An effective visualization technique for determining co-relations in high-dimensional medieval manuscripts data

Swati Chandna, Danah Tonne, Rainer Stotzka; Karlsruhe Institute of Technology, Germany; Hannah Busch; Trier University, Germany Philipp Vanscheidt, Celia Krause; Technical University of Darmstadt, Germany

Abstract

Digital libraries have played a very important role in preserving the cultural heritage be it medieval manuscripts, audio material or paintings. Currently, standard manual techniques exist to analyze medieval manuscripts. Analysis performed in this way is very tedious and labor intensive which restricts the number of medieval manuscripts that can be analyzed. Digital methods, tools and techniques are reaching its technological limits due to lack of visual interfaces.

This paper presents a visualization concept called CodiVis to facilitate explorations of co-relations in the abstract feature space of large sets of digitized medieval manuscripts. As a starting point, CodiVis uses the medieval manuscripts digitized within the scope of the project "Virtual Scriptorium St. Matthias". It combines two visualization techniques in order to overcome the shortcomings of the single visualization technique. In the first technique, manuscripts are clustered according to the bibliographic metadata and represented in radial tree. This gives a quick overview of the whole dataset. In the second technique, bibliographic metadata is further linked to the macro- and micro-structural layout features in the parallel coordinate view. Interactive changes in radial tree are automatically reflected in the parallel coordinate view. The proposed visualization concept shows the potential of analysis by enabling quick exploration of big humanities data. Furthermore, the evaluation tests and feedback from humanities scholars and other users showed that CodiVis is capable of identifying co-relations in arts and humanities.

Introduction

During the last years, digital libraries have played a very important role in preserving the cultural heritage be it medieval manuscripts, audio material, paintings etc. The major challenge lies in addressing some of the fundamental questions as "How to facilitate the exploration of cultural heritage"?, "How can the traditional skills used by humanities scholars be reshaped into digital world"? [1]. The heterogeneous nature and size of the humanities data and the need to analyze such large data sets is the main challenge for arts and humanities. Currently, standard manual techniques exist to analyze medieval manuscripts for example, analyzing the size of the text space which is changing throughout the manuscript. As the humanities get more and more data intensive the research becomes considerably hindered. Analysis performed in this way is very tedious and labor intensive which restricts the number of medieval manuscripts that can be analyzed. Digital methods, tools and techniques are also reaching their technological limits due to lack of visual interfaces.

Information Visualization [2] has the potential of overcoming the challenges of manual analysis methods for humanities scholars. It focuses mainly on abstract information. Information visualization, at its core, have two main components: representation and interaction [2]. The representation component, whose roots lie in the field of computer graphics, concerns the mapping from data to representation. The interaction component involves dialog between the user and system as the user explores the data to uncover insights. Without interaction, an InfoVis technique becomes static. Thus, usefulness becomes more limited as the data set that they represent grows larger with static images. A key research challenge is to discover a visualization technique for digital humanities such that analysis of humanities data becomes easier and enables research on large data sets. It should enable humanities scholars to get the information they want, make sense of that information, determine co-relations in the data and to reach decisions in short period of time.

In this context, a concept of visualization framework called CodiVis is presented to analyze the layout features like page size, written space and pictorial space as shown in Figure 1. CodiVis is part of the BMBF-funded project "eCodicology" [3] which aims to establish a workflow for automatic identification and analysis of macro- and micro-structural layout features of medieval manuscripts.

Requirements and Related Work

In order to design the visualization framework for analysis of medieval manuscripts multidimensional data and to determine the co-relations in large datasets of manuscripts the following requirements should be met

1. extract important variables
2. detect various outliers
3. discover underlying structure

In principle, it is feasible to use existing well established visualization techniques for the representation of multidimensional data. Over the years the visualization community has researched and developed various techniques suitable for specific data and information types.

Palladio [4] is a web-based visualization tool for complex humanities data. It is a package that includes number of tools, each of which allows to get different angle on same data. One big limitation of Palladio is that it is not able to analyze large datasets. Analyzing with simple charts like pie charts, bar charts is not possible.

The most common visualization techniques like Scatterplot matrices [5] combines all combination pairs of all dimensions and organize them by matrix. Parallel coordinates [6] maps the k-dimensional space onto the two dimensional display by using k equidistant axes which are parallel to one of the display axes. The axes corresponding to the dimensions are linearly scaled from

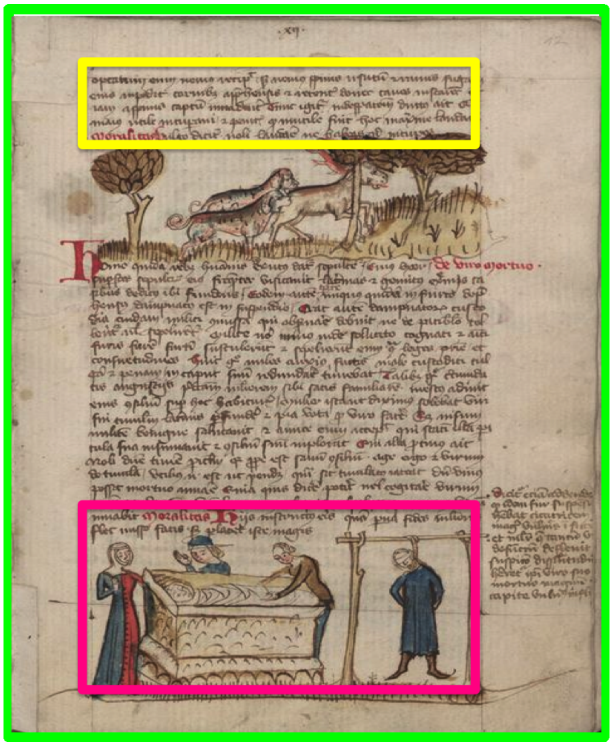


Figure 1. Examples from medieval manuscripts showing various layout features 🟩 page size, 🟨 written space, 🟥 pictorial space.

minimum to the maximum value of the corresponding information. Each data item is presented as a polygon line intersecting each of the axes at that point.

Another class of visual exploration technique are iconic displays [6] like stick figure icons, star icons. They basically map attribute values of the multi-dimensional data item to the features of the icon. Dense pixel display [6] like circle segment technique maps each dimension value to a colored pixel and groups the pixels belonging to each dimension in adjacent areas.

Interaction techniques like brushing and linking [7] helps the data analyst to directly interact with visualizations and dynamically change the visualization according to exploratory objectives. The idea of brushing and linking is to combine different visualizations methods to overcome the shortcomings of single techniques.

Fua et al. [8] supported hierarchical clustering using extended parallel coordinates. They developed a multiresolutional view of the data via hierarchical clustering, and use a variation on parallel coordinates to convey aggregation information for the resulting clusters. Users can navigate the resulting structure until the desired focus region and level of detail is reached. Lex et al.[9] proposed focus and context visualization for clustering biomolecular data. The main goal of this work was to support experts in the process of hypotheses generation concerning the roles of genes in diseases. They employed two different multiple-view approaches. But, their approach is limited to particular dataset.

These methods do not provide expected results when applied to high dimensional medieval manuscripts data with heterogeneous layout structures. These methods are not able to detect outliers, extract important variables and answer other humanistic research questions. The user can only upload data which is similar to the standard patterns accepted by these methods. The usage of existing techniques alone is not an option for humanities data as above mentioned requirements are not met. Thus, a new methodology has been specified namely CodiVis which is presented in the following chapter.

Overview of CodiVis

The basic idea is to design a visualization framework for analyzing and determining co-relations in the abstract feature space of large sets of medieval manuscripts. We designed a framework which comprises three main stages as detailed below and illustrated in Figure 2.

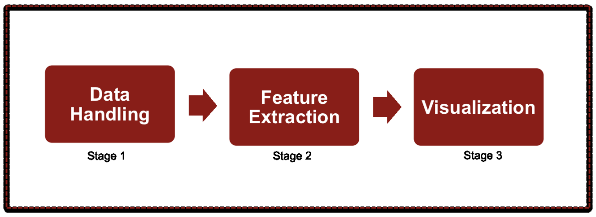


Figure 2. Workflow of visualization framework

Stage 1: Data Handling

The first stage deals with uploading and downloading of the data from data repositories. It consists of two steps described as follows:

Data Ingest

This step uploads a library stock of roughly 440 medieval manuscripts, mostly written between eighth and sixteenth century and collected in the library of the Benedictine Abbey of St. Matthias in Trier (Germany), into data repositories. These medieval manuscripts were digitized and enriched with bibliographic metadata within the scope of the project “Virtual Scriptorium St. Matthias”.The image data and metadata can be accessed from the project homepage (http://stmatthias.uni-trier.de) using a user application called DFG-Viewer [10]. The St. Matthias data is ingested into the so called CodiStore repository using the service stack of the data repository software KIT Data Manager[11]. The data is ingested in order to manage and allow data intensive computing.

Data Access

This is the second step under data handling which allows to access the St. Matthias database from CodiStore for further processing and analyzing the image data. The data can be easily downloaded using services of KIT Data Manager.

Stage 2: Extraction of Manuscript Layout Features

The second stage is to extract the layout features of the medieval manuscripts with the help of image processing methods. SWATI (Software Workflow for the Automatic Tagging of the Medieval Manuscript Images) [12] is used to extract various layout features shown in Table 1.

The stage also involves extraction of bibliographic metadata from TEI [13] file shown in Table 2. It was added during digitization of manuscripts. TEI is a collectively developed standard for the representation and encoding of texts in the digital form. It is widely used in the field of digital humanities to store information about any kind of textual data. For this a TEI P5-conformant ODD-based metadata schema has been designed that allows to store metrical data in the manuscript description. ODD - which stands for One Document Does It All - files can be converted easily into various XML schema languages by using a tool developed for generating customization for TEI called ROMA [14].

Stage 3: Visualization

The third stage helps to find co-relations between bigger group of manuscripts. Bibliographic metadata is taken in combination with automatically detected layout features. This combination is visualized using radial tree and linked parallel coordinates which is described in following subsections:

Table 1: Layout features of the medieval manuscripts extracted by SWATI

S.No	Features
1.	Number of Pages
2.	Mean Color Value
3.	Page Width
4.	Page Height
5.	Upper Left Corner Coordinates of Page
6.	Relative Measurements of the Page
7.	Text Width
8.	Text Height
9.	Text Areas
10.	Upper Left Corner Coordinates of Text
11.	Relative Measurements of the Text
12.	Pictorial Width
13.	Pictorial Height
14.	Number of Pictorial Areas
15.	Upper Left Corner Coordinates of Pictures
16.	Relative Measurements of the Pictures

Table 2: Bibliographic metadata of the medieval manuscripts

S.No	Bibliographic metadata	Values
1.	Format	2°, 4°, 8°, 12°, 16°
2.	Material	Paper, Parchment, Both, None
3.	Century	8 AD., 9 AD., 10 AD., 11 AD., 12 AD., 13 AD., 14 AD., 15 AD., 16 AD.,

Radial Tree to Overview Bibliographic Metadata

Various hierarchical levels of bibliographic metadata is represented by radial tree [15] as shown in Figure 3. The radial tree is a node-link tree with transformations in polar coordinates. This visualization technique is a better usage of space if few hierarchy levels and many bottom nodes exist. From the bibliographic metadata shown in Table 2, "Century" is taken and represented in the radial tree as the first prototype of visualization metaphor. The cluster hierarchy is also used for color assignment. Keeping color blindness of some users in mind, white color is used for the nodes belonging to top and root level. The nodes at the bottom level of hierarchy are assigned different colors with varying brightness using a qualitative color map. Different colors help to distinguish between classes of items and increase the appeal of the visualization.

Parallel Coordinates

Various layout features extracted by SWATI workflow are represented using parallel coordinates [15]. It is a way of visualizing high dimensional data and analyzing multivariate data. The bibliographic metadata from TEI file is represented in radial tree and corresponding layout features are represented in parallel coordinates. Here page height, page width and text areas are selected as the sample features out of 16 features shown in Table 1 to be represented in parallel coordinates. These layout features are mapped onto a vertical axis and each data value from CSV file is represented along a line. It is scaled to lie between minimum and maximum at the top. A pure collection of points would not be useful, so the points belonging to the same record are connected with lines. This creates a characteristic jumble of lines which parallel coordinates are famous for. The color assignment is done similarly as in the radial tree.

Tableview

In the table view, each row and column of the CSV file is represented in this prototype. This table view is linked to both radial tree and parallel coordinate.

Brushing and Linking in Radial tree, Parallel Coordinates and Tableview

Two kinds of brushes are provided for analyzing the medieval manuscripts data and exploring the co-relations for example, determining similar manuscripts and their corresponding features or detecting outliers and improbable information.

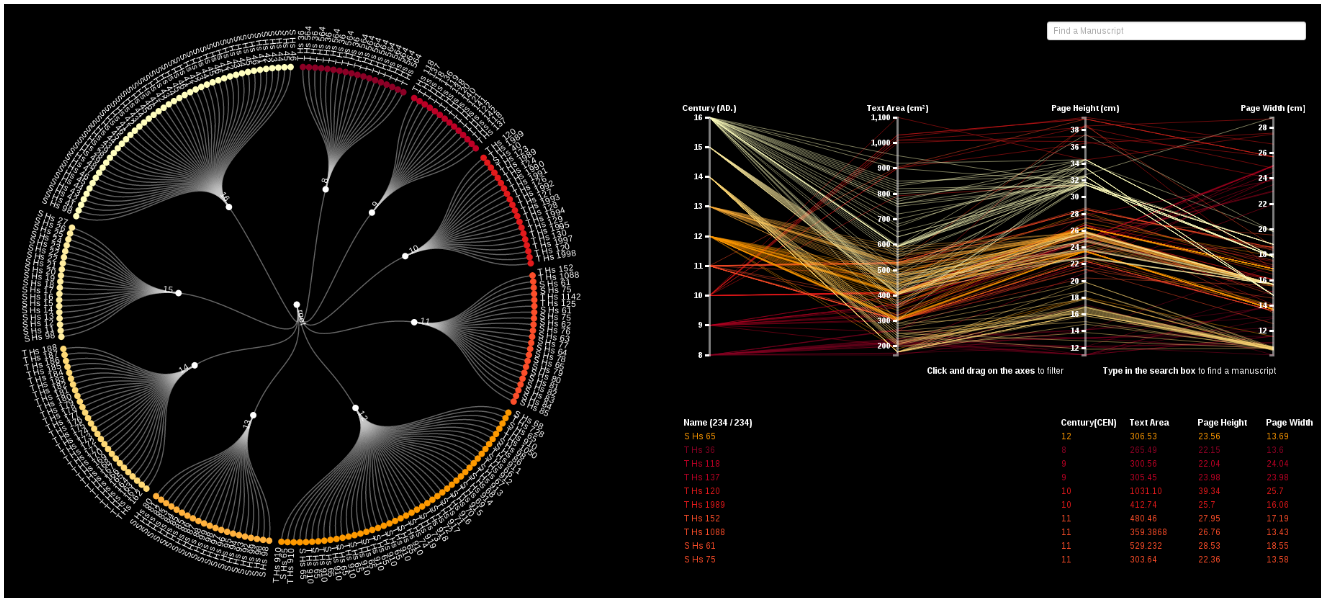
1. Polar brush: Users can brush in polar coordinates and select different nodes which either represent manuscripts at the bottom level of hierarchy or a cluster of manuscripts at the top level of hierarchy as shown in Figure 4. When users brush various nodes of the radial tree, corresponding data gets automatically reflected in the parallel coordinate view and the table view.

2. Vertical axis brush: Users can brush any of the vertical axis of the parallel coordinate view. When user brushes on vertical axis, corresponding line in parallel coordinate and node in radial tree gets highlighted.

A table is linked to both radial tree and parallel coordinates. Users can mouseover each row in the table and link the corresponding line in parallel coordinates and node in radial tree.

Results and Evaluation

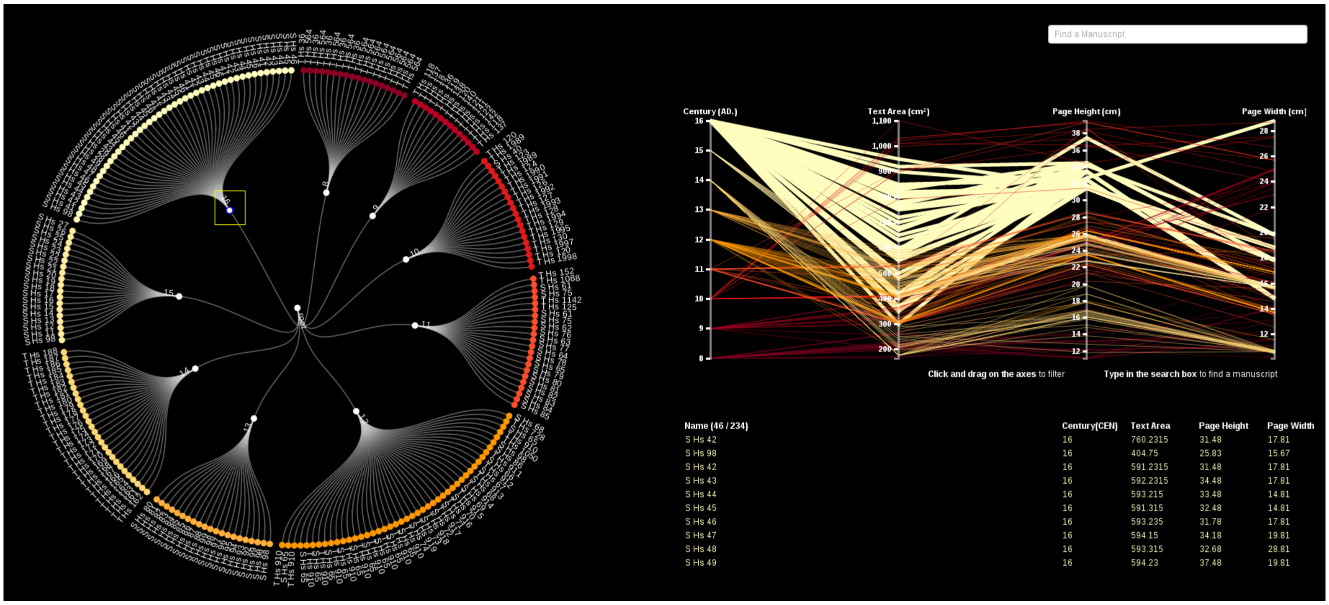
For the first evaluation of CodiVis prototype 20 participants of different age groups and from different backgrounds like humanities, medical sciences, physics, electrical engineering, finances were asked to participate in the evaluation. Firstly, Co-



Click and drag on the axes to filter **Type in the search box** to find a manuscript

Name (234 / 234)	Century(CEN)	Text Area	Page Height	Page Width
S Hs 65	12	306.53	23.56	13.69
T Hs 36	8	265.49	22.15	13.6
T Hs 118	9	300.56	22.04	24.04
T Hs 137	9	305.45	23.98	23.98
T Hs 120	10	1031.10	39.34	25.7
T Hs 1989	10	412.74	25.7	16.06
T Hs 152	11	480.46	27.95	17.19
T Hs 1088	11	359.3868	26.76	13.43
S Hs 61	11	529.232	28.53	18.55
S Hs 75	11	303.64	22.36	13.58

Figure 3. *This figure shows a radial tree view on left side which represents century of the manuscripts and the corresponding layout features are represented in parallel coordinate view on the right side*



Click and drag on the axes to filter **Type in the search box** to find a manuscript

Name (46 / 234)	Century(CEN)	Text Area	Page Height	Page Width
S Hs 42	16	760.2315	31.48	17.81
S Hs 98	16	404.75	25.83	15.67
S Hs 42	16	591.2315	31.48	17.81
S Hs 43	16	592.2315	34.48	17.81
S Hs 44	16	593.215	33.48	14.81
S Hs 45	16	591.315	32.48	14.81
S Hs 46	16	593.235	31.78	17.81
S Hs 47	16	594.15	34.18	19.81
S Hs 48	16	593.315	32.68	28.81
S Hs 49	16	594.23	37.48	19.81

Figure 4. *This figure shows application of polar brush in radial tree and automatic reflections in parallel coordinate view and table view*

diVis was explained in detail to the participants followed by a live demonstration. Afterwards, prototype was provided to the participants to see and interact with. These participants used Firefox as a web browser. Screen resolution was set to 1920*1280. Then we asked the participants to define the difficulty level and to grade the techniques used according to usefulness and usability.

User Tasks

User tasks are divided into simple and complex category i.e. primary user tasks and secondary user tasks. These tasks are described in detail as follows:

Primary User Tasks

In primary user tasks, firstly CodiVis and various interaction techniques like brushing and linking are described in detail. We worked through two sample questions to make the participants comfortable with various interaction techniques. After that the prototype was given to each participant with 10 analytical questions to answer. These analytical questions were made with help of humanities scholars. Some examples of these analytical questions are 1) Which are the manuscripts with largest page size (Width, Height, Area)? 2) Which are the manuscripts with smallest page size (Width, Height, Area)? 3) Which manuscript has maximum and minimum number of pages? 4) Which manuscript has largest and smallest text area? 5) In which era were the maximum number of manuscripts written?

Participants used brushing and linking technique to explore the co-relations in medieval manuscripts data sets. They appreciated the combination of two visualization techniques as now it was possible to get an overview of the whole dataset on the one hand and simultaneously get the details on the other hand. The majority of participants solved the primary user task correctly as shown in Figure 5. 8 out of 10 questions were answered correctly by 80-100 percent of the participants.

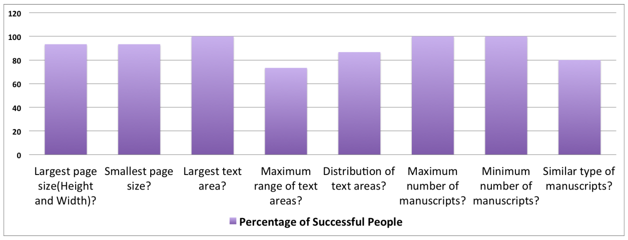


Figure 5. *Evaluation based on primary tasks*

Secondary User Tasks

In secondary user tasks, participants were asked to solve various activities as follows:

- **Finding outliers and improbable information**: In this task, participants needed to find examples of improbable information. As each of the medieval manuscripts is different with respect to color, shape and size, automatic layout extraction can produce unwanted or improbable results. The participants were asked to find such errors and their corresponding information. Participants determined many outliers in 10th and 11th century. In principle, statistical analysis is widely used for outlier detection but evaluation showed that it is also possible to detect outliers with CodiVis.
- **Finding similar type of medieval manuscripts and their corresponding information**: The participants were asked to provide an overview of similar manuscripts and to identify the corresponding information. The participants interacted with CodiVis to generate radial tree according to "centuries of the manuscripts" to provide an overview of the whole dataset. All the different centuries were listed and represented in top level hierarchy and at the bottom level all the similar manuscripts were grouped and linked to corresponding top level hierarchy. After clustering the similarity of the manuscripts was easily studied.
- **Determining the effect of various features on analysis**: The participants were asked to determine the effect of various features. They were asked to sort out most influential of all the considered features. To determine the effect of various features participants studied different clusters of radial tree and linked parallel coordinates. They learned that manuscripts belonging to 9th century have the highest average page area but the smallest text area. Thus, page area and the text area are the most influential features of the manuscripts belonging to 9th century.

Participants solved the first secondary task by exploring various radial tree generated according to selected bibliographic metadata and their corresponding information in parallel coordinate view. They were able to provide an overview with bird's eye view.

They also solved the secondary user task 2 while solving the user task 1. For finding the effect of various features, they explored the clusters which covered largest or smallest feature space.

Figure 6 shows that majority (60-80 percent) of the participants were able to solve this task. 20-40 percent of the inexperienced participants shared the opinion, that solving secondary tasks would require much more time and understanding of the data.

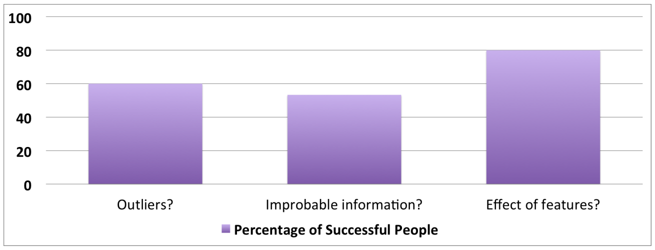


Figure 6. *Evaluation based on secondary tasks*

Overall feedback from participants

The participants liked the use of brushing and linking and other interaction techniques to combine radial tree and parallel coordinates. They saw the approach as very intuitive way to facilitate the exploration and visualization of cultural heritage. Using such ways of visualizing large data sets, traditional skills used by humanities scholars to analyze the humanities data can be overcomed. A small reaction test was taken in order to evaluate the appearance and functionality of CodiVis. As shown in Figure 7, reaction test consisted of set of certain adjectives describing the visualization. Results show that 80-90 percent of the participants found the visualization technique efficient, useful, fun and intuitive.

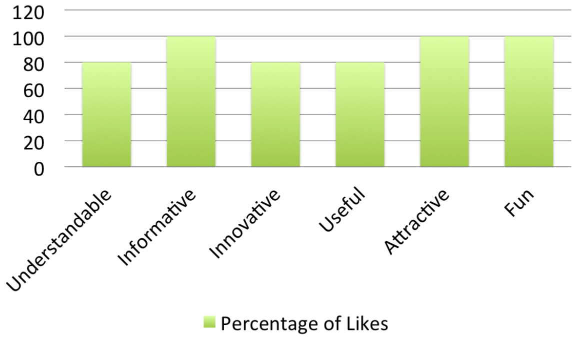


Figure 7. *Results of reaction test*

Conclusions and Future Scope

This paper presents a novel approach for the visualization of multidimensional datasets of digitized medieval manuscripts i.e CodiVis. It shows great potential for exploring co-relations between many manuscripts. CodiVis combines two different visualization techniques in order to overcome the shortcomings of single techniques. Interactive changes made in one visualization are automatically reflected in the second visualization. It uses radial tree to provide an overview and to avoid visual clutter. Radial tree is used to visualize different types of bibliographic metadata of the manuscripts. Additionally, a linked parallel co-ordinate view is provided to analyze the manuscripts data in more detail.

Furthermore, the evaluation tests shows that 80-100 percent of the participants solved the primary and secondary tasks correctly. CodiVis opened up many new questions for humanities scholars. This will lead humanities scholars back to digital libraries to explain such patterns. Such a visualization framework is one example to show an ability to answer the fundamental questions like how to reshape the traditional skills used by humanities into the digital world. It can help humanities scholars to explore big humanities data using dynamic components of information visualization. Participants found the visualization framework useful, attractive and innovative. However, characteristics like visual exploration of single page is still missing. Considering further research activities, future work will focus on the visualizations to explore the details of 170,000 pages of medieval manuscripts. Finally, new humanistic research questions will be explored as the next step. The visualization framework will be integrated as a service into eCodicology and DARIAH infrastructure to make it adaptable for wider range of documents.

Acknowledgments

This work has been funded by Federal Ministry of Education and Research (BMBF) under the agreement O1UG1350. Additionally this work is supported by Helmholtz Portfolio Extension “Large Scale Data Management and Analysis” with contributions of the Data Life Cycle Lab “Key Technologies”. We would also like to thank Prof. Dr. Carsten Dachsbacher for providing us with his ideas of brushing and linking.

References

[1] Humanities, D.(July 2015), http://sites.library.northwestern.edu/dh/.

[2] Gershon, Nahum, and Stephen G. Eick. "Information visualization." IEEE Computer Graphics and Applications 4 (1997): 29-31. Visualization and Computer Graphics, IEEE Transactions on 13, no. 6 (2007): 1224-1231.

[3] eCodicology(July 2015), http://www.ecodicology.org/.

[4] Palladio(July,2015), http://hdlab.stanford.edu/projects/palladio/.

[5] Carr, Daniel B., et al. "Scatterplot matrix techniques for large N." Journal of the American Statistical Association 82.398 (1987): 424-436.

[6] Inselberg, Alfred, and Bernard Dimsdale. "Parallel coordinates." Human-Machine Interactive Systems. Springer US, 1991. 199-233.

[7] Yi, Ji Soo, et al. "Toward a deeper understanding of the role of interaction in information visualization." Visualization and Computer Graphics, IEEE Transactions on 13.6 (2007): 1224-1231.

[8] Fua, Ying-Huey, Matthew O. Ward, and Elke A. Rundensteiner. "Hierarchical parallel coordinates for exploration of large datasets." Proceedings of the conference on Visualization’99: celebrating ten years. IEEE Computer Society Press, 1999.

[9] Lex, A., Streit, M., Kruijff, E., Schmalstieg, D. (2010, March). Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context. In Pacific Visualization Symposium (PacificVis), 2010 IEEE (pp. 57-64). IEEE.

[10] DFG-Viewer(July 2015). http://dfg-viewer.de/ueber-das-projekt/.

[11] KIT Data Manger(July 2015). http://ipelsdf1.lsdf.kit.edu/index.php/nav-kit-dm-resources/nav-kit-dm-resources-binary.

[12] Chandna, Swati, et al. "Software workflow for the automatic tagging of medieval manuscript images (SWATI)." IST/SPIE Electronic Imaging. International Society for Optics and Photonics, 2015.

[13] TEI (July 2015). http://www.tei-c.org/index.xml.

[14] ROMA (July 2015). //www.tei-c.org/Roma/.

[15] Shneiderman, Ben. "The eyes have it: A task by data type taxonomy for information visualizations." Visual Languages, 1996. Proceedings., IEEE Symposium on. IEEE, 1996.

[16] Data-Driven Documents(July 2015), http://d3js.org/.