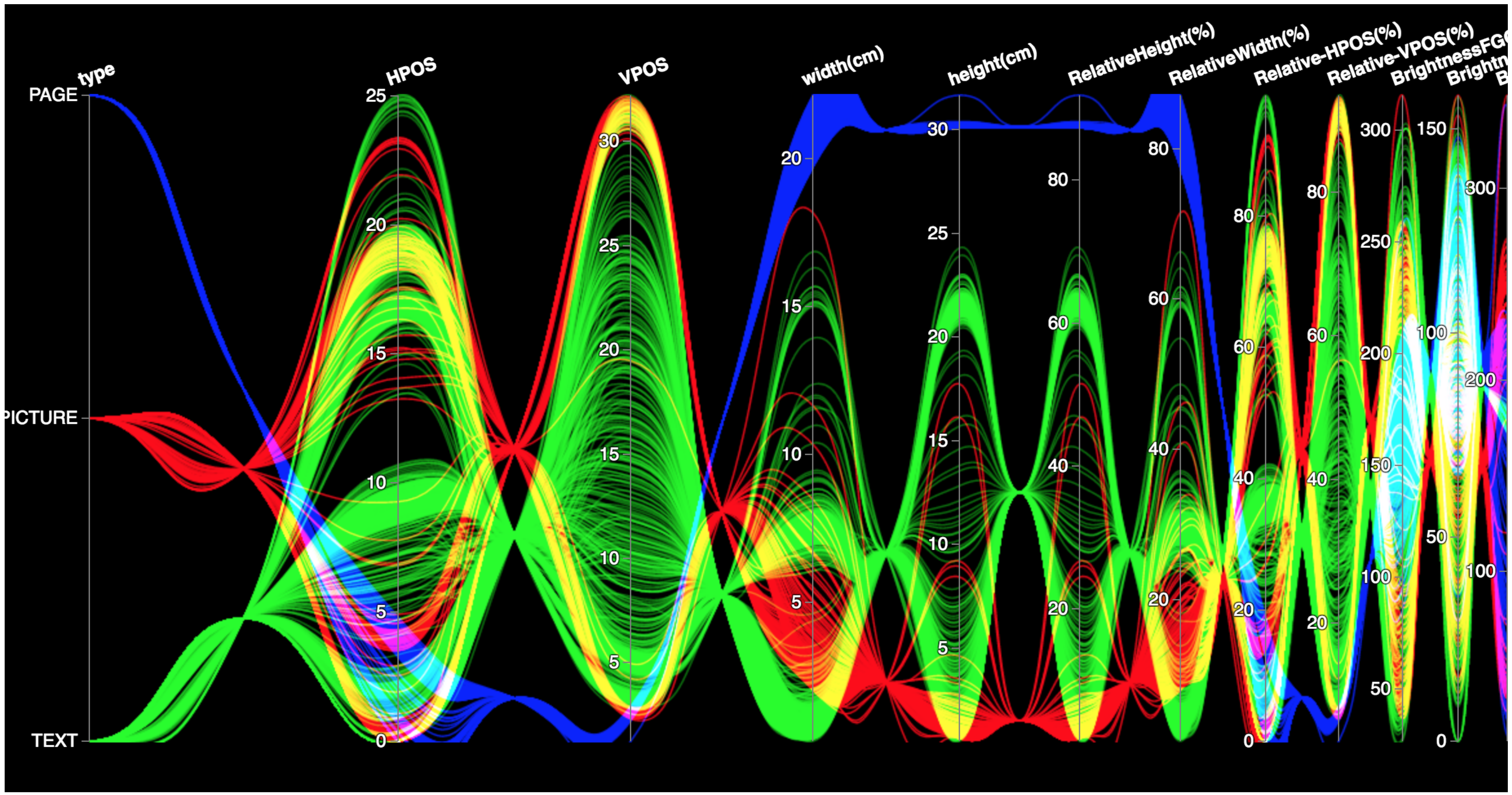


Figure 5.13 – The multidimensional layout features data displayed as bundled plots. Data is clustered by type of physical region (first vertical axis). In the bundled plot, bundling is $\alpha = 0.25$ and $\beta = 1$.

5.2.2.2 Superimposition plot

At the bottom left of the document page explorer view (see Figure 5.11), the bounding boxes elements representing the page regions, text regions, and picture regions are superimposed on top of each other to create an overall image effect. Figure 5.14 shows the bounding boxes representing the page regions, text regions, and picture regions separately. Figure 5.15 shows the superimposition of all physical regions (i.e., page regions, text regions, and picture regions) on top of each other.

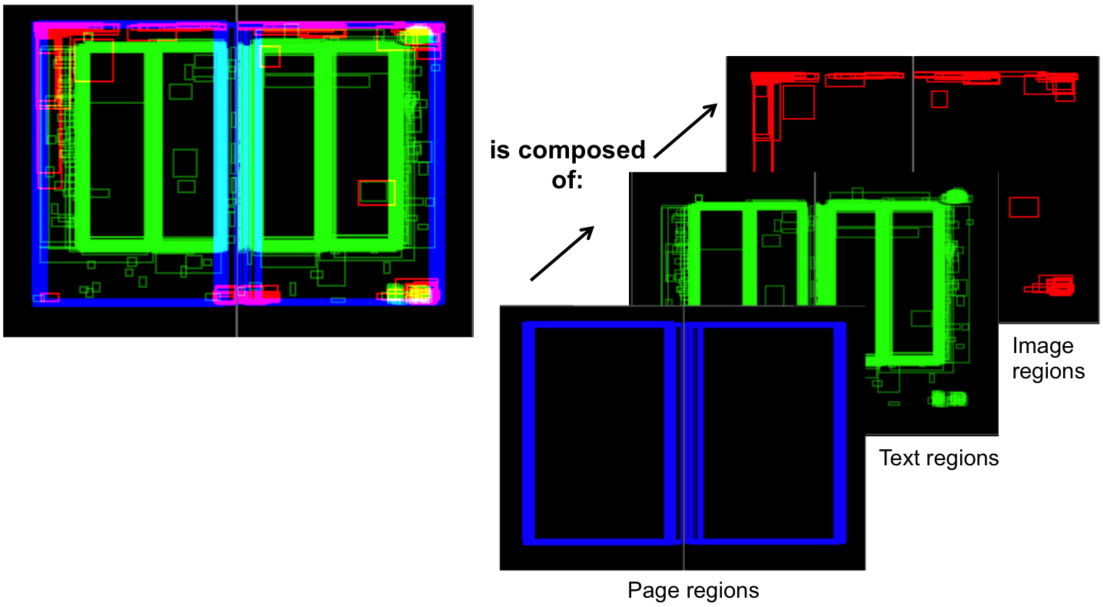


Figure 5.14 – Superimposition of bounding boxes representing (a) page regions, (b) text regions, and (c) picture regions.

The bounding boxes representing various physical regions are superimposed according to the document page enumeration: the odd pages are overlaid on the left-hand side of the superimposition component, while the even pages are superimposed on the right-hand side. Here, a canvas, which is an element used to draw graphics via scripting is initialized for three types of bounding boxes representing: page regions, text regions, and picture regions. In this case, primary colors, i.e., red, green, blue are also used to represent image, text, and page regions respectively. Similar to the parallel coordinate plot described in the previous section, the commutative blending technique is used to handle the overdrawn regions in the superimposition plot. For example, as shown in Figure 5.15 a and Figure 5.15 b, the regions where the text and picture regions are overdrawn on each other are enhanced by a yellow color, which is a result of the blending of green (text regions) and red color (picture regions). Such a visualization technique can help to identify the physical structure of the complete document at a single glance.

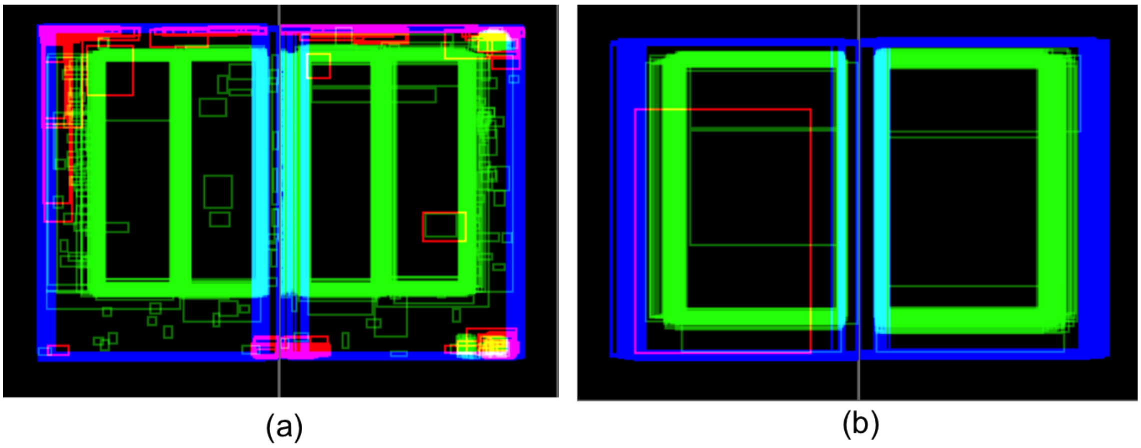


Figure 5.15 – Superimposition of all physical regions, i.e., page regions, text regions, and picture regions (a) two column layout (b) one column layout.

For example, in Figure 5.15 a, it can be seen that the document includes a two-column layout with some overlapping picture regions. Another example is shown in Figure 5.15 b, which shows the superimposition of the bounding boxes in the document with the one-column layout.

5.2.2.3 Document page montage plot

At the bottom right of Figure 5.7, all the document pages of the document selected from the document explorer view are juxtaposed to create a document page montage plot. The bounding boxes elements representing various physical regions of the document are also overlaid on top of each document page itself, as shown in Figure 5.16. Such a visualization has a potential to see the bounding box measurements or the size features of the physical regions directly reflected on the document page, and it can also identify any measurement errors generated by the automatic analysis approach.

5.2.2.4 Interaction techniques used in the document page explorer view

Various interaction techniques are employed within the three plots, such as "brushing + linking" and "focus + context". Brushing can be applied to any of the vertical axes in the parallel coordinate plot to highlight the selected line segments and fade out all the other line segments. It will also highlight the corresponding bounding boxes representing the physical regions in the superimposition plot and the document montage plot. For example, if the user brushes or selects the data points from any vertical axis, the corresponding lines in the parallel coordinate plot, the relevant bounding boxes in the superimposition plot, and the corresponding pages in document page montage get

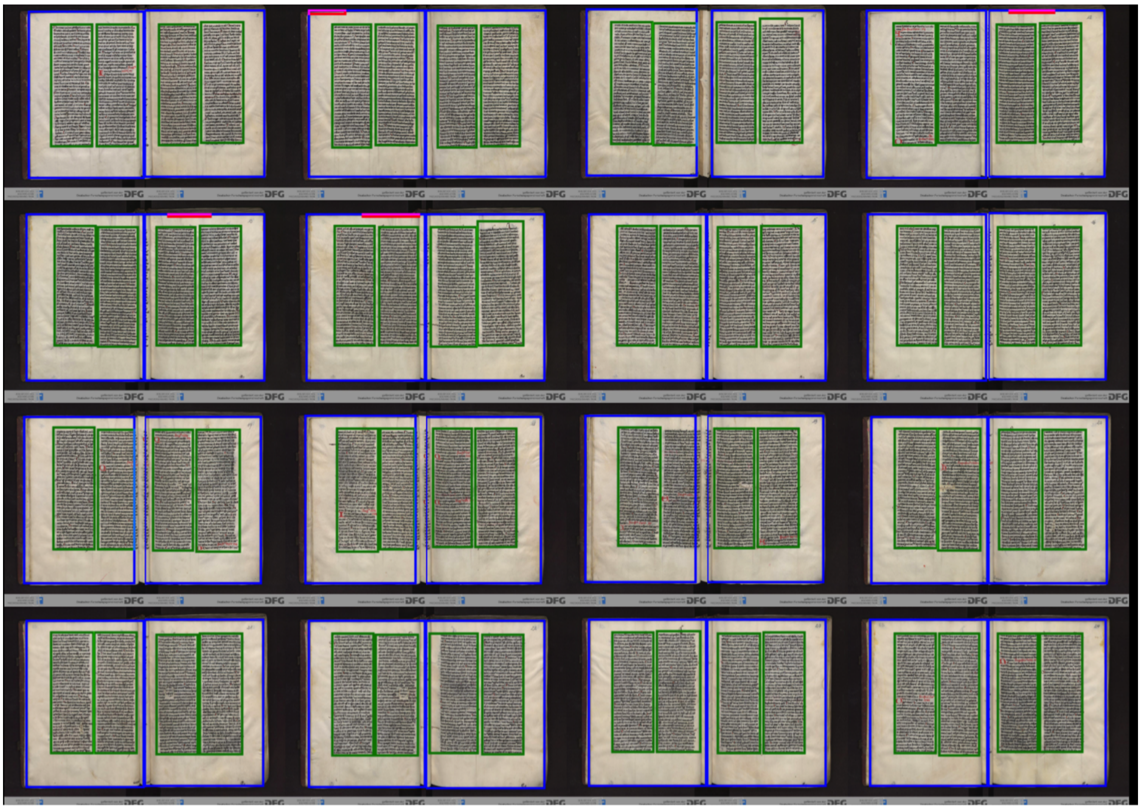


Figure 5.16 – Document page montage plot: In this view, all the document pages are combined into the composite view, and the size features extracted from the automatic layout analysis approach are plotted on top of each respective page.

highlighted (see Figure 5.17). Multiple brushing is also possible, where users can select multiple data values in the single vertical axis. Furthermore, to compare two or more different manuscripts, users can open the manuscript page explorer view in separate windows.

Fisheye distortion is also applied to magnify or focus the region of interest in the parallel coordinate plot while leaving the other part of the visualization unaffected. This distortion allows viewing small areas of the plot in greater detail.

5.2.3 Single document page explorer view

There exist some smaller physical regions that are also part of the document page represented as red-colored regions. These red-colored regions can be part of either the text regions or the picture regions. Therefore, these regions are displayed on the third level of the exploration.

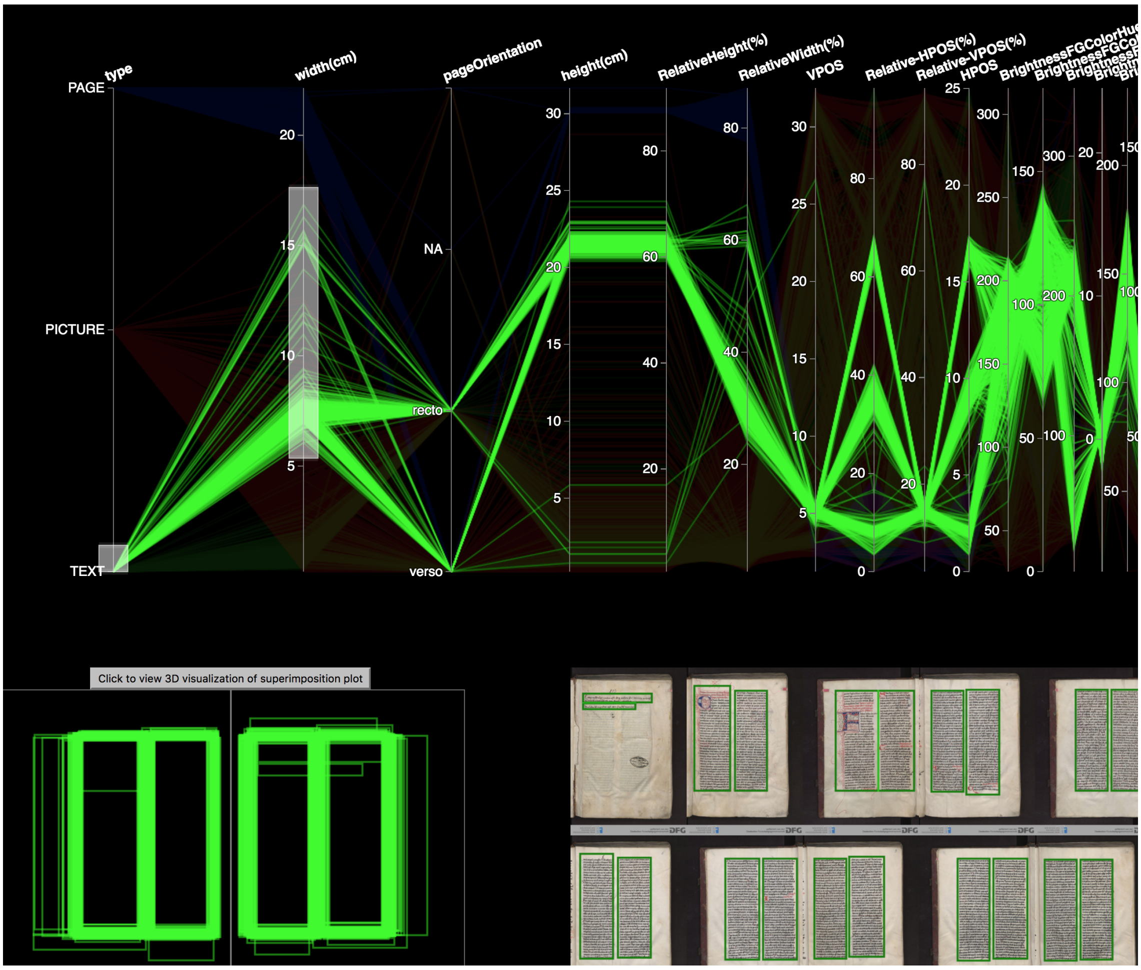


Figure 5.17 – Brushing and linking technique effect between the parallel coordinate plot, the superimposition plot, and the document page montage plot.

After selecting the document page from the document page montage plot present in the document page explorer view, a single document page is shown (see Figure 5.18). This shows all the physical regions, i.e. page, text, image, and red-colored regions overlaid on the document page itself. Such a view can help to see very smaller physical regions present in the document page with greater detail.

5.2.4 Preliminary qualitative user study

In order to validate and access how well the domain experts can use the CodiViz-II visualization interface to explore, and discover interesting findings and find any usability issues, a qualitative evaluation (laboratory study) was conducted. A qualitative user evaluation is preferred also in this case as compared to quantitative evaluation mainly

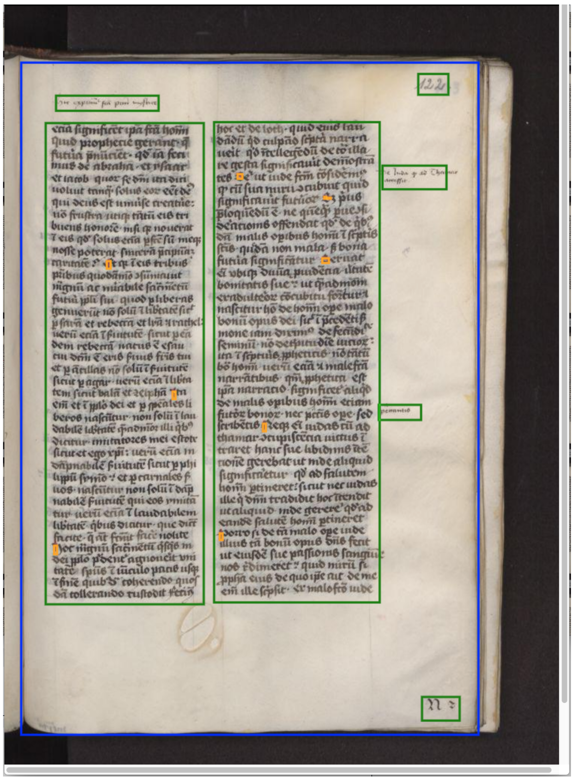


Figure 5.18 – Single document page explorer view.

again because of the sample size (number of users required to validate the visualization interface).

Participants and apparatus

For performing the user study, two domain experts were recruited from the author's professional network. These domain experts were research staff who had digital humanities background. The idea behind selecting the domain experts was to assess that despite having the limited technical expertise how well they could understand and explore the multidimensional layout features using CodiViz-II. The experiment was conducted on the domain experts' desktops, which had Windows 10 or macOS operating system and were connected to a 24-inch LCD and 27-inch, 5k retina display, respectively. Participants used a computer mouse and keyboard to interact with the prototype of CodiViz-II.

Procedure and user tasks

Each of the domain expert was called individually and at first all the features and levels of CodiViz-II were explained in detail and demonstrated to the users for approximately

10 minutes. After the demonstration, the participants were requested to perform a series of tasks belonging to the primary and the secondary user tasks. During this whole process, the participants were asked to perform each user task by thinking aloud. They could ask for detailed explanations when they faced any difficulty while completing the task. The whole study lasted for around 1.5 to 2 hr with each participant.

Primary user tasks

The primary user tasks (PT) are described as follows.

1. **Find distribution**:
   - **PT1.** What is the number of the historical documents written on parchment as compared to those written on paper?
   - **PT2.** Compare the number of the historical documents written on paper, parchment, or both by the century of production.
2. **Determine deviation**:
   - **PT3.** What is the approximate deviation of the text height/text width of the ninth page of the historical documents with the identifier "T0151" with respect to its mean text height/text width respectively ?
3. **Determine correlation**:
   - **PT4.** What is the correlation between the number of historical documents written on parchment as compared to those written on paper with respect to the century of production?
   - **PT5.** What is the correlation between the leaves count with respect to the century of production?
   - **PT6.** What is the correlation between the number of leaves count in a particular historical document written on parchment as compared to those written on paper with respect to the century of production?
   - **PT7.** What is the correlation between the binding format of the with respect to the century of the production?
4. **Get overview**:
   - **PT8.** Can you prepare an overview of the historical documents written on parchment and those written on paper in terms of the average page height and average page width?

5. **Determine range of values**:

    - **PT9.** What is the range of the text width of the historical document titled "T0283"?
    - **PT10.** What is the range of the century of production of the documents currently found in the City Library of Trier, Germany?

6. **Determine maximum and minimum**:

    - **PT11.** In which century of production were the most/least documents written?
    - **PT12.** Which historical document has the maximum number of pages?
    - **PT13.** What is the maximum brightness color value of the document page of the document with the identifier "S0042"?

7. **Find outliers**:

    - **PT14.** Are there any outliers in the historical documents written on paper with respect to those written after the 15th century?
    - **PT15.** Are there any outliers in the brightness color values of the historical documents written on parchment?

8. **Find clusters**:

    - **PT16.** Identify similar historical documents having approximately same leaves count and page height?
    - **PT17.** Identify the number of the historical documents currently located in three different libraries: the City Library of Trier, Abbey of Saint Matthias in Trier, and the Diocese's Archive of Trier.

9. **Extraction of important variables**:

    - **PT18.** In which century of production do the historical document have the largest page size?
    - **PT19.** What is the average text height of the historical document with the identifier "S0047"?

10. **Find errors**:

    - **PT20.** Are there any errors in the text width for the historical documents having largest leaves count?

- **PT21.** Observe detailed view of some of the historical documents. Do you see any error regarding automatically identified physical regions present in the historical documents?

Secondary user tasks

The secondary user tasks (ST) are described as follows.

1. **ST1**: Choose two different historical documents from the document explorer view and state three different facts about the documents.
2. **ST2**: Report any interesting finding, error, or outlier for the above-chosen documents.

5.2.5 Experimental results

General feedback

After the preliminary user study, general feedback was requested from the domain experts. Based on this input, it was indicated that CodiViz-II was easy to learn and that the design of CodiViz-II is aesthetic. The participants felt comfortable with the key components of CodiViz-II, such as the superimposition plot, the document montage plot, and the detailed views between the document explorer view and the document page explorer view. The participants were able to select the document based on its bibliographic features in the first level, and in the second level, the participant were able to get into the details of the layout features. The other participants suggested that "*adding a search textbox in the document explorer would be helpful to find the document directly instead of hovering over each document separately*." They also suggested adding the "*possibility to filter out the vertical axis from the parallel coordinate plot that is not required for the analysis would be helpful, as it would reduce overall cluttering and give more information to the users*." In the end, the participants desired to use the CodiViz-II for further exploration, which was very encouraging.

Observations

Primary user tasks

Both the participants were able to complete each primary user task in between 1.5 to 2 minutes. Some of the complex tasks, such as PT8, PT13, PT14, PT15, PT16, and PT21 took more than 2 minutes to solve because in these tasks domain experts required some assistance with the interactions. Figure 5.19 shows the average time taken by both the domain experts to address the primary tasks.

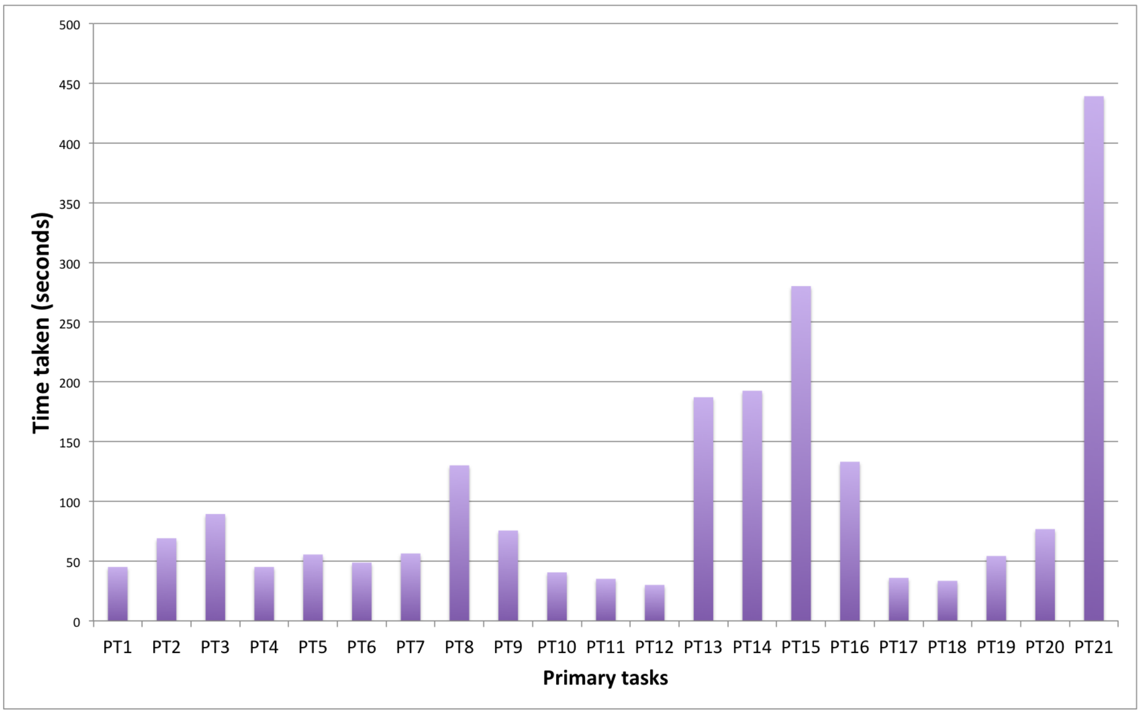


Figure 5.19 – Time taken by domain experts to solve primary user tasks.

Find distribution For the PT1, both the participants used the parallel coordinate plot in the document explorer view and reported that handwritten documents written on parchment are more in number in comparison to those written on paper. However, they were not able to tell the exact number because CodiViz-II did not have the counting measure. Participants suggested that it would be beneficial for them if they could get the accurate count. Similarly, for PT2, they used the *century* and the *material* axes from the parallel coordinate plot and reported that the maximum number of historical documents were written in 15th century.

Determine deviation In order to solve PT3, both the participants opened the document page explorer view of the document with the identifier "T0151" and used brushing and linking to estimate the mean text width and height. After that, they were able to report that how much was the text width of the ninth page was deviating from the mean text width.

Determine correlation To answer PT6, one participant found the document written on parchment with more than 500 pages and made a direct hypothesis that this document was either written over the course of different centuries or multiple document pages were combined to form one document. Another participant observed as a result of

PT6 that the number of leaves count for the documents which were written on parchment ranged between 100 and 300 while for the documents which were written on paper the leaves count was more than 300 pages. The participant reported that this could be true because the parchment documents were costly and documents with a large number of leaves were not easily affordable.

One of the participants also reported as a result of PT7 that the three most common binding formats for St. Matthias dataset were 2°, 4°, and 8°.

Get overview To get the overview, both participants reordered the page height, page width, and material vertical axes of the parallel coordinate plot and placed them parallel to each other. They observed that page height and page width were varying a lot for the handwritten documents written on parchment as compared to the documents which were written on paper.

Determine range of values The participant reported for PT10 that the City library of Trier has the maximum number of documents as compared to other two libraries. The City library of Trier contains the manuscripts from the 15th till the 18th century but most of them are from the 14th century.

Determine maximum and minimum As a result of PT11, participants reported that the maximum number of documents were written during the 15th century because paper was introduced into Europe at that time.

Participants spent much more time on the task PT13 and had requested assistance to solve PT13 because solving this question involved focus+context interaction (fisheye distortion). One prevalent issue where participants got frustrated was when they had to find the vertical axis from the document page explorer view. They found it difficult to find the right vertical axis, but after some assistance and after being told that the vertical axes were grouped according to the category of layout features, i.e., first size features, second color features, and third count and other derived features, they were able to find the required axes. But as an improvement, they suggested adding the possibility to remove the axis which is not necessary for future developments.

Find outliers One of participant discovered an outlier in the historical documents which were written on paper. However, they spent much more time to solve and asked for assistance again to solve PT15 as they were confused in how to choose the vertical axes which represented the brightness color value for the document. After some initial help, they were able to find the outliers in the brightness color values of the historical document which were written on parchment but had few document pages made of paper in them.

Secondary user tasks

After performing the primary tasks, the participants got familiar with the CodiViz-II design, and they were able to explore much more deeply while performing the secondary user tasks. Each of the two secondary user tasks took around 15-20 minutes to execute. As a part of the task, the participants were requested to select any two arbitrary documents from the document explorer view and identify three different facts about those documents. Most of the participants told these three facts from the document explorer view itself, and the facts were related to the bibliographical features of the document. They also found some exciting things in the detailed document page explorer view which they also confirmed from the catalog of the document described on the "Virtual Scriptorium Saint Matthias" website. For instance, they found that for most of the documents which were written on parchment the binding was redone after 15th century and these documents have a paper binding in the first and last two pages of the document. The paper binding was much brighter than the parchment and was easily spotted in the parallel coordinate plot and document page montage plot where the brightness value was more for the first and last two pages as compared to other pages present in the document which were written on parchment.

Another outlier which was explored was that one document which was written on parchment and had the binding format of 8° had the page height and page width of more than 35 and 25 cm respectively. The participant instantly reported that this was indeed an outlier because parchment books with 8° binding format cannot be of such a large size. The participant concluded that either the parchment was made from the skin of a calf or the document has been assigned a wrong binding format. More detailed case-studies are explained in Section 5.5.

5.3 Visual exploration results

The prototype of CodiViz-II has seen an early use in answering various domain-specific questions. This section describes multiple case-studies where CodiViz-II has helped domain experts to gain insights into their data by solving primary and secondary user tasks as defined in the previous section. They also detected various measurement uncertainties which helped them to make informed decisions. However, as the domain experts are still in the early stages of analysis, many other cases are yet to be identified. However, some of the exploration results are described below:

1. **Exploration of hidden relationships**: While solving various primary and secondary user tasks, participants discovered some interesting findings in their data. Some of these interesting findings are listed on the next page:

- The domain experts discovered the presence of a fixed cluster of page height, page width, and binding format for the documents written on paper between 14th and 18th century as shown in Figure 5.20. This provided an insight of paper production techniques or size of the frames used in historical times to produce paper. With this domain experts also discovered a case as shown in Figure 5.20 which indicated that the document is written on paper and is written during the 9th century. This is an outlier because the document written on paper during the 9th were not found in Europe [22]. The domain experts concluded that either the information is wrongly entered in the bibliographic features or the document under question was not from Europe.

- The domain experts could easily distinguish between the hair side of the parchment as compared to the flesh side of the parchment based on clustering of the brightness color value in the parallel coordinate plot in the document page explorer view. For example, Figure 5.21 shows the clustering of brightness color value of the document pages and Figure 5.22 shows the document page in the single document page explorer view clearly distinguishing between the hair side and the flesh side.

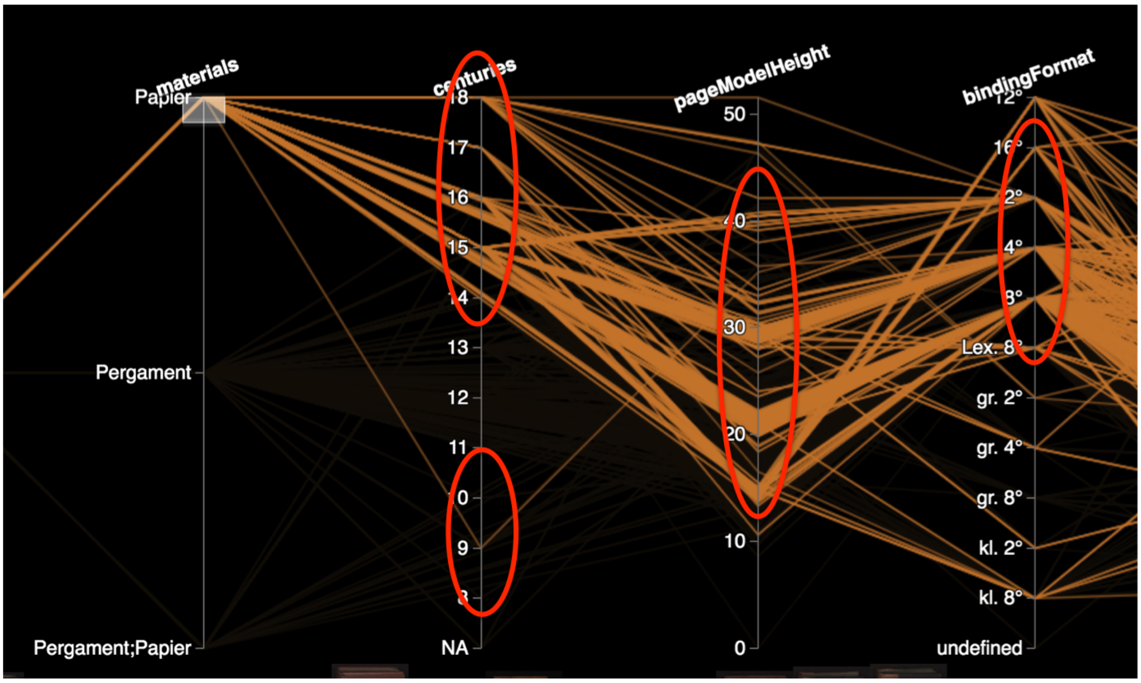


Figure 5.20 – Exploration of page width, page height, and binding format for the document written on paper after 14th century.

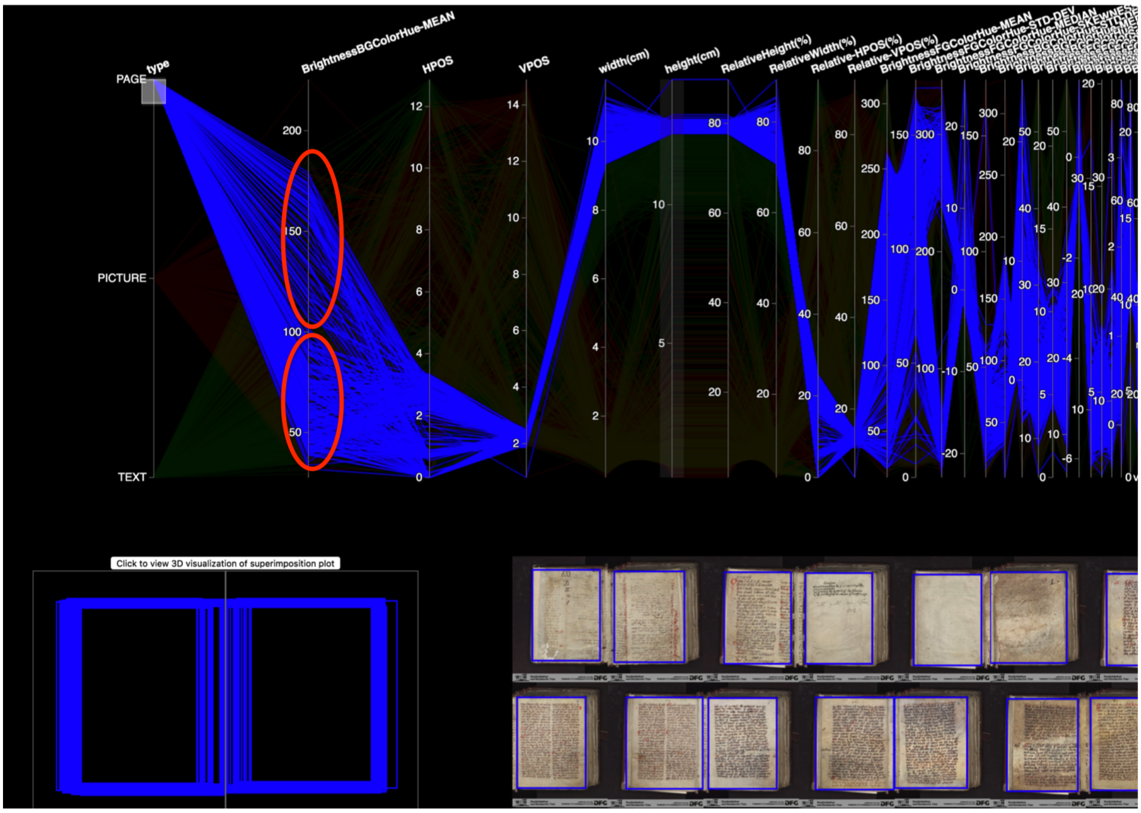
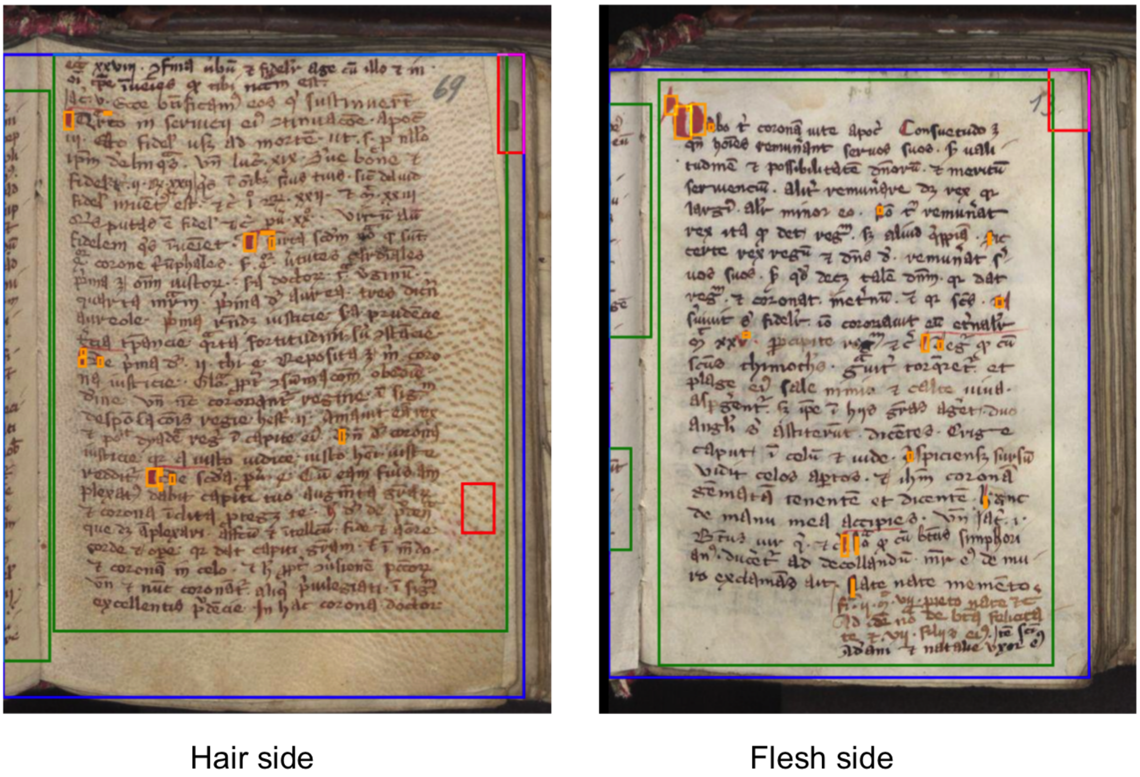

type
PAGE
PICTURE
TEXT
BrightnessBGColorHue-MEAN
HPOS
VPOS
width(cm)
height(cm)
RelativeHeight(%)
RelativeWidth(%)
Relative-HPOS(%)
Relative-VPOS(%)
BrightnessFGColorHue-MEAN
BrightnessFGColorHue-STD-DEV
BrightnessFGColorHue-MEDIAN
Click to view 3D visualization of superimposition plot

Figure 5.21 – Exploration of clustering of the brightness color value in the parallel coordinate plot.

- The historical documents which were written before the 12th century consisted of single columns as shown in Figure 5.23 whereas after 12th century the physical layout moved to one or two columns as shown in Figure 5.24. This indicated usage of a ruling pattern followed after the 12th century.

2. **Determination of measurement uncertainty**: Extraction of various layout features, such as size features, color features, and count features automatically raised a question about whether these measurements were telling the truth or not. These values were uncertain and required a quality check to prove their correctness. For example, Figure 5.25 shows the superimposition plot of only text regions where all the text regions have varying text height and text width. Such a plot shows the presence of measurement uncertainty in the data which is due to the irregular nature of the superimposition plot.

   Therefore, to get certain or sure about the measurements of layout features, domain experts visually compared the parallel coordinate plot, superimposition plot, and document page montage plot to perform a quality check where bounding boxes representing the physical regions were superimposed on each other and on the document images itself. This visual comparison allowed to verify whether the values of the extracted size features were correctly measured or whether there were any anomalies in the automatically extracted layout features. The combination of superimposition plot and the document page montage plot enabled domain experts to identify the document pages which had anomalies and need to be reprocessed by different set of parameters or algorithms.


Hair side

Flesh side

Figure 5.22 – Exploration of document written on hair side and flesh side of the parchment.

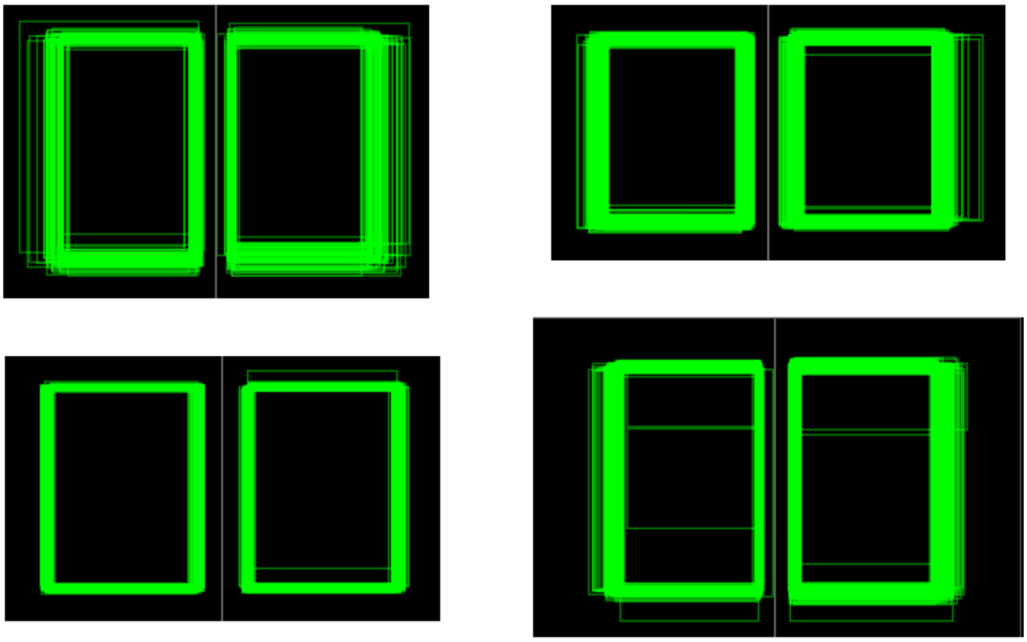


Figure 5.23 – Exploration of document written before 12th century consisting of single columns.

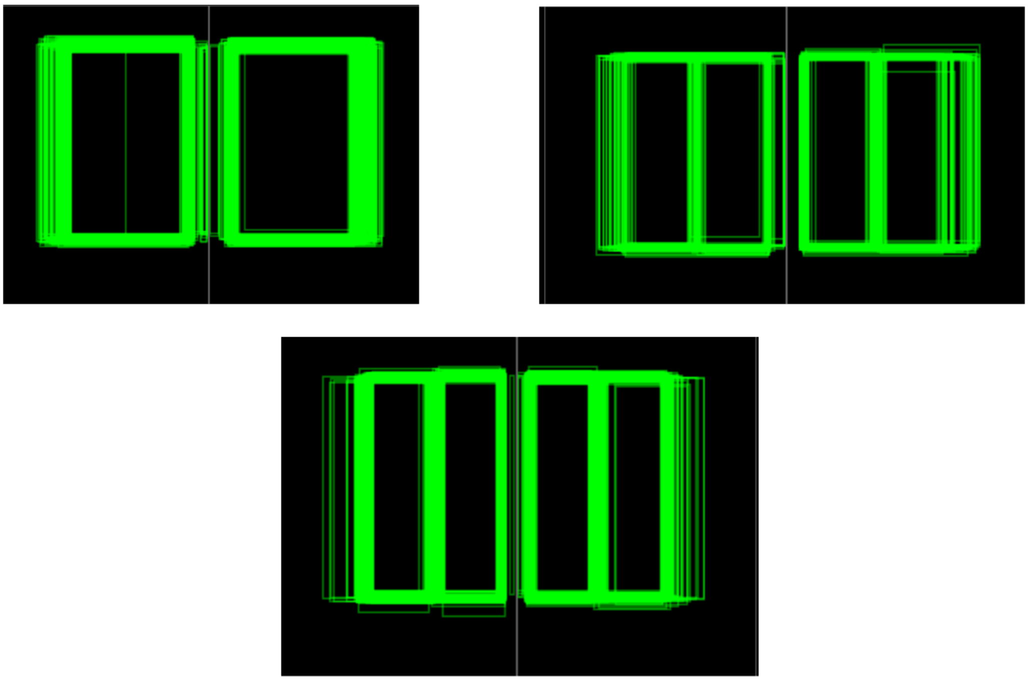


Figure 5.24 – Exploration of document written after 12th century consisting of two columns.

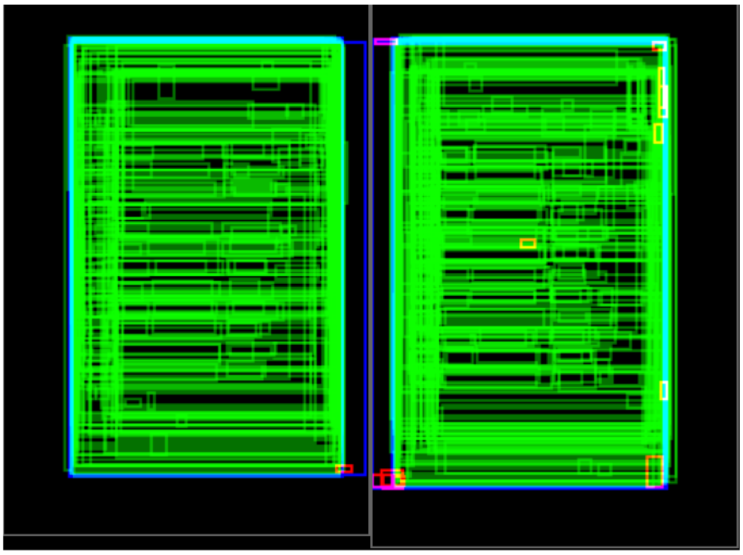


Figure 5.25 – Exploration of the superimposition plot of only text regions where all the text regions have irregular text height and text width.

5.4 Summary

This chapter describes the proposed designs and implementation of the visualization interfaces assisting the domain experts to find interesting relationships in their data. The interface has approached the challenge of enabling the exploration possibility in the multidimensional datasets in the digital humanities domain. The CodiViz-II visualization interface allows domain experts to find clusters, correlations, outliers, errors in their data by using flexible browsing capabilities. A preliminary qualitative evaluation with domain experts has shown the usefulness of CodiViz-II in exploring complex layout features. It is worth noting that this design is not only limited to the handwritten documents present in the St. Matthias dataset but can be easily adapted for other handwritten and printed documents present in other digital libraries.

Chapter 6

Discussion and conclusion

This thesis answers the research questions mentioned in chapter one which focus on the identification of the physical regions from the historical handwritten documents, and the visual exploration of these identified physical regions to gain better insights. This thesis is part of the research project eCodicology which started in 2013 and was funded by Federal Ministry of Education and Research (BMBF). It was initiated with the aim of developing, testing, and optimizing new algorithms for the identification and exploration of the physical regions embedded in the historical documents. It should enable the humanities researchers to answer the research questions such as connections of dislocated manuscripts, and writing proportions, which medieval artisans followed to write the historical documents, etc. A short summary of this thesis is given below.

6.1 Summary

In this thesis, a generic, extensible, and a fully automated approach for identification of the physical regions and extraction of their corresponding layout features was presented for handwritten as well as printed documents with overlapping layout. It utilizes methods such as preprocessing, region segmentation, and feature extraction.

Preprocessing involves the processing steps that are needed to make the acquired image less dependent on the data acquisition hardware and more suitable for the automatic document layout analysis. As the handwritten historical document images were scanned with different scanners they have different representations of the same color. This made the digitized historical document images very much dependent on the data acquisition hardware, i.e., scanner hardware. Therefore, a color calibration process was required to eliminate this hardware dependency by calibrating the colors of the document images to color representations which are similar to each other. A linear transformation model was used for this color calibration. Other preprocessing steps such as spatial conversion were performed to correlate the pixels of the document image to real-world units. Also, resizing to smaller scale was applied to decrease the processing time of the document images. Region segmentation was applied to divide the whole document image into the constituent physical regions. This process was one of the most challenging tasks due to various factors such as non-uniform background variations in the document images which makes it difficult to distinguish between the background and the foreground regions.

The segmentation of the physical regions enclosed in the handwritten historical document images was performed by utilizing a projection profile and a machine learning approach. The projection profile approach is mainly used for segmentation of page regions. The machine learning approach was applied for segmentation of text and picture regions using the Trainable Weka Segmentation (TWS) plugin. This plugin uses the machine learning process provided by WEKA (Waikato Environment for Knowledge Analysis). It includes various state-of-the-art machine learning algorithms which could be easily compared and trained. The TWS plugin allows to efficiently train the machine learning algorithms and directly evaluate the results of segmentation. The evaluation of various machine learning algorithms showed that the Random Forest Classifier outperformed the rest of the machine learning algorithms such as the NaiveBayes or Support Vector Machines algorithm. The HSB (Hue, Saturation, Brightness) color model was used to segment physical regions which were written with red-colored ink as text color is an important distinction for the humanities scholars.

Relevant quantitative information describing the physical characteristics of the segmented physical regions was extracted. This quantitative information included a) size features such as bounding box measurements of the physical regions, and area measurements, b) color features such as image statistics for hue, saturation, and brightness channel, c) image-specific features like the number of text regions, number of picture regions enclosed in the historical document image, and d) other derived features such as the text region to left margin region ratio and text region to right margin region ratio. These layout features contain further information about each of the physical regions and act as the differentiators for each of the document images.

The application of feature extraction resulted in extraction of approximately 162,800,000 features for 150,000 handwritten historical document images. These results were stored in an XML file corresponding to the standard PAGE format proposed in 2017.

This thesis also presents a generic design strategy to aid visualization designers to choose a visualization technique suitable for a particular kind of domain. It involved identification of the problem and understanding of the data to extract the first set of requirements from the domain experts. Based on these requirements, the advantages and the disadvantages of various information visualization techniques were evaluated on the basis of visualization and analysis tasks, type of data and number of records and dimensions which were difficult to define because each of the visualization techniques was subjected to display screen size and individual perception skills. However, various interaction techniques such as selecting, zooming, distorting, and filtering can be applied to each of the visualization technique to overcome the limitation of scalability and dimensionality in large datasets.

As a result of applying this generic design strategy on the real-world humanities data, it was found that the parallel coordinate plot and the radial tree visualization techniques could be applied to historical documents because these visualization techniques are capable of detecting clusters, outliers, and trends for qualitative and quantitative data. These visualization techniques and other visual composition methods such as juxtaposition and superimposition were used to create a visual exploration of multidimensional extracted layout features in two iterations. The second iteration resulted in the final design of the visualization interface which made use of the parallel coordinate plot and the visual composition methods, i.e., juxtaposition and superimposition. A user study with the domain experts showed that the second iteration was capable of providing an overview, detecting clusters, outliers, and anomalies, or retrieving a value.

The second iteration is an interactive multi-level visualization interface. On the first level, a parallel coordinate plot is combined with a document montage plot to provide an overview of the entire document collection and employs a multiple coordinated view to show the details of various aspects such as the century of the production and the materials. Users can interact with the document montage plot to navigate to the second level of the exploration hierarchy. On the second level, a parallel coordinate plot is fused together with a superimposition plot and a document layout montage plot to form a multiple coordinated view which is used to show the details of multidimensional layout features extracted from automatic methods such as size and color features. Commutative blending modes are applied to the superimposition and parallel coordinate plot to resolve the occlusion problem, which occurs because of the overlap, or overdraw of two or more data points. Furthermore, various state-of-the-art methods such as edge bundling, smoothing, and fisheye distortion were also utilized to optimize the parallel coordinate plot. This design allows users to explore, compare, and detect clusters and outliers in the multidimensional information to gain knowledge. And, at the third level, each of the document images is shown in detail.

6.2 Discussion and future directions

Physical layout analysis

The first research question in chapter one is **"How to identify the physical regions from the heterogeneous and irregular handwritten historical document images, and also how to extract their corresponding layout features?"**

To answer this research question a layout analysis approach to identify the physical regions from handwritten document images having overlapping layout is presented in this thesis. This approach was designed with generic abstractions, which allows applying the layout analysis to the vast variety of handwritten as well as the printed document having an overlapping physical layout. Also, it is entirely automated which means it can apply to a large number of document images without any user intervention and can produce reproducible and deterministic results. The extracted features for each of the physical region act as the differentiators for the document images to assist domain experts in answering their research questions. The precise information about the physical regions and layout features are stored in the widely accepted PAGE format so that results can be shared and reused by other research communities. Furthermore, the accuracy tests on the sample dataset showed that the automatic and manual analysis is comparable.

The results presented in this thesis were used in the context of research project eCodicology which also gave a chance for further collaboration. For example, the automatic layout analysis approach was integrated with the interactive manual annotation tool called Semantic Topological Notes (SemToNotes). This tool was developed in the context of DARIAH-DE to offer the possibility of not only analyzing the results of automatic layout analysis but also providing the opportunity to deal with the semantics of the physical regions. It allowed annotating the physical regions by using humanities terminology. As a part of its additional collaborative contributions, the eCodicology research project was also approached to apply its results to printed documents, e.g., early modern printed Spanish magazines. Currently, the methods developed in the context eCodicology are being used and enhanced for the Aristoteles documents in the context of the project of collaborative research center 980 "Episteme in Motion - transfer of knowledge from the ancient world" [17]. It is also being used for the research project OCR-D [151] which deals with the advancement of Optical Character Recognition (OCR) techniques for the historical prints.

However, in the current state of implementation, the automatic layout approach is limited to each layout region as a circumscribed bounding box. This could be further extended to a polygon level which would provide more detailed analysis. Furthermore, this approach generates metadata which is currently being modeled in the standard PAGE format. But, if any of the region segmentation or feature extraction is replaced by another improved version of region segmentation or feature extraction, the metadata in the PAGE format is also subjected to change. This evolving metadata is termed dynamic metadata [152]. In future, the domain experts want to add and share further descriptions about the physical regions in the form of annotations. Therefore, a flexible data model called Web Annotation Data Model (WADM) [153] is required in order to analyze, query as well as modify the annotations to add an interoperability feature as one of the significant contributions of CRC-980.

Additionally, the layout analysis approach can also be divided into individual web services such as preprocessing, region segmentation, or feature extraction service. Afterwards, each of these web services can be orchestrated with the help of workflow management system which would also contribute shareability and reusability. Additionally, provenance for each of the execution can also be captured which will enable domain experts to compare the results of multiple versions of individual web services [154].

Visual exploration

The second research question of chapter one is **"How to design and to enable interactive visual exploration of multidimensional datasets with real-world applications in digital humanities?"**

To answer this research question, a generic design strategy to identify the information visualization technique which could be applied to a particular domain is designed. Such a design strategy can assist visualization designers who are working in interdisciplinary research projects to identify the problem of domain experts and choose a list of possible information visualization techniques which can be applied to a particular domain. Furthermore, to enable visual exploration of multidimensional datasets a multi-level visualization interface is proposed for the real-world data of documents. This interface represents the multidimensional information of layout features extracted from the physical layout analysis process. It enables the domain experts to get an overview of their data and dive into the selected subset of data to further explore the region of interest. It uses interaction techniques such as "brushing + linking", "focus + context" to explore the clusters, outliers, and anomalies in their dataset.

The first level of the visualization interface combines a parallel coordinate plot with a document montage plot in a multiple coordinated view. This level can enable domain experts to get an overview of their data and explore interesting patterns between the dimensions which are independent of each other. By taking into account the benefits of the parallel coordinate plot, the line segments which represent the multidimensional data make it easier to identify the clusters or patterns in the dataset. The interaction techniques such as axis reordering overcome the limitation of overlaying of lines for same data values. The document montage plot is useful to gain a physical overview of all the documents easily. The interaction technique, i.e., brushing and linking combines the parallel coordinate plot and document montage plot. This enables the domain experts to explore the subset of documents in detail by selecting the desired documents with the help of the parallel coordinate plot and thus giving them the feel of a virtual digital library.

The second level of design enables the domain experts to focus on different parts of the data without losing an overview of the dataset. It combines the parallel coordinate plot, the superimposition plot, and the document page montage plot to explore different parts of data in detail. Here, the parallel coordinate plot assists the domain experts to get an overview of all the layout features which are extracted from the layout analysis approach. Interaction techniques such as "focus + context" and axis reordering in the parallel coordinate plot can help to check out the correlations in two different axes. The superimposition plot enables the domain experts to see the physical structure of the complete document with a bird's eye view. The document page montage plot can enable to explore the sequences of documents themselves in a rapid succession.This plot can help to quickly find errors which occur during physical layout analysis process. Each component of this second level of visualization interface complements each other.

The third level of visualization interface uses the document itself to help the domain experts to explore the layout features of each document individually. A preliminary user study with the domain experts and the first exploration results showed that the visualization interface proposed in the thesis is capable of providing an overview, detecting clusters, outliers, and anomalies, or retrieving a value and is also capable of answering some of their research questions. However, the validation of the visualization interface presented in this thesis is a preliminary user evaluation. Thus it cannot be stated that this interface is capable of providing answers to all the questions that the domain experts may come up with as a result of the presented visualization interface. In order to answer more complex problems, the proposed visualization interface will have to be adapted and validated according to new requirements of domain experts. For example, it can be extended by adding more intelligence to the visualization interface where the interface itself makes suggestions for the domain experts to dive in the data. These suggestions can be learned by recording the history of interactions performed by the domain expert.

Also, the visualization interface proposed in this thesis can be considered as the first generation multi-level visualization interface for exploration of physical layouts of the documents. This may have some usability issues which were already discovered using the user evaluation. Hence, it is strongly recommended to the visualization designers to consider the usability aspects before conducting any new design study.

6.3 Conclusion

As the complexity and the amount of the data continue to increase, the need for developing efficient computational methods helping researchers to explore and analyze their data also increases. The research in this thesis is defined within the scope of an interdisciplinary research domain, i.e., digital humanities, document layout analysis, and information visualization. If the bond between these research domains will become stronger in the future, more intelligent layout analysis and visualization systems will arise.

This thesis has investigated a physical layout analysis approach for identifying the physical regions and extracting their corresponding layout features. It has also investigated a multi-level interactive visualization interface with applications in digital humanities, which facilitates the domain experts to actively engage themselves in the exploration process by detection of correlations, outliers, clusters, and so forth.

As a result of this research work, it is now possible to identify physical regions and extract layout features of heterogeneous and irregular handwritten historical documents precisely and automatically in a reproducible and deterministic manner. Furthermore, this research work has enabled domain experts in the field of digital humanities to explore the identified physical regions, and their corresponding layout features more engagingly to gain better insights and discover knowledge in their data. The work presented in this thesis may act a gateway to new scientific research, tools, and possibilities.

Appendix A

Page model

The following shows the XML schema used for defining the page model width, page model height, and the input RGB values of the color patches found on the color calibration chart models used for color calibration and page region segmentation.

Listing A.1 – XML Schema for representing page model

```
<?xml version="1.0" encoding="UTF-8"?>
  <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
     elementFormDefault="qualified" attributeFormDefault="
     unqualified">
       <xs:element name="pageModel">
            <xs:complexType>
                 <xs:sequence>
                      <xs:element name="id" type="xs:string"><
                         /xs:element>
                      <xs:element name="width" type="xs:double
                         "></xs:element>
                      <xs:element name="height" type="
                         xs:double"></xs:element>
                      <xs:element name="unit" type="xs:string"
                         ></xs:element>
                      <xs:element name="resolution">
                           <xs:complexType>
                                <xs:sequence>
                                     <xs:element name="
                                        xResolution" type="
                                        xs:int"></
                                        xs:element>
                                     <xs:element name="
                                        yResolution" type="
```

```
                                       xs:int"></
                                       xs:element>
                        </xs:sequence>
                    </xs:complexType>
                </xs:element>
                <xs:element name="inputColors" maxOccurs
                   ="unbounded">
                    <xs:complexType>
                        <xs:sequence>
                            <xs:element name="blue
                               " type="xs:double">
                               </xs:element>
                            <xs:element name="
                               colorType" type="
                               xs:string"></
                               xs:element>
                            <xs:element name="
                               green" type="
                               xs:double"></
                               xs:element>
                            <xs:element name="red"
                                type="xs:double"><
                               /xs:element>
                        </xs:sequence>
                    </xs:complexType>
                </xs:element>
            </xs:sequence>
        </xs:complexType>
    </xs:element>
</xs:schema>
```

Appendix B

Color model

The following shows the XML schema used for target RGB values used for the process of color calibration.

Listing B.1 – XML Schema for representing page model

```
<?xml version="1.0" encoding="UTF–8"?>
  <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    elementFormDefault="qualified" attributeFormDefault="
    unqualified">
      <xs:element name="colorModel">
          <xs:complexType>
              <xs:sequence>
                  <xs:element name="targetColors"
                    maxOccurs="unbounded">
                      <xs:complexType>
                          <xs:sequence>
                              <xs:element name="
                                colorType" type="
                                xs:string"></
                                xs:element>
                              <xs:element name="red"
                                 type="xs:int"></
                                xs:element>
                              <xs:element name="
                                green" type="xs:int
                                "></xs:element>
                              <xs:element name="blue
                                " type="xs:int"></
                                xs:element>
                          </xs:sequence>
                      </xs:complexType>
                  </xs:element>
              </xs:sequence>
          </xs:complexType>
      </xs:element>
  </xs:schema>
```

List of Figures

List of Tables

Bibliography

[1] D. Tonne, R. Stotzka, T. Jejkal, V. Hartmann, H. Pasic, A. Rapp, P. Vanscheidt, B. Neumair, A. Streit, A. Garcia, *et al.*, "A federated data zone for the arts and humanities," in *Proceedings of 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pp. 198–205, IEEE, 2012.

[2] D. Humanities, "What are Digital Humanities." `https://www.uni-trier.de/index.php?id=52422&L=2`. Accessed: 2017-07-30.

[3] D. M. Berry, "Introduction: Understanding the digital humanities," in *Understanding digital humanities*, pp. 1–20, Springer, 2012.

[4] P. Svensson, "The landscape of digital humanities," *Digital Humanities*, 2010.

[5] "DARIAH-EU." `https://www.dariah.eu/`. Accessed: 2017-11-23.

[6] "ESFRI." `https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri`. Accessed: 2017-11-23.

[7] "DARIAH-DE." `https://de.dariah.eu/`. Accessed: 2017-11-23.

[8] M. Mehri, *Historical document image analysis: a structural approach based on texture*. PhD thesis, Université de La Rochelle, 2015.

[9] J. Tukey, *Exploratory Data Analysis,*. Addison-Wesley Series in Behavioral Science: Quantitative Methods, Addison-Wesley, 1977.

[10] A. M. Namboodiri and A. K. Jain, "Document structure and layout analysis," in *Digital Document Processing*, pp. 29–48, Springer, 2007.

[11] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999. ISBN - 13: 978-1-4665-0893-4.

[12] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[13] C. Ware, *Information Visualization: Perception for Design*. Information Visualization: Perception for Design, Morgan Kaufmann, 2013.

[14] C. Tominski, *Event based visualization for user centered visual analysis.* PhD thesis, University of Rostock, 2006.

[15] S. Chandna, D. Tonne, T. Jejkal, R. Stotzka, C. Krause, P. Vanscheidt, H. Busch, and A. Prabhune, "Software workflow for the automatic tagging of medieval manuscript images (SWATI).," in *In Proceedings of SPIE, Document Recognition and Retrieval XXII*, vol. 9402, p. 940206, 2015.

[16] F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[17] "CRC-980 Episteme in Motion." `http://www.sfb-episteme.de/en/`. Accessed: 2017-11-23.

[18] N. Rißler-Pipka, S. Chandna, and D. Tonne, "Automatische bild-textanalyze: Chancen für die zeitschriftenforschung jenseits von reinen textdaten," in *Proceedings of Digital Humanities im Deutschsprachigen Raum*, pp. 94–99, DHd, 2017.

[19] S. Chandna, D. Tonne, R. Stotzka, H. Busch, P. Vanscheidt, and C. Krause, "An effective visualization technique for determining co-relations in high-dimensional medieval manuscripts data," in *In Proceedings of Electronic Imaging*, vol. 2016, pp. 1–6, Society for Imaging Science and Technology, 2016.

[20] S. Chandna, F. Rindone, C. Dachsbacher, and R. Stotzka, "Quantitative exploration of large medieval manuscripts data for the codicological research," in *Large Data Analysis and Visualization (LDAV), 2016 IEEE 6th Symposium on*, pp. 20–28, IEEE, 2016.

[21] R. Gonzalez and R. Woods, *Digital Image Processing*. Pearson/Prentice Hall, 2008.

[22] K. Kise, "Page segmentation techniques in document analysis," in *Handbook of Document Image Processing and Recognition*, pp. 135–175, Springer, 2014.

[23] G. Nagy and S. Seth, "Hierarchical representation of optically scanned documents," in *Proceedings of Seventh International Conference on Pattern Recognition (ICPR)*, vol. 7, pp. 347–349, Digital Commons, 1984.

[24] H. Baird, S. Jones, and S. Fortune, "Image segmentation by shape-directed covers," in *Proceedings of tenth International Conference on Pattern Recognition*, vol. I, pp. 820 – 825 vol.1, IEEE, 1990.

[25] K.-C. Fan, C.-H. Liu, and Y.-K. Wang, "Segmentation and classification of mixed text/graphics/image documents," *Pattern Recognition Letters*, vol. 15, no. 12, pp. 1201–1209, 1994.

[26] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1162–1173, 1993.

[27] K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area voronoi diagram," *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 370–382, 1998.

[28] F. Shafait, D. Keysers, and T. Breuel, "Performance evaluation and benchmarking of six-page segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 941–954, 2008.

[29] F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block segmentation and text extraction in mixed text/image documents," *Computer graphics and image processing*, vol. 20, no. 4, pp. 375–390, 1982.

[30] H.-M. Sun, "Page segmentation for manhattan and non-manhattan layout documents via selective crla," in *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, (Washington, DC, USA), pp. 116–120, IEEE Computer Society, 2005.

[31] M. Agrawal and D. S. Doermann, "Voronoi++: A dynamic page segmentation approach based on voronoi and docstrum features," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, pp. 1011–1015, IEEE, 2009.

[32] S. S. Bukhari, A. Azawi, M. I. Ali, F. Shafait, and T. M. Breuel, "Document image segmentation using discriminative learning over connected components," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 183–190, ACM, 2010.

[33] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana, "Layout analysis for arabic historical document images using machine learning," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 639–644, IEEE, 2012.

[34] R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, "Robust text and drawing segmentation algorithm for historical documents," in *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp. 110–117, ACM, 2013.

[35] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page segmentation of historical document images with convolutional autoencoders," in *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, (Washington, DC, USA), pp. 1011–1015, IEEE Computer Society, 2015.

[36] D. Jurafsky and J. H. Martin, *Speech and language processing*, vol. 3. Pearson London:, 2014.

[37] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: a literature survey," in *Proceedings of Document Recognition and Retrieval X*, vol. 5010, pp. 197–208, International Society for Optics and Photonics, 2003.

[38] T. A. Tran, K. Oh, I.-S. Na, G.-S. Lee, H.-J. Yang, and S.-H. Kim, "A robust system for document layout analysis using multilevel homogeneity structure," *Expert Systems with Applications*, vol. 85, pp. 99–113, 2017.

[39] F. Le Bourgeois and H. Emptoz, "Debora: Digital access to books of the renaissance," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 193–221, 2007.

[40] F. Le Bourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz, "Document images analysis solutions for digital libraries," in *Proceedings of first International Workshop on Document Image Analysis for Libraries*, pp. 2–24, 2004.

[41] Xenetix, "DocuWorks." `http://xenetix.com.sg/docuworks/`. Accessed: 2017-10-30.

[42] J.-Y. Ramel, S. Busson, and M.-L. Demonet, "Agora: the interactive document image analysis tool of the bvh project," in *Proceedings of Second International Conference on Document Image Analysis for Libraries*, pp. 11–pp, IEEE, 2006.

[43] J.-Y. Ramel, S. Leriche, M. Demonet, and S. Busson, "User-driven page layout analysis of historical printed books," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 243–261, 2007.

[44] B. Couasnon and I. Leplumey, "A generic recognition system for making archives documents accessible to public," in *Proceedings of Seventh International Conference on Document Analysis and Recognition*, pp. 228–232 vol.1, 2003.

[45] A. Gordo, D. Llorens, A. Marzal, F. Prat, and J. M. Vilar, "State: A multimodal assisted text-transcription system for ancient documents," in *Proceedings of the eighth IAPR International Workshop on Document Analysis Systems*, pp. 135–142, IEEE, 2008.

[46] "Abbyy Fine Reader." `https://www.abbyy.com/en-eu/`. Accessed: 2017-10-30.

[47] C. Reul, U. Springmann, and F. Puppe, "Larex: A semi-automatic open-source tool for layout analysis and region extraction on early printed books," in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pp. 137–142, ACM, 2017.

[48] A. Fischer, H. Bunke, N. Naji, J. Savoy, M. Baechler, and R. Ingold, "The hisdoc project. automatic analysis, recognition, and retrieval of handwritten historical documents for digital libraries," *InterNational and InterDisciplinary Aspects of Scholarly Editing*, 2012.

[49] I. B. Messaoud, H. Amiri, H. El Abed, and V. Märgner, "A multilevel text-line segmentation framework for handwritten historical documents," in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pp. 515–520, IEEE, 2012.

[50] K. Chen, M. Seuret, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, "Ground truth model, tool, and dataset for layout analysis of historical documents.," in *Proceedings of SPIE, DRR*, p. 940204, 2015.

[51] M. Alberti, M. Bouillon, R. Ingold, and M. Liwicki, "Open evaluation tool for layout analysis of document images," *arXiv preprint arXiv:1712.01656*, 2017.

[52] M. Cheriet, R. F. Moghaddam, and R. Hedjam, "Visual language processing (vlp) of ancient manuscripts: Converting collections to windows on the past," in *Proceedings of GCC Conference and Exhibition (GCC)*, pp. 407–412, IEEE, 2013.

[53] C. Grana, D. Borghesani, S. Calderara, and R. Cucchiara, "Inside the bible: segmentation, annotation and retrieval for a new browsing experience," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pp. 379–386, ACM, 2008.

[54] J. W. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on computers*, vol. 100, no. 5, pp. 401–409, 1969.

[55] W. Playfair, *Playfair's commercial and political atlas and statistical breviary*. Cambridge University Press, 2005.

[56] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings of IEEE Symposium on Visual Languages, 1996.*, pp. 336–343, IEEE, 1996.

[57] D. A. Keim, "Information visualization and visual data mining," *IEEE transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[58] S. Liu, W. Cui, Y. Wu, and M. Liu, "A survey on information visualization: recent advances and challenges," *The Visual Computer*, vol. 30, no. 12, pp. 1373–1393, 2014.

[59] J. Dykes, A. M. MacEachren, and M.-J. Kraak, *Exploring geovisualization*. Elsevier, 2005.

[60] J. Zhao, *Interactive Visual Data Exploration: A Multi-Focus Approach*. PhD thesis, University of Toronto (Canada), 2015.

[61] A. Cockburn, A. K. Karlson, and B. B. Bederson, "A review of overview+ detail, zooming, and focus+ context interfaces.," *ACM Comput. Surv.*, vol. 41, no. 1, pp. 2–1, 2008.

[62] J. Zhao, F. Chevalier, and R. Balakrishnan, "Kronominer: using multi-foci navigation for the visual exploration of time-series data," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1737–1746, ACM, 2011.

[63] G. G. Robertson and J. D. Mackinlay, "The document lens," in *Proceedings of the 6th annual ACM symposium on User interface software and technology*, pp. 101–108, ACM, 1993.

[64] R. Rao and S. K. Card, "The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 318–322, ACM, 1994.

[65] J. Bertin, *Semiology of Graphics*. University of Wisconsin Press, 1983.

[66] M. S. T. Carpendale, "Considering visual variables as a basis for information visualisation," tech. rep., University of Calgary, Calgary, AB, 2003.

[67] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: foundations, techniques, and applications*. CRC Press, 2010.

[68] K. Kucher and A. Kerren, "Text visualization browser: A visual survey of text visualization techniques," *Poster Abstracts of IEEE VIS*, vol. 2014, 2014.

[69] C. Collins, S. Carpendale, and G. Penn, "Docuburst: Visualizing document content using language structure," in *Proceedings of Computer graphics forum*, vol. 28, pp. 1039–1046, Wiley Online Library, 2009.

[70] A. Thudt, U. Hinrichs, and S. Carpendale, "The bohemian bookshelf: supporting serendipitous book discoveries through information visualization," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1461–1470, ACM, 2012.

[71] D. A. Keim and D. Oelke, "Literature fingerprinting: A new method for visual literary analysis," in *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, (Washington, DC, USA), pp. 115–122, IEEE Computer Society, 2007.

[72] V. Setlur, C. Albrecht-Buehler, A. A Gooch, S. Rossoff, and B. Gooch, "Semanticons: Visual metaphors as file icons," in *Proceedings of Computer Graphics Forum*, vol. 24, pp. 647–656, Wiley Online Library, 2005.

[73] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information visualization*, vol. 7, no. 2, pp. 118–132, 2008.

[74] H. Strobelt, D. Oelke, C. Rohrdantz, A. Stoffel, D. A. Keim, and O. Deussen, "Document cards: A top trumps visualization for documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1145–1152, 2009.

[75] G. Robertson, M. Czerwinski, K. Larson, D. C. Robbins, D. Thiel, and M. Van Dantzich, "Data mountain: using spatial memory for document management," in *Proceedings of the 11th annual ACM symposium on User interface software and technology*, pp. 153–162, ACM, 1998.

[76] J. N. Vilaplana and M. Pérez-Montoro, "Diggersdiaries: Using Text Analysis to Support Exploration and Reading in a Large Document Collection," in *Proceedings of EuroVis - Posters* (A. P. Puig and T. Isenberg, eds.), The Eurographics Association, 2017.

[77] Y. Berzak, M. Richter, C. Ehrler, and T. Shore, "Information retrieval and visualization for the historical domain," *Language Technology for Cultural Heritage*, pp. 197–212, 2011.

[78] S. Jänicke and D. Joseph Wrisley, "Visualizing mouvance: Toward a visual analysis of variant medieval text traditions," *Digital Scholarship in the Humanities*, vol. 32, no. suppl_2, pp. ii106–ii123, 2017.

[79] T. team at University of Innsbruck, "Transkribus." `https://transkribus.eu/Transkribus/`. Accessed: 2017-09-28.

[80] P. Becker, *Die Benediktinerabtei St. Eucharius-St. Matthias vor Trier*, vol. 8. Walter de Gruyter, 1996.

[81] T. Jejkal, A. Vondrous, A. Kopmann, R. Stotzka, and V. Hartmann, "Kit data manager: the repository architecture enabling cross-disciplinary research," *Large-Scale Data Management and Analysis-Big Data in Science,*, 2014.

[82] B. Matthews, S. Sufi, D. Flannery, L. Lerusse, T. Griffin, M. Gleaves, and K. Kleese, "Using a core scientific metadata model in large-scale facilities," *International Journal of Digital Curation*, vol. 5, no. 1, pp. 106–118, 2010.

[83] A. M. Hesham, M. A. Rashwan, H. M. Al-Barhamtoshy, S. M. Abdou, A. A. Badr, and I. Farag, "Arabic document layout analysis," *Pattern Analysis and Applications*, vol. 20, no. 4, pp. 1275–1287, 2017.

[84] N. Instuments, "Spatial Calibration." `http://www.ni.com/white-paper/2907/en/`. Accessed: 2017-10-30.

[85] H. Prashanth, H. Shashidhara, and B. M. KN, "Image scaling comparison using universal image quality index," in *International Conference on Advances in Computing, Control, & Telecommunication Technologies, 2009*, pp. 859–863, IEEE, 2009.

[86] T.-Y. Chang, Y. Takiguchi, and M. Okada, "Physical structure segmentation with projection profile for mathematic formulae and graphics in academic paper images," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition, ICDAR.*, vol. 2, pp. 1193–1197, IEEE, 2007.

[87] A. G. Shanbhag, "Utilization of information measure as a means of image thresholding," *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 5, pp. 414–419, 1994.

[88] J. N. Kapur, P. K. Sahoo, and A. K. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer vision, graphics, and image processing*, vol. 29, no. 3, pp. 273–285, 1985.

[89] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[90] I. Arganda-Carreras, V. Kaynig, C. Rueden, K. W. Eliceiri, J. Schindelin, A. Cardona, and H. Sebastian Seung, "Trainable weka segmentation: a machine learning tool for microscopy pixel classification," *Bioinformatics*, vol. 33, no. 15, pp. 2424–2426, 2017.

[91] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "Knime-the konstanz information miner: version 2.0 and beyond," *AcM SIGKDD explorations Newsletter*, vol. 11, no. 1, pp. 26–31, 2009.

[92] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "Yale: Rapid prototyping for complex data mining tasks," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 935–940, ACM, 2006.

[93] U. Köthe, "Reusable software in computer vision," *Handbook of computer Vision and Applications*, vol. 3, pp. 103–132, 1999.

[94] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, *et al.*, "Fiji: an open-source platform for biological-image analysis," *Nature methods*, vol. 9, no. 7, pp. 676–682, 2012.

[95] I. H. Witten, E. Frank, L. E. Trigg, M. A. Hall, G. Holmes, and S. J. Cunningham, "Weka: Practical machine learning tools and techniques with java implementations (working paper 99/11)," 1999.

[96] M. A. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, University of Waikato Hamilton, 1999.

[97] D. Koller and M. Sahami, "Toward optimal feature selection," tech. rep., Stanford InfoLab, 1996.

[98] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[99] S. Kalmegh, "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news," *International Journal of Innovative Science, Engineering and Technology*, vol. 2, no. 2, pp. 438–46, 2015.

[100] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Proceedings of Advances in neural information processing systems*, pp. 1473–1480, 2006.

[101] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of biomedical informatics*, vol. 35, no. 5, pp. 352–359, 2002.

[102] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Machine learning: ECML-98*, pp. 137–142, 1998.

[103] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[104] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[105] M. W. Schwarz, W. B. Cowan, and J. C. Beatty, "An experimental comparison of rgb, yiq, lab, hsv, and opponent color models," *ACM Transactions on Graphics (TOG)*, vol. 6, no. 2, pp. 123–158, 1987.

[106] S. Pletschacher and A. Antonacopoulos, "The page (page analysis and ground-truth elements) format framework," in *Proceedings of 20th International Conference on Pattern Recognition (ICPR*, pp. 257–260, IEEE, 2010.

[107] P. K. Robertson and L. De Ferrari, "Systematic approaches to visualization: is a reference model needed," *Scientific Visualization*, vol. 18, pp. 287–305, 1994.

[108] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, "Manyeyes: a site for visualization at internet scale," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, 2007.

[109] "Tableau." `https://public.tableau.com/en-us/s/`. Accessed: 2017-11-23.

[110] S. Ribecca, "Data visualization catalogue." `https://datavizcatalogue.com/`. Accessed: 2017-10-30.

[111] R. M. Pickett and G. G. Grinstein, "Iconographic displays for visualizing multidimensional data," in *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, vol. 514, p. 519, 1988.

[112] H. Chernoff, "The use of faces to represent points in k-dimensional space graphically," *Journal of the American Statistical Association*, vol. 68, no. 342, pp. 361–368, 1973.

[113] A. Inselberg, "Parallel coordinates," in *Encyclopedia of Database Systems*, pp. 2018–2024, Springer, 2009.

[114] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield, "Scatterplot matrix techniques for large n," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 424–436, 1987.

[115] E. Kandogan, "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions," in *Proceedings of the IEEE Information Visualization Symposium*, vol. 650, p. 22, 2000.

[116] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu, "Pixel bar charts: a visualization technique for very large multi-attribute data sets," *Information Visualization*, vol. 1, no. 1, pp. 20–34, 2002.

[117] M. Ankerst, D. A. Keim, and H.-P. Kriegel, "Circle segments: A technique for visually exploring large multidimensional data sets," in *Proceedings of Visualization*, 1996.

[118] B. Johnson and B. Shneiderman, "Tree-maps: A space-filling approach to the visualization of hierarchical information structures," in *Proceedings of the 2nd conference on Visualization*, pp. 284–291, IEEE Computer Society Press, 1991.

[119] W. H. Smith, *Graphic statistics in management*. McGraw-Hill, 1924.

[120] C. Garcıa-Osorio and C. Fyfe, "Visualization of high-dimensional data via orthogonal curves," *Journal of Universal Computer Science*, vol. 11, no. 11, pp. 1806–1819, 2005.

[121] M. A. Hearst, "Tilebars: visualization of term distribution information in full text information access," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 59–66, ACM Press/Addison-Wesley Publishing Co., 1995.

[122] C. Weaver, "Building highly-coordinated visualizations in improvise," in *Proceedings of IEEE Symposium on Information Visualization*, pp. 159–166, IEEE, 2004.

[123] J.-D. Fekete, "The infovis toolkit," in *Proceedings of IEEE Symposium on Information Visualization*, pp. 167–174, IEEE, 2004.

[124] J. Heer, S. K. Card, and J. A. Landay, "Prefuse: a toolkit for interactive information visualization," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 421–430, ACM, 2005.

[125] M. Bostock and J. Heer, "Protovis: A graphical toolkit for visualization," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, 2009.

[126] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[127] M. O. Ward, "Xmdvtool: Integrating multiple methods for visualizing multivariate data," in *Proceedings of the Conference on Visualization*, pp. 326–333, IEEE Computer Society Press, 1994.

[128] M. Theus *et al.*, "Interactive data visualization using mondrian," *Journal of Statistical Software*, vol. 7, no. 11, pp. 1–9, 2002.

[129] M. M. Dias, J. K. Yamaguchi, E. Rabelo, and C. Franco, "Visualization techniques: Which is the most appropriate in the process of knowledge discovery in data base?," in *Advances in Data Mining Knowledge Discovery and Applications*, InTech, 2012.

[130] W. Javed and N. Elmqvist, "Exploring the design space of composite visualization," in *Proceedings of Pacific Visualization Symposium (PacificVis)*, pp. 1–8, IEEE, 2012.

[131] G. Andrienko and N. Andrienko, "Making a gis intelligent: Commongis project view," *AGILE99*, pp. 19–24, 1999.

[132] N. Boukhelifa, J. C. Roberts, and P. J. Rodgers, "A coordination model for exploratory multiview visualization," in *Proceedings of International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pp. 76–85, IEEE, 2003.

[133] K. Matkovic, W. Freiler, D. Gracanin, and H. Hauser, "Comvis: A coordinated multiple views system for prototyping new visualization technology," in *Proceedings of 12th International Conference Information Visualisation*, pp. 215–220, IEEE, 2008.

[134] I. Lokuge and S. Ishizaki, "Geospace: An interactive visualization system for exploring complex information spaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 409–414, ACM Press/Addison-Wesley Publishing Co., 1995.

[135] X. Yuan, P. Guo, H. Xiao, H. Zhou, and H. Qu, "Scattering points in parallel coordinates," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1001–1008, 2009.

[136] N. Henry, J.-D. Fekete, and M. J. McGuffin, "Nodetrix: a hybrid visualization of social networks," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1302–1309, 2007.

[137] S. Carpendale, "Evaluating information visualizations," *Information visualization*, pp. 19–45, 2008.

[138] M. D. Lee, R. E. Reilly, and M. E. Butavicius, "An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data," in *Proceedings of the Asia-Pacific symposium on Information visualisation-Volume 24*, pp. 1–10, Australian Computer Society, Inc., 2003.

[139] NIST, "Engineering statistics handbook." `http://www.itl.nist.gov/div898/handbook/eda/section3/eda33qb.htm`. Accessed: 2017-10-30.

[140] S. Wang, Y. Yang, J.-S. Chang, and F.-P. Lin, "Using penalized regression with parallel coordinates for visualization of significance in high dimensional data," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 4, no. 10, 2013.

[141] J. Heinrich and D. Weiskopf, "State of the art of parallel coordinates.," in *Proceedings of Eurographics (STARs)*, pp. 95–116, 2013.

[142] D. A. Keim, "Visual exploration of large data sets," *Communications of the ACM*, vol. 44, no. 8, pp. 38–44, 2001.

[143] G. J. Myatt, *Making sense of data: a practical guide to exploratory data analysis and data mining*. John Wiley & Sons, 2007.

[144] M. H. Shimabukuro, *Visualizações temporais em uma plataforma de software extensível e adaptável*. PhD thesis, Universidade de São Paulo, 2004.

[145] K.-P. Yee, D. Fisher, R. Dhamija, and M. Hearst, "Animated exploration of dynamic graphs with radial layout," in *Proceedings of IEEE Symposium on Information Visualization*, pp. 43–50, IEEE, 2001.

[146] R. A. Becker and W. S. Cleveland, "Brushing scatterplots," *Technometrics*, vol. 29, no. 2, pp. 127–142, 1987.

[147] C. Dunne, N. Henry Riche, B. Lee, R. Metoyer, and G. Robertson, "Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history," in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1663–1672, ACM, 2012.

[148] B. Lee, G. Smith, G. G. Robertson, M. Czerwinski, and D. S. Tan, "Facetlens: exposing trends and relationships to support sensemaking within faceted datasets," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1293–1302, ACM, 2009.

[149] J. Heinrich, Y. Luo, A. E. Kirkpatrick, H. Zhang, and D. Weiskopf, "Evaluation of a bundling technique for parallel coordinates," *Xiv:1109.6073*, 2011.

[150] D. Holten and J. J. Van Wijk, "Evaluation of cluster identification performance for different pcp variants," in *Computer Graphics Forum*, vol. 29, pp. 793–802, Wiley Online Library, 2010.

[151] "OCR-D." `http://ocr-d.de/`. Accessed: 2017-11-23.

[152] J. Graybeal, S. Miller, and K. Stocks, "The MMI guides: navigating the world of marine metadata." `http://uop.whoi.edu/techdocs/presentations/MMI_Guides.pdf(2010)`. Accessed: 2017-11-23.

[153] "WADM." `https://www.w3.org/TR/annotation-model/`. Accessed: 2017-11-23.

[154] A. Prabhune, A. Zweig, R. Stotzka, M. Gertz, and J. Hesser, "Prov2one: An algorithm for automatically constructing provone provenance graphs," in *Proceedings of International Provenance and Annotation Workshop*, pp. 204–208, Springer, 2016.

List of Abbreviations

ARFF	Attribute-Relation File Format
CRLA	Constrained Run-Length Algorithm
CSMD	Core Scientific Metadata Model
CSV	Comma Separated Value
CVVs	Composite visualization views
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DEBORA	Digital AccEss to BOoks of the RenAissance
DMOS	Description and MOdification of Segmentation
DPI	Dots Per Inch
HSB	Hue, Saturation,Brightness
HTR	Handwritten Text Recognition
IPE	Institute for Data Processing and Electronics
KIT	Karlsruhe Institute of Technology
METS	Metadata Encoding Transmission Standard
MLP	Multilayer Perceptron
OCR	Optical Character Recognition
PAGE	Page Analysis and Ground-truth Elements
PT	Primary user Tasks

RGB	Red, Green, Blue
ST	Secondary user Tasks
SVM	Support Vector Machine
TEI	Text Encoding Initiative
TIFF/TIF	Tagged Image File Format
TWS	Trainable Weka Segmentation
WEKA	Waikato Environment for Knowledge Analysis

Publications related to this thesis

Conference Papers and Abstracts

- **Software workflow for the automatic tagging of medieval manuscript images (SWATI)**
  S. Chandna, D. Tonne, T. Jejkal, R. Stotzka, C. Krause, P. Vanscheidt, H. Busch, A. Prabhune
  SPIE/IS & T Electronic Imaging, 2015
- **An effective visualization technique for determining co-relations in high-dimensional medieval manuscripts data**
  S. Chandna, D. Tonne, R. Stotzka, H. Busch, P. Vanscheidt, C. Krause
  Electronic Imaging, 2016
- **Quantitative exploration of large medieval manuscripts data for the codicological research**
  S. Chandna, F. Rindone, C. Dachsbacher, R. Stotzka
  In proceedings of Large Data Analysis and Visualization (LDAV), 2016 IEEE 6th Symposium
- **Quanticod revisited. Neue möglichkeiten zur analyse mittelalterlicher handschriften**
  H. Busch, **S. Chandna**, C. Krause, P. Vanscheidt
  Book of Abstracts DHd, 2015
- **Visualisierung mittelalterlicher handschriften im projekt eCodicology**
  H. Busch, **S. Chandna**, D. Tonne, C. Krause, P. Vanscheidt, O. Schmid
  Konferenzabstracts DHd, 2016
- **Automatische bild-textanalyze: Chancen für die zeitschriftenforschung jenseits von reinen textdaten**
  N. Rißler-Pipka, **S. Chandna**, D. Tonne.
  Digitale Nachhaltigkeit, DHd 2017
- **The computer and the medieval library**
  H. Busch, **S.Chandna**
  Codicology and Palaeography in the Digital Age 4, Books on Demand, 2017

Conference Talks and Workshops

- **From Image to Text and vice-versa. Quantitative research with the medieval manuscripts in the project eCodicology (Vom Bild zum Text und wieder zurück. Quantitative Forschung mit mittelalterlichen Handschriften im Projekt eCodicology)**
  S. Chandna , C. Krause.
  Formen der digitalen Edition und Präsentation beschrifteter Artefakte, Heidelberg, Germany, 2014.

- **Development of New Technologies for the Automatic Analysis of Medieval Manuscripts**
  H. Busch, **S. Chandna**, C. Krause, P. Vanscheidt.
  Möglichkeiten der automatischen Manuskript Manuskriptanalyse, Trier, Germany, 2014

- **Software workflow for the automatic tagging of medieval manuscript images (SWATI)**
  S. Chandna
  SPIE/IS & T Electronic Imaging, San Francisco, California, USA, 2015

- **The Technical Perspective of eCodicology**
  S. Chandna
  Möglichkeiten der automatischen Mustererkennung und Analyse historischer Dokumente, Karlsruhe, Germany, 2015.

- **An effective visualization technique for determining corelations in high-dimensional medieval manuscripts data**
  S. Chandna
  Electronic Imaging, San Francisco, California, USA, 2016

- **Quantitative exploration of large medieval manuscripts data for the codicological research**
  S. Chandna
  Large Data Analysis and Visualization (LDAV), Baltimore, Maryland, USA, 2016

- **eCodicology-Algorithms for the Automatic Tagging of Medieval Manuscripts**
  S. Chandna
  Forschung mit Schriftquellen in digitalen Zeitalter, Darmstadt, Germany, 2016

- **Automatic image-text analysis: Opportunities for magazines research beyond text data (Automatische Bild-Text-Analyse : Chancen für die Zeitschriftenforschung jenseits von reinen Textdaten)**
  S. Chandna, N. Rißler-Pipka
  DHd, Digitale Nachhaltigkeit, Bern, Switzerland, 2017

Dissemination

- Organized eCodicology Conference “Maschinen und Manuskripte II - Möglichkeiten der automatischen Mustererkennung und Analyse historischer Dokumente”, Karlsruhe, Germany, 2015
- Presented CodiViz to general public in exhibition Maschinen und Manuskripte - Digitale Erschließung der Handschriften von St. Matthias, Trier, Germany, 2015

Acknowledgements

I would like to thank my advisor Dr. Rainer Stotzka for providing me the opportunity to work as a Ph.D. student at Karlsruhe Institute of Technology. He has been a great resource of support, inspiration, and motivation. Without his constant support my doctoral studies would not have been possible. It was mostly from him that I learned how to be a good researcher and how to perform critical thinking.

Also many thanks to my supervising professors Prof. Dr.-Ing. Carsten Dachsbacher and Prof. Dr. Marc Weber for regularly taking out their valuable time and providing constructive feedback.

Special thanks to all my colleagues (Thomas Jejkal, Volker Hartmann, Danah Tonne, and Germaine Götzelmann) at the Steinbuch Center for Computing (SCC) and my ex-colleagues at the Institute for Data Processing and Electronics (IPE) at the Karlsruhe Institute of Technology for a nice and a friendly working environment. Especially I would like to thank Dr. Nicole Ruiter who has been a great support at IPE, all the colleagues of the software methods group and the USCT group.

I would also like to thank my students Joseline Samago, Kayrat Saginaev, Shubham Kapoor, and Kevin Geggus for their valuable work and support, our partners at Technical University of Darmstadt and Trier Center for Digital Humanities from eCodicology project. This work would not have been possible without the handwritten historical document images contributed by the City Library of Trier. Many thanks to Prof. Dr. Andrea Rapp and Prof. Dr. Claudine Moulin for motivating collaboration.

Finally, I would like to thank my family and friends, especially to my parents, who supported me in every sense throughout my life, and my husband Manish Gulati for his support, encouragement, and patience.

This work was mainly funded by Federal Ministry of Education and Research and German Research Foundation. Additionally this work is supported by Helmholtz Portfolio Extension "Large Scale Data Management and Analysis" with contributions from the Data Life Cycle Lab "Key Technologies".