Chapter 1

Introduction

In different scientific disciplines, from biomedical imaging to microscopy and remote sensing, researcher and engineers gather an ever growing amount of image data. During the last years, especially the joint use of data obtained by means of multiple imaging techniques, varying parameterizations or in temporal studies has received significant attention. The synergistic combination of complementary and redundant image data can provide insight into an application domain, the underlying phenomena, and facilitate more detailed and reliable analyses.

Multispectral and hyperspectral imaging techniques based on air- and satellite-borne sensor systems are employed in remote sensing for the acquisition of terrain information at different wavelengths and for the monitoring of temporal changes. Tomographic imaging techniques in medicine such as Magnetic Resonance Imaging (MRI), Computer Tomography (CT), and Positron Emission Tomography (PET) allow not only for the three-dimensional imaging of the human body, but the combination of these techniques can provide valuable anatomical and functional information for diagnostic purposes. In breast cancer research and studies of brain activities, Dynamic Contrast Enhanced Magnetic Resonance Imaging (DCE-MRI) and functional Magnetic Resonance Imaging (fMRI) are used for the generation of image sequences over time. Commonly, such imagery could be thought of in terms of multiple scalar fields defined over the same spatial and temporal domain and might be referred to as *multivariate image data*.

The analysis of *abstract* and *scientific* data is often considered in the context of Knowledge Discovery in Databases (KDD). KDD provides a comprehensive theoretical framework for data analysis processes with the final aim to extract useful information and knowledge from the data. In this context, *data mining* can be thought of as the main component of the KDD process in which patterns and models are derived by means of a variety of techniques with roots in *machine learning*, *artificial neural networks*, and *statistical pattern recognition* in general.

In spite of the effort that has been devoted to the development of techniques for an automated analysis of multivariate imagery, these approaches often are only applicable when the data and the underlying phenomena are fully understood. Nevertheless, due to the amount and complexity of the data involved, a manual analysis of such imagery quickly becomes prohibitive. Therefore, it is not surprising that many of the most efficient frameworks for the analysis of multivariate imagery in remote sensing are based on the combination of the capabilities of humans and computers and could be considered as human-assisted machine or machine-assisted human schemes [Landgrebe, 2000].

Even though knowledge discovery itself is a highly interactive and user guided process, for a long time the role of human reasoning and visualization within the KDD framework was rather limited. Techniques for data visualization have been employed mainly to convey the results of a data mining process. Nowadays it becomes more and more accepted that the integration of visualization and data mining techniques as well as of concepts from *exploratory data analysis* is essential.

Exploratory data analysis is a synonym for open-minded data examination without any prior assumptions or hypotheses. Techniques of exploratory data analysis are especially useful for providing first insight into the underlying structure of the data. To this end, traditional approaches to exploratory data analysis make use of a broad range of simple visualization techniques for the investigation of elementary statistical properties and relationships in multivariate data sets. As a consequence of this development, *visual data mining* has evolved as a new paradigm in knowledge discovery. Visual data mining stands for the tight coupling of visualization techniques and data mining methods in order to utilize the exceptional human visual capabilities with respect to pattern recognition and generalization as well as to take advantage of expert knowledge. Therefore, concepts of visual data mining are of high value for the purpose of exploratory data analysis.

Even though the analysis of multivariate imagery has become increasingly important, the notion of visual data mining has been discussed mostly in studies of abstract concepts and relationships. While a variety of *information visualization* techniques has been developed for the representation of high-dimensional multivariate data sets, the visualization of multivariate image data is a still a top research problem. Typically, information visualization techniques are not well suited for the generation of an integrated graphical representation of such data while the majority of *scientific visualization* techniques are only applicable for the representation of a single scalar, vector or tensor valued entity defined over a two- or three-dimensional domain.

In this context, *image fusion* gives rise to an important methodology for the extraction and integration of information from multivariate imagery. This applies especially to *pixel level* image fusion techniques based on concepts of statistical pattern recognition. Using *dimension reduction* techniques by means of method for *feature selection* and *feature extraction*, lower dimensional data representations can be generated according to different objective functions. This way, fused images and other data representation can be generated with a minimum loss of significant information as well as those revealing interesting structures in the data. The combination of these concepts with scientific visualization techniques provides powerful options for the exploratory analysis of multivariate imagery. Hence, using these techniques, spatial relationships and hidden regularities can be detected visually.

This thesis is concerned with methods for the exploratory analysis of multivariate image data. It covers selected aspects from the field of visual data mining for scientific data with a special emphasis on visualization and image fusion techniques. It provides an overview of important scientific visualization techniques and color mapping methods. With respect to image fusion several linear and non-linear techniques from statistical pattern recognition based on unsupervised learning are discussed.

The techniques presented in this thesis have been implemented and evaluated in several biomedical research projects which have been carried out at Bielefeld University over the last years. Aside from studies of image data from DCE-MRI of the female breast, the analysis of fluorescence images from the screening of synthetic *stationary-phase promoters* using Green Fluorescent Protein (GFP) as a reporter protein is presented. While techniques for the exploratory data analysis are inherently application dependent, the focus of this work therefore lies on the development of generic concepts which are widely applicable and could be employed even in combination with irregular sampled and non-spatial data.

Next to color based visualization techniques, established image fusion methods such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), and the Self-Organizing Map (SOM), as well as the application of two recently developed approaches, the Hyperbolic Self-Organizing Map (HSOM) and Hyperbolic Topographic Mapping for Proximity Data (HTMP) is discussed.

1.1 Organization of this Thesis

Chapter 1 - Introduction

This chapter provides an introduction to the aims and scope of this work and gives an outline of the thesis.

Chapter 2 - Exploratory Data Analysis

Basic principles of visual data mining for the exploratory analysis of multivariate image data are discussed. For this purpose visual data mining is introduced as a part of a KDD process. This is followed by an overview of scientific visualization methods for image and volume data such as ray-casting, iso-surface, and non-photorealistic rendering techniques as well as a discussion concerning color in visualization. Besides color mapping techniques for uni-, bi,- and trivariate data a new approach for the generation of perceptually optimized color mapping techniques based on the CIELab color model and sRGB color specification is presented [cf. Saalbach et al., 2004].

Finally, in the context of image fusion a survey of techniques from statistical pattern recognition based on unsupervised learning is given.

Chapter 3 - Linear Techniques

An introduction to the theory and application of linear methods for the exploratory analysis of multivariate image data is given. Next to PCA, *Generalized Procrustes Analysis* for the comparison of subspaces and ICA are discussed. PCA is one of the most commonly employed techniques of multivariate statistics and can be used for the deviation of low-dimensional data representations preserving a maximum of the data variability. The applicability of PCA based methods is illustrated based on image sequences from breast cancer research. After an introduction to fundamental aspects of DCE-MRI, PCA is employed for image fusion and temporal change detection using data from the Department of Radiology of Munich University [cf. Saalbach et al., 2004, Twellmann et al., 2004a]. Visual data representations are generated by means of *score-plots* and *score-images* and a quantitative analysis of the results is carried out based on Receiver Operating Characteristic (ROC) statistics. Following the Generalized Procrustes Analysis, the stability of the PCA based analysis and the deviation of a consensus subspace for all data sets is discussed. Thereafter, ICA is introduced for the computation of statistically independent variables and as a complement to PCA. After a discussion of methods for stability assessment in the context of ICA, the applicability of this technique to DCE-MRI data is investigated.

Chapter 4 - Non-Linear Techniques

The SOM and the Topographic Mapping for Proximity Data (TMP) are introduced as well as the hyperbolic extensions of these techniques. After a consideration of theoretical aspects of the SOM algorithm, the method is employed in a framework for the screening of synthetic stationary-phase promoters which was developed in cooperation with the Fermentation Engineering Group, Bielefeld University [cf. Bettenworth et al., 2004, 2005, Miksch et al., 2005a, 2006]. Following this, a new approach for the *dynamic* visualization of multivariate image data based on a generalization of the SOM, the HSOM is presented. The Hyperbolic Data Explorer (HyDE) is demonstrated in the context of data from DCE-MRI, and is quantitatively evaluated using ROC statistics and measures of topology preservation [cf. Saalbach et al., 2005a]. Finally, it is demonstrated how the previously introduced concepts can be applied for image fusion at segment level based on proximity data. In a corresponding case study the Earth Mover's Distance (EMD) is employed as a dissimilarity measure for highly suspicious tissue segments from DCE-MRI. Based on HTMP a framework for the exploratory analysis of such data is introduced [Saalbach et al., 2005b].

Chapter 5 - Conclusion

In the last chapter the thesis is summarized and an outlook on possible extensions and further work is given.

Chapter 2

Exploratory Data Analysis

In recent years, multivariate imagery arises in an ever increasing number of domains as a challenging but also promising data source. While the acquisition of such data in remote sensing, medicine, or microscopy is often linked to specific questions, new imaging setups and novel application domains raise the need for techniques which allow for an undirected, i.e. exploratory data analysis. However, the special character of multivariate imagery limits the usefulness of many visualization and pattern recognition techniques and there is a desperate need for effective combinations of methods from both domains.

This chapter provides an overview of important concepts for the exploratory analysis of multivariate imagery. Therefore, the notion of exploratory data analysis is introduced within a process model for Knowledge Discovery in Databases (KDD). A process which ultimately aims at the extraction of useful knowledge from abstract and scientific data.

In the context of KDD the necessity for a close integration of data mining and visualization components has become apparent for quite some time. Here, more recently the term *visual data mining* has been coined to describe a paradigm for visual data exploration which takes advantage of both visualization and data mining techniques. However, these concepts have so far been considered mainly for the analysis of abstract rather than for scientific data.

The development of visualization techniques for scalar, vector, and tensor fields is traditionally the discipline of *scientific visualization*. Nevertheless, the visualization of multifield data is among the top research problems in this domain and the utility of these techniques for the data discussed in this thesis is rather limited. Following a general discussion about scientific visualization, a classification scheme for scientific data, based on the concept of *dependent* and *independent variables* is discussed. Thereafter, an overview of visualization techniques for (multiple) scalar fields will be given. In a section concerning color in visualization, color fundamentals, different device dependent and independent color spaces are discussed. This is followed by an overview of uni-, bi-, and trivariate color scales. Based on the CIELUV color space and sRGB color specifications a device independent and perceptually optimized color scale will be introduced [Saalbach et al., 2004].

Finally, different levels of image fusion are identified. Based on *dimension reduction techniques* such as *feature selection* and *feature extraction*, low-dimensional representations of even high-dimensional data could be generated for visualization purposes. The combination of such representations with established scientific visualization techniques empowers the development of new visual exploration techniques for scientific data.

2.1 Knowledge Discovery

Knowledge Discovery in Databases is a field of research which is concerned with the development of data driven methods to support analysts in the extraction of information from large data sets and in decision making processes. Formally, KDD refers to the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayyad et al., 1996a]. Concepts for knowledge discovery have been addressed in different scientific fields ranging from machine learning, artificial neural networks, and statistical pattern recognition in general to data visualization. Therefore, terms like data mining, information discovery, or information harvesting have been established to denote methods which allow for the detection of useful patterns [Fayyad et al., 1996a].

KDD is a process that consists of multiple phases which are employed in an interactive and iterative fashion. KDD begins with an investigation of the application domain under consideration and the identification of objectives and requirements. The process ends with the interpretation of the discovered patterns and the consolidation of the derived knowledge which often triggers new questions. Hence, KDD is a cyclic process that involves an extensive preparation of the data, the application of data mining methods and often visualization techniques for the representation of the results. Although often used interchangeably, in the context of KDD, the term data mining denotes more specific the application of a method for the detection of patterns and models. Therefore, data mining is usually considered as the most prominent phase in KDD. An overview of a basic KDD process according to [Fayyad et al., 1996a] is given in figure 2.1.

The first phase of a KDD process comprises the generation of a target data set by selecting variables and data samples. Preprocessing techniques for the purpose of noise removal and the handling of missing values result in a preprocessed data set. The data transformation component serves for the computation of features depending on the goals of the KDD process, e.g. by means of dimension reduction techniques. Yet, considered as another phase of data preparation, data transformation is closely related to the data mining component as it can involve the application of a broad range of methods from statistical pattern recognition. According to the goal of the process, data mining techniques for, e.g. classification, regression, clustering, summarization, dependency modeling, and change and deviation detection can be employed. During the last phase an interpretation and consolidation of the derived patterns takes place.

Basically, two different goals in KDD can be distinguished [Fayyad et al., 1996a]. In discovery oriented studies the focus is on the automated detection of patterns or the prediction of future behavior of certain entities. In less common verification oriented scenarios the evaluation of hypotheses is considered. While not directly depicted in the model of Fayyad et al. [1996b], concepts of exploratory data analysis play an important role in KDD [cf. Vesanto, 2002].

The roots of exploratory data analysis are commonly attributed to the work of Tukey [1977]. As a complement to the common statistical practice of testing proposed hypotheses, Tukey [1977] advocated the use of techniques which allow for a data driven

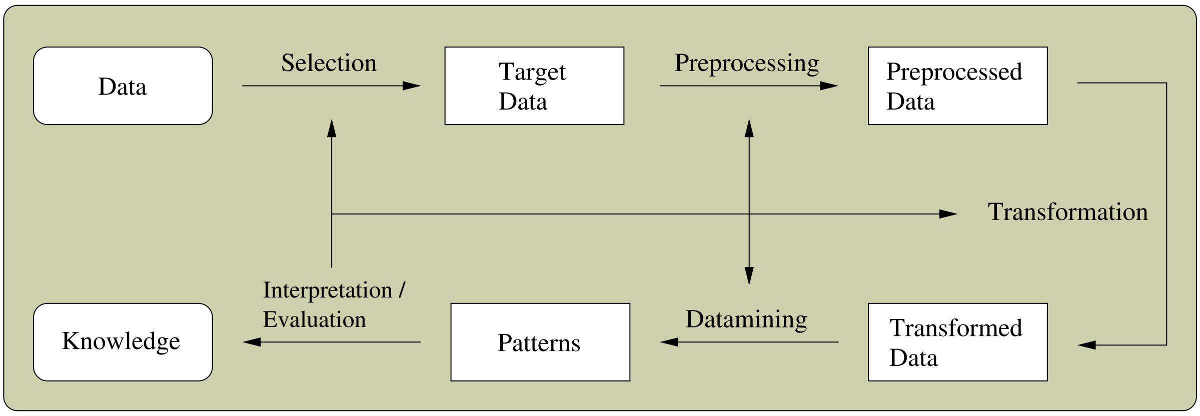

Figure 2.1: KDD process model [Fayyad et al., 1996a].

hypothesis generation. Traditionally, the notion of exploratory data analysis is closely related to the use of graphical techniques and is sometimes referred to as statistical graphics [NIST/SEMATECH, 2005]. While these techniques are most suitable for small and low-dimensional data sets, nowadays exploratory data analysis has become a synonym for an open-minded data examination without any prior assumptions or hypotheses. The employed techniques typically are interactive and visual.

Even though KDD is oriented towards the deviation of knowledge, techniques of exploratory data analysis are important for the purpose of data understanding. This relation becomes more obvious in recent KDD frameworks such as CRoss Industry Standard Process model for Data Mining (CRISP-DM). In this model, *data understanding* represents one of six fundamental phases in a knowledge discovery process. Here, understanding data involves the initial data collection, the assessment of the fundamental data properties, and the data quality. It is a process which serves the purpose of getting familiar with the data. It can be used for the identification of quality problems and for discovering first insights and developing hypotheses.

Despite of its strong relation to knowledge discovery in health and business applications, KDD has a long tradition in the analysis of scientific, for instance, spatial data [Fayyad et al., 1996b]. However, while predominantly employed for data reduction purposes in terms of object classification and detection, the utility of human abilities with respect to pattern recognition in KDD were obviously limited.

2.2 Visual Data Mining

The examination of huge amounts of patterns which can be generated by means of data mining techniques is one of the most challenging tasks in KDD. In order to enhance the efficiency of this process, visualization techniques have been routinely employed as an interface between analysts and data mining algorithms. However, over the course of time it has become more and more apparent that visual feedback is an important aspect within the entire KDD process [Fayyad et al., 2002].

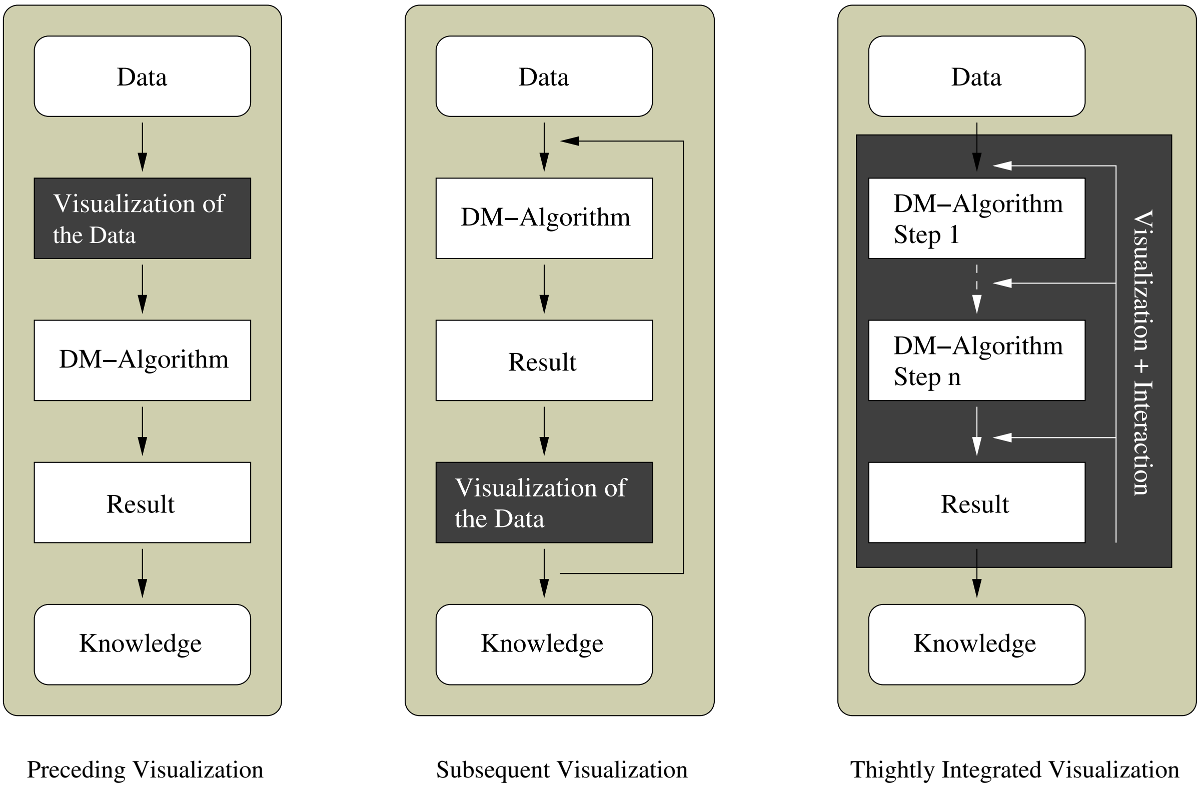

Figure 2.2: Overview of human involvement in visual data mining [Keim et al., 2005].

While visualization techniques have been employed in KDD predominantly for the presentation of the results, visual data mining aims at the integration of visualization techniques in the whole process of knowledge discovery. In the data analysis process, visualization techniques provide valuable options with respect to data understanding and hypothesis formulation as well as to confirmatory analyses. Since visualization in this context affects not only the representation of the data mining results but may provide an enhanced understanding of the employed algorithms. The more comprehensive term visual data mining has been adopted for the integration of visualization into the data mining process [Wong, 1999, Hinneburg et al., 1999].

According to [Keim et al., 2005], three different integration schemes can be distinguished (see figure 2.2). The application of a visualization method prior to a data mining technique can be referred to as *preceding visualization*. This can be carried out for an estimation of an appropriate parameterization of a data mining technique and for the restriction of the search space. In the context of *subsequent visualization*, the interpretation of the patterns which were generated by means of a data mining technique is facilitated using visualization techniques. The examination of the results can trigger the renewed application of an algorithm with a different parameterization. Furthermore, visualization techniques can be integrated even more closely into the data mining process. Thus, visualization techniques are not only applied prior to the application of a data mining algorithm or for the presentation of the results but throughout the entire process. Visualizations of intermediate results are employed for the selection of appropriate data mining techniques and parameterization based on the identified patterns and the domain knowledge of analysts. This approach can be referred to as *tightly integrated visualization.*

Following the visual information seeking mantra *Overview first, zoom and filter, then details-on-demand* [Shneiderman, 1996], the close integration of these techniques might be most suitable for the purpose of exploratory data analysis [Keim et al., 2004]. In visual data exploration the analysts start with getting an overview of the data. The identification of patterns leads to the application of visualization techniques which promote the representation of a subset of the data. Finally, an analysis of the data details is carried out. Some advantages of visual data exploration have been pointed out in [Keim et al., 2004]:

- Visual data exploration can easily deal with highly heterogeneous and noisy data.
- Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.
- Visualization can provide a qualitative overview of the data, allowing data phenomena to be isolated for further qualitative analysis.

Over the course of the last years, a large number of information visualization techniques for the representation of abstract data has been developed and much attention in visual data mining has focused on the analysis of such data [for an overview see Keim et al., 2004, de Oliveira and Levkowitz, 2003]. With an increasing complexity of scientific data sets a growing demand comes up for corresponding data exploration techniques. However, only a rather small number of scientific visualization techniques for the representation of multiple scalar fields exist. Furthermore, unlike for abstract data, spatial objects are often interrelated, i.e. objects are influenced by other, nearby objects [Keim et al., 2004]. Therefore, the combination of data mining and visualization techniques seems to be very promising for the detection of complex patterns within the data.

2.2.1 Scientific Visualization

According to the Oxford English Dictionary [Simpson and Weiner, 1998], visualization is the generation of a mental vision, image, or picture of (something not visible or present to sight, or of an abstraction); to make visible to the mind or imagination. In computational science visualization has a more specific meaning and may be defined as the use of computer-supported, interactive, visual representations of data in order to amplify cognition [Card et al., 1999]. Within this broad field of research two different areas can be identified: information visualization and scientific visualization. Historically, information visualization techniques are concerned with the representation of abstract data, typically given as a set of multivariate samples. Contrary, for visualizing measurements sampled from a multidimensional domain, e.g. from a physical phenomenon, techniques