# Visual Analytics for Improving Exploration and Projection of Multi-Dimensional Data

**Dissertation zur Erlangung des akademischen Grades eines Doktor der Ingenieurwissenschaften (Dr.-Ing.)**

**vorgelegt von**

**Matthias Jörg Schäfer**

**an der**

Universität Konstanz

**Mathematisch-Naturwissenschaftliche Sektion**
**Informatik und Informationswissenschaft**

Tag der mündlichen Prüfung: 4. August 2014

Referenten: Prof. Dr. Daniel A. Keim, Universität Konstanz
Jun.-Prof. Dr. Tobias Schreck, Universität Konstanz

# Abstract

In the last years visual analytics got an important research topic to keep track of the vast amounts of electronically stored data and gain new information out of the data. This thesis arose from several real application areas and deals with visual analytics of two data types, multi-dimensional time related event based data and multi-dimensional data without time stamp, which are very heterogeneously.

In the first part of the thesis a flexible approach to find significant events, event clusters and event patterns is introduced. The system has built-in functions for ordering of event groups according to the similarity of their event sequences, temporal gap alignments and stacking of co-occurring events. Three different case studies dealing with business process events, news articles and time related 3D data demonstrate the flexible capabilities of this approach.

In the second part an automatic and interactive approach for improving the quality of projections in terms of both structural preservation and class separation by feature selections and transformations is introduced. Quality measures for assessing the structural preservation quality and the visual quality of the projections are proposed. The effectiveness of the approach is evaluated by applying it to several widely used projection techniques using a set of benchmark data sets.

A data example for which it can be shown how well the two parts fit together analyzes a common data set. It shows the combination of both approaches and the benefit that can be achieved with them in a sequential visual analytics process. Furthermore there exists a close interaction between the visual and the algorithmic parts of the approaches and a combination of an algorithmically optimization with user interaction guides the user to find an optimal projection in terms of user satisfaction and the quality measures. This results in a task defined better projection via user interaction as a step-wise optimization.

But the approaches also cover other benefits like a descriptive real-time presentation of the measures visually and by numbers at once. Furthermore a selectable stress value visualization leads to a better understanding of the data exploration and the projection techniques.

**Keywords:** Visual Analytics, Visualization, Visual Analysis, Information Visualization, Exploration, Interaction, Search, Projection-based Data Analysis, Feature Transformation, Feature Selection, Event Data, Multi-Dimensional Data, Multimedia Data, Quality Measures

# Zusammenfassung

In den letzten Jahren wurde „Visual Analytics“ zu einem wichtigen Forschungsthema, um über die riesigen Mengen elektronisch gespeicherter Daten den Überblick zu behalten und neue Informationen aus den Daten zu gewinnen. Diese Doktorarbeit entstand aus verschiedenen realen Anwendungen und befasst sich mit „Visual Analytics“ zweier Datentypen, nämlich multi-dimensionalen Zeit-bezogenen Eventdaten und multi-dimensionalen Daten ohne Zeitstempel, die sehr heterogen sind.

Im ersten Teil der Arbeit wird ein flexibler Ansatz vorgestellt, um signifikante Events, Event-Cluster oder Event-Muster zu finden. Das System enthält Funktionen, um Event-Gruppen nach der Ähnlichkeit ihrer Event-Sequenzen zu ordnen. Außerdem können Zeitlücken auf unterschiedliche Weise und gleichzeitig auftretende Events gestapelt angeordnet werden. Drei verschiedene Fallbeispiele mit Business-Prozess Eventdaten, Nachrichten-Artikeln und Zeit-bezogenen 3D Daten zeigen die flexiblen Einsatzmöglichkeiten dieses Ansatzes.

Im zweiten Teil der Arbeit wird ein automatischer und interaktiver Ansatz vorgestellt, um die Qualität von Projektionen mit Feature Selektionen und Transformationen in Bezug auf die Erhaltung der Struktur und der Trennung der Klassen zu verbessern. Um die strukturelle Erhaltung und die visuelle Qualität der Projektionen zu messen, werden Gütemaße vorgestellt. Zur Evaluierung der Effektivität des Ansatzes werden verschiedene Benchmark Datensätze mit mehreren weithin bekannten Projektionstechniken untersucht.

Um zu zeigen wie gut die beiden Teile zusammenpassen, wird eine Anwendung mit einem gemeinsamen Datensatz betrachtet. Diese zeigt die Kombination der beiden Ansätze und den Nutzen, der aus einem sequentiellen „Visual Analytics“ Prozess gewonnen werden kann. Außerdem besteht eine enge Interaktion zwischen den visuellen und den algorithmischen Teilen der Ansätze und eine Kombination einer algorithmischen Optimierung mit der Benutzerinteraktion führt den Benutzer zu einer optimalen Projektion bezogen auf die Benutzerzufriedenheit und die Gütemaße. Das Ergebnis ist eine anwendungsbezogene Projektion, die durch Benutzerinteraktion schrittweise verbessert werden kann.

Die Ansätze zeigen aber auch weitere Vorteile wie eine anschauliche real-time Darstellung der Gütemaße gleichzeitig durch Zahlen und visuell. Außerdem führt eine auswählbare Stresswert-Visualisierung zu einem besseren Verständnis der Datenexploration und der Projektionstechniken.

**Schlagwörter:** Visual Analytics, Visualisierung, Visuelle Analyse, Informationsvisualisierung, Exploration, Interaktion, Suche, Projektion-basierte Datenanalyse, Feature Transformation, Feature Selektion, Eventdaten, Multi-Dimensionale Daten, Multimedia Daten, Gütemaße

# Contents

# Chapter 1

# Introduction

We are living in an information society with highly increasing data volume. Most of this data is saved electronically or is changed to be stored electronically. A white paper published from IDC [32] amounts the electronic data existing in 2006 to a total of 180 exabytes. By 2011, the amount of electronic data created and saved growed to 1,800 exabytes or by 10 times. That means an annual growth rate of nearly 60% respectively a doubling every two years. Assuming that the data is growing on by this factor we will get measureless electronic data in the future which results in “Data is the new (s)oil” (David McCandless [24]). Another important issue are the new possibilities in storing and processing of the data: On the one hand a vast amount of data can be stored easy and cheap these days and on the other hand this data bulk can be processed fast end efficient by standard computers with distributed systems without expensive high performance computers of former times. By these developments nowadays the expression “Big Data” has managed it to move from specific professional articles not only to the general technical press but also to daily newspapers.

All these trends make it extremely important to keep track of this high amount of data without losing the overview to get access and learn from the data to put the data into information and harvest new insights. Humans are not able to overview this vast amount of data and produce subjective errors in finding new information, for which reason automatic methods are essential. But humans have the ability to detect patterns, an asset that automatic methods lack. Hence a combination of automatic methods to visualize data and human interaction is preferable. The term visual analytics represents this kind of analysis and information mining from the data.

In this thesis I put the focus on showing that visual analytics can be used for improvements of exploration and projection approaches for multi-dimensional data types. The thesis is based on several papers, which have been written by me as author or coauthor during the last few years. The content that is taken of these papers is not specifically labeled by quotations.

***In the following I list the papers in which parts of the thesis were published in:***

1. M. Schaefer, F. Wanner, F. Mansmann, C. Scheible, V. Stennett, A. T. Hasselrot and D. A. Keim. Visual Pattern Discovery in Timed Event Data. In Proceedings of Conference on Visualization and Data Analysis, 2011, see [85].
   ***The contributions:***
   The main contribution is a novel and flexible system for analyzing timed event data, that includes advanced features such as similarity ordering, temporal gap alignment and stacking of co-occurring events. The effectivness of this system is demonstrated on two characteristically different case studies.

2. M. Schaefer, L. Zhang, T. Schreck, A. Tatu, J. A. Lee, M. Verleysen and D. A. Keim. Improving projection-based data analysis by feature space transformations. In Proceedings of VDA 2013, 2013, see [86].
   ***The contributions:***
   The main contributions are an improved projection-based data analysis framework which transforms the feature vector space by extending the identified relevant features, as well as a new quality measure to automatically evaluate projection displays, integrating structure preservation and clutter avoidance. An evaluation of the effectiveness of different feature space transformations strategies, as a guideline for further development demonstrates the usefulness of the concept.

3. D. Perez, L. Zhang, M. Schaefer, T. Schreck, D. A. Keim and I. Diaz. Interactive Visualization and Feature Transformation for Multidimensional Data Projection. Proc. EuroVis Workshop on Visual Analytics Using Multidimensional Projections, 2013, see [77].
   ***The contributions:***
   The main contribution is a novel visual analytics approach for improving the quality of multi-dimensional data projection, including a quality evaluation. It works with a combination of dimension selection and feature transformation steps with an interactive visualization, in particular using a parallel coordinates view for the dimensions of the data.

***I also contributed to two student theses, which include contents of this thesis:***

1. A. Tatu. Multimedia Datenbank Retrieval: Suche in Bilddatenbanken mit Hilfe klassifikations-basierter Featureselektion, 2009, see [95].

2. M. Regenscheit. Multimedia Datenbank Retrieval: Visuelle & Interaktive Analyse von Multimedia Daten, 2010, see [79].

***Furthermore I contributed to the following papers, that where published but not part of the thesis:***

1. J. Krause, M. Spicker, L. Wörteler, L. Zhang, M. Schaefer and H. Strobelt. Interactive Visualization for Real-time Public Transport Journey Planning. In Proceedings of SIGRAD 2012, 2012.

2. B. Bustos, T. Schreck, M. Walter, J. M. Barrios, M. Schaefer and D. A. Keim. Improving 3D Similarity Search by Enhancing and Combining 3D Descriptors. Multimedia Tools and Applications, 2012.

3. M. Schaefer, F. Wanner, R. Kahl, L. Zhang, T. Schreck and D. A. Keim. A Novel Explorative Visualization Tool for Financial Time Series Data Analysis. International UKVAC Workshop on Visual Analytics, 2011.

4. F. Wanner, M. Schaefer, F. Leitner-Fischer, F. Zintgraf, M. Atkinson and D. A. Keim. DYNEVI - DYnamic News Entity VIsualization. In Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010), 69–74, 2010.

5. D. A. Keim, J. Kohlhammer, G. Santucci, F. Mansmann, F. Wanner and M. Schaefer. Visual Analytics Challenges. In Proceedings of eChallenges 2009.

6. P. Bak, M. Schaefer, A. Stoffel, D. A. Keim and I. Omer. Density Equalizing Distortion of Large Geographic Point Sets. Journal of Cartographic and Geographic Information Science (CaGIS), 36 (3): 237–250, 2009.

7. D. A. Keim, P. Bak and M. Schaefer. Dense pixel displays. In Encyclopedia of Database Systems, Springer-Verlag, 789–795, 2009.

8. P. Bak, D. A. Keim, M. Schaefer, A. Stoffel and I. Omer. Visual Analytics Using Density Equalizing Geographic Distortion. In Geospatial Visual Analytics Workshop at Giscience, 2008.

# Chapter 2

# Visual Analytics

## 2.1 Visual Analytics Framework and Definitions

The field of information visualization and visual analytics became more and more a hot research topic in the last 20 years. It developed from pure visualization systems, see [16] for an early collection of classic information visualization papers, to real user interaction visual analytics systems. Nowadays the expression "Big Data" got a buzz word and the amount of data being collected and stored by commercial organizations is increasing at a fast rate; hence implementing intelligent and flexible information visualizations, see [57], and visual analytics systems is important in many business cases.

According to Shneiderman's information seeking mantra [91]:

*Overview first - Zoom and Filter -*
*Details on Demand*

as well as the visual analytics paradigm from Keim et al. [55]:

*Analyze First - Show the Important - Zoom, Filter and Analyze Further -*
*Details on Demand*

a visual analytics framework has to be build up like shown in Figure 2.1 which is an extension of the visual analytics process that can be found in [54].

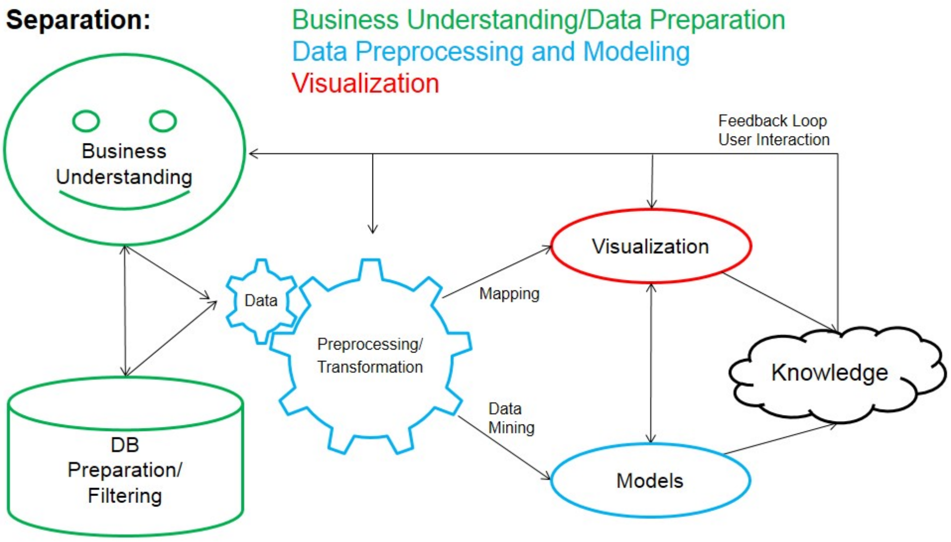


Figure 2.1: This schematic diagram shows the steps of the visual analytics framework. The steps in this framework are separated in business understanding and data preparation (green), algorithms for aggregation, data mining, clustering in the data preprocessing and modeling (blue) and the visualization (red) to get new knowledge out of the data. A feedback loop flows back to all steps which is important to allow and combine user feedback with interaction in all steps to improve the analysis result.

A framework built up including the ideas of this figure contains a user guided feedback loop and interaction between the models and the visualization as postulated in the visual data exploration pipeline from Keim et al. [53]. This means first to allow the user an overview and global analyzing, with showing the important characteristics of the data. Based on that the user can zoom, filter and analyze further to get details on demand. It is important to allow and combine user feedback with interaction in all steps to judge and improve the analysis result. Examples for interactive data visualizations, designs and real-world use cases can be found in [93] and [109]. In this thesis flexible systems for visual analytics of different types of multi-dimensional data with multiple features and an automatic feature selection have been implemented and will be discussed.

Figure 2.1 highlights the different parts, affecting such a system: First, the business understanding and data preparation (green), e.g., database extracts. This implies the main difference to the visual analytics process in [54]. For a good result in the end it is extremely important to add this step for a good data basis because the results can only be good when the data has a good quality. Mostly different people or departments with different expert knowledge in companies are responsible for that and it is a big challenge to coordinate them. Second, algorithms for aggregation, data mining, clustering in the data preprocessing, transformation and modeling (blue) and third, the data mapping and visualization (red) to get new knowledge out of the data. The challenges rising from this tasks will be discussed particularly in Chapter 2.2. Throughout this thesis the challenges for different areas of applications, with different data types, respectively different data of the same type and different visualization and interaction techniques will be shown. All tasks have the same goal and result in the same output: new insights and knowledge.

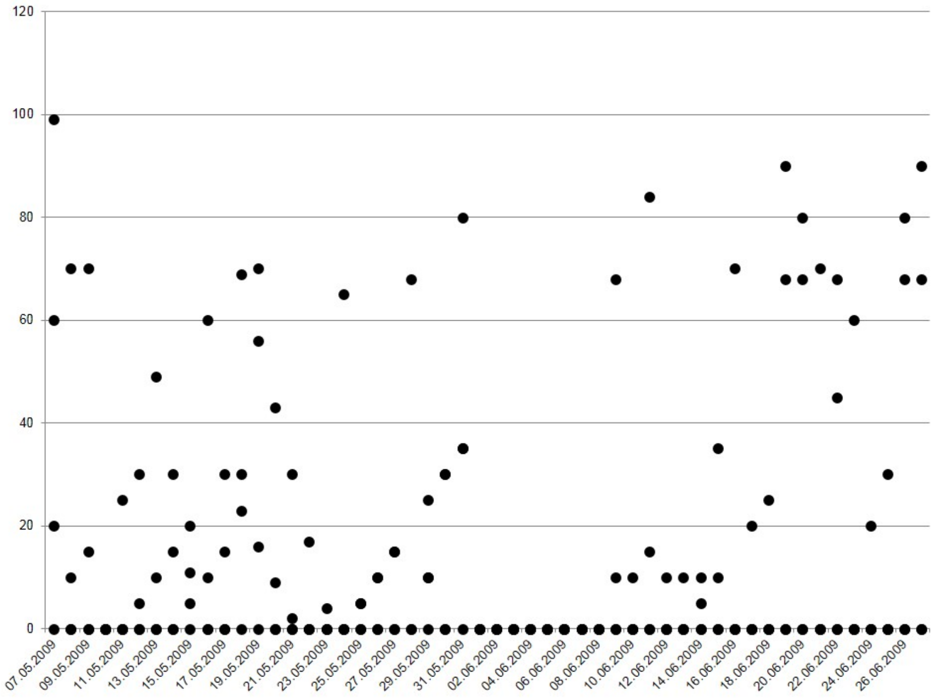


Figure 2.2: The figure shows an example for time related event based data. Visualized is a sequence of events and 0 to N events occur with multiple values per time stamp. The challenges in visualization are display-wasting, overplotting and low information.

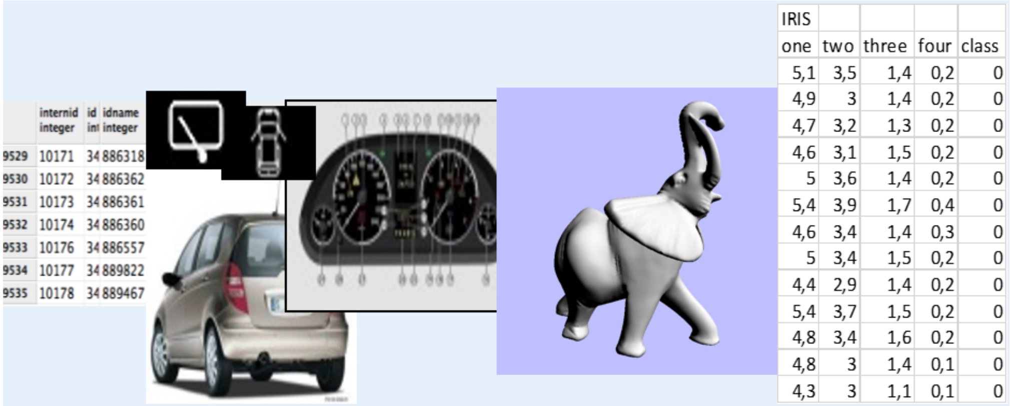
The first specified research project focuses on analyzing multi-dimensional event time series data and the second on multimedia and multi-dimensional feature data. The definition of event data turned out diverse in science over the years. Events are very task and domain dependent and can be designated abstractly as a change of a certain status. In this thesis an event is defined as a single, time-stamped item. Event time series data is a sequence of events which occurs 0 to N times with multiple values per time stamp, see Figure 2.2 as an example.

The challenges in visualization are the low information content through display-wasting, which can be seen through the predominant white space that demands the bigger part in Figure 2.2. In addition to that overplotting is a problem when events occur with multiple values per time stamp because they are plotted on the same point in the visualization. In Chapter 3 these challenges are overcome with using a pixel based visualization that takes the advantage of the whole space on screen and can view events and different events that are occurring at the same point in time with a different coloring and different alignments. It will be shown that the information content rises through the presented visualization approach.

The multimedia and multi-dimensional data analyzed in this thesis include audio, image, 3D, bioinformatics or video data. The common characteristic of these data types is that it is possible to calculate high-dimensional content-based feature vectors that represent the multimedia data, see Figure 2.3. The proposed approaches can be applied to all kinds of data, which have this high-dimensional format.

All these types of data are difficult to visualize. So the challenge is to analyze and visualize this data in such a way that it is possible to gain new insights and extract knowledge from the data. The goal is to achieve a close interaction between the algorithmic and the visual parts of the visual analytics approaches and a high flexibility so that the user can delve deep into the data to get details on demand and can also give feedback to steer new analysis tasks.

Applications in fraud detection, finance, multimedia and 3D projects will be used to demonstrate the flexibility of the approaches in establishing a separation between the data and the visualization and also the ability to adapt them to several domains. Visual analytics algorithms have been developed to



| | internid<br>integer | id<br>int | idname<br>integer |
|---|---|---|---|
| 9529 | 10171 | 34 | 886318 |
| 9530 | 10172 | 34 | 886362 |
| 9531 | 10173 | 34 | 886361 |
| 9532 | 10174 | 34 | 886360 |
| 9533 | 10176 | 34 | 886557 |
| 9534 | 10177 | 34 | 889822 |
| 9535 | 10178 | 34 | 889467 |

| IRIS | | | | |
|---|---|---|---|---|
| one | two | three | four | class |
| 5,1 | 3,5 | 1,4 | 0,2 | 0 |
| 4,9 | 3 | 1,4 | 0,2 | 0 |
| 4,7 | 3,2 | 1,3 | 0,2 | 0 |
| 4,6 | 3,1 | 1,5 | 0,2 | 0 |
| 5 | 3,6 | 1,4 | 0,2 | 0 |
| 5,4 | 3,9 | 1,7 | 0,4 | 0 |
| 4,6 | 3,4 | 1,4 | 0,3 | 0 |
| 5 | 3,4 | 1,5 | 0,2 | 0 |
| 4,4 | 2,9 | 1,4 | 0,2 | 0 |
| 5,4 | 3,7 | 1,5 | 0,2 | 0 |
| 4,8 | 3,4 | 1,6 | 0,2 | 0 |
| 4,8 | 3 | 1,4 | 0,1 | 0 |
| 4,3 | 3 | 1,1 | 0,1 | 0 |

Figure 2.3: The figure shows examples for multi-dimensional feature-based data and how this is stored in a high-dimensional data format.

overcome challenges such as the visualization of different alignment strategies to deal with temporal gaps, aggregation and change calculations of the data, pattern detection as well as automatic and interactive feature engineering.

## 2.2 Scopes and Challenges in Visual Analytics of Multi-Dimensional Data

The two parts of this thesis in Chapter 3 and 4 differ in the needs and requirements of the specified two data types. This allows to generate very specific solutions. But there will also be shown an application that illustrates the fruitful common usage of the two parts and how the two parts can be used to perform a visual analytics process on a common data set to visualize it and get new insights.

Therefore an example which deals with both kind of data shows the combination of both approaches and the benefit that can be achieved with them in a sequential visual analytics process. It will also be shown that the approaches are very modular and flexible because of the separation between data and the visualization. There is also a close interaction between the visual and the algorithmic parts of the systems.

To put focus on the challenges of each of the data types and the related tasks the particular chapters start with related business questions and needs with "Challenges and Needs" chapters. The similarity of the solutions to deal with these needs is to combine different algorithms, visualizations and interaction techniques to get new insights and show the usefulness as well as the assignability to other domains. This is arranged via evaluations and use cases, at which the research challenge and scientific benefit of this work is mostly shown in the transfer to several domains and their data. This means the systems can visualize and get new findings for any data fullfilling the above specifications. But nevertheless the scopes and challenges in dealing with this multi-dimensional data types as explained result from the same starting point, where vast amounts of data exist from which in information should be extracted or better new so far unknown findings should be detected. The visual analytics approaches proposed in this thesis are motivated through business questions.

Both works are combinations between research and real applications. Because the visual analytics expert is not mandatory an expert in the real application domain, this always should include getting a business understanding of the specific domain for the visual analytics expert, as well as an intensive exchange with the domain experts. After that a collective data preparation and preprocessing has to be done before the analysis and the visual analytics part can start. For these analyses the role of the visual analytics expert is to select, develop and provide appropriate visualization and analysis methods for the representation and investigating of the multi-dimensional data. This can be via scatter plots, parallel coordinates, pixel-based systems, projection methods, Euler diagrams etc. as well as clustering algorithms, dimension reduction techniques and so on. But beyond that insider knowledge from the domain expert is important, too, because the domain expert knows the data and the specific characteristics in his familiar domain. Therefore the visual analytics expert only can provide the appropriate approaches and tools and then work closely together with the domain experts or enable them to use the developed systems on their own.

*I compare this to an electrician who first puts on the light in a big room full of old books, enabling the antiquarians to see the real treasures, ideally with highlighting them in a special manner without knowing them before.*

# Chapter 3

# Visual Exploration of Multi-Dimensional Event Data

Parts of this chapter are based on the following paper [85]. I took the lead and responsibility of the text and adapted it for my thesis. Chapters with a high description portion of specific authors are named explicitly.

- M. Schaefer, F. Wanner, F. Mansmann, C. Scheible, V. Stennett, A. T. Hasselrot and D. A. Keim. Visual Pattern Discovery in Timed Event Data. In Proceedings of Conference on Visualization and Data Analysis, 2011.

## 3.1 Challenges and Needs

In this chapter the focus is on analyzing all kinds of timed event based data, Chapter 3.1.1 gives a detailed defintion of this kind of data. The goal was to implement a visual exploration system specifically targeting this data type. Case studies and examples in Chapter 3.4 will show the flexibility of the approach that enables the user via a visual interface to find significant events, event clusters and event patterns. The characteristically different case studies dealing with business process events, news articles and other data, namely time related 3D data, demonstrate the capabilities of the system to explore event data.

The basic system was implemented during a research project with *Lloyds*

*Banking Group, Wolverhampton, England* (see Chapter 3.4.1) and extended afterwards. The project's needs showed clearly that business processes have tremendously changed the way large companies conduct their business: The integration of information systems into the workflows of their employees ensures the company to guarantee a high service level for the customers and thus a high customer satisfaction. One core aspect of business process engineering are the events that steer the workflows and trigger internal processes. Strict requirements on interval-scaled temporal patterns, which are common in time series, are thereby released through the ordinal character of such events. It is this additional degree of freedom that opens unexplored possibilities for visualizing event data.

### 3.1.1 Definition Event Data

Temporal events occur in an extremely wide range of applications in business, government, and science. While some of these events can be aggregated over time in a meaningful way and thus be presented in time series visualizations, other application scenarios require each event to be visible. In addition to that, events often do not uniformly spread over time, but tend to be strongly biased. If any or both of these two characteristics are in the data, time series visualizations typically degrade, which means that a lot of display space is wasted or/and not all events can be displayed due to an overlap problem.
To systematically study event data first some related basic terminology for events, their properties and associated analysis tasks are defined and then the solution to the above problems is outlined.

**Event:** *An event is a single, time-stamped item.*

A data point in time is considered as an event, which can be a time-stamped news article, a system event or any measured value at specific points in time. This coincides with the definition in EventSummarizer [58] or Mannila et al. [68]. Galton and Augusto call such kind of event an atomic event [31]. Guralnik and Srivastava define an (atomic) event as a change of behavior of a dynamic phenomenon [35]. For the visualization only the change of the time-reference of an event is relevant.

Different event data sets display different properties. Thus, for a more systematic analysis, they are categorized as follows:

**Event Sequence:** *An event sequence is a set of events that are ordered in time.*

**Event Episode:** *An event episode is a set of events that are time-stamped.*

In [68] there is a distinction between event sequences and event episodes. These notions are used but comprehended in another way: An event sequence is a set of events that are ordered in time. Thereby, the ordering is the important property. Whereas an event episode is a set of events that are time-stamped and therefore the distance between the atomic events matters.
Under the assumption that every event has an assigned value for some dimensions of its metadata, event data can be further refined into a) *time-synchronous event data*, in which an accurate time-stamp is important, b) *ordinal event data*, where the ordering of the events according to time or metadata plays an important role, c) *aggregateable event data*, which can be summarized for a particular interval, and d) *hierarchical event data*, where the grouping is defined based on a hierarchical structure in the metadata.

### 3.1.2 Analysis Tasks

This chapter looks at the special needs for the relevant analysis tasks, which are performed and shown in the evaluation in Chapter 3.4. To foster a better understanding of this analysis tasks for event data, the terms significant event, event cluster, and event pattern are defined.

**Significant Event:** *A significant event is a single event that is interesting for some reason.*

**Event Cluster:** *An event cluster is a set of events that are considered as being similar to each other. This may, but not necessarily, include similarity in time.*

**Event Pattern:** *An event pattern is an event sequence or episode that shows some interesting regularity with respect to certain properties.*

The specific visualization is designed to support an analyst in his task to search for event clusters, event patterns and significant events. Other work, such as [68] focused on finding frequent episodes. An event pattern is a sequence or an episode that shows some interesting regularity with respect to a certain property.

Time series visualizations heavily depend on the fact that the displayed data can be aggregated or are spread sufficiently in time so that no overlap occurs. However, for many practical applications neither of these properties hold since many events occur at the same time or long periods elapse without event activity. In such a case, time series visualizations typically degrade, which means that a lot of display space is wasted while still not all events can be displayed due to an overlap problem. The proposed event data visualization tackles exactly these two shortcomings by rendering each atomic event and by abstracting or leaving out long temporal gaps in the representation. Thereby, the method has proven to be a flexible approach for finding significant events, event clusters and event patterns.

In particular, the first case study in Chapter 3.4.1 deals with business process events in fraud detection where the ordinal character of the events is of importance. In this case, the approach's capability to deal with event sequences is demonstrated, which are ordered but whose absolute temporal reference is irrelevant for the analysis. Based on real data from a bank's mortgage fraud database, it was possible to find several event patterns, such as potential fraud cases of suspicious solicitors, a suspicious bank account shared by several fraudulent customers, a systematic mortgage application pattern of one customer and potential future risks on book.

The second case study in Chapter 3.4.2 is about sentiment analysis in news blogs. Hereby, time-synchronized event episodes and search for significant events, clusters and patterns therein are considered. Using political RSS news feeds about the U.S. presidential election in 2008, it will be shown that significant events, such as a positive denial of an obvious scandal, event clusters such as feeds reporting very similar about one candidate, and event patterns like emotional debates can be identified. In another analysis task for evaluating the usefulness of the approach in Chapter 3.4.3 a visual cluster analysis on time related 3D data is performed, thus event clusters and patterns are relevant.

This will also show the approach's potential of interacting of the user with the system and use it for further analysis in projection an cluster analysis like performed in Chapter 4.

## 3.2 Related Work

The related work in this chapter is divided in two parts: first, it will be discussed for data with temporal aspects that was analyzed in time series visualizations, second, the younger field of visual event analysis will be presented.

### 3.2.1 Time Series Visualization

Time series are an important type of data encountered in almost every application domain. The field has been intensely studied and received considerable research attention, especially with respect to financial and business applications [2, 3, 52, 56]. Concerning particular analysis tasks, not only highlighting patterns is an important aspect, but also arrangement of multiple time series to support comparison between several monitored items as studied in [38]. Hochheiser and Shneiderman's *Time Searcher* system [42] uses traditional line graphs, which can be analyzed using a dynamic query interface. It includes specification of ranges of values and time intervals, query-by-example, queries over multiple time-varying attributes, query manipulation, pattern inversion, similarity search, and graphical bookmarks.
Other application scenarios deal with the problem of identifying patterns on larger time scales by using traditional metaphors for visualization, such as clocks [7, 110] or calendars [4, 102]. Yet another common approach to cope with time are small multiples (e.g., [69, 78]) or multi-resolution representations [39,62,70]. A broader overview of visualization methods for time-oriented data can be found in [1]. A lot of this work in time series visualization only represents aggregated values, whereas each atomic event is important in many applications of timed event data. In sentiment analysis of news, for example, an averaged sentiment score has only little meaning since it can hide important characteristics of the underlying event data, such as a controversial debate with very negative and positive opinions at the same time.

### 3.2.2 Visual Event Analysis

Event-based systems have a broad application range in research and the industry with an application scope varying from genome research to business intelligence and analysis. *Event Tunnel* [94] is one such event analysis system for business processes. In these tunnel plots, the inner circles contain old events, whereas new events are plotted larger on the outer circles. A single business process is thus represented through a chain of connected dots from the inner to the outer circles. The angular axis can be used for assigning an additional data dimension of the business process. Alternative layouts are tunnel plots with two assignable axis and scatterplots. Other variables of an event, such as the type, status, etc. can be encoded using the dots' color and border, or by altering the shapes of the event representations. *WireVis* [18] introduces a system which also deals with fraud detection in the bank sector like one of the case studies later does. The authors present a tool with different visualizations based on identifying specific keywords within wire transactions. It is very useful for advanced investigators in the bank who are able to detect accounts and transactions with suspicious behavior. The tool was implemented to deal with this very specific task and it was planned to integrate it in the bank's daily work flow. *Gapminder* [80] comes quite close to an event analysis system. Its animated scatterplot visualization displays a snapshot of two preset variables for each country in each time interval. Single countries can be marked in order to track the event episode of a country's development over time. This is visually encoded through a number of connected dots in the scatterplot. While old events of unmarked countries disappear in the animation, the marked country's events are maintained throughout the animation. The geographic research community defines events through both temporal and spatial references, which results in special requirements for geographic visualization. One example in this field is the space-time cube [33], which maps spatiotemporal events using geographic coordinates on the first two dimensions and time on the third dimension. Atomic events are then connected with connecting lines and form event episodes. Animation can be used as an alternative visual representation as shown for telecommunication network and service events in *SWIFT-3D* [60]. Animation can be discarded as a visualization option for event data since it is hard to track large quantities of events appearing and disappearing. While these systems and publications have demonstrated some of the potential that visual event analysis can have in specialized application domains, specific visualizations for that vast amount of data are still “in their infancy” [94].

Newer approaches can be found in the *LifeFlow* visualization tool [112], that was developed for analyzing point-based process log data. It combined the list-based display of its predecessor, *LifeLines2* [107], with an aggregated display that shows a summarization of the whole data set in a single view. Recently a system was published in “Temporal Event Sequence Simplification” [72], in which the *EventFlow* system [73] was refined because it was visually so crowded when it was loaded with big data. It should be an important capability for all systems dealing with timed event data to be able to deal with big data sets, as well as different data sets. The approach proposed in this thesis therefore presents a more general way for timed event data, which is demonstrated on two characteristically different event data sets.

## 3.3 Multi-Dimensional Event Data Exploration System

With the availability of large storage devices, huge memory chips and multi-core CPUs, computers for capturing and storing massive amounts of data have become an affordable commodity even for small businesses. Likewise, running resource intensive data mining algorithms is mostly not a problem anymore. However, drawing the correct conclusions and gaining insight into raw data and results of data mining algorithms is still an essential and often unsolved challenge. Visual analytics aims at bridging this gap between automated analysis techniques and the human analyst by combining the former with human-interpretable visual interfaces.

In this chapter it is demonstrated how the system supports the interaction between the data mining and the visualization techniques on the way from data in the database to new insights. By solving real application problems using both automated and visual techniques, it will be shown how significant events, clusters and patterns can be identified.

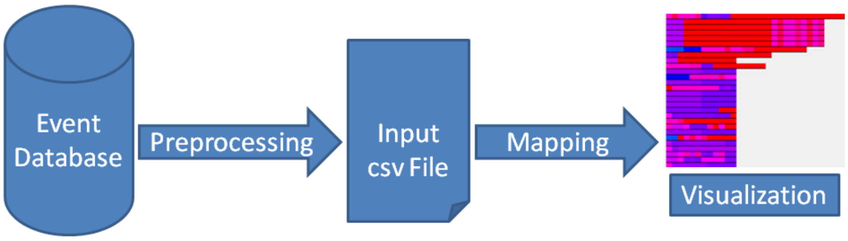


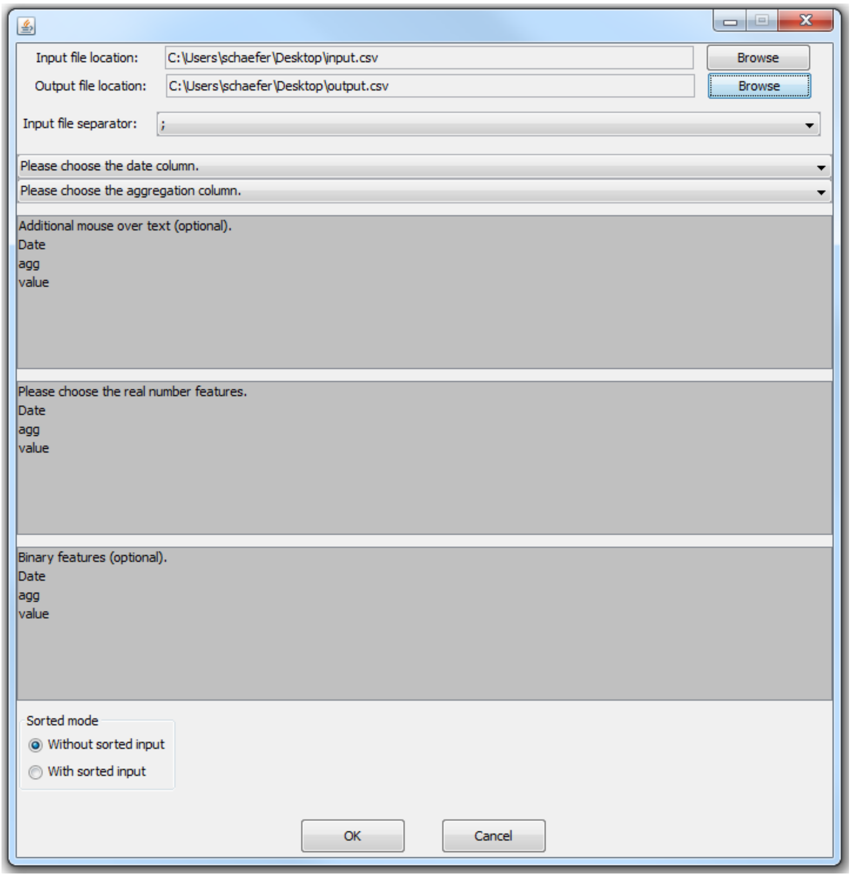
Figure 3.1: System pipeline with a preprocessing and a mapping step to achieve a high flexibility of the approach (published in [85]).

Figure 3.1 shows the system pipeline, in which events and their associated metadata are stored in a database. There are two user-driven processes: a) first the preprocessing step defines which attributes of the data set are used for grouping event into sequences or episodes, and b) the visual mapping step, which assigns visual properties of the representation to dimensions of the event data. With this approach a high flexibility in processing and visualizing different kinds of event data is received. Note that visual analytics approaches, and in particular this system, are often developed for advanced users who have domain knowledge and the use cases show its successful application for visual pattern discovery in event data.

### 3.3.1 Data Preprocessing and Mapping

The basis for the preprocessing algorithm is data which is stored in a database. As shown later this provides great flexibility for creating data for the preprocessing step with database tools and methods. The strength of this approach is that solid database functions, such as ordering, filtering etc. can be used. In the system's preprocessing step the relevant columns are chosen and the data is aggregated and transformed by the system into input files for the visualization. It is also possible to reload the preprocessed data into the database, processing the data in the database and setting new flags, which can be visualized afterwards.

A sequential processing of the data allows it to deal with large volumes of data. The data is aggregated for each entry, which is user defined like bank accounts, news feeds/entities or time related 3D data in the examples. The resulting file contains only the entries and their related events in the lines in a flat-file csv format. This is used as input for the mapping step to create the system's visual output and allows a fast processing of large amounts of data in the mapping step, too. Through this sequential processing in both steps, the preprocessing and the mapping, limitations only depend on the assigned memory.



Input file location: C:\Users\schaefer\Desktop\input.csv

Output file location: C:\Users\schaefer\Desktop\output.csv

Input file separator: ;

Please choose the date column.

Please choose the aggregation column.

Additional mouse over text (optional).
Date
agg
value

Please choose the real number features.
Date
agg
value

Binary features (optional).
Date
agg
value

Sorted mode
● Without sorted input
○ With sorted input

Figure 3.2: Flexible preprocessing interface of the multi-dimensional event data exploration system.

With this settings the system can be configurated very flexibly. Figure 3.2 shows the preprocessing interface, in which the database input and output file locations have to be selected. The system can deal with all kinds of input file seperators (tab, semicolon, @, ...). The date column selects the time stamp of the event out of the database file, whereas the aggregation column specifies the attribute for the aggregation. Detail levels on demand can be chosen in an additional mouse over text and there can be selected unlimited attribute columns for real number features as well as binary features, that will be shown in the visualization afterwards via the described visual variables and flags that show the binary features.

An optional sorting possibility mode completes the preprocessing. This means the entries are sorted according to a similarity algorithm that groups together entries with similar events. For getting a fast and at the same time very beneficial result, the algorithm searches step by step for the next similar entry since he passed through the whole data set. The in this manner generated outputfile is used in the visual analysis interface as described in Chapter 3.3.2.

### 3.3.2 Visual Analysis Interface

The events are represented through rectangles that are colored according to categorical, ordinal or interval-scaled metadata. As stated, the system is supplemented by an automated ordering, which places similar event groups next to each other in order to support correlation analysis. Further features are temporal gap alignment and stacking of co-occurring events.

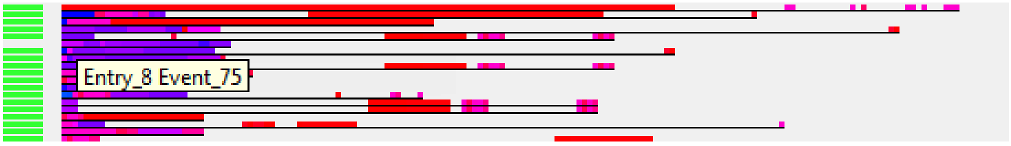


Figure 3.3: Multi-dimensional event data exploration system with unordered entries (published in [85]).

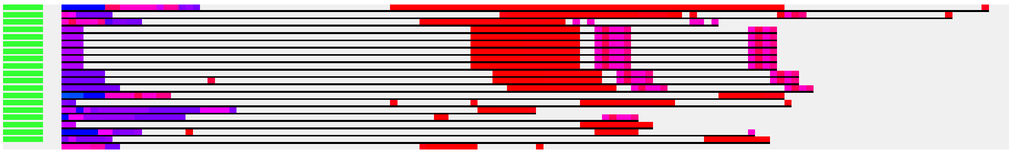


Figure 3.4: Multi-dimensional event data exploration system with an ordering and clustering of entries according to the similarity algorithm (published in [85]).

The system's visual output is shown for the 20 first entries of an event database and their related events in Figure 3.3. Each line starts with a flag (green or no flag in this example) and represents an entry and its related events. The events are colored according to their defined value. This flexible user-controlled mapping can be easily adjusted to the application and task. For coloring, several different color maps have been implemented, so the most convenient coloring scheme can be chosen for a specific analysis task. Hovering the mouse over an event in the visualization triggers a yellow box with text describing the event as shown in Figure 3.3. The displayed text can be defined flexible in the preprocessing step using metadata from the database. In addition to that the user can add special flags to the entries for faster identification. In Figure 3.3 all entries except the sixth one have a flag, which can be seen by the green coloring at the beginning of each line. Flags can be defined easily and flexible in any number in the preprocessing step and help the user to classify the entries. They also can be used for the visual cluster analysis, which is shown later in Chapter 3.4.3.

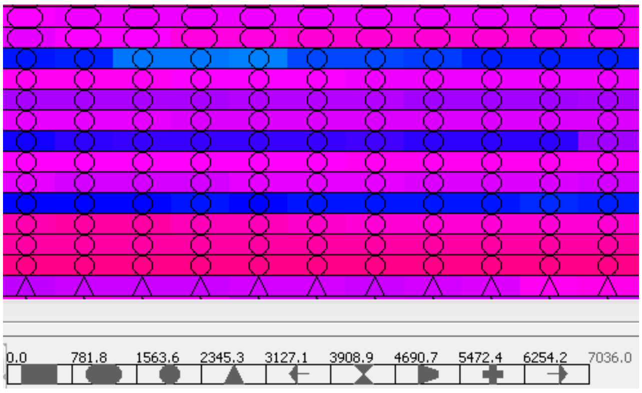


Figure 3.5: Multi-dimensional event data exploration system: An additional attribute is mapped to shapes (circles, triangles, etc.).

To show the flexibility of the system in terms of visualization theory, more visual variables, referred to Bertin [6], have been implemented and more detail levels on demand have been included. The user can select these levels according to his needs via the visual analysis interface. Figure 3.5 shows an example of a visualization using the additional visual variable shape for an extra attribute.

Once loaded in the system the user can interactively change the mapping of the attributes to the visual variables.

### 3.3.3 Advanced Features

On top of the basic visual analysis system advanced features which support the user in his visual analysis task have been implemented. First of all the ordering and a clustering of the entries based on similarity of the event patterns is provided. Figure 3.4 has the same data basis as Figure 3.3 but an ordering step is included in the preprocessing algorithm, which groups together entries with similar event patterns.

The algorithm runs linearly through the entries starting with the first one and then searching for the most similar entry using the Euclidian distance measure. A not equal length of the entries is penalized, so that entries with similar length are also placed together in a certain degree. Then the next similar entry to the one that was found by this prodeeding is detected and this procedure is repeated until all of the entries are ordered. This proceeding is very fast but of course alternatives are possible, that search for similarities locally and group together entries in another way. But the described practice worked well and it can be seen in Figure 3.4 that similar entries are placed together. Again the first 20 results are shown. Events with the same patterns are clustered together. Another effect is that entry 6 without the green flag as shown in Figure 3.3 is not in the first 20 entries in Figure 3.4 anymore. This is because of the dissimilarity of its event pattern to the others. Therefore the last entry with no green flag appears in the result set of Figure 3.4. This feature helps the user to find entries with similar event data but different flags respectively classes.

Another feature of the system are different alignments of the events. This is important because, as stated before, event data often does not uniformly spread over time, but tend to be strongly biased. For dealing with this problem the user has three different options to handle temporal gaps between the events. Figure 3.6 shows them: The top visualization shows one gap for each point in time where no events occurs. The middle visualization reduces this sequence of gaps to only one gap, independent of the sequence's length. The bottom visualization excludes gaps completely.

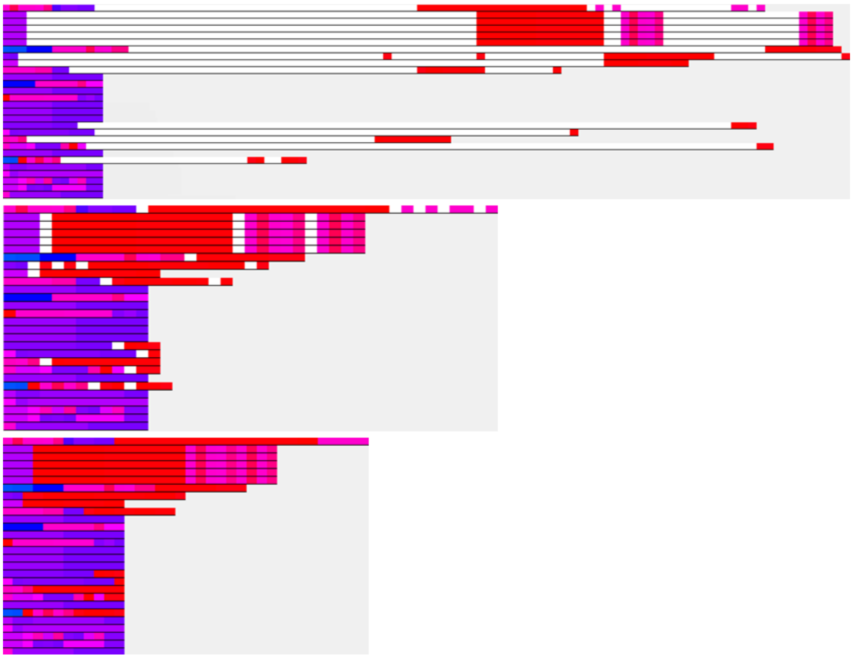


Figure 3.6: Three different alignments strategies to deal with temporal gaps in event data. Top: visualize all gaps, middle: visualize only one gap, bottom: visualize no gaps (published in [85]).

Stacked events in y-direction are an alternative alignment to deal with the occurrence of more than one event at the same time. So far the previous shown visualizations placed all events of one entry one after each other in one line. This leads to lines with arbitrary length. Figure 3.7 shows an alternative approach, in which all events at the same point in time are stacked over each other. This is very useful in some applications, since it conveys additional information, such as that many events occurred in one day.

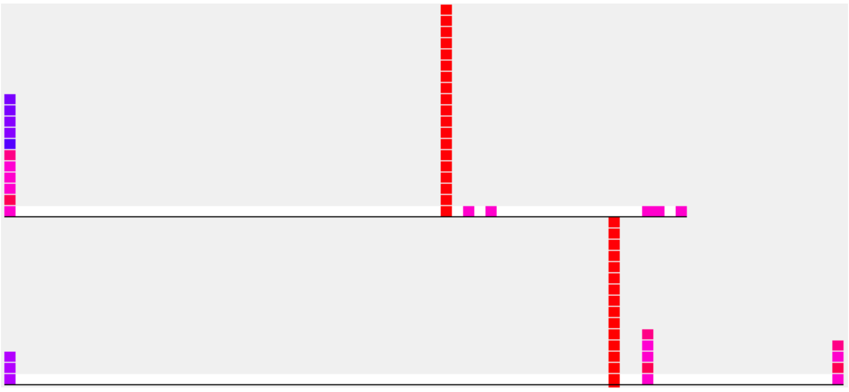


Figure 3.7: Vertical alignment with stacked events on top of each other when occurring at the same point in time (published in [85]).

## 3.4 Evaluation

In this chapter the capabilities of the visual exploration tool on the basis of the two event data sets are presented, as shown in the case studies in [85] and continuative with another data set. The two characteristically very different case studies in [85] are dealing with business process events and news feeds and demonstrate the capabilities of the approach to explore the event data. The results from the third other data set will be picked up in Chapter 4.5.4 again. Traditional time series data analysis methods for event time series data have limitations, such as a poor use of the display space and over plotting, so that the new knowledge often remains hided and cannot be visualized, like Figure  2.2 showed as an example. Therefore the flexible system like described in Chapter 3.3 was designed to find significant events, event patterns and event clusters.

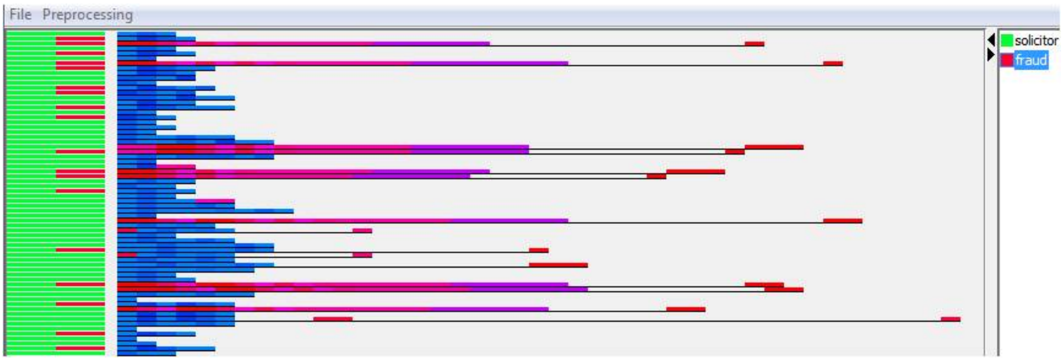


Figure 3.8: Fraud detection in a bank's time series and sequence data: each line visualizes bank customers who have a mortgage contract. During the contract period a black box system is firing rules based on the behavior of the customer, e.g., address or income changes, money transfers etc. These rules or events are indicators for fraudulent behavior and are colored according to the degree to which they indicate fraud (red represents a high fraud rule and blue a low fraud rule). The events are fired at irregular intervals for each customer and hence the gaps (represented by white spaces) between the events differ. The Figure is visualizing data for one particular solicitor (indicated by the green flag in front of each line) who is connected with 18 known fraud cases (additionally marked with a red flag).

The first case study deals with event sequences, where the ordering of events is more important than their absolute temporal references. By visualizing fraud detection events from a bank's mortgage department, it is possible to find a number of event patterns. Figure 3.8 and 3.9 describe the application. Similar

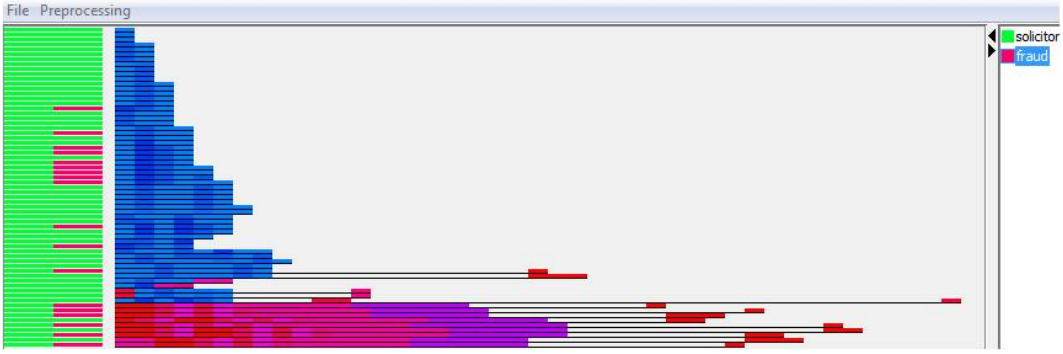


Figure 3.9: Fraud detection in a bank's time series and sequence data: Based on the same data as in Figure 1 but the data is sorted according to the similarity of the rule patterns, which leads to visually identifying other cases of fraudulent behavior.

patterns in the visualization are grouped together by applying algorithms for pattern detection and different alignment strategies to deal with the temporal gaps in the event time series data. Algorithmic similarity and visual pattern detection goes hand in hand.

The second case study analyses political news feeds from the debates of the U.S. presidential election in 2008. In this application, the absolute temporal reference plays an important role. In this data set, significant events, event clusters, and event patterns have been detected.

The third case study shows the flexibility of the system with a completely other data set of time related 3D feature data and demonstrates interaction possibilities for a visual cluster finding.

### 3.4.1 Fraud Detection in a Bank's Database

*Explored and described together with Verity Stennett and Anders T. Hasselrot, Lloyds Banking Group, Wolverhampton, England.*

The first case study is about fraud detection in a bank's database, where events are defined as system alerts triggered by customer behavior. To show the system's effectiveness experts in the operational, strategy and specialist fraud areas who could recommend how best to rank the data and assist in identifying real fraud cases in the event data have been brought in. Therefore four experts have been asked to analyze their own well known and daily used data with the new multi-dimensional event data exploration system. They have not been visualization experts before, so that they needed an introduction how to load in and visualize the data. Therefore there has been an intensive communication about the business issues and common data processing steps, as well as an introduction and training of the system. After that the experts were able to run the system and find so far unknown cases on their own. This was very important since not all bank data was accessible due to security reasons. The flexible preprocessing interface of the system helped them to be able to use the system autonomously to find analysis results on their own.

All cases were exposed by combining the visualizations with user input and obtaining additional data from the bank's database. The experts should think about how to group the data in a way to raise suspicions of possible fraudulent behavior. With this proposed task they grouped together events for one bank account number or for one solicitor to identify cases for unknown fraudulent behavior visually. Their feedback was used to improve the system, too. For example, the flags have been added in an improving step at a later date to make the grouping of events easier to identify. The goodness of this procedure became apparent because new fraudulent events have been detected that have been unknown to the experts before, despite their operating experience. All of them stated that they learned more about their own data with the system and got new insights and knowledge which are elementary goals of a visual analytics system.

### Data Set

The data was extracted from a stand alone fraud database used by the bank's fraud teams and combined with internal customer application and performance data. The fraud database contains external information in the form of rules that indicate the possibility of fraud and flags identifying whether applications were investigated internally and found to be fraud or clear. The internal data brought in includes application data such as name, date of birth, bank account details, address information and third party details such as solicitors and brokers. These details are used to rank the event data for visualization. Internal data clarifies if a mortgage has completed successfully, it highlights whether post completion any elements associated with fraud have become apparent and sets out how the mortgage is being maintained, i.e. whether the borrower has fallen behind on their mortgage repayments, if they have defaulted (3+ missed monthly payments) or if the property has been repossessed.

Once the data is collated and ranked appropriately it is read into the system. This data can sometimes include hundreds of rows and several columns per application. In this case study data with 550000 entries and up to 1000 events each was examined. The processing for that lasts about 1 minute for the preprocessing and 30 seconds for the mapping (Intel Core 2 Duo SP9400 (2.4 GHz, 1066 MHz, 6 MB Second Level Cache)).

The strength of the system is to condense, group and visualize both fixed and time series information on customers in one compact image, allowing the user to identify suspicious individuals and groups that could indicate collusive fraud.

The main concern in using the external rule data is that the rules information does not confirm fraud, it only gives indications and information to assist in investigations. Therefore even if accounts match against rules which typically indicate fraud, investigation must be performed and the application could be cleared if no hard evidence is found to the contrary. Experimenting with different selections and grouping of the data has exposed a number of uses and cases that required further investigation. These included:

- Assisting in better understanding of rules that indicate fraudulent or non-fraud behavior.

- Identifying new fraud on book accounts by ranking/clustering via names, brokers & solicitors etc., postcodes/demographics, and bank account numbers.

- Questions around policies and procedures used within the bank dealing with customers applying for several mortgages.

- Identifying new targets for fraud models - for example, rules which identify fraud in other banks.

### Visualization of Event-Rules for Fraud Detection

Each line in the visualizations represents one account from the bank's database and aims to assist in fraud detection. The information on each account includes fixed data in the form of flags, such whether the account got a mortgage with the bank on the left hand side of the visualizations and time series events in the form of rules on the right hand side of the visualizations which indicate the possibility of fraud. The rules data are colored according to the colormap in Figure 3.10 with a rising fraudulent probability from blue to red. The fixed flags include whether accounts got on book, how they are performing and whether they have been found to be fraudulent post completion; see the colors at the beginning of each line in Figures 3.11 or 3.12 (Green for "Case on book" (obtained a mortgage with the bank) and Red for "Case allocated a fraud flag post obtaining a mortgage with the bank").

Figure 3.10: The colormap shows how the events, defined as system alerts in the form of rules, are colored: From blue to red the rules indicate more and more fraudulent behavior (published in [85]).

### Findings

The first case was identified when visualizing and ranking the data by solicitors. Figure 3.11 shows a solicitor that at first sight was linked to a number of known fraud on book entries (18 red flags) and several other suspicious entries matching against rules post completion that typically indicate fraud. Further investigations revealed the solicitor had already been removed from the panel but visualizing the solicitor's business has instigated investigations of 14 cases for fraudulent behavior. In Figure 3.11 this can be seen in the entries without a red flag but with a red ending event.

The next case shown in Figure 3.12 was identified when visualizing and ranking the data by bank account numbers. It exposed a number of cases where the same bank account number had been entered at application stage. Of the cases which had successfully completed (green flags), a proportion had already been identified as fraudulent post completion (red flags), importantly, the vi-

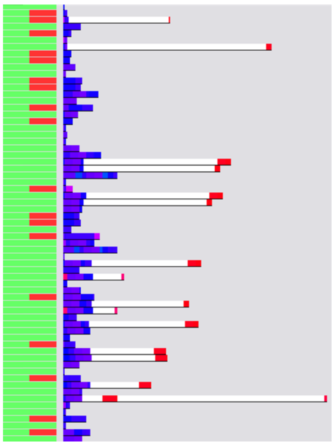


Figure 3.11: Figure is visualizing data for one solicitor with 18 known fraud on book cases with a red flag and 14 visually identified cases for fraudulent behavior (published in [85]).

sualization tool was able to flag a number of linked accounts. The fraud team had previously flagged these as fraud after identifying income fraud collusion between these customers using the same bank account number. The group of individuals in question were all part of the same family and owned a property business together. The other 7 entries linked to this bank account number are presently being investigated and are likely to be assigned fraud flags.

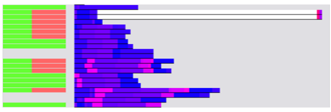


Figure 3.12: Figure is visualizing one bank account number used by several cases and customers. 10 cases out of the 17 on book have been flagged as fraudulent but the Fraud team were not aware of the other 7 using the same bank account number. These are presently being investigated (published in [85]).

The final case shown in the fraud detection application area was identified when visualizing the data ranked again via solicitors. At first sight in Figure 3.13 the solicitors business is all clear and performing well (no red or blue flags). But the matched time series rules data shows seven of the cases linked to this solicitor raised some suspicions. In Figure 3.13 they can be seen in the entries with a green flag and with a red ending event. Further investigations exposed that the solicitor was being monitored and the seven suspicious cases were split between two customers. All entries were performing well but the volume of mortgages and type of rules being fired raised suspicion and further investigations on these two individuals are being carried out.

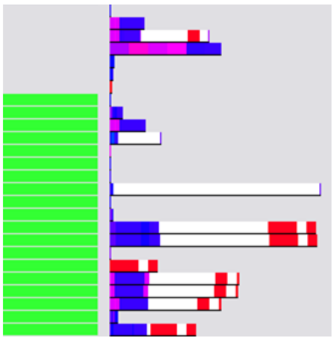


Figure 3.13: Visualized data for one solicitor for whom business is all clear and performing well (no red or blue flags) but several cases are matching against fraud rules post completion raising suspicions of possible fraudulent behavior (published in [85]).

### 3.4.2 Sentiment Analysis in News Feeds

*Explored and described together with Franz Wanner, University of Konstanz. Parts of the analysis are a further development of "Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008" [108].*

In the second case study in [85] a data set about the sentiment of RSS news postings about the U.S. Presidential Election in 2008 was analyzed. This case study explores online news with respect to emotional debates about selected entities. In particular, the visualization technique is used to display event episodes, in which the absolute temporal reference plays an important role.

#### Data Set and the Visualization of Event Episodes

The data in this case study was gathered from 50 different RSS news feeds that mainly dealt with the 2008 US presidential elections. The RSS feeds were retrieved every 30 minutes during a time interval of one month (10/09/2008 - 11/10/2008). For every news event in each feed date, title and description, as well as the id of the feed was saved. Next, noise was eliminated out of the title and description. With noise strings that do not carry any relevant content with respect to the sentiment annotation are referred, such as URLs or strings consisting of special characters. The concatenation of title and description was then considered to be the content of the news posting. Finally, those documents that contained none of the following signal words: “Obama”, “McCain”, “Biden”, “Palin”, “Democrat” and “Republican” were filtered out. More than 23000 news postings contained at least one of the six strings.

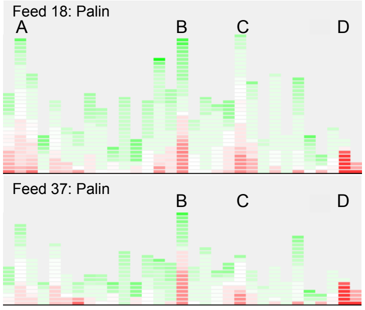


Figure 3.14: Sarah Palin in a negative context in Feed 18 and 37: (A) Only one positive green news event sticks out in Feed 18: “Palin acted [...] within law...” and didn’t abuse her power (B) she abused her power, (C) further news regarding “abusing her power by charging the state when her children traveled with her”, (D) Palin bought a too expensive wardrobe. You can see in the cross-feed analysis between Feed 18 and 37 above, that both reported very similar (published in [85]).

Since emotional debates are the point of interest, each event was enriched with a sentiment score. Therefore a freely available list of words that evoke positive or negative associations, see [15], was used. The number of positive and negative words was counted and the whole news event as rather positive if it contains in total more positive than negative words was evaluated. Likewise, the event is evaluated as rather negative if it contains more negative than positive words. The absolute relation of positive against negative words normalized by the event’s length, provides the sentiment score. Finally, for the visualization task the sentiment score was normalized to a score between 0 and 100, where 0 means very bad sentiment, 50 marks a neutral event and 100 denotes very positive news. One important point to mention here is that the appearance of a candidate, e.g., in a negative context, does not necessarily mean, that the event contains negative publicity for the candidate, but simply that he appears in a negatively connoted context. This becomes clear when the example of news telling that racists planned to assassinate Obama was considered, which was bad news for Obama not about Obama, with a visibly negative connotation.

The visualization aims to provide a meaningful representation of the data and serves as an appropriate starting point for interactive exploration and discovery of interesting patterns. Figure 3.15 shows a screenshot of one of the 50 monitored news feeds. Each horizontal black line represents the baseline of the news for the respective entity. In total six entities are showed: Obama, McCain, Biden, Palin, the Democratic party and Republican party. Based on the first black line all the news belonging to Obama posted by the feed with ID 37 can be seen. Every news posting is represented through a red, white or green rectangle. All events of one day are sorted according to their sentiment score and arranged in a vertical stacked bar.

In contrast to the previous case study, it is not aimed at displaying event sequences with only relative temporal references, but rather event episodes with an absolute daily temporal reference. Each day is represented by one vertical bar of events, which enables to do cross-entity and cross-feed comparisons since the temporal alignment is fixed. Furthermore, for better visibility of the proportions between positive and negative events, the events are sorted according to their sentiment score within each day.

### Findings

Already on the second day of the data collection many negative news postings occurred about *Sarah Palin* as shown in Figure 3.14. Almost all red marked articles deal with the topic whether she had *abused her power in Alaska* or not. Only one exceptionally positive green news event sticks out on top of Feed 18 (A). A closer look at this significant event reveals that it is a response from the McCain-Palin presidential campaign: "Sarah Palin acted "within proper and lawful authority" in removing the state's public safety commissioner". The same topic reappeared on another day: on Saturday, 10th October, many negative news postings occurred about Sarah Palin. Cluster (B) of intensively red shapes symbolizes bad news coverage of Palin. Five days later, cluster (C) displays further negative news turning up: "A new ethics complaint has been filed against Sarah Palin, accusing the Alaska governor of abusing her power

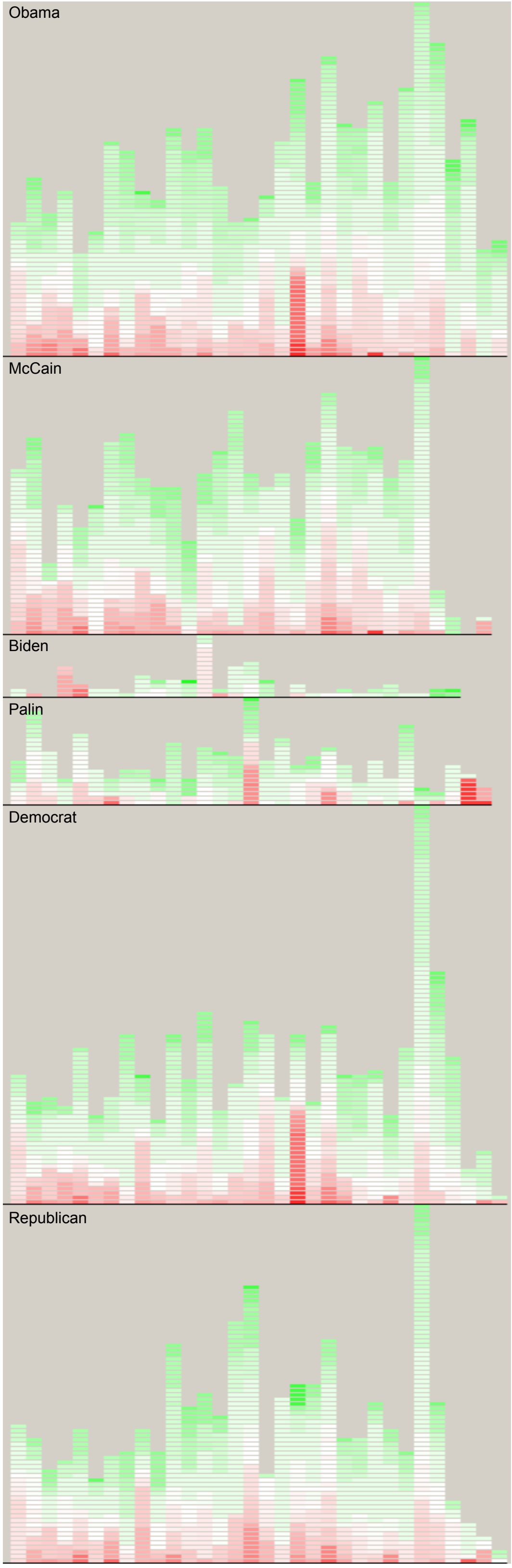


Figure 3.15: Cross-entity analysis showing more news about Palin than Biden and a high number of mostly positive postings on the election date November 4, 2008 (published in [85]).

by charging the state when her children traveled with her". After the election some very negatively rated events stick out in cluster (D). These news deal with some critical notes about the *expensive wardrobe*, which was bought by Sarah Palin for her campaign, and her inappropriate use of language describing her critics.

Cross-feed analysis in Figure 3.14 shows that both Feeds 18 and 37 reported very similar on the topic, which is due to the fact that they both used postings of the same news agency as the basis for their articles.

Cross-entity analysis as shown in Figure 3.15 enables comparison of different entities. In this case, through interpretation of the two diagrams in the middle, it immediately becomes obvious that the Republican vice presidential candidate Palin was a lot more in the news than her Democratic counterpart Biden, whereas the total amount of news about each of the two parties in the lower two diagrams is comparable. Approximately one week before the US presidential election day a high appearance of news which included "Obama" (see Figure 3.15) was detected. The sentiment scores of these postings were mainly negative and dealt with a plot to assassinate Barack Obama and 102 blacks. These news are bad for him but not about him, meaning that a negative incident is related to him in the news postings although the negative opinion words do not refer to him as a person.

A further remarkable event pattern is the extremely high number of mostly positive postings on the election date November 4, 2008 as seen for all entities in Figure 3.15. This is followed by a steep drop of news about the unsuccessful Republican presidential candidate McCain.

Note that although each RSS posting only consist of a few sentences, the few contained positive or negative opinion words are sufficient to provide clear results.

### 3.4.3 Visual cluster analysis on time related 3D data

*Explored and described together with Robert Gregor, University of Konstanz.*

This chapter will show and validate an extension of the approach via visualizing and analyzing time related 3D data. This will also demonstrate how the approach can be used for a visual clustering as a preprocessing step to the following Chapter 4. Therefore a heat kernel signatures data set for 3D models is used. Figure 3.16 shows an example for such a 3D model.

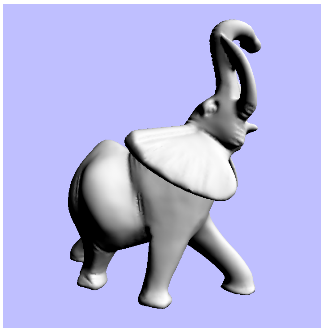


Figure 3.16: Example for 3D model. The figure shows an elephant shape. On this shape heat kernel signatures are calculated what results in time related 3D data.

This 3D model has 24955 points. The kernel signatures are calculated at 101 time values for every point. The values are measured and arranged logarithmically. That means the first values of each line follow faster in time than the final ones. Every value implies the portion of the heat amount that does not flow off the point at a specific time. This results in a 101x24955 Matrix and can be treated like the multi-dimensional data processed in Chapter 4. The data is not clustered yet, but it is assumed that specific parts of the 3D shape, e.g., the extremities like the trunk or the feet of the elephant, dispense heat extremely fast or slow. Figure 3.17 left shows the visualization of parts of the data in the original order. Patterns are distinguishable but it is hard to find similar behaving points, respectively entries.

Thus, in the first stage, the sorting algorithm as proposed Chapter 3.3.1 is applied on the data to get a better visual pattern detection. This can be seen in Figure 3.17 right. Now clear patterns are visible.

To find clusters in the data the system automatically suggests entries with similar patterns as clusters. Therefore in the first analysis step clusters in the data are detected from the proposed visual analytics system. The big advantage then is the visual refinement and user interaction the system allows in the next step of this visual analytics procedure.

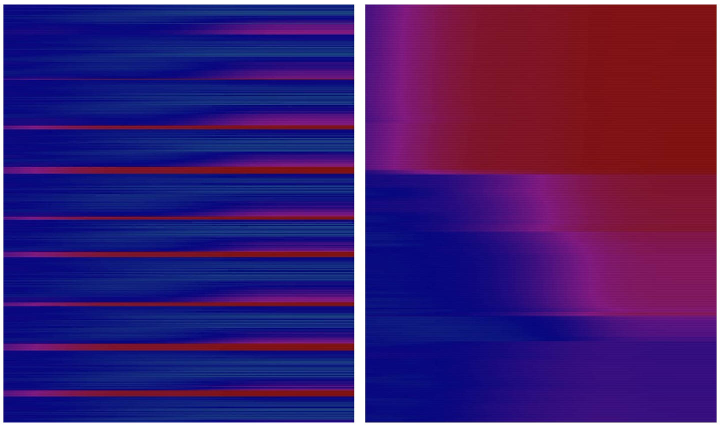


Figure 3.17: Left: Visualization of parts of the time related 3D data in the original order in the multi-dimensional event data exploration system. Right: Sorted visualization of parts of the time related 3D data.

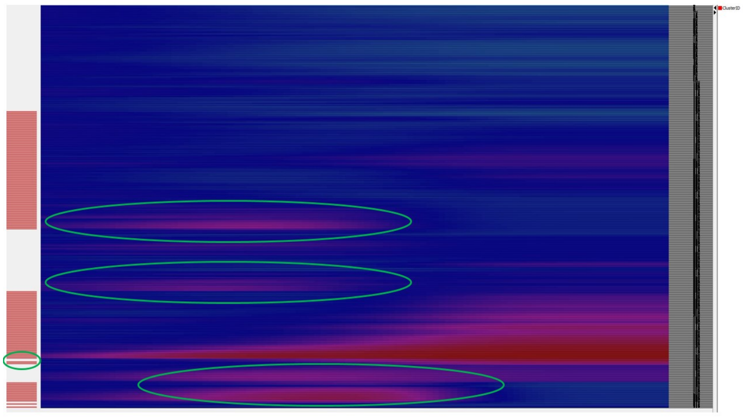


Figure 3.18: Sample visualization and the automatically detected clusters of the time related 3D data. The flags in front of each entry show the clusters. The green circles show regions where a user could prefer another clustering.

Both visualizations in Figure 3.17 show only an extract of the whole visualization of the data because it is too big (24955 lines) to overview it at a glance (scrolling is required). Therefore a sampling algorithm reduces the data, that is visualized, to a smaller amount. Dependent of the screen size the user can select the percentage of the data, that should be visualized. To ensure that every cluster is kept and the sizes of the classes are still representing the initial distribution a stratified sampling is adopted. Figure 3.18 shows this sample visualization and the automatically detected clusters, marked with flags in front of each entry.

Figure 3.19 shows the projection of the heat kernel signatures on the 3D elephant shape with a coloring of the automatically detected clusters. Now the user can validate this interactively.

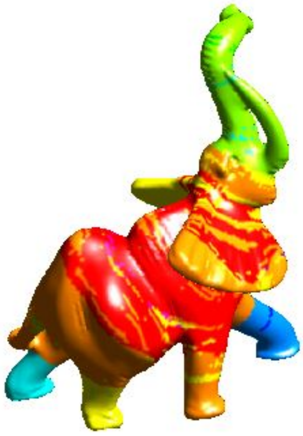


Figure 3.19: 3D elephant shape with classified and colored heat kernel signatures from the system's automatically detected clusters.

The green circles in Figure 3.18 show regions where a user could prefer another clustering. In this case the user grouped together entries where the lines have a similar coloring, which means a similar behavior of the entries. Via user validation and interaction the entries highlighted with the green circles can be grouped together and the flags can be resetted in front of the entries. After this resetting the cluster flags in front of the entries can be seen in Figure 3.20.

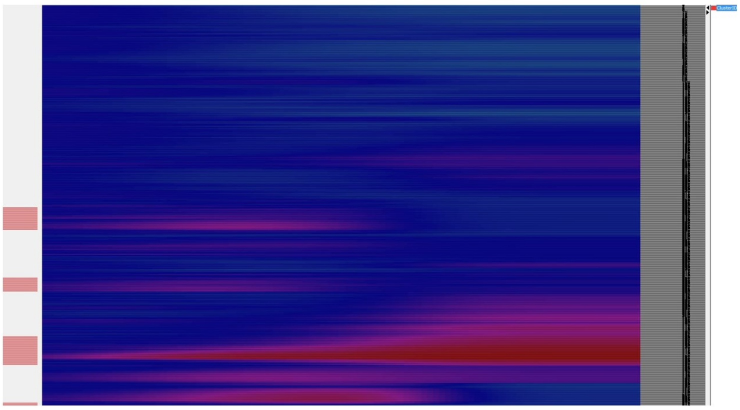


Figure 3.20: Sample visualization and the resetted clusters of the time related 3D data. The flags in front of each entry show the new cluster assignments according to the user interaction (indicated by the green circles in Figure 3.18.

Figure 3.21 shows the projection of the heat kernel signatures on the 3D elephant shape with a coloring of the user's resetted clusters. It can be seen that the extremities like the trunk or the feet of the elephant and its body belong to separate clusters, so that the visible improvement is clearly.

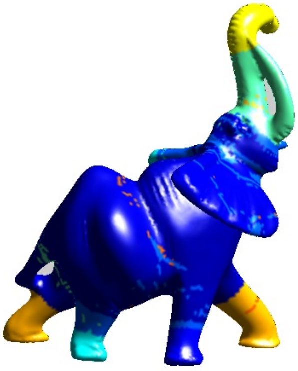


Figure 3.21: 3D elephant shape with classified and colored heat kernel signatures from the user's resetted clusters.

The results of this visual clustering will be discussed at the end of this thesis again. Chapter 4.5.4 will show that with the approaches of the second part of this thesis the visual impression will be confirmed by the numbers of the quality measures. This will show the link between the visual exploration approaches of multi-dimensional event data und the visual exploration approaches of multi-dimensional feature data. The latter is presented in the following chapter.

# Chapter 4

# Visual Exploration of Multi-Dimensional Feature Data

Parts of this chapter are based on [86]. I took the lead and responsibility in the whole project and the text of the paper and adapted it for my thesis.

- M. Schaefer, L. Zhang, T. Schreck, A. Tatu, J. A. Lee, M. Verleysen and D. A. Keim. Improving projection-based data analysis by feature space transformations. In Proceedings of VDA 2013, 2013.

Continuing work in visual exploration of multi-dimensional data flows in a collaboration with Daniel Perez, University of Oviedo, and is also part of this chapter, see [77]:

- D. Perez, L. Zhang, M. Schaefer, T. Schreck, D. A. Keim and I. Diaz. Interactive Visualization and Feature Transformation for Multidimensional Data Projection. Proc. EuroVis Workshop on Visual Analytics Using Multidimensional Projections, 2013.

## 4.1 Challenges and Needs

In this chapter another type of data - multi-dimensional data - will be examined which is often the basis for information gaining and therefore needs to be explored. The usual procedure for analyzing high-dimensional data is to reduce and map it to the 2D display and get a scatter-plot visualization [88]. This is called projection-based data analysis with visual embeddings. But it is very difficult to generate effective visual embeddings of high-dimensional data, because the analyst wants to see the high-dimensional structure of the data and patterns and relations of the data. Given the high dimensionality, noise and imperfect embedding techniques, it is hard to come up with a satisfactory embedding that preserves the data structure well, whilst highlighting patterns and avoiding visual clutters at the same time.

A large number of Dimension Reduction (DR) techniques exist for performing such a task, however it is hard to come up with a satisfactory embedding (projection) that provides both good reflection of the data structure and clear indication of class boundaries at the same time. This is due to many factors. First of all, DR techniques only provide approximations of the distances between data items in the data space. It is nearly impossible to get a "perfect" embedding which shows the exact structure of the data in the visual space. Secondly, high-dimensional data often contains irrelevant attributes which obscure the real distance between data items, and the noise introduced by such irrelevant information may lead to a cluttered visual display where class boundaries are blurred. Such limitations hinder the human analyst from understanding relationships between data items and identifying patterns in the data. Although supervised DR techniques aim at providing better group separation based on class labels in the data, methods that maintain good comprise between structural preservation and visual cluttering avoidance (i.e., classes are well separated and easy to identify in the visual display) are lacking. Furthermore, there is a need in measures that can be directly applied to evaluate the quality of a given embedding based on all the above mentioned criteria.

## 4.2 Related Work

As stated, the analysis of high-dimensional data is a challenge and yet a hot research topic, as the high dimensionality induces problems in the automatic analysis and visualization as well as in the interaction possibilities. In addition, measures have to be defined to evaluate the results. This related work chapter covers all three aspects: DR for automatic analysis and its visualization, interaction techniques and quality measures.

### 4.2.1 Dimension Reduction for High-dimensional Data Visualization

On the automatic side, redundant and noisy dimensions may degrade the performance of the analysis algorithms [40] and values in these dimensions create noise in the distance measure. This noise can be filtered out by reducing the influence of noisy dimensions via feature transformation or by detecting irrelevant dimensions manually or automatically for dimensionality reduction [49]. On the visualization side, one has to cope with the conflict between the limited number of visual dimensions and the large number of data dimensions. To visualize the structure of high-dimensional data effectively, the data items have to be mapped in such a way that similar items are close to each other and dissimilar ones are far apart. This is usually achieved by a DR technique which tries to approximate the distances between data items in data space to the corresponding Euclidean distances in the visual display.

These DR techniques are a well-studied topic by the machine learning community. The idea behind all DR techniques is that they should produce low-dimensional representations that preserve meaningful structural properties of the data. In general, these properties formalize proximity relationships. They can be similarities (adjacencies, dot products) or dissimilarities (distances, angles). DR was achieved by feature selection or transformation to improve the performance of many applications in several research fields [11, 27, 36, 67]. A large number of DR techniques exist, ranging from classical approaches such as *Principal Component Analysis (PCA)* [51], numerous variants of *Multidimensional Scaling (MDS)* [12,29,98], to more recent extensions such as *Curvilinear Component Analysis (CCA)* [26], *Isomap* [97], *Generative Topographic Map (GTM)* [10], and *Stochastic Neighbor Embedding (SNE)* [41].

For generating a visual embedding, the training data can be either *labeled* or *unlabeled*, leading the development of *supervised* and *unsupervised* DR algorithms, which are often studied separately. While unsupervised DR aims at representing high-dimensional data in lower-dimensional spaces in a faithful way, supervised DR tends to emphasize features relevant for a given labeling of the data in the final embedding such that the visualization provides better class separation. However many DR techniques have variations for both, supervised and unsupervised learning. The algorithm proposed in this thesis works in a supervised manner, however, it can be applied to improve an embedding generated by any DR technique as long as trustworthy class labels are provided. Next, a few representative DR techniques as a basis for understanding the fundamental principle of DR are discussed. Comprehensive surveys of DR techniques can be found in [63, 76, 101, 105, 111].

PCA and classical MDS are probably the two linear DR techniques for embedding high-dimensional data most widely used in data visualization. Also known as Karhunen-Loève transform, PCA reduces the dimensionality of the data to summarize the most important parts and simultaneously filters out noise. The algorithm de-correlates variables and selects those that bear most of the data variance for projection. PCA was later extended to Classical MDS which starts from either a Gram matrix or a matrix of pairwise Euclidean distances instead of the sample covariance to compute the projections along the principle components.

One inherent limitation of linear approaches is that they cannot take into account nonlinear structures consisting of arbitrarily shaped clusters or curved manifolds. Among many nonlinear extensions of MDS which are designed to overcome such limitation [5, 26, 81], Isomap is a interesting variation. Instead of using pairwise input-space distances as simple Euclidean distances, Isomap uses geodesic distances along the manifold of the data (technically, along a graph formed by connecting all $k$-nearest neighbors) to recover certain types of manifolds. SOM is one of the earliest nonlinear techniques which trains a discretized map representation of the input space of the training samples. A neighborhood function is used to preserve the topological properties of the input space. A notable recent extension of SOM is GTM, which is a generative version of SOM. GTM represents the probability density of high-dimensional data in a smaller number of latent or hidden variables using latent variable models [50]. The SNE method and its variant, t-SNE [100], have recently received much attention. They are based on similarity preservation instead of distance preservation. They involve specific similarity definitions that show interesting invariance properties and make them robust against the phenomenon of norm concentration.

### 4.2.2 Interactive Dimension Reduction for High-dimensional Data Visualization

Automatic DR techniques help to show underlying structure and relationships in high-dimensional data. However, with the increasing size and complexity of data, it becomes more and more difficult to generate meaningful transformations without interactive analysis based on integrating human knowledge and feedback during the learning process. This leads to the development of *interactive dimension reduction* techniques. Examples include the iPCA system [47] which supports the analysis of multivariate data sets through extensive interaction with the PCA output. The iVisClassifier system [22] facilitates the interpretability of the computational model applied to the data via interaction and multiple views projections. Furthermore, the DimStiller framework [44] defines an interactive workflow that guides users through the process of finding suitable dimension subsets. In [28] the importance of integrating interactions with statistic methods (in particular, DR techniques) to support explorative analysis of high-dimensional data is discussed. In [13] interactive selection of sets of features was proposed. The approach generates a projection for each candidate feature set. Using color-coding, a comparison matrix of the individual projections is provided which supports identification of similar and complementary feature sets. A related problem was addressed in [14]. There, Dendrogram structures were extracted from alternative feature sets, and applied for interactive comparison and selection of feature sets.

The general problem is that different clustering or projection algorithms impose different clusters. The algorithmic clustering or the projection step can be wrong, this can be found in [37] or [43]. In this thesis the idea of a visual and an algorithmical validation for an automatically judging of the algorithms to get better results is suggested. To improve automatically produced visualization, in this case projections, both methods mostly used in literature, relevance feedback and user interaction, are combined. Relevance feedback was developed in information retrieval where measures are calculated, like precision, recall etc. Relevance feedback techniques can also be used for visual displays as supporting tools, examples can be found in [32]. User interaction for projection visualization can be found in [75], where the Projection Explorer [76] is enriched with the possibility that the user can label classified objects or that he can mark them for automatic labeling.

Most of researchers have focused on automatic clustering algorithms, but very few have addressed the human factor in the clustering process, like [19] or [20]. The above mentioned techniques show that a rich body of research exists on high-dimensional data visualization. However, how to appropriately compute and visualize structure and patterns in high-dimensional data remains a tough challenge due to the nature of the data and DR techniques, and there do not exist many related studies, which compare different DR methods in matters of class separation and human judgment. One study to mention is [66], that investigates, how professionals and beginners rate the quality of 2D scatterplot projections from different DR methods.

A work which deals with the evaluation of different visual encodings (2D Scatterplots, interactive 3D Scatterplots, or Scatterplot Matrices) and different DR techniques can be found in [89]. One finding in this paper is that 2D scatterplots are often good enough, and if not, the most promising way is to try another DR method but keep the visual encoding. For this thesis this means with analyzing different embeddings with different methods and their extensions in looking at the 2D scatterplot is an appropriate way. There is no user study with many participants executed in [89], but a so called data study performed, that investigates many data sets by a small number of expert users. This thesis traces a similar approach, but the results of different data sets will be evaluated by state of the art and self-implemented quality measures. The next chapter gives a survey of related work concerning quality measures.

### 4.2.3 Quality Measures

Despite the fact that DR research started long ago and a large number of DR techniques have been developed, the question of quality assessment of a given embedding remains mostly unanswered until recent years. Two ways of measuring the resulting projections are possible: taking into account structural preservation or developing a set of visual quality measures. A combination of both is preffered.

The first possible way to measure the quality of a embedding is to use a so called *stress* or *strain* measure [25, 61, 82]. These measures often come with nonlinear DR methods, and are typically used as objective functions for measuring the quality of structural preservation in terms of how well do the Euclidean distances between pairwise data items in a low-dimensional embedding approximate the corresponding distances in high-dimensional data space. While strain and stress measures check the preservation of global structure of data with respect to distance/similarity preservation, in recent years, more and more research has been devoted to designing of new criteria for quality assessment with a broader applicability, taking into consideration also the small neighborhood preservation. Examples include the trustworthiness and continuity measure [103], the local continuity meta-criterion [21], and the k-ary neighborhoods measure [64]. In the case of labeled data, the classification error is a typical choice, see for instance [83] and other references in [104]. The integration of classification error measures in the DR technique leads to better group separation in the final embedding.

As mentioned previously, apart from studying the structure of the data, analysts also expect to see patterns in the embedding. Such patterns include grouping information, such as classes (with labeled data) and clusters (with unlabeled data) and outliers. Although it is often intuitive to tell the visual quality of an embedding by seeing how cluttered it is and how badly groups overlap, a qualitative measure is still preferred by analysts for quality assessment. In recent years some research has been devoted to the evaluation of visual quality of graphical representations of high-dimensional data in a broader sense [8, 48, 92, 96]. Such graphical representations include not only visual embeddings of high-dimensional data generated by DR techniques, but also other forms of visualizations such as parallel coordinates [45] or scatter plot projection of data values over a subset of dimensions (usually two or three dimensions, mapped to x-, y- and z- axes in the visual display). The visual quality measures can be based on different user tasks, for example, outliers, clusters, correlation, and abstraction level. Two existing visual quality measures are closely related to the proposed problem [96]. The first measure is the *Histogram Density Measure* which was designed for ranking scatter plot visualizations. The second one is the *Class Density Measure* which is based on an image processing algorithm that transforms each group in a continuous,

smooth density function based on local neighborhoods and measures the mutual overlap between pairwise data items in the scatter plot. An overview of approaches that use quality measures in high-dimensional data visualization and a systematization based on a literature review is presented in [9].
None of these above mentioned measures however provides the facility to assess the quality of a given embedding in terms of both structural preservation and class separation, as well as visual cluttering avoidance. Therefore a measure was designed that takes into account both, structural preservation quality and visual quality to make sure the embedding reflects the original structure of the data as well as provides clear (low-cluttered) visualizations for pattern analysis.

## 4.3 Feature Engineering

### 4.3.1 Feature Selection

*Explored and described together with Andrada Tatu and Michael Regenscheit, both University of Konstanz, in parts of their student theses, see [79, 95].*

Given a database with multi-dimensional data, the first step in the feature engineering task is to give the possibility to select features as well as giving suggestions for the most appropriate feature to visualize the multi-dimensional data. This proceeding helps to overcome the following challenges:

- Get a deeper understanding of the multi-dimensional database.
- Provide an appropriate system for selecting classifications.
- Enable a visual evaluation of different features for different classifications with a clustering according to the content level.
- Calculate measures to give automatically suggestions for selections and transformations.

This part of the thesis and the developed system was motivated from a practical problem of analyzing an image database. Therefore a tree visualization was chosen and with an interactive system it is possible to get a new visual access to the data which allows to give detailed insights into data in the database. It is possible to select classifications and test features for different classifications. In the image database, e.g., color is the most dominant and distinguishing feature for most of the images, but it is not suitable for black and white images. Using the image descriptions and the meta-information of the images one could find out which features work best for which image. The goal was to build up the whole process from understanding the database to visualize classes and select and compare the features in a 2D projection.

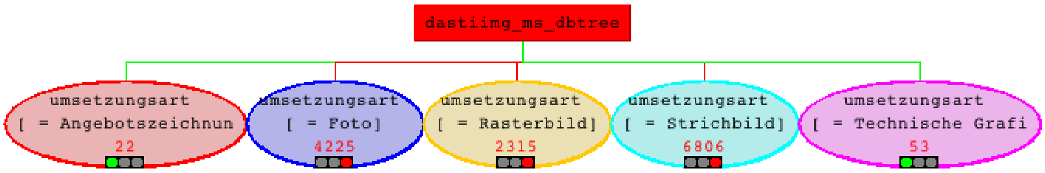


Figure 4.1: Tree with feature "*umsetzungsart*" and its different values as nodes.

In Figure 4.1 the tree visualization is shown for the table column "*umsetzungsart*". The user can interactively choose the data and analyze different features and see the impact on the resulting visualizations. For an automatic feature selection a high-dimensional overlap measure is calculated to select the best separating feature for a chosen classification. This measure enables an evaluation of the calculated features numerically. Equation 4.1 shows the calculation of the overlap value (ov):

$$ov(i,j) = \frac{d - sd_i - sd_j}{min(sd_i, sd_j)} \tag{4.1}$$

Dependent of the position of the classes in the high-dimensional space 4 cases to calculate the overlap exist. This is shown in the 2D examples in Figure 4.2, where $i$ and $j$ indicate the classes in each case.
The evaluations later will show the experimental results on the image database. The evaluations are done on image data, but the system can handle all kinds of feature-based multimedia data which results in high-dimensional data.

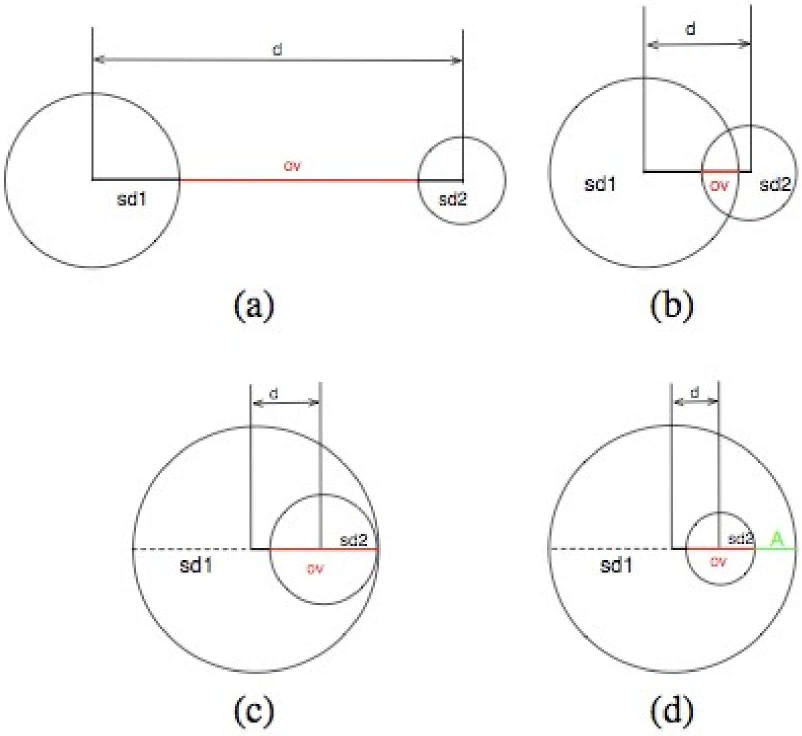


Figure 4.2: 2D example for high-dimensional overlap measure. (b) (c) and (d) cause negative ov values, so it is good to have high overlap values (ov).

### 4.3.2 Feature Space Transformation

A generic framework that tackles the problem of high-dimensional data embeddings and improves the quality of an existing embedding in terms of both structural preservation and class separation by feature space transformations was implemented. This is achieved by extending relevant features in the feature vector space. The framework first identifies relevant features in the data and then extends them in the original feature space using various transformation strategies. Given an initial embedding, users can interactively update the configuration by enhancing the influence of relevant features using different transformation strategies. A compound quality measure, which takes into consideration both structural preservation and visual clutter avoidance is designed to assess the quality of the embeddings. The structural preservation is evaluated by existing structural quality measures. The visual clutter avoidance is evaluated by a density function that measures the overlap between classes and an area measure that calculates the size of overlap regions between classes.

The combined measure provides a clear indication of the trustworthiness of an embedding in terms of both structural preservation and class separation. With an appropriate graphical user interface users can achieve the best compromise between the two according to their preferences. The proposed approach works in a supervised manner, that is, class labels are used to evaluate the visual quality of an embedding in terms of class separation. Unlike most of the existing supervised embedding techniques, which are based on specific DR algorithms, the approach is DR technique independent. It can be applied to improve an initial embedding generated by any DR algorithm.

Most projection-based analyses involve two steps: Given a high-dimensional data set, a distance or similarity matrix that records pairwise distances (similarities) between objects is first calculated using a preselected measure; a DR technique is then applied, aiming to approximate in the projection space the pairwise distance (similarity) measured in the data space. The visual embedding is meant to help analysts to understand the data structure as well as identify meaningful patterns in the visual display, in particular, arbitrary shaped clusters. However, high-dimensional data often contains irrelevant dimensions which obscure the real distance between objects and even with carefully chosen DR techniques, the grouping information may still be hidden in the visual representation due to the noise. To reduce noise and preserve grouping information as well as structure of data, a feature vector transformation approach is proposed which first designs a transformation strategy and then transforms the original feature space by extending corresponding feature vectors. Such transformation is expected to provide better group separation in the final embedding. The resulting embedding can be evaluated by a quality measure to make sure the final embedding shows clear separation of classes, whilst still preserving data structure well. This is achieved by a quality measure combining class-related overlap measures on the one hand, and stress-based measures on the other. Now the method for feature vector space transformation is introduced and the approach is exemplified with two simple transformation strategies.

The main idea of the feature space transformation is to extend the relevant features (e.g., mean values of selected dimensions) for better group separation in the final embedding. This can be achieved by adding additional feature

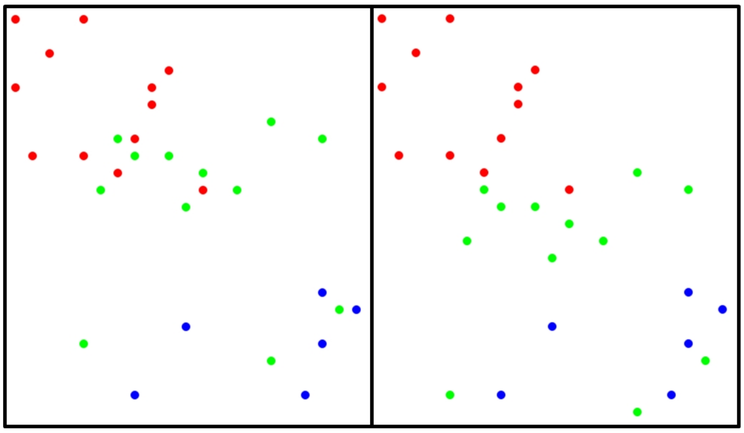


Figure 4.3: Example for a mapping of two 2-dimensional data sets with three classes. Left: data set A, right: data set B. Classes 0, 1 and 2 are colored in red, green and blue (published in [86]).

vectors to the original feature space to leverage the noise introduced by irrelevant dimensions. For selecting dimensions to extend, there are a number of intuitive guidelines, for example, one can check the *range* of data values over a dimension - typically the smaller the range, the less likely the class is going to overlap with others in the dimension. Therefore dimensions that have high range may be less relevant to the class separation. Another choice is the *spread* - if all the class members share similar values in one dimension, it is likely this dimension is discriminative to the class label. In other words, generally speaking dimensions that have high spread may be selected for extension. Next these two measures are used for relevant feature selection.

To illustrate the feature space transformation approach, two simple 2-dimensional data sets, A and B, are generated. Each contains 30 objects that belong to three different classes. Figure 4.3 shows the 2D mapping of both data sets. Colors are used to indicate class labels.

One simple feature space transformation strategy is the addition of mean values. Although a feature space can be transformed in many different ways, besides this simple strategy. For example, *median* or *mode* can be applied instead of mean values (depending on the nature of data). Also, the number of extensions can vary. The maximum number of extensions can be the total number of dimensions $n$, in which case all dimensions are extended. The experimental results show that this maximum extension leads to a good group separation but loss of similarity preservation between group objects. Next, the mean-value extension strategy using the two data sets A and B is illustrated. First of all, the mean values mv$_{dc}$ of each dimension $d \in \{1, \ldots, n\}$ for each class $c$ are calculated (see Table 4.1). A simple heuristic approach based on mean spread of dimensions is already suitable for selecting which dimensions to extend. Ideally the selected dimensions should be discriminative to the class labels. In this case, the *range* and *spread* are taken (see below for details).

| Class | mv$_{1c}$ data A | mv$_{2c}$ data A | mv$_{1c}$ data B | mv$_{2c}$ data B |
|---|---|---|---|---|
| c=0 | 10 | 10 | 10 | 10 |
| c=1 | 16 | 16 | 16 | 19 |
| c=2 | 20 | 24 | 20 | 24 |

Table 4.1: Mean values $mv_{dc}$ for each dimension $d \in \{1, 2\}$ for each class $c \in \{0, 1, 2\}$ of data set A and B.

First, the range r$_d$ for each dimension $d \in \{1, \ldots, n\}$ with class labels c is calculated (here $c \in \{0, 1, 2\}$):

$$r_d = max(mv_{dc}) - min(mv_{dc}) \tag{4.2}$$

The results for the range values for data sets A and B are shown in Table 4.2.

| | d=1 data A | d=2 data A | d=1 data B | d=2 data B |
|---|---|---|---|---|
| min | 10 | 10 | 10 | 10 |
| max | 20 | 24 | 20 | 24 |
| r | 10 | 14 | 10 | 14 |

Table 4.2: Range values r$_d$ for each dimension $d \in \{1, 2\}$ of data set A and B.

Next the spread$_d$ measure for each dimension is calculated, which is defined as:

$$spread_d = \begin{cases} r_d^2/sd_d & if\ sd_d \neq 0 \\ r_d^2 & else \end{cases} \tag{4.3}$$

where sd$_d$ is the standard deviation of the differences between the ordered mean values of the classes within dimension *d*. With this heuristic equally spread mean values are rewarded, the higher the spread measure the better. Table 4.3 shows the spread values for the data sets A and B.

In the next chapter, quality measures are defined that can be applied to evaluate the effectiveness of the above mentioned heuristic approaches.

| | d=1 data A | d=2 data A | d=1 data B | d=2 data B |
|---|---|---|---|---|
| difference between c0 and c1<br>difference between c1 and c2 | 6<br>4 | 6<br>8 | 6<br>4 | 9<br>5 |
| standard deviation | 1.4 | 1.4 | 1.4 | 2.8 |
| range | 10 | 14 | 10 | 14 |
| spread | 70.7 | 138.6 | 70.7 | 69,3 |

Table 4.3: Spread-measure spread$_d$ for each dimension $d \in \{1, 2\}$ of data set A and B.

### 4.3.3 Quality Measures

As mentioned previously, to visualize high-dimensional data a good embedding is expected to approximate the data structure well and highlight patterns. There are a number of quality measures for evaluating these properties (see 4.2.3), however not many of them take both aspects into consideration and allow the user the define the best compromise. Therefore a quality measure based on structural preservation and visual clutter avoidance is proposed to access the quality of embeddings. The effectiveness of the approach is evaluated by applying it to several widely used embedding techniques using a set of benchmark data sets and the result looks promising. This new quality measure combines three score functions for evaluating structural and visual aspects of an visual embedding respectively and gives users the freedom of adjusting the weight for each score. First of all, for measuring the structural preservation, two alternative quality measures are used, which assess how well the structure of data is preserved in the embedding. Secondly, for measuring overlapping between groups, an area-based overlapping measure and a density-based overlapping measure are combined. The former calculates the size of the overlapping regions between groups. The latter measures how objects between different groups overlap in a particular region. The combination gives a good indication of how well groups are separated in the projection.

Note that a wide range of measures exist for calculating the structure preservation, the area of overlap between two regions and the density of overlapping objects inside a region. The measures presented in this work are used to illustrate the basic principle of the approach and can be replaced by other measures of similar nature.

### Stress measure

To evaluate structuring preservation *Sammon's stress* [82] and *k-ary neighborhoods measure* [65] are applied. While Sammon's stress focuses on the quality of global pairwise similarity preservation [82], k-ary neighborhood measure also shows the quality of local neighborhood preservation of an embedding. The k-ary neighborhood measure records two types of neighborhood preservation errors in the embedding, neighborhood intrusion error and neighborhood extrusion error [64].
Given a data with $N$ objects, Sammon's stress computes an error $E$, which represents how well the present configuration of $N$ points in the lower dimensional space fits the $N$ points in the high-dimensional space:

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \tag{4.4}$$

The distance between data item $i$ and data item $j$ in the original high-dimensional space is defined by $\mathrm{d}^*_{ij}$, and the distance between their lower dimensional projections by $\mathrm{d}_{ij}$. The stress value is calculated from the data of the projection resulting from the transformed features, against the initial feature vector data. With this procedural method it is ensured that the distance preservation is measured to the initial feature space.
The *k-ary neighborhoods measure* is defined as

$$Q_{\mathrm{NX}}(K) = \sum_{i=1}^{N} \frac{|n^*_i(K) \cap n_i(K)|}{KN} \quad , \tag{4.5}$$

where $n^*_i(K)$ is the set of indices of the K nearest neighbors of the i-th datum in the HD space whereas $n_i(K)$ corresponds to the set of indices of the K nearest neighbors in the LD space.

### Overlap measures

For pattern search, it is important to avoid visual cluttering in the projection such that patterns can be easily identified and groups can be easily perceived. Visual cluttering in an embedding can be caused by either overlaying of objects in the display or overlap between group boundaries. E.g., Figure 4.4 shows 3 different projections of the same data set. While the left projection has a cluttered region on the top left corner where objects are plotted on top of each other and the middle one has a big overlap region between purple and orange group, the right projection shows much clearer view of the groups. To achieve a less cluttered embedding, two overlap measures to evaluate the visual cluttering level of an embedding are designed. The first measure calculates the size of overlapping region between groups and the second measure sums up the density of objects in overlapping regions.

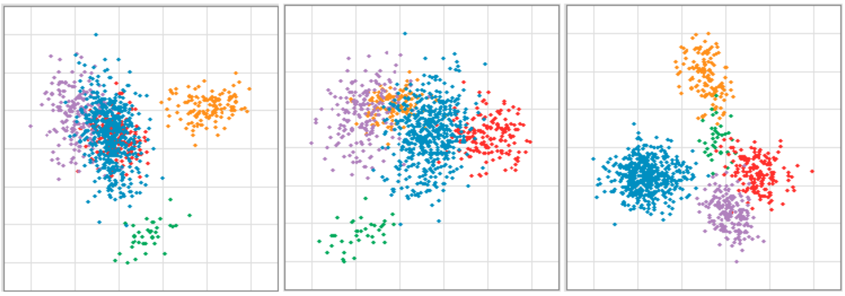


Figure 4.4: 2D projections of high-dimensional data with 5 classes, color of objects represent class labels (published in [86]).

**Overlap area measure.** To calculate the size of overlapping regions between groups of labeled points, the region of each of the groups has to be defined. The basic idea is to describe the group boundary by drawing a representative, enclosing hull of the objects that belong to the same group. A number of methods exist for computing boundaries for point sets, e.g., various convex hull generation methods [34, 46, 59] and the isocontour approach as proposed in [23]. In [87] the overlap between convex hulls for the visual comparison of the discrimination capabilities in alternative feature spaces is computed. For the overlap measure used here a different and less known hull formation method proposed by Moreira and Santos [74] is applied, which computes the region of a set of points as a concave hull based on a *k-nearest neighbors* approach. The advantage of this approach is that the generated region is usually more compact and better reflects arbitrary shaped groups in the embedding (see Figure 4.5). For each point that has to be connected to the next point, the algorithm first searches the best connection among its $k$ nearest neighbors to make sure the result hull is as compact as possible. Compared to the isocontour approach, the parameter setting of Moreira and Santos' method is much simpler and can be automated. Only one parameter needs to be defined, which is the $k$ parameter. It controls the smoothness of the computed hull. When $k$ is set to 3, the algorithm automatically looks for the smoothest possible envelop, i.e., the most compact concave hull. In the worst case, when $k$ is equal to *number of points-1*, the algorithm will output a convex hull.

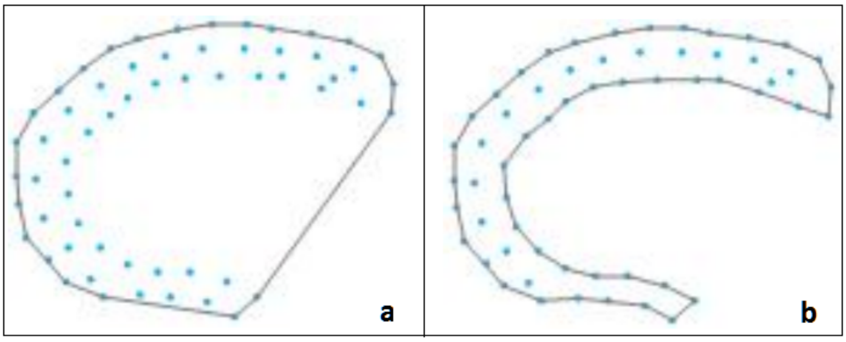


Figure 4.5: Regions of a set of points generated by: a) convex hull approach, b) concave hull approach (published in [86]).

Once the region of each group is defined, the overlap region $intersect(i, j)$ for each pair of groups $i$ and $j$ is calculated. The overlap area measure sums up the area of all the overlap regions between pairwise groups for the set $g$ of groups:

$$ov_{reg} = \sum_{i=1}^{|g|-1} \sum_{j=i+1}^{|g|} intersect(i, j) \tag{4.6}$$

Figure 4.6 shows an example of the overlap area measure of the projections of the two simple data sets A an B (Figure 4.3). Given three classes colored in red, green and blue, the method first computes the boundary of each class using the concave hull approach. The size of overlap region(s) between pairwise concave hulls is then calculated and summed up as the area overlap measure (black surrounded areas in Figure 4.6).

**Overlap density measure.** The hull-based approach does not consider the possibly, non-uniform density of points. Therefore, it is complemented by an overlap density measure that evaluates how strongly points are over-plotted in the visual display. The display area is divided into grid units (where the resolution of the grid can be adjusted). A Gaussian function $G$ is used to

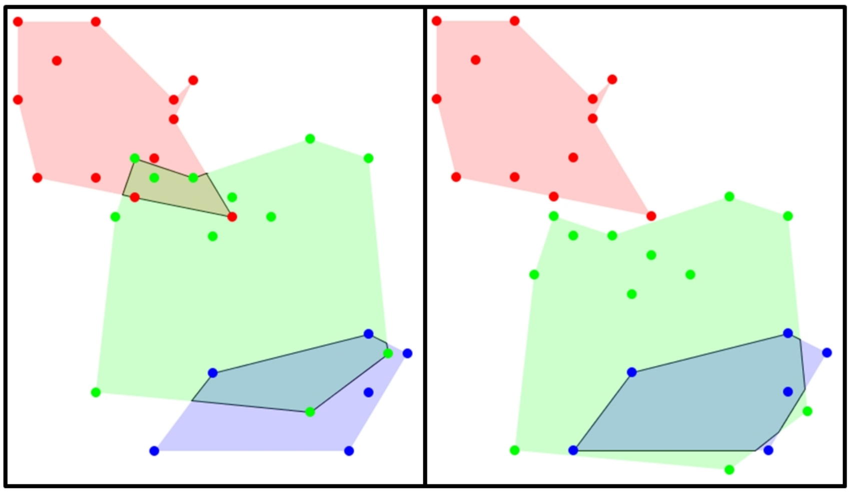


Figure 4.6: The black surrounded areas show the overlap of the concave hulls for data set A(left) and data set B(right) (published in [86]).

determine whether a grid unit is occupied by a particular class depending on the density of objects inside the grid square. Once the Gaussian model for each class is computed, the approach checks pairwise classes to see how many grid units are occupied by more than two classes. The count will be summed up as overlap density measure.

Equation (4.7) defines the overlap measure for a data set with $K$ classes and an image with $P$ pixels.

$$ov_{density} = \sum_{i=1}^{|K|-1} \sum_{j=i+1}^{|K|} \sum_{p=1}^{|P|} f(G_{ip}, G_{jp}) \tag{4.7}$$

The Gaussian function is defined in Equation (4.8) where $f$ is an indicator function which gives “1” in case when a grid square $k$ is occupied, according to the Gaussian model, by a pair of classes, and “0” otherwise.

$$f(G_{ip}, G_{jp}) = \begin{cases} 1 & if\ G_{ip} > 0\ and\ G_{jp} > 0 \\ 0 & else \end{cases} \tag{4.8}$$

In the following examples, the grid resolution is set to 3 pixels and the $\sigma$ value of the Gaussian model to 12 pixels. However the grid resolution can be adjusted and normalized to fit in the size of the display and the $\sigma$ value can also be changed.

Figure 4.7 shows the density overlap of the projections of the two examples for the data sets A an B (Figure 4.3) and visualizes the overlap regions between classes. The overlap grid squares are shaded in gray, and the scale of the gray color indicates the density level.

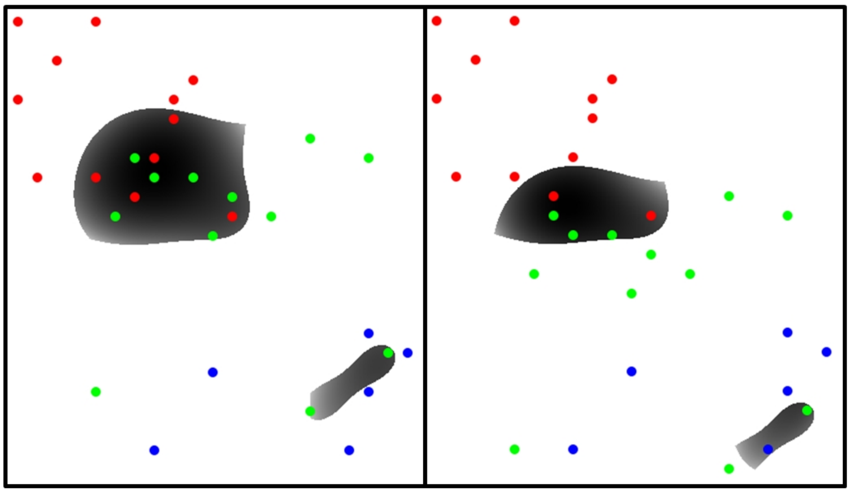


Figure 4.7: The gray shaded regions show the density overlap for data set A(left) and data set B(right) (published in [86]).

## 4.4 Automatic and Interactive Feature Engineering

In the previous chapter the possibilities of an improvement of the quality of low-dimensional embeddings of high-dimensional data by feature selections and feature space transformations were shown. The problems of these projections of high-dimensional data analysis are the large possibilities to project the data and therefore the large amount of resulting projections. In projection-based tasks for high-dimensional data analysis it is often unclear and unknown beforehand what is an optimal sight into data. What the user perceives to be optimal cannot be defined algorithmically and can differ from user to user and task to task.

As discussed former a lot of projection methods that deal with that problem and solve specific tasks very good have been developed in the past and can be adjusted with various parameters. This is not always feasible for the user. Additionally the defined measures to judge the resulting projections are dependent on their definition and influnce each other. E.g., a low stress value reflects preserving high-dimensional structure in the-low dimensional projection on one side, but closed clusters or separated classes are traded as good on the other side.

A solution to deal with this challenges is the combination of an automatic and an interactive feature engineering, which leads to a step-wise optimization. In the first step, an automatic search in the space of possible transformations brings up a number of candidate views for the user to inspect. This means the user gets an output that results out of a global optimization of the defined measures, namely the stress value and the overlap measures that can be parameter-weighted. Second, the system allows a refinement of this algorithmically global optimal projection, which is fixed and not flexible by the parameter setting, by user interaction. The user can select and deselect dimensions und his interaction takes into account his personal experience and requests, which leads to the user's personal local optimum.

The goal is an interactive HD data visualization system that supports a user-driven optimization of the projection. In this system the user, starting with an algorithmically optimal projection, can change the projection result him-self according to his perceives via adjustments like dimension selections. The real-time calculated measures help him to see the impact of his changes immediately. The main contribution is this combination of an algorithmically optimization with user interaction to guide the user quickly to find an optimal projection in terms of user satisfaction and the measures. This results in a task defined better projection via user interaction as a step-wise optimization like described above.

But the system is also able to cover other benefits: A descriptive presentation of the measures by numbers and visual at once und therefore a selectable stress value visualization, which results in a better understanding of the embedding methods. The latter can be used for all embedding methods to detect regions in the projection that are responsible for a high amount of the stress value.

### 4.4.1 Automatic Dimension Selection

The contradictory effect on the stress and overlap measures brings up the challenge in finding the optimal number of extended dimensions as well as the selection of the transformation strategy. In general, the space of possible feature transformations is huge. This means on the one hand an automatic dimension selection and feature extension and on the other hand an automatic dimension selection and dimensionality reduction. The first is done via the described feature extensions and a calculation of which of the dimension is used in terms of getting better measures. The second leads to a dimensionality reduction of the extended feature vector in two ways: First the automatic selection of only the most discriminative extensions, and second, the reduction of the initial features to the important ones.

Therefore the following values are calculated. First the range sum is defined as sum of the ranges of the values of all classes for a dimension x. Equation 4.9 shows the calculation:

$$rangesum_{\mathrm{d_x}} = \sum_{i=1}^{m} range(c_i, d_x) \tag{4.9}$$

Figure 4.8 illustrates the calculation of the range sum for dimension d1 over all clusters or classes from 1 to m.

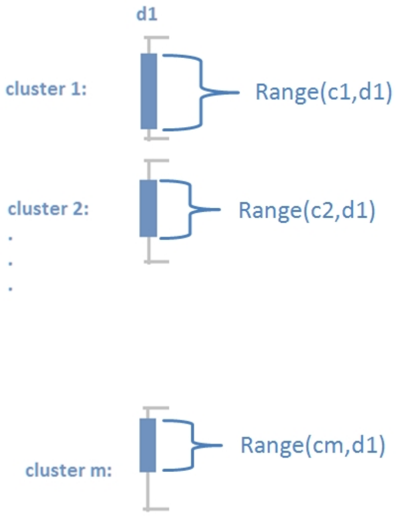


Figure 4.8: Illustration of range sum calculation for dimension d1 over all clusters or classes from 1 to m.

Next, the mean range as difference between the highest and lowest mean of the ranges for each dimension x is calculated, see Equation 4.10:

$$meanrange_{\mathrm{d_x}} = mean(range_{max}) - mean(range_{min}) \tag{4.10}$$

With this two measures the system can automatically reduce the feature space to important features via an automatically dimensionality reduction. It is also possible to find the most discriminative extensions of the dimensions. Therefore the following heuristics are implemented:

**Reduction to the important features:** The measure for the reduction to the important features is calculated as range sum divided through mean range. If a so calculated value is small is positive for this feature, because for descriptive features the range sum should be small and the mean range high.

**Selection of the most discriminative extensions:** For the extensions like defined in Chapter 4.3.2 the mean values within the classes are all the same. This results in range sum = 0 for all extended dimensions. The most discriminative extensions should have high differences, that means that the mean range should be high. Therefore extensions with a high mean range are used for the visual output the user gets for his interaction.

As a result, this means the system automatically selects a number of discriminative extensions and reduces the feature space to get a good trade off between overlap and stress value. The user gets a clearly represented view of the dimensions and their distributions as well as their usage for the resulting projection. The user can adjust this via interaction which is described in the next chapter.

### 4.4.2 Interactive Dimension Selection and Feature Extension for High-dimensional Data Projection

The indication of the projection quality helps the user to understand the trustworthiness of the projection. This means on the one hand that the quality indicators, respectively the defined measures, are mapped to the background to show the trustworthiness of the projection. And on the other hand that the used dimensions and their distributions are shown clearly.

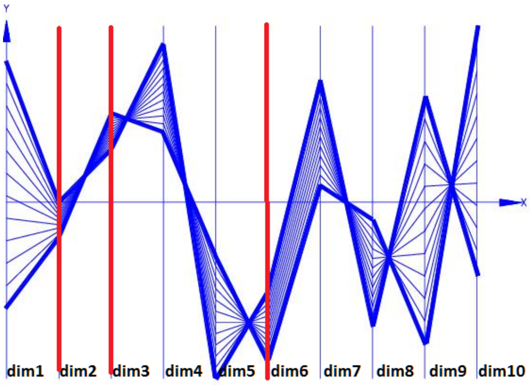


Figure 4.9: Important dimensions refer to a subset of dimensions that have the smallest value range (or variation), for example, assume all the data items in the parallel coordinates belong to the same cluster, it can be seen that dim2, dim3 and dim6 are the principal dimensions, because the data items have similar values in these dimensions.

Therefore for a clearly represented view additionally to the visualization window which displays the projection result, several visualizations that show the dimensions are displayed. Figure 4.9 shows an example for a parallel coordinates view and explains how such a view can help the user to identify important dimensions.

Embedded in the system such a parallel coordinates visualization is produced to show the distribution of data values over all dimensions of the input data. The interactive visualization panel allows the analysts to select dimensions for feature extensions based on the data distribution and their knowledge about the data. The quality of the projections is evaluated with the quality measures and is compared to select the one that has better quality. The analysts can iteratively repeat the process until a satisfactory projection is achieved.

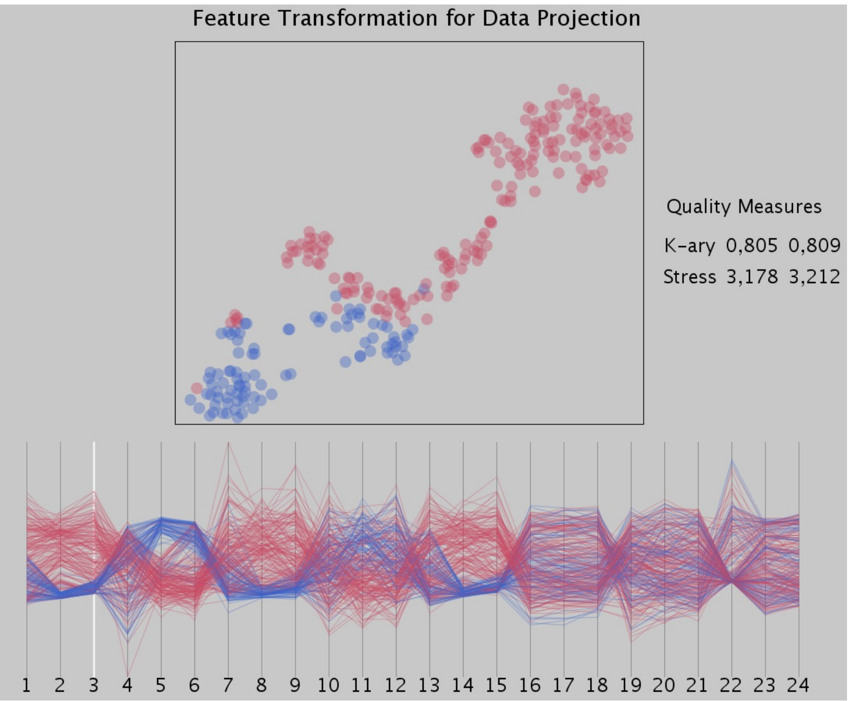


Figure 4.10: Example for interactive parallel coordinates visualization with a big distinction in the 5th and 6th dimension (published in [77], reproduced by kind permission of the Eurographics Association, (c) Eurographics Association 2013).

The parallel coordinates visualization shows the data distribution over all dimensions with different colors for each class. This view can be used to help the analyst identify dimensions that provide clear distinctions between different classes. For example, in Figure 4.10 from the parallel coordinates visualization it is not difficult to find out that there is a big distinction in the 5th and 6th dimension.

Another way for the visualization of the dimensions are bar charts. Figure 4.11 shows an example with the bar charts of the dimensions of each class displayed on the left. The system shown in this figure was implemented to integrate all the approaches described so far.

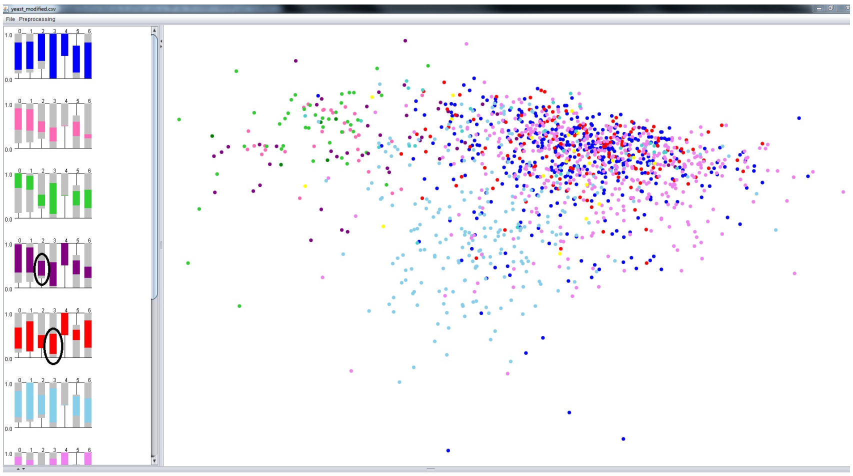
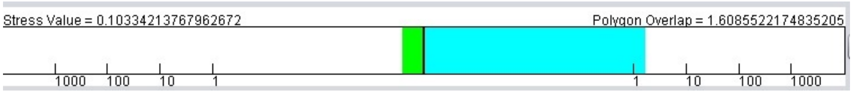


Figure 4.11: Example for interactive bar chart visualization. The bar charts of the dimensions of each class are displayed on the left. Circles indicate how a user can select specific dimensions in a specific class.

In this system it is also possible for the user to watch the effect of his interaction on the measures immediately, see Figure 4.12. This means the trade-off between the stress value and the overlap value is calculated in real-time and is shown by horizontal bars.


Stress Value = 0.10334213767962672

Polygon Overlap = 1.6085522174835205

Figure 4.12: Example for visualizing the trade-off between stress and overlap values with horizontal bars. The stress value is exposed to the left in green and the polygon overlap is exposed to the right in blue.

Examples will illustrate this in the following evaluation chapter in detail.

## 4.5 Evaluation

### 4.5.1 Evaluation Feature Engineering - Feature Selection

*Explored and described together with Michael Regenscheit, University of Konstanz, in parts of his student thesis, see [79].*

For the first evaluation in the field of feature engineering, the feature selection in the domain of multimedia database retrieval is examined. Multimedia database retrieval is an important issue in computer science since the data volume is increasing more and more. The definition of multimedia data is wide, mostly it consists of digital media data:

- Audio
- Images
- Graphics
- Text
- Video

These types of data are often stored as multi-dimensional data and therefore serve representative for this kind of data in this chapter. This part of the work is motivated from a very practical external project with industry, where a big image database with images and meta information about the images was the starting point for the analysis. A visual and interactive tool was implemented, which helps to select the semantically classified multimedia data, described by content-based numeric features, as well as to evaluate the goodness of these high-dimensional features visually in the 2D-space.

The characteristics they have in common is that it is possible to calculate content-based features for finding classifications and similarities in the data. This results in multi-dimensional databases, supplemented with metadata information, which are difficult to overview or it is hard to find similarities.

The database was more or less well-kept which means there was wrong, no or not complete data in it. The goal was to support the search in this image database with the extraction of content-based, numeric features which characterize the multimedia database objects. For example, with a color histogram and so on. That means that the challenge was to add a content level and use this content level to build up classes in contrast and comparison to the classes of the metadata level, which was not always correct filled in the database. This was a hard problem and there was not a solution for all cases, but for some it worked very well and the consistency of the metadata could be checked through this procedure.

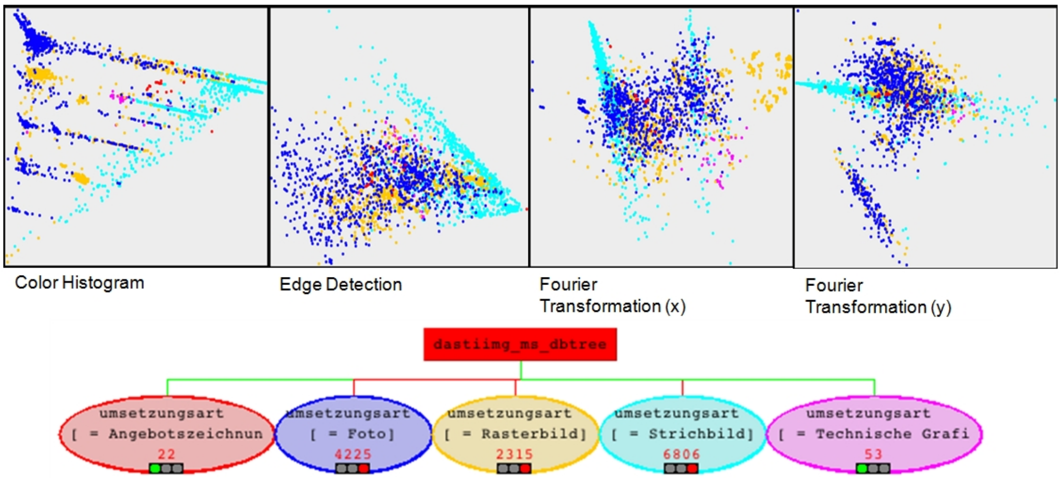


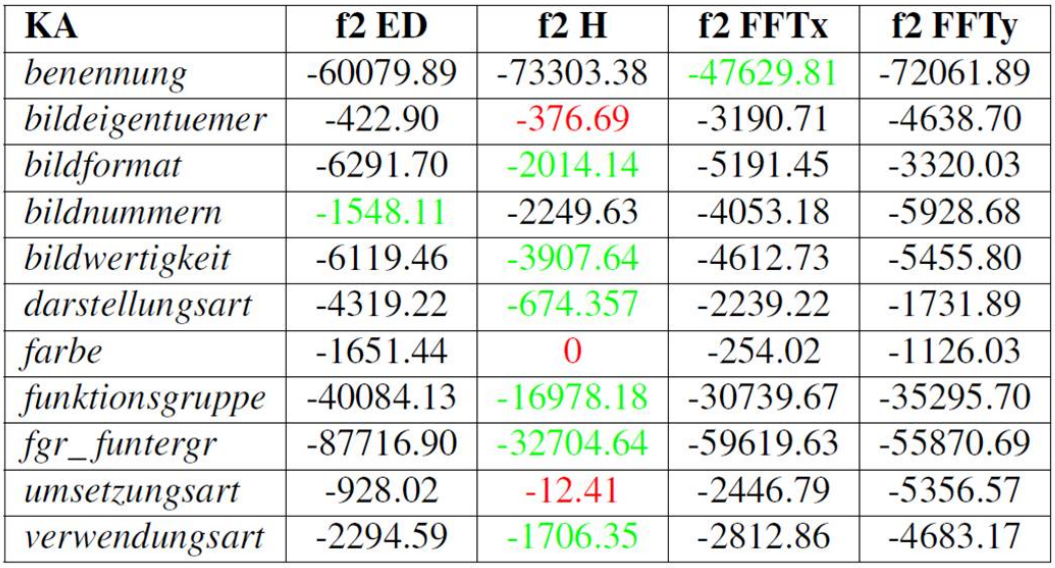
Figure 4.13: Feature engineering in multimedia databases: different features for a classification based on metadata: Five classes are shown at the bottom (e.g., red for “umsetzungsart=Angebotszeichnung”). Above is shown the 2D projection of four different calculated features. Visually the best is a color histogram because it separates the two biggest classes blue (“Foto”) and turquoise (“Strichbild”) at the best.

Although low-level feature extraction algorithms are well understood and are able to capture subtle differences between colors, textures, global color layouts, dominant color distributions, and so forth, the link between such low-level primitives and high-level semantic concepts remains an open problem. To narrow this so called semantic gap is a challenge that has drawn the attention of researchers in computer science in recent years and many low-level feature extraction techniques have been developed to retrieve relevant information from image databases.

Besides, the motivation was to give with the help of a constructible tree a new access to the data which is able to raise the understanding about the data.

The evaluations are done on image data, but the system can handle all kinds of feature-based multimedia data. The high-dimensional vectors and the results are visualized by a mapping to the 2D feature vector space with a MDS projection algorithm with a clustering according to the content level. Figure 4.13 shows an example of this approach.

Because of the fact that it cannot be expected that every content-based feature separates the semantic classes, different features for different classifications are needed, e.g., a color histogram algorithm for separating black/white images from color images or an edge detection algorithm for separating classes with images with different edges. To select the feature that works best for arbitrary selected classes, an automatic feature selection method was implemented using the calculated high-dimensional overlap measure like described in Chapter 4.3.1. The measure enables the calculated features to be evaluated numerically. The table in Figure 4.14 shows the results.



| KA | f2 ED | f2 H | f2 FFTx | f2 FFTy |
|---|---|---|---|---|
| *benennung* | -60079.89 | -73303.38 | -47629.81 | -72061.89 |
| *bildeigentuemer* | -422.90 | -376.69 | -3190.71 | -4638.70 |
| *bildformat* | -6291.70 | -2014.14 | -5191.45 | -3320.03 |
| *bildnummern* | -1548.11 | -2249.63 | -4053.18 | -5928.68 |
| *bildwertigkeit* | -6119.46 | -3907.64 | -4612.73 | -5455.80 |
| *darstellungsart* | -4319.22 | -674.357 | -2239.22 | -1731.89 |
| *farbe* | -1651.44 | 0 | -254.02 | -1126.03 |
| *funktionsgruppe* | -40084.13 | -16978.18 | -30739.67 | -35295.70 |
| *fgr_funtergr* | -87716.90 | -32704.64 | -59619.63 | -55870.69 |
| *umsetzungsart* | -928.02 | -12.41 | -2446.79 | -5356.57 |
| *verwendungsart* | -2294.59 | -1706.35 | -2812.86 | -4683.17 |

Figure 4.14: High-dimensional overlap measures for different classifications for different features (edge detection ED, color histogram H and two Fourier transformations FFT): Green is the best in each line for each classification and red highlights the three best ones separately. The best feature for a classification based on “umsetzungsart” is the color histogram again. So the visual impression from Figure 4.13 is supported.

### 4.5.2 Evaluation Feature Engineering - Feature Space Transformation and Quality Measures

In this section three different simple transformation strategies are proposed and their effectiveness by applying each of them to transform eight different data sets is demonstrated. For each setting and each data set, four different DR techniques to both the original and the transformed data are applied and the quality of the generated embeddings is compared. Two types of existing structural preservation measures are used, as well as the two overlap measures proposed in Chapter 4.3.3 to evaluate the quality of each embedding. Now first the transformation strategies, then the data, the embedding methods, the quality measures, and the results of the experiments are explained in detail.

#### Transformation Settings

Given a labeled data set with $n$ dimensions and $m$ items, the feature space can be transformed in many ways. As stated three very simple transformation strategies are applied to demonstrate the effectiveness and generalizability of the approach. The basic idea is to extend with mean values to leverage noise in the data and thus avoid visual cluttering in the embedding. In the first and second strategy, the dimension that has the highest range or spread is selected. The feature vector space is extended by adding an additional dimension with mean values of a relevant dimension assigned to all the data items to reduce the noise introduced by irrelevant dimensions. To further analyze the compromise between structural preservation and visual clutter avoidance, the third strategy extends all $n$ dimensions in the data by adding an additional dimension for each of them to the original feature space, with mean values of the dimension assigned to all data items. More specifically, the experiment extends a given feature space in the following ways:

1. Extend the initial feature vector with the mean value of the dimension that has the highest range.

2. Extend the initial feature vector with the mean value of the dimension that has the highest spread.

3. Extend the initial feature vector with mean values over all dimensions.

### Data

To evaluate the effectiveness of the different transformation settings, each approach is tested with eight data sets (see Table 4.4 for more details). The chosen data consists of four synthetic data sets, created by Gaussian functions with a grid, and four benchmark data sets that have been used by various recent visualization publications, including the *ecoliProteins* data which encodes amino acid proteins sequences from the *E.colie bacteria* [106], the *yeast* data set which denotes cellular localization sites of proteins citeuci, the *tse300* data set which records the weekly price history of 300 TSE index stocks in the year 2002 [106], and the *bbdm13* which is a subset from a hospital based case-control study designed to examine the epidemiology of fibrocystic breast disease [99].

| Name | Type | Points | Dimensions | Classes | Provenance |
|---|---|---|---|---|---|
| twoSquare | synth. | 968 | 3 | 4 | [90] |
| gauss-d5-3c | synth. | 500 | 5 | 3 | [90] |
| gauss-d5-5c | synth. | 500 | 5 | 5 | [90] |
| ecoliProteins | real | 332 | 7 | 8 | visumap [106] |
| yeast | real | 1452 | 8 | 9 | uci [30] |
| tse300 | real | 244 | 9 | 8 | visumap [106] |
| gauss-d10-5c | synth. | 500 | 10 | 5 | [90] |
| bbdm13 | real | 200 | 13 | 5 | umass [99] |

Table 4.4: List of data sets, ordered by number of dimensions (published in [86]).

### Embedding Methods

2D embeddings of both, the original data and the transformed data, are generated. Four DR techniques (PCA, Classical Scaling, Sammon's mapping and IDMAP) are used, that are implemented in the PEx (Projection Explorer) System [76], using the default parameter settings. PEx is a widely visualization tool for creating and exploring visual representations of high-dimensional data. PCA and Classical Scaling (MDS) are two of the most widely used linear DR techniques by existing visualization systems. Sammon's mapping is one of the most known nonlinear multi-dimensional scaling methods [82]. IDMAP is another nonlinear extension of MDS. It is an improved embedding technique for supporting visual exploration of multi-dimensional data sets [71].

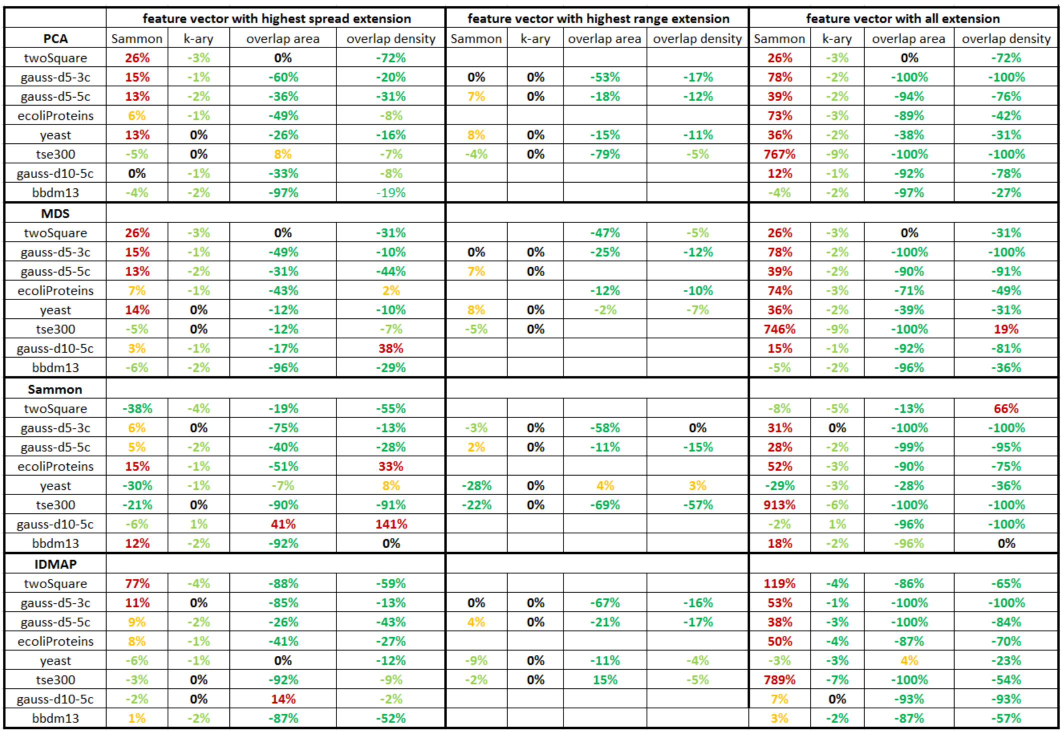
### Quality Measures

For each embedding the quality of its structural preservation is evaluated by two measures, namely *Sammon's stress* and *k-ary neighborhood measure*. The quality of clutter avoidance is evaluated by the two overlap measures, *overlap area* and *overlap density* as defined in Chapter 4.3.3.

### Results

The evaluation results are detailed in the table of Figure 4.15. Note that the Sammon's stress and k-ary neighborhood measure are always based on the distance matrix computed from the original data. First the difference between the stress measure and overlap measure before and after each transformation is compared. The colored columns show the difference. Red indicates the transformation worsened the quality of the embedding significantly (the value is more than 10 percent higher after transformation), orange indicates slightly worse performance (the value is higher, but less than 10% higher after transformation). Similarly, light green shows slight improvement ($< 10\%$ ) after the transformations and dark green indicates significant improvement ($> 10\%$). Empty fields in the table indicate the choice of range value and spread value do not make a difference to the result.

From the result, it is not surprising to see that most the studied feature vector transformations lead to a higher Sammon's stress value. This is not necessarily a bad sign, because the Sammon's stress records how well the lower dimensional embedding approximates the pairwise distances between data items in the high-dimensional data space. As mentioned previously, in high-dimensional data the real distances between data items are often obscured in the projection by the irrelevant dimensions and noise in the data. The feature space transformations aim to leverage noise in the data by extending relevant features, which naturally change the distances between pairwise data items in the transformed data space. With respect to neighborhood preservation, which is measured by the k-ary neighborhood measure, the result is significantly better than Sammon's stress. In majority of the cases the measure improves after



| | feature vector with highest spread extension | | | | feature vector with highest range extension | | | | feature vector with all extension | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PCA** | Sammon | k-ary | overlap area | overlap density | Sammon | k-ary | overlap area | overlap density | Sammon | k-ary | overlap area | overlap density |
| twoSquare | 26% | -3% | 0% | -72% | | | | | 26% | -3% | 0% | -72% |
| gauss-d5-3c | 15% | -1% | -60% | -20% | 0% | 0% | -53% | -17% | 78% | -2% | -100% | -100% |
| gauss-d5-5c | 13% | -2% | -36% | -31% | 7% | 0% | -18% | -12% | 39% | -2% | -94% | -76% |
| ecoliProteins | 6% | -1% | -49% | -8% | | | | | 73% | -3% | -89% | -42% |
| yeast | 13% | 0% | -26% | -16% | 8% | 0% | -15% | -11% | 36% | -2% | -38% | -31% |
| tse300 | -5% | 0% | 8% | -7% | -4% | 0% | -79% | -5% | 767% | -9% | -100% | -100% |
| gauss-d10-5c | 0% | -1% | -33% | -8% | | | | | 12% | -1% | -92% | -78% |
| bbdm13 | -4% | -2% | -97% | -19% | | | | | -4% | -2% | -97% | -27% |
| **MDS** | | | | | | | | | | | | |
| twoSquare | 26% | -3% | 0% | -31% | | | -47% | -5% | 26% | -3% | 0% | -31% |
| gauss-d5-3c | 15% | -1% | -49% | -10% | 0% | 0% | -25% | -12% | 78% | -2% | -100% | -100% |
| gauss-d5-5c | 13% | -2% | -31% | -44% | 7% | 0% | | | 39% | -2% | -90% | -91% |
| ecoliProteins | 7% | -1% | -43% | 2% | | | -12% | -10% | 74% | -3% | -71% | -49% |
| yeast | 14% | 0% | -12% | -10% | 8% | 0% | -2% | -7% | 36% | -2% | -39% | -31% |
| tse300 | -5% | 0% | -12% | -7% | -5% | 0% | | | 746% | -9% | -100% | 19% |
| gauss-d10-5c | 3% | -1% | -17% | 38% | | | | | 15% | -1% | -92% | -81% |
| bbdm13 | -6% | -2% | -96% | -29% | | | | | -5% | -2% | -96% | -36% |
| **Sammon** | | | | | | | | | | | | |
| twoSquare | -38% | -4% | -19% | -55% | | | | | -8% | -5% | -13% | 66% |
| gauss-d5-3c | 6% | 0% | -75% | -13% | -3% | 0% | -58% | 0% | 31% | 0% | -100% | -100% |
| gauss-d5-5c | 5% | -2% | -40% | -28% | 2% | 0% | -11% | -15% | 28% | -2% | -99% | -95% |
| ecoliProteins | 15% | -1% | -51% | 33% | | | | | 52% | -3% | -90% | -75% |
| yeast | -30% | -1% | -7% | 8% | -28% | 0% | 4% | 3% | -29% | -3% | -28% | -36% |
| tse300 | -21% | 0% | -90% | -91% | -22% | 0% | -69% | -57% | 913% | -6% | -100% | -100% |
| gauss-d10-5c | -6% | 1% | 41% | 141% | | | | | -2% | 1% | -96% | -100% |
| bbdm13 | 12% | -2% | -92% | 0% | | | | | 18% | -2% | -96% | 0% |
| **IDMAP** | | | | | | | | | | | | |
| twoSquare | 77% | -4% | -88% | -59% | | | | | 119% | -4% | -86% | -65% |
| gauss-d5-3c | 11% | 0% | -85% | -13% | 0% | 0% | -67% | -16% | 53% | -1% | -100% | -100% |
| gauss-d5-5c | 9% | -2% | -26% | -43% | 4% | 0% | -21% | -17% | 38% | -3% | -100% | -84% |
| ecoliProteins | 8% | -1% | -41% | -27% | | | | | 50% | -4% | -87% | -70% |
| yeast | -6% | -1% | 0% | -12% | -9% | 0% | -11% | -4% | -3% | -3% | 4% | -23% |
| tse300 | -3% | 0% | -92% | -9% | -2% | 0% | 15% | -5% | 789% | -7% | -100% | -54% |
| gauss-d10-5c | -2% | 0% | 14% | -2% | | | | | 7% | 0% | -93% | -93% |
| bbdm13 | 1% | -2% | -87% | -52% | | | | | 3% | -2% | -87% | -57% |

Figure 4.15: Stress, area and density overlap for initial feature vectors and their transformations for 4 synthetic and 4 real data sets (ordered by number of dimensions) for projection techniques PCA, MDS, Sammon's Mapping and IDMAP. The changes are colored in (red) orange, if the transformations perform (very) badly with respect to the initial feature vector. If the transformations perform better or much better, the percentages for the changes are colored in light or dark green (published in [86]).

transformation and there is no negative sign in the result. As a matter of fact, even with Sammon's stress, there are some positive cases. For example, with both *tse300* and *bbdm13* data sets, the quality of structural preservation actually improved after transformation in most of the cases. Given the fact that nearly all the overlap area and density measures decreased significantly after the transformation (as indicated by the dark green units in the table), this shows that transformations exist which can improve class separation without scarifying structure preservation.

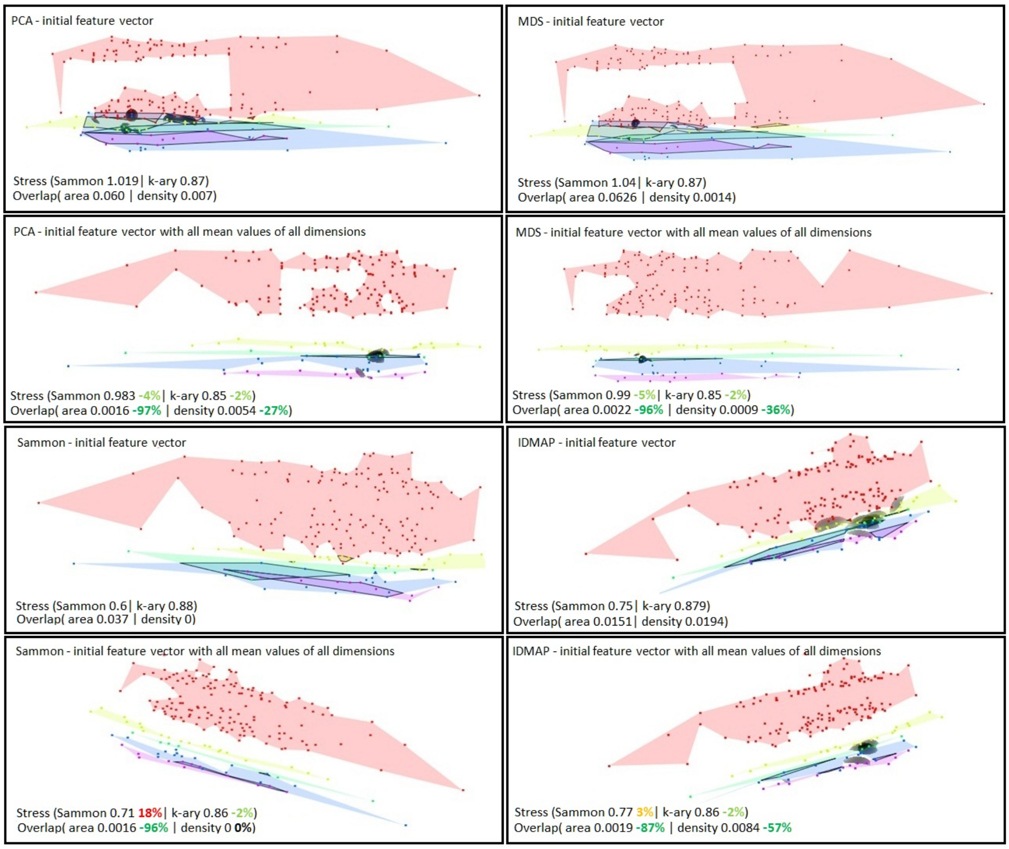


Figure 4.16: Embeddings of original and transformed "bbdm13" data using 4 different DR technique (PCA, Classical Scaling, Sammon's Mapping, and IDMAP) (published in [86]).

The top two quarters in Figure 4.16 show the embeddings of the *bbdm13* data on 2D display. I can be seen, that there is no big change in the structure in terms of positioning of classes and data items in the display, and on the other hand the classes are better separated after the transformation. This is also the case with the third transformation setting which simply adds all mean values over all dimensions.

The results of the evaluation are very encouraging and there is no DR technique or data set for which the transformation would not work at all. Even the trade-off between the structural preservation and the class separation is not as critical as anticipated. A big advantage of the proposed approach is its flexibility. The framework is DR technique independent - it can be applied to improve the quality of an embedding generated by any DR techniques. Also it is easy to extend the framework by integrating user interactions and feedbacks which will be shown later. But next the experimental results are assessed in more detail, before discussing some limitations of the approach.

### Detailed Assessment of the Experimental Results

For higher-dimensional data sets Sammon's stress values increase in total numbers, what is obvious, e.g., from comparing (*twoSquare*) (3 dimensions).to (*bbdm*) (13 dimensions). But in Figure 4.15 the results show that the higher the dimensionality, the less the stress value increases due to the transformations. In other words, the transformation performs better for data sets with higher dimensionality. The same is true for the number of classes the data set contains: The more classes the data set contains, the less increase of the stress value after transformation, but the trend is not very obvious.

In terms of class separation the opposite effect can be seen by looking at the overlap measures: The overlap measures decrease more for data with fewer dimensions. There is no obvious correlation between the number of classes and the increasing overlap measures. Nearly all overlap measures in the table decreased after transformation, indicated by the predominant green numbers in Figure 4.15. That means the transformations perform well on avoiding overplotting and visual cluttering. A decrease can also be seen in the example in Figure 4.16: The areas highlighted in black contours (overlap areas) and the gray shades (overlap density) are much smaller after the transformation. The area overlap measures decrease around 87% to 97%, and overlap density measures decreased around 27% to 57%.

Now the relationship between the number of extended dimensions and the quality of class separation is studied in terms of both overlap area and overlap density. First transformation setting 1 and 3 are applied as discussed in

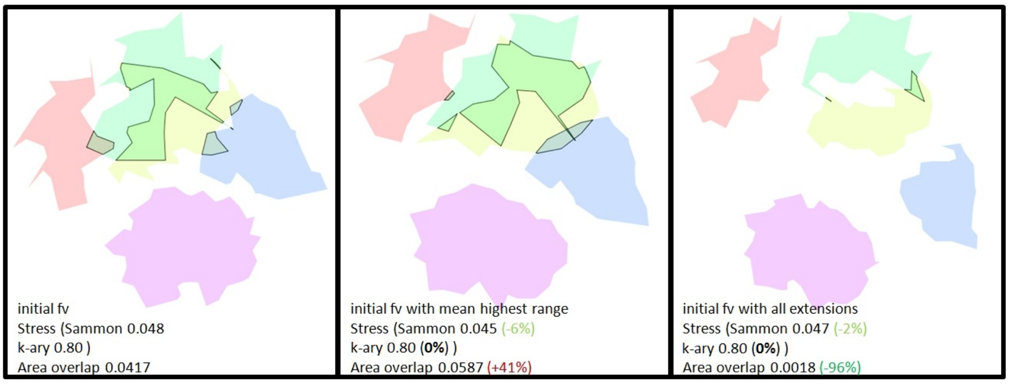


Figure 4.17: Overlapping areas in Sammon's Mapping of "gauss-d10-5c" data set. Left: with the initial feature vector; middle: with the extension of the feature vector with the mean value of the dimension that has the highest range; right: with the extension of the feature vector with all mean values of all dimensions (published in [86]).

Chapter 4.5.2 to the *gauss-10d-5c* data, and embeddings from both the original data set and transformed data sets are generated, using the same DR technique. The result is shown in Figure 4.17 where overlapping regions are highlighted in black regions. From the 3 figures it can be seen that the 3 embeddings have nearly the same Sammon's stress value, however, the overlap area measure is worsened by the 1st transformation strategy (+41%), and improved substantially after applying the 3rd transformation strategy (-96%). It can be assumed that extending more dimensions reduces the overall overlap area.

Next the same transformation settings are performed to the *twosquare* data set. Here a different correlation between the number of extended dimensions and the overlap density can be seen. As Figure 4.18 shows, the 1st transformation strategy gives the smallest overlap density measure (-55%) and at the same time reduced Sammon's stress measure substantially (-38%). The 3rd transformation strategy on the other hand worsened the over-plotting problem to a large extend (+66%) although it slightly improved the structural preservation (Sammon's stress -8%). It appears that adding too many dimensions brings negative effects to the overlap density measures.

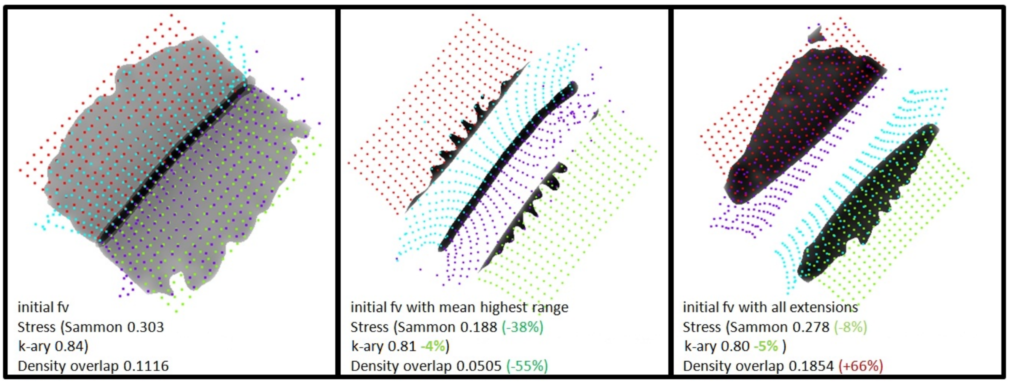


Figure 4.18: Area and density overlap of Sammon's Mapping of data set "twosquare". Left: with the initial feature vector; middle: extension of the feature vector with the mean value of the dimension that has the highest range; right: extension of the feature vector with all mean values of all dimensions (published in [86]).

### Limitations

The quality measures cover important aspects of the projections (stress and overlap). Defining overlap is subtle, and requires the definition of a hull model. The proposed concave hulls form intuitive and compact shapes, but other shapes are possible. The quality measures could serve as objective criteria for an in-depth search over the space of possible transformations, identifying the best result. This is why the possibility to integrate it in an interactive analysis environment where the user can explicitly change the extensions method, and select dimensions for extension, and interact with the embedding results will be presented in Chapter 4.5.3. The user is allowed to set a trade-off to weight the different quality measures, arriving at an application- and user-dependent best choice, which will be shown.

Note that the approach is based on availability of class labels. This means that a necessary preprocessing step for non-classified data could be an applying of a clustering or classification algorithm to classify the data. This possibility is also integrated in the final approach shown in Chapter 4.5.3.

### 4.5.3 Evaluation - Automatic and Interactive Feature Engineering

*This part of the work is partly explored and described together with Daniel Perez, University of Oviedo.*

As stated the approach was extended in terms of an automatic and interactive component for the feature engineering tasks, namely selections and transformations. The user is supported by visual assistance, which helps him to see the impact of his interaction immediately. To show this, first the visualizing of the measures will be shown with a consumption data set, see [77]. The visualized data used there exists of the measurement of several electrical and environmental dimensions, which are collected at a university building. Figure 4.19 shows the visual result with bar charts for the dimension visualization like explained in Chapter 4.4.2. It is possible to select and deselect dimensions and trigger a recalculation of the projection. The user can see the effect of his changes in the displayed numbers and the horizontal bars like explained in Chapter 4.4.2, as well as in the visualization itself.

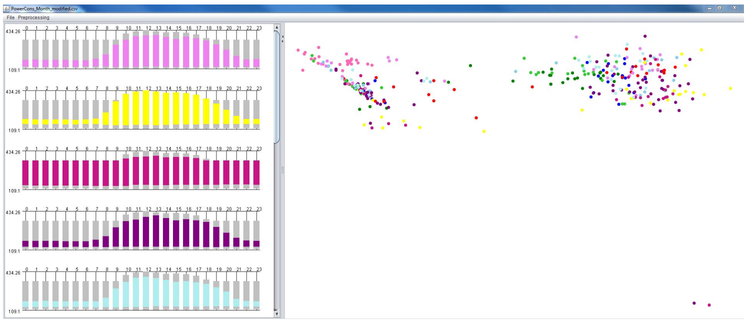


Figure 4.19: Example for visualization of electrical and environmental consumption data, collected at a university building with 10 classes that are colored accordingly in the visualized projection.

The following figures show how the measures are visualized. On the one hand Figure 4.20 left shows an illustration of the stress value with a circle visualization. Bigger circles around the points indicate a higher stress value causation. On the other hand Figure 4.20 right shows an illustration of the stress value with a Gaussian density function. Regions that cause a high stress value are colored in deep black.

Figure 4.21 shows the overlap visualizations: On the left side the polygon overlap and on the right side the point respectively the density overlap.
This means additional to the numbers and the complementary horizontal bars, like presented in Figure 4.12, the user can see visually the impact of his changes directly in the projection visualization. This makes it possible to detect areas in the projection that have a high influence on the measures. This is a main contribution for the task of developing the pure visualization of data to a real visual analytics system.

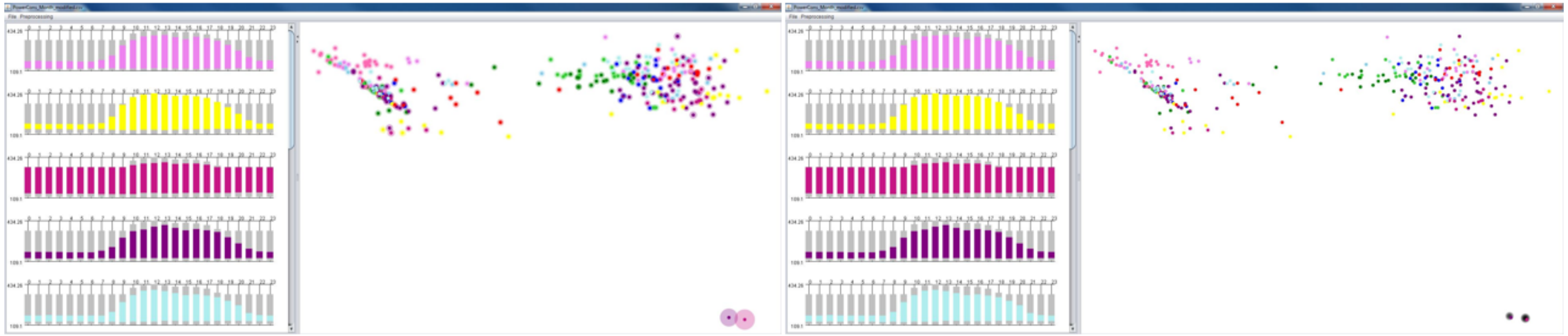


Figure 4.20: Left: Example for a stress value illustration with circle visualization: Bigger circles around the points indicate a higher stress value causation. Right: Example for a stress value illustration with a Gaussian density function: Dense areas that mean high stress value causation are colored in deep black. It is possible to select and deselect dimensions in each visualization on the left side via the bar charts to trigger a recalculation of the projection and the stress value.

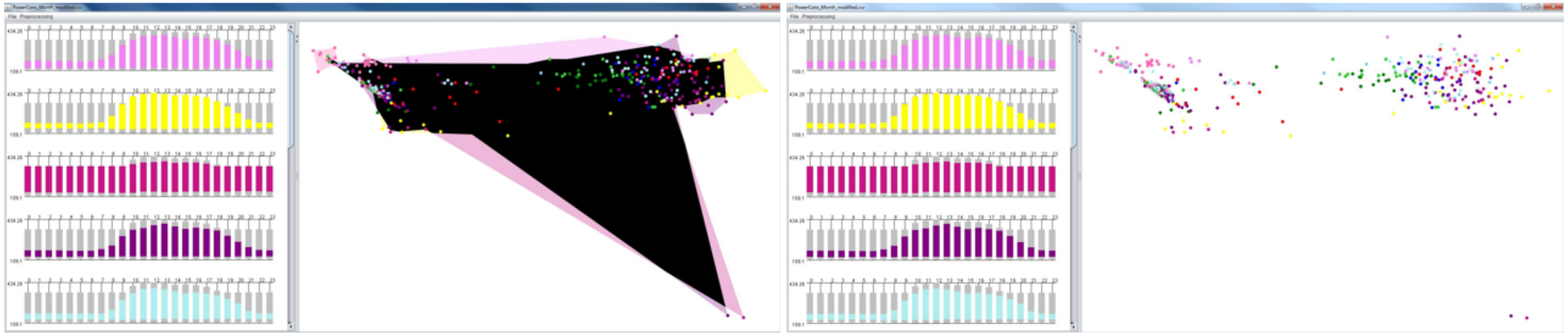


Figure 4.21: Left: Example for overlap value visualization with polygon overlaps: The polygon overlap is shown by the black area. Right: Example for overlap value visualization with a Gaussian density function: Dense areas mean a high overlap and are colored in black. It is possible to select and deselect dimensions in each visualization on the left side via the bar charts to trigger a recalculation of the projection and the overlap values.

### 4.5.4 Evaluation - Visual Cluster Analysis

In this final chapter the loop is drawn to Chapter 3 and the link is shown between the different types of multi-dimensional data, namely event data and feature data considered in this thesis. The approaches proposed here can be combined and used for a common analysis that supports and validates the results of each other, as well as it complements each other and shows possible other perspectives of the data. Therefore the visual clustered 3D data evaluated in Chapter 3.4.3 is projected. Figure 4.22 left shows the projection before the visual clustering and Figure 4.22 right shows the same data after the visual clustering.

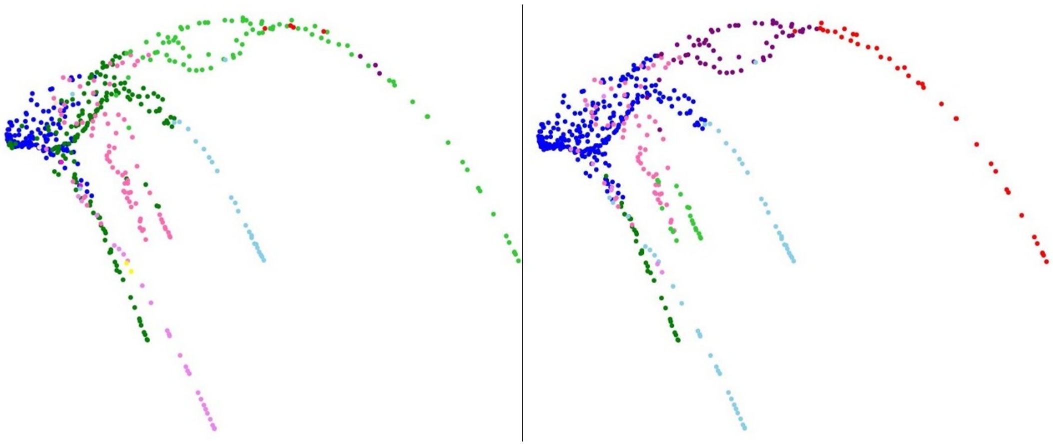


Figure 4.22: Projections of the time related 3D data. Left: With colored class labels before the visual clustering. Right: With colored class labels after the visual clustering.

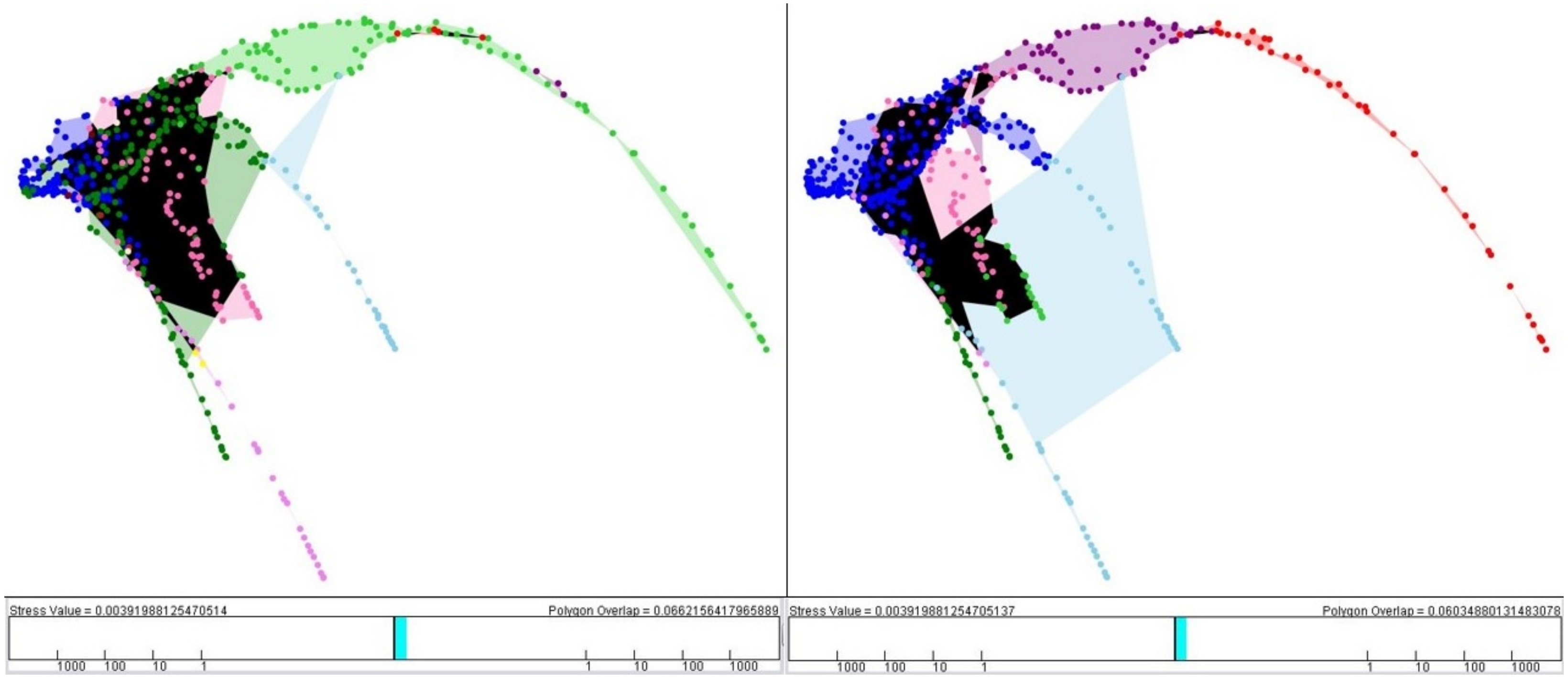
Of course stress values are the same because of the same underlying data, but the overlap values decrease clearly, the polygon overlap by 9% and the point overlap by 45%.

Table 4.5 shows the detailed numbers:

| *data* | *stress value* | *polygon overlap* | *point overlap* |
|---|---|---|---|
| time related 3D data before visual clustering | 0.0039 | 0.066 | 0.0108 |
| time related 3D data after visual clustering | 0.0039 | 0.060 | 0.0059 |
| change in % | 0% | -9% | -45% |

Table 4.5: Stress and overlap values for the time related 3D data before and after the visual clustering.

This means the findings of Chapter 3.4.3, a better clustering through the visual clustering are supported, by numbers as well as visually. This can be seen in the following Figures 4.23 and 4.24 that face the projections and the visualizations of the measures before and after the visual clustering. The polygon an the point overlaps are visualized by black regions and the numbers are displayed on the bottom of each visualization.



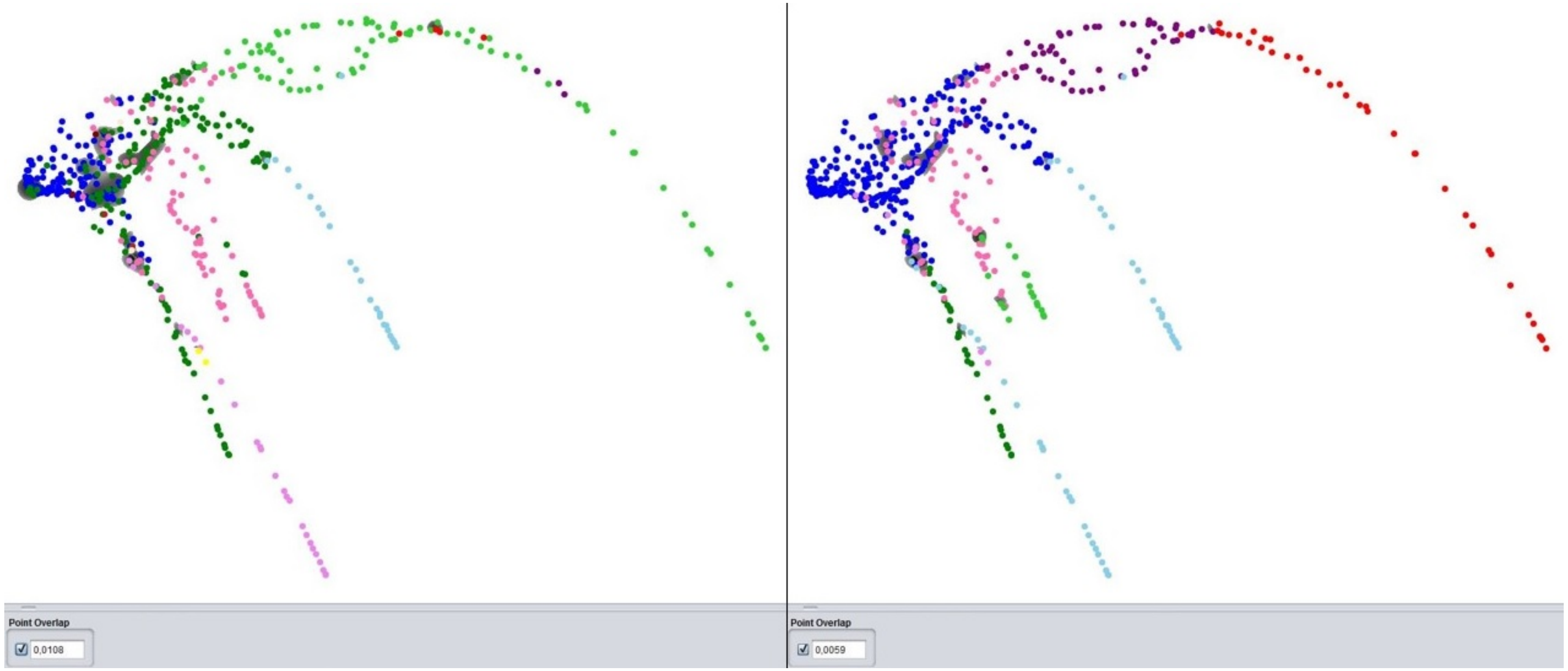
Stress Value = 0.00391988125470514

Polygon Overlap = 0.0662156417965889

Stress Value = 0.003919881254705137

Polygon Overlap = 0.06034880131483078

Figure 4.23: Left: Projection of the time related 3D data before the visual clustering. Right: Projection of the time related 3D data after the visual clustering. The polygons are colored according to the classes and the polygon overlap is shown by the black area. The polygon overlap decreases from 0.066 to 0.060, which means 9%.



Point Overlap

☑ 0,0108

Point Overlap

☑ 0,0059

Figure 4.24: Left: Projection of the time related 3D data before the visual clustering. Right: Projection of the time related 3D data after the visual clustering. The point overlap is shown by the black areas and decreases from 0.0108 to 0.0059, which means 45%.

# Chapter 5

# Concluding Remarks and Perspectives

In this thesis two approaches have been developed and evaluated, that provide the ability to perform visual analytics of multi-dimensional data types in terms of improving the exploration and projection of this kind of data. The approaches can be combined and it was shown which additional benefit is provided. But there is still the opportunity to further improve the developed approaches, which will also enable them to be used for more projects and applications.

One application which is not part of the thesis but should be followed up is the visual analytics in combination with economic models. The paper [84] shows first ideas in this direction and describes a visual data analysis tool that enables an interactive analysis of stock time series data globally and locally. A pixel-based visualization technique, which can display a large volume of data without overlaps, and a line graph visualization, which provides an intuitive understanding of patterns and trends of this stock time series data, are used in this approach. It gets clear that the stock prices over time can be globally visualized with pixel-based techniques. For further analysis economic factor models should provide an explanation for the majority of the time series and deviations from the model due to special events should be highlighted. Then the user is able to delve deeper into special events and see more levels of detail, such as the line chart itself or relevant news articles about the stock.

To deal with this tasks further improvements are possible to the pattern detection algorithm and also in guiding the user automatically to interesting patterns.

Further future work includes user feedback and evaluations in the feature engineering system. The user could be given the possibility to indicate the correctness of the results, perhaps via drag-and-drop of a falsely classified object to its correct class in the 2D space. This could then be used as input for the system and for refinement of future feature engineering. As a result, the similarity measure changes with this user interaction.

Other future work can be done regarding the measures. For example, the proposed concave hulls approach for the polygon overlap form intuitive and compact shapes, but other shapes are possible.

At last the user interaction could to be evaluated in a bigger style. One way of evaluation done in this work via use and business cases is surely helpful but always specific for respected case. This does not lower the value but is fragile to generalize the results. Ideas of a neutral evaluation therefore would be to inspire more people to play around with visual analytic systems, like done in the google image labeler or the gopher game [17]. This paper introduces Gophers, a social game for mobile devices that utilizes task oriented gameplay to create a novel entertainment experience. The study combines a number of key research themes: mobile social gaming, acquiring useful data through gameplay and content sharing in mobile settings. The experience of trialing the game in the real world discussed there and the findings from the study look very promising.

In conclusion, it can be said that the rising amount and complexity of multi-dimensional data makes it necessary to find automated and interactive approaches to get new insights and knowledge of this kind of data via visual analytics systems. This thesis has described some relevant research questions in the field of improving the exploration and projection of different kinds of multi-dimensional data. There are still interesting and practically relevant research questions in this area that will be further developed in the future.

# Bibliography

[1] W. Aigner, S. Miksch, W. Muller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):47–60, 2008.

[2] M. Ankerst, D. A. Keim, and H.-P. Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In *Proc. Visualization '95, Atlanta, GA*, pages 279–286, 1995.

[3] M. Ankerst, D. A. Keim, and H.-P. Kriegel. Circle segments: A technique for visually exploring large multidimensional data sets. In *Visualization '96, Hot Topic Session, San Francisco, CA*, 1996.

[4] Benjamin B. Bederson, Aaron Clamage, Mary P. Czerwinski, and George G. Robertson. Datelens: A fisheye calendar interface for pdas. *ACM Trans. Comput.-Hum. Interact.*, 11(1):90–119, 2004.

[5] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591. MIT Press, 2001.

[6] Jacques Bertin. *Sémiologie graphique : les diagrammes, les réseaux, les cartes.* Bordas, Paris, 1967.

[7] E. Bertini, P. Hertzog, and D. Lalanne. SpiralView: towards security policies assessment through visual correlation of network resources with evolution of alarms. In *IEEE Symposium on Visual Analytics Science and Technology, 2007. VAST 2007*, pages 139–146, 2007.

[8] Enrico Bertini and Giuseppe Santucci. Visual quality metrics. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel*

*evaluation methods for information visualization*, BELIV '06, pages 1–5. ACM, 2006.

[9] Enrico Bertini, Andrada Tatu, and Daniel Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *Proceedings of the IEEE Symposium on IEEE Information Visualization (InfoVis)*, 17:2203–2212, 2011.

[10] Christopher M. Bishop, Markus Svensén, and Christopher K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[11] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97(1-2):245–271, December 1997.

[12] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[13] S. Bremm, T. von Landesberger, J. Bernard, and T. Schreck. Assisted descriptor selection based on visual comparative data analysis. *Wiley-Blackwell Computer Graphics Forum*, 30(3):891–900, 2011. (Proceedings of Eurographics / IEEE-VGTC Symposium on Visualization 2011).

[14] S. Bremm, T. v. Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacher. Interactive comparison of multiple trees. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 31–40. IEEE Computer Society, 2011.

[15] Vanja Buvac. Internet General Inquirer, 2008. http://www.webuse.umd.edu:9090/ as retrieved on Nov. 14.

[16] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[17] Sean Casey, Ben Kirman, and Duncan Rowland. The gopher game: a social, mobile, locative game with user generated content and peer review. In Masa Inakage, Newton Lee, Manfred Tscheligi, Regina Bernhaupt,

and Stéphane Natkin, editors, *Advances in Computer Entertainment Technology*, volume 203 of *ACM International Conference Proceeding Series*, pages 9–16. ACM, 2007.

[18] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. WireVis: Visualization of categorical, time-varying data from financial transactions. In *IEEE Symposium of Visual Analytics Science and Technology*, pages 155–162, 2007.

[19] Keke Chen and Ling Liu. Validating and refining clusters via visual rendering. In *ICDM*, pages 501–504. IEEE Computer Society, 2003.

[20] Keke Chen and Ling Liu. ivibrate: Interactive visualization-based framework for clustering large datasets. *ACM Trans. Inf. Syst.*, 24(2):245–294, 2006.

[21] L. Chen. *Local multidimensional scaling for nonlinear dimensionality reduction, graph layout, and proximity analysis.* PhD thesis, University of Pennsylviana, July 2006.

[22] Jaegul Choo, Hanseung Lee, Jaeyeon Kihm, and Haesun Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *IEEE VAST*, pages 27–34, 2010.

[23] Christopher Collins, Gerald Penn, and M. Sheelagh T. Carpendale. Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Trans. Vis. Comput. Graph.*, 15(6):1009–1016, 2009.

[24] David McCandless. Information Is Beautiful, 2010, timestamp = 2010.02.04. http://www.davidmccandless.com/.

[25] J. de Leeuw and W. Heiser. Theory of multidimensional scaling. In *Handbook of Statistics*, chapter 13, pages 285–316. North-Holland Publishing Company, Amsterdam, 1982.

[26] P. Demartines and J.Hérault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of datasets. *IEEE Transactions on Neural Networks*, 8(1):148–154, 1997.

[27] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley-Interscience, New York, 2nd edition, 2001.

[28] Alex Endert, Chao Han, Dipayan Maiti, Leanna House, Scotland Leman, and Chris North. Observation-level interaction with statistical models for visual analytics. In *IEEE VAST*, pages 121–130. IEEE, 2011.

[29] Stephen L. France and J. Douglas Carroll. Two-way multidimensional scaling: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(5):644–661, 2011.

[30] A. Frank and A. Asuncion. University of California Irvine (UCI) Machine Learning Repository, 2010.

[31] Antony Galton and Juan Carlos Augusto. Two approaches to event definition. In *DEXA '02: Proceedings of the 13th International Conference on Database and Expert Systems Applications*, pages 547–556, London, UK, 2002. Springer-Verlag.

[32] John F. Gantz, Christopher Chute, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgang Schlichting, and Anna Toncheva. The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011. An idc white paper - sponsored by emc, IDC, March 2008.

[33] P. Gatalsky, N. Andrienko, and G. Andrienko. Interactive analysis of event data using space-time cube. In *International Conference on Information Visualisation*, volume 8, pages 145–152, 2004.

[34] R.L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, 1(4):132–133, 1972.

[35] Valery Guralnik and Jaideep Srivastava. Event detection from time series data. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–42, New York, NY, USA, 1999. ACM.

[36] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research,Special Issue on Variable and Feature Selection*, (3):1157–1182, 2003.

[37] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: part i. *SIGMOD Rec.*, 31(2):40–45, June 2002.

[38] M. Hao, D. Keim, U. Dayal, and T. Schreck. Importance-driven visualization layouts for large time series data. In *IEEE Symposium on Information Visualization (InfoVis 2005)*, 2005.

[39] M. Hao, D. Keim, U. Dayal, and T. Schreck. Multi-resolution techniques for visual exploration of large time-series data. In *Eurographics/IEEE-VGTC Symposium on Visualization, Norrkoping, Sweden*, 2007.

[40] A. Hinneburg, C. Aggarwal, and D. Keim. What is the nearest neighbor in high dimensional spaces? In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 506–515, 2000.

[41] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[42] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[43] Zhexue Huang, Michael K. Ng, and David Wai-Lok Cheung. An empirical study on the visual cluster validation method with fastmap. In *DASFAA*, pages 84–91. IEEE Computer Society, 2001.

[44] Stephen Ingram, Tamara Munzner, Veronika Irvine, Melanie Tory, Steven Bergner, and Torsten Möller. DimStiller: Workflows for dimensional analysis and reduction. In *IEEE VAST*, pages 3–10. IEEE, 2010.

[45] Alfred Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[46] R.A. Jarvis. On the identification of the convex hull of a finite set of points in the plane. *Information Processing Letters*, 2(1):18–21, 1973.

[47] Dong Hyun Jeong, Caroline Ziemkiewicz, Brian D. Fisher, William Ribarsky, and Remco Chang. iPCA: An interactive system for PCA-based visual analytics. *Comput. Graph. Forum*, 28(3):767–774, 2009.

[48] Jimmy Johansson and Matthew D. Cooper. A screen space quality method for data abstraction. *Comput. Graph. Forum*, 27(3):1039–1046, 2008.

[49] Sara Johansson and Jimmy Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):993–1000, 2009.

[50] C Loehlin John. *Latent variable models: an introduction to factor, path, and structural analysis.* L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1986.

[51] I. Jolliffe. *Principal Component Analysis.* Springer, 3rd edition, 2002.

[52] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu. Pixel bar charts: A visualization technique for very large multi-attribute data sets. *Visualization, San Diego 2001, extended version in: Information Visualization Journal, Palgrave*, 1(2), 2002.

[53] Daniel A. Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Joern Kohlhammer, and Guy Melancon. Visual Analytics: Definition, Process, and Challenges. In *Dagstuhl group report*, 2007.

[54] Daniel A. Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. *Mastering The Information Age - Solving Problems with Visual Analytics.* Eurographics, 2010.

[55] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual Analytics: Scope and Challenges. In Simeon Simoff, Michael H. Boehlen, and Arturas Mazeika, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics.* Springer, 2008. Lecture Notes in Computer Science (LNCS).

[56] Daniel A. Keim, Tilo Nietzschmann, Norman Schelwies, Jörn Schneidewind, Tobias Schreck, and Hartmut Ziegler. FinDEx: A spectral visualization system for analyzing financial time series data. In *EuroVis 2006: Eurographics/IEEE-VGTC Symposium on Visualization, Lisbon, Portugal, 8-10 May*, 2006.

[57] Stasko-J.T. Fekete J.-D. Kerren, A. and North C. *Information Visualization.* Springer, 2008.

[58] Jerry Kiernan and Evimaria Terzi. Eventsummarizer: a tool for summarizing large event sequences. In *EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology*, pages 1136–1139, New York, NY, USA, 2009. ACM.

[59] J. Koplowitz and D. Jouppi. A more efficient convex hull algorithm. *Information Processing Letters*, 7(1):56–57, 1978.

[60] Eleftherios E. Koutsofios, Stephen C. North, Russell Truscott, and Daniel A. Keim. Visualizing large-scale telecommunication networks and services (case study). In *VIS '99: Proceedings of the conference on Visualization '99*, pages 457–461, Los Alamitos, CA, USA, 1999. IEEE Computer Society Press.

[61] J.B. Kruskal. Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation". In R.C. Milton and J.A. Nelder, editors, *Statistical Computation*, pages 427–440. Academic Press, New York, 1969.

[62] Nitin Kumar, Nishanth Lolla, Eamonn Keogh, Stefano Lonardi, and Chotirat Ann Ratanamahatana. Time-series bitmaps: a practical visualization tool for working with large time series databases. In *SIAM 2005 Data Mining Conference*, pages 531–535. SIAM, 2005.

[63] J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction.* Springer, 2007.

[64] J.A. Lee and M. Verleysen. Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods. In Y. Saeys, H. Liu, I. Inza,

L. Wehenkel, and Y. Van de Peer, editors, *JMLR Workshop and Conference Proceedings (New challenges for feature selection in data mining and knowledge discovery)*, volume 4, pages 21–35. September 2008.

[65] J.A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7–9):1431–1443, 2009.

[66] J. M. Lewis, L. van der Maaten, and V. de Sa. A behavioral investigation of dimensionality reduction. In *Proc. 34th Conf. of the Cognitive Science Society (CogSci)*, pages 671–676, 2012.

[67] H. Liu and H. Motoda. *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007.

[68] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Min. Knowl. Discov.*, 1(3):259–289, 1997.

[69] Florian Mansmann, Daniel A. Keim, Stephen C. North, Brian Rexroad, and Daniel Sheleheda. Visual Analysis of Network Traffic for Resource Planning, Interactive Monitoring, and Interpretation of Security Threats. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 2007.

[70] Peter McLachlan, Tamara Munzner, Eleftherios Koutsofios, and Stephen North. Liverac: interactive visual exploration of system management time-series data. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1483–1492, New York, NY, USA, 2008. ACM.

[71] Rosane Minghim, Fernando V. Paulovich, and Alneu de Andrade Lopes. Content-based text mapping using multi-dimensional projections for exploration of document collections. In *Vis. and Data Analysis 2006*, Proc. SPIE-IS&T Electronic Imaging, pages 259–270, San Jose, California, USA, 2006. SPIE.

[72] Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2227–2236, 2013.

[73] Megan Monroe, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman, J. Millstein, and S. Gold. Exploring point and interval event patterns: Display methods and interactive visual query. *University of Maryland Technical Report*, 2012.

[74] Adriano J. C. Moreira and Maribel Yasmina Santos. Concave hull : a k-nearest neighbours approach for the computation of the region occupied by a set of points. pages 61–68, 2007.

[75] J. G. S. Paiva, L. Florian-Cruz, H. Pedrini, G. P. Telles, and R. Minghim. Improved similarity trees and their application to visual data classification. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2459–2468, 2011.

[76] Fernando V. Paulovich, Maria Cristina F. Oliveira, and Rosane Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing*, SIBGRAPI '07, pages 27–36, Washington, DC, USA, 2007. IEEE Computer Society.

[77] Daniel Perez, Leishi Zhang, Matthias Schaefer, Tobias Schreck, Daniel A. Keim, and Irene Diaz. Interactive Visualization and Feature Transformation for Multidimensional Data Projection. In *Proc. EuroVis Workshop on Visual Analytics Using Multidimensional Projections*, 2013.

[78] Doantam Phan, John Gerth, Marcia Lee, Andreas Paepcke, and Terry Winograd. Visual analysis of network flow data with timelines and event plots. In *VizSEC 2007*. Springer, 2008.

[79] Michael Regenscheit. Multimedia datenbank retrieval: Visuelle & interaktive analyse von multimedia daten, 2010.

[80] H. Rosling. Gapminder. Available from http://www.gapminder.org/. Accessed on Mar. 30, 2010.

[81] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[82] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, 18(5):401–409, May 1969.

[83] L.K. Saul and S.T. Roweis. Think globally, fit locally: Unsupervised learning of nonlinear manifolds. *Journal of Machine Learning Research*, 4:119–155, June 2003.

[84] Matthias Schaefer, Franz Wanner, Roman Kahl, Leishi Zhang, Tobias Schreck, and Daniel A. Keim. A Novel Explorative Visualization Tool for Financial Time Series Data Analysis. In *Proc. UKVAC Workshop on Visual Analytics*, September 2011.

[85] Matthias Schaefer, Franz Wanner, Florian Mansmann, Christian Scheible, Verity Stennett, Anders T. Hasselrot, and Daniel A. Keim. Visual Pattern Discovery in Timed Event Data. In *Proceedings of Conference on Visualization and Data Analysis*. SPIE, 2011.

[86] Matthias Schaefer, Leishi Zhang, Tobias Schreck, Andrada Tatu, John A. Lee, Michel Verleysen, and Daniel A. Keim. Improving projection-based data analysis by feature space transformations. In *In Proceedings of VDA 2013*, 2013.

[87] T. Schreck and C. Panse. A new metaphor for projection-based visual analysis and data exploration. In *IS&T/SPIE Conference on Visualization and Data Analysis*, pages 64950L.1–64950L.12. SPIE Press, 2007.

[88] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner. Dimensionality reduction in the wild: Gaps and guidance. *Technical report, Dept. of Computer Science, University of British Columbia*, 2012.

[89] M. Sedlmair, T. Munzner, and M. Tory. Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2634–2643, December 2013.

[90] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31(3):1335–1344, 2012. (Proceedings of Eurographics / IEEE-VGTC Symposium on Visualization 2012).

[91] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *In IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[92] Mike Sips, Boris Neubert, John P. Lewis, and Pat Hanrahan. Selecting good views of high-dimensional data using class consistency. *Comput. Graph. Forum*, 28(3):831–838, 2009.

[93] Robert Spence. *Information Visualization: Design for Interaction (2Nd Edition).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2007.

[94] Martin Suntinger, Hannes Obweger, Josef Schiefer, and M. Eduard Groeller. Event tunnel: Exploring event-driven business processes. *IEEE Computer Graphics and Applications*, 28:46–55, 2008.

[95] Andrada Tatu. Multimedia datenbank retrieval: Suche in bilddatenbanken mit hilfe klassifikations-basierter featureselektion, 2009.

[96] Andrada Tatu, Georgia Albuquerque, Martin Eisemann, Joern Schneidewind, Holger Theisel, Marcus Magnor, and Daniel Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)*, pages 59–66, 2009.

[97] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):2319–2323, December 2000.

[98] Warren Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, December 1952.

[99] University of Massachusetts. Statistical Data and Software Help, 2011. http://www.umass.edu/statdata/statdata/, last accessed Nov. 2011.

[100] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[101] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. *Tilburg University Technical Report*, 2009.

[102] Jarke J. Van Wijk and Edward R. Van Selow. Cluster and calendar based visualization of time series data. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 4–9. IEEE Computer Society, 1999.

[103] J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Proceedings of ICANN 2001*, pages 485–491. Springer, Berlin, 2001.

[104] J. Venna and S. Kaski. Nonlinear dimensionality reduction as information retrieval. In M. Meila and X. Shen, editors, *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, pages 568–575. San Juan, Puerto Rico, March 2007.

[105] Jarkko Venna and Samuel Kaski. Comparison of visualization methods for an atlas of gene expression data sets. *Information Visualization*, 6(2):139–154, May 2007.

[106] VisuMap Technologies Inc. VisuMap Data Repository, 2011. http://www.visumap.net/, last accessed Nov. 2011.

[107] Taowei David Wang. Interactive visualization techniques for searching temporal categorical data, 2010.

[108] Franz Wanner, Christian Rohrdantz, Florian Mansmann, Daniela Oelke, and Daniel A. Keim. Visual sentiment analysis of rss news feeds featuring the us presidential election in 2008. In *In Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW)*, 2009.

[109] Matthew O. Ward, Georges Grinstein, and Daniel A. Keim. *Interactive Data Visualization: Foundations, Techniques, and Application.* A. K. Peters, Ltd, May 2010.

[110] Marc Weber, Marc Alexa, and Wolfgang Muller. Visualizing time-series on spirals. *Information Visualization, IEEE Symposium on*, 0:7, 2001.

[111] Axel Wismüller, Michel Verleysen, Michaël Aupetit, and John Aldo Lee. Recent advances in nonlinear dimensionality reduction, manifold and topological learning. In *ESANN*, 2010.

[112] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. Lifeflow: Visualizing an overview of event sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1747–1756, New York, NY, USA, 2011. ACM.