COMENIUS UNIVERSITY BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS

PROJECT OF DISSERTATION

2006

Matej Novotný

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS, PHYSICS AND INFORMATICS
INSTITUTE OF APPLIED INFORMATICS

PROJECT OF DISSERTATION

Information Visualization of Large Data

AUTHOR: MATEJ NOVOTNÝ
SUPERVISOR: DOC. DR.-TECHN. ING. MILOŠ ŠRÁMEK

BRATISLAVA, 2006

Abstract

The scope of the hereby presented dissertation project is the effective information visualization of large data. This thesis declares the aims and goals of the project along with documenting the previous work on this topic. Several original results, which have already been published, are presented as well.

In its early stages the project investigates the issues introduced to an information visualization environment by the increase of the respective data volume. These are either bottlenecks inside the visualization pipeline handled by the computer or perception issues within the human visual system on the side of the observer. Both of these aspects are considered.

After a successful identification of the key impact of large data on visualization, the project employs several means of modifying the original techniques (along with presenting several original ones) to improve the situation. These concepts include visual abstraction, data reduction, output-sensitive rendering and (to a certain extent) also hardware acceleration.

Contents

Chapter 1

Introduction

One of the most sophisticated senses of a human is the sense of vision. The amount and the complexity of information that the human visual system is able to process is great. Graphical depiction of real world phenomena has therefore accompanied the human civilization since its very beginning. Whether it is artistic imagery or technical drawing, either of them has the potential to communicate complicated information. In large number of cases it also surpasses the verbal or numerical form of storing and presenting information. Hence the saying *"a picture is worth a thousand words."*

Due to the great presentational value of pictures and the powerful comprehension skills of the human the visualization has become an invaluable companion to almost every field of science. With the help of computers and interactive computer graphics the power of this cooperation is even strengthened.

Considering the different uses of visualizing information, three main categories of tasks can be declared:

1. **Presentation** – the knowledge that is already present has to be explained or presented to another target audience. For example the poll results are presented by a bar chart to emphasize the relative proportion of voters' preferences – Figure 1, left.

2. **Confirmation** – a phenomenon or model is visualized to prove or disprove a certain hypothesis about it. For example the hypothesis about correlation between a car's weight and the power of its engine – Figure 1, right.

3. **Exploration** – similarly to confirmation, the observed model is visualized. In contrast to it, the graphical depiction is observed to search for interesting relations not known beforehand. Such exploratory-oriented data visualization is now the leading visualization task in many areas of science and together with statistical and data mining methods they form what is now called *the visual analytics* [43].

This project focusses on the exploratory task of visualization. Using the visual analytics language, it means to *discover the unexpected*. That means to reveal relations and structures that were not anticipated before or that can not be easily located using computational approaches of statistics and data mining.

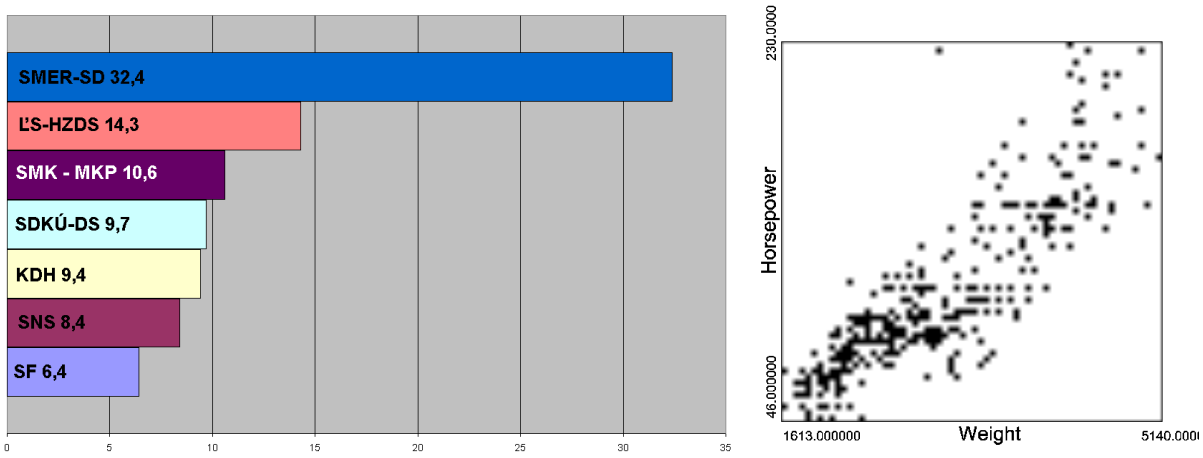

Figure 1.1: Basic visualization for presentation and confirmation purposes.

Left: A bar chart visualization of the voters's preferences.
Right: A scatterplot visualization of car performance versus car weight. A clear relation can be seen that confirms the hypothesis that powerful cars are heavier.

1.1 Visualization

The main differences between computer visualization and the rest of computer graphics fields lie within the priorities of each of these domains. Computer graphics, as is known to most of the public and even to many of the scientists, aims to synthesize an artificial world that mimics the behavior and looks of the real one. This is usually achieved by using photo-realistic rendering, sophisticated physical lighting and shading models as well as detailed geometric modeling. The main application targets of this approach are virtual environments for medical or engineering purposes, special effects for entertainment industry and similar.

On the other hand, computer visualization, be it scientific or medical or information visualization, strives to create imagery that describes a certain model using the most comprehensible graphical techniques. The photo-realism of such a rendition is the least aim. The overall information value of the display is the top priority over features like special effects and eye-candy. Going even further, we can say that computer graphics tries to create a virtual world in a way that we believe. Visualization tries to present the actual world in a way that we understand it.

Even though the visualization domain is united in its aims and goals, the techniques to achieve them vary among different subdomains. This is due to the fact that different types of data have to be visualized and different target application requirements have to be met. Three specific groups of similar properties can be formed to subdivide the visualization domain:

1. **Medical visualization** – data obtained by computer tomography or similar technology is visualized using three-dimensional visualization or by two-dimensional slices. Medical visualization helps the doctors to plan operations, assess the condition of the treated subject and in general to take actions that are not invasive with respect to the patient. Compared to

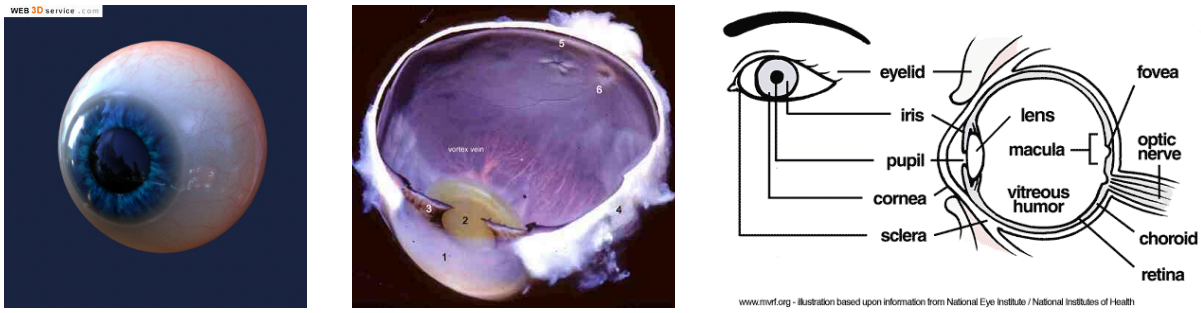

Figure 1.2: Different views of an eye

Left – photorealistic computer graphics [8]. **Middle** – medical or scientific visualization [4]. **Right** – information visualization [5]

other visualization applications, medical visualization strives the most for a photo-realistic imagery. However the priority remains on the side of a clear and understandable image.

2. **Scientific visualization** – the sources of data for scientific visualization are much more diverse compared to medical visualization. Physical simulations and real world measurements model real world phenomena such as flow, dispersion, magnetic or electric fields etc. Many of the visualized attributes and properties are abstract and have no graphical form in the real world – e.g. turbulent energy inside a flow or temperature of the environment. Therefore the graphical methods of scientific visualization introduce virtual elements such as arrows, pathlines, isolines and different glyphs. However, similarly to medical visualization, these data contain certain underlying spatial organization. Usually a two-dimensional or three-dimensional real world space is modeled and the source data is provided in the form of a two-dimensional or three-dimensional vector field. These spatial properties define the geometric cornerstones of the visualization and the spatial organization of the resulting rendition tends to correspond to our perception of the organization of the three-dimensions in the real world.

3. **Information visualization** – unlike medical or scientific visualization, the information visualization does not rely on the underlying spatial structure of the data. At the very least because the source data often does not have any given geometry. The projection from the model space to the screen is therefore not intuitive and it is much harder for the observer to build mental models visually. But, as proven by numerous applications, in many cases it is the best, and often the only, option to communicate information from the numerical or textual representation to the abstract human thinking.

The different approaches to visualization operate on various types of data and target various subtasks. However in the exploration of real world it often happens that the observed phenomenon has many various properties that can not be purely assigned to a certain kind of visualization. Therefore visualization methods are often combined to create a multi-view environment in which different techniques cooperate to create the effective visualization of the given data. For instance,

Figure 1.3: John Snow's map of the cholera spread in the Soho quarter. The black dots mark the deaths of cholera. In the visual center of the dots lies a black cross marking the position of the Broad Street pump, which was indeed the local center of the epidemic.

meteorological applications combine a quasi photo-realistic visualization of satellite images with scientific visualization of dynamic atmospheric features and with information visualization of the measured variables.

1.2 Information visualization

Exploratory data analysis through graphics has a history of its own. The legendary case of cholera outbreak in London describes the use of visualization to determine the cause and the source of the cholera spread in the Soho quarter of London [44]. Since then, information visualization has gone a long way. Now it is heavily supported by computers and computer graphics in order to create sophisticated interactive multi-view environments that support data analysis and decision making. Using the definition from [12], information visualization (or shortly *infovis*) is

> "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition."

The advantage of visualization lies in the ability of human brain to make effective judgements about a large number of items just from their visual appearance – the position, color, shape or motion. Groups of similar properties, or on the other hand – noise patterns and outliers, can (in most cases) be discovered instantly [45]. Compared to automated machine-based statistical or data mining methods, which have the same aim, the human brain has the advantage of non-linear thinking, uncertainty, intuition and domain knowledge. All of them can be employed at once

during the human cognitive process which results in discovering structures that are hidden to the rigid, formal and domain-ignorant mathematical methods.

The wide variety of infovis techniques ranges from the simplest visualization to sophisticated designs of complicated structure and meaning. The different information communicated through infovis is very diverse. But understanding the absence of a greater common denominator of the source data and target application reveals the motivation for creating different infovis techniques and modifying them to fit particular needs.

1.2.1 Data of interest

The source data that is usually depicted by infovis techniques originates in wide range of application or research domains. It is obtained using different data acquisition techniques and it describes different models. Nevertheless, there are few common properties that can characterize this data or categorize the types of the observed attributes.

Types of values

The type of variables of the model can be considered using different aspects of it. To illustrate the different data types we can consider the cars data set [1] which describes a set of 392 car models by recording different properties of them. These properties define the dimensions of the data space and the semantics of the particular dimension – Miles per gallon, Number of cylinders, Horsepower, Weight, Acceleration (the time to accelerate from zero to a certain velocity), Year of manufacture (1970-1990), Country of origin (America, Europe, Japan).

With respect to the number of values the attributes within this data set can be divided into

- **continuous** – Miles per gallon, Horsepower, Weight, Acceleration
- **discrete** – Cylinders, Year, Country

The discrete attributes can further be divided into

- **ordinal** – Cylinders, Year
- **categorical** – Country

The categorical attributes are also often referred to as nominal. They form a special type of data since no ordering or a comparison relation is defined within such an attribute. Even though some categorical values are numerical (e.g. credit card numbers) they can not be treated as ordinal values. This project does not consider categorical values, since the visualization of such attributes is a complicated self-standing issue itself [21].

Dimensionality and volume

As the observed models have multiple interesting attributes, the resulting data also contains multiple attributes or data dimensions. For example a single observation in a census, say a household, can be described by such properties as number of family members, male/female ratio, average income, mean age of the children, distance from town center, average commuting time and so forth [3]. The dimensionality of such a data can range from low-dimensional data sets containing three to ten dimensions through multi-dimensional data sets of tens of dimensions up to very high-dimensional data sets of even hundreds of dimensions [48].

Another property describing the data set is the number of data samples or observations. Each observation represents a single entity in the observed space. It may be an individual in the census, a stock in market data or a certain position in space within a physical simulation.

We can picture the data set, as stored inside the numerical domain of a computer, as a table of a spreadsheet. The observations are the rows of the table and the different data dimensions are the columns of the table. Hence the extreme cases of data sets can be described as ”tall” – having a large number of observations but only a small number of dimensions, or ”wide” – having a small number of observations but each with a high number of attributes. Due to the limitations of the data storage and data acquisition processes it rarely happens that a data sets has both large number of observations and large number of dimensions.

1.2.2 Interaction

Information visualization benefits from interaction by several means. First of all it makes the visualization more flexible and available to fit the needs of the user without needs to redesign the actual visualization. This is especially important in exploratory data analysis where the goal of the analysis is not known beforehand and the visualization must be able to provide different options of insight on the observed data [31].

Another, and often the most important, advantage of interaction is that by manipulating the view and observing the changes, the user builds a mental image of the model inside his/her brain. If this change happens within a fraction of a second (i.e. in nearly real time) the user immerses into the visualization and gets the feeling of actually touching the data [15].

The interaction options of a display involve changing parameters of the visualization (displayed axes, data value ranges, type of visualization), changing parameters of the data projection (rotating, zooming, panning) and manipulating the data (selecting and highlighting samples, hiding them, creating semantics and relations [47].) There are numerous interaction tools and metaphors to visualization. Compared to other visualization domains, the information visualization handles the most of them [42].

1.2.3 Multiple views

Often a single visualization method does not suffice to effectively present the observed data and multiple different methods or different instances of the same method are used. This creates a multiple view environment. Such environments are widely used and often incorporate displays

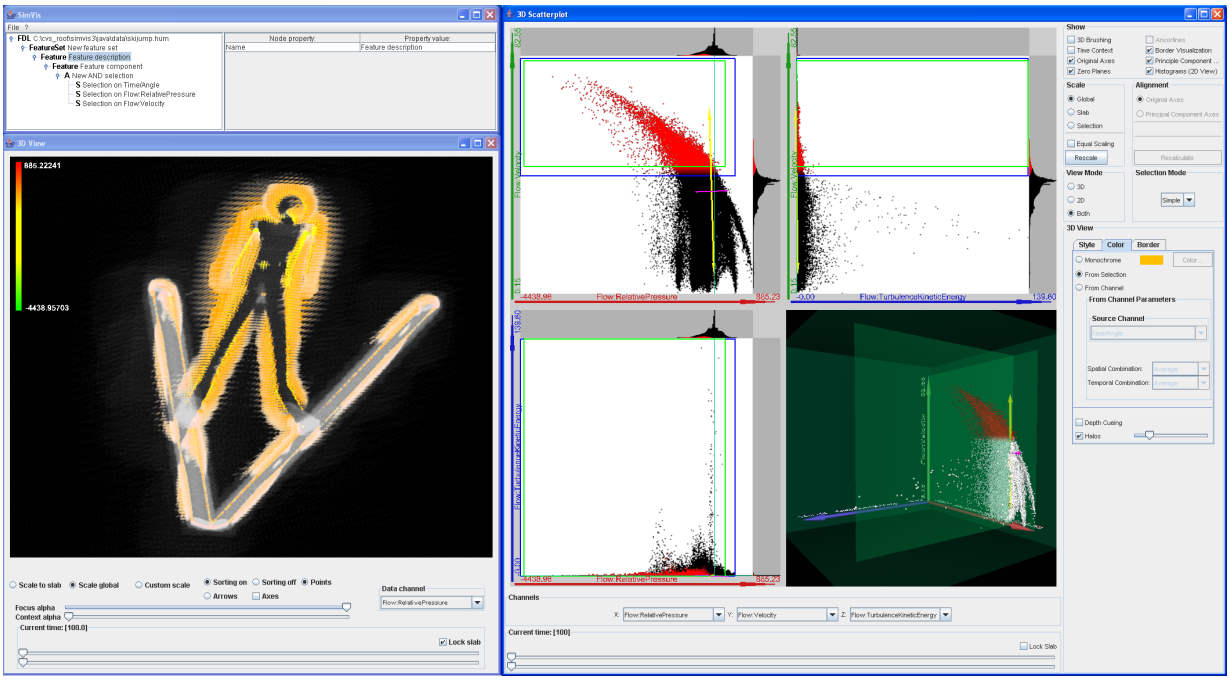

Figure 1.4: SimVis [6] – an example of a multiple view environment combining different visualization techniques

of scientific visualization as well. If these views are interlinked via sharing the same interaction metaphors and data access approach, the user receives the information by different presentation channels. The connections between multiple views, as described in [10], help to understand a complex model and to gain combined information that would be hard to perceive if only a single visualization or a set of unlinked visualizations would be present.

The linking between different views mostly involves sharing the set of selected samples – the focus. If a user is interested in a certain portion of the data set this portion is usually selected in one of the views and this selection is then adopted by all the other views so that the user can observe the behavior of the selection from different perspectives.

Current visualization applications usually use multiple linked views and the concept of multi-view visualization has become a standard in the visualization domain. The connection of scientific visualization and information visualization into a unified framework also enlarges the application scope of the visualization and improves either part of the visualization by presenting the data from another point of view [14].

Chapter 2

Large data in information visualization

The development of information technology and scientific processes implies the improvement of data acquisition and data storage methods. The information increase steeply accelerated by the end of the 20th century through many novel effective and powerful tools in science and commerce – computer hardware, the Internet, measuring technology etc. It is estimated that one exabyte of unique data is produced every year [26]. The large size of the data produces a significant pressure on the limits of information visualization. Thus the topic of this project and the solutions in it consider this issue in order to improve real situations and to contribute to actual application needs.

2.1 Origins of large data

The application domains that take advantage of infovis are different and include for instance economy, natural sciences (biology, chemistry), engineering, marketing and many others. There are several causes of the great information increase in these areas. Natural sciences benefit from cheaper and more precise measuring devices that enable the scientist to conduct a larger number of observations and to gather information about a larger number of attributes such as concentration of chemical elements.

Improvements in data storage are the leading cause for information increase in areas of marketing and economy. The current technology for automatic data processing is capable of recording millions of financial transactions or stock operations every day [27]. The commercial companies maintain large databases of shopping habits of their customers to increase their own profit by choosing an effective marketing strategy.

Engineering experts are capable of simulating their experiments with better precision and to sample the real world in much finer details. All these improvements of technology together with the accelerating growth of the human civilization itself are the source of the large data that is the motivation for this project's research.

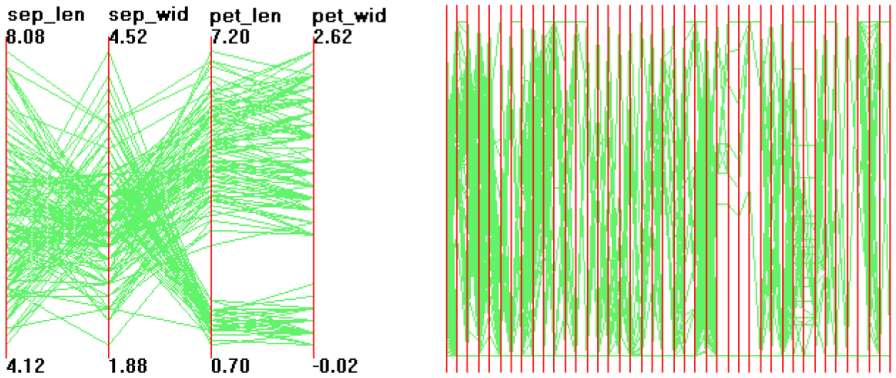

Figure 2.1: Parallel coordinates visualization of 5 dimensions (left) and 42 dimensions (right), as described in [48]

2.2 Large number of dimensions

One aspect of the information increase is the increasing number of data attributes. New ways of observing real world and new real world phenomena cause that the number of dimensions that a certain model involves might grow to very high values. For example geochemical measurements now consider the concentrations of elements and substances resulting in over 500 attributes in some of the cases.

Even though there are several visualization techniques that are primarily designed to display multidimensional data (e.g. parallel coordinates [23], worlds within worlds [17], projection pursuit [20]) they do not scale well for really high number of dimensions (Figure 2.1) For data sets of high dimensionality even these techniques soon reach their limits and their contribution to the information cognition diminishes due to various problems [48].

Although the large number of dimensions is not the primal target of this project, there are several techniques to be recognized. For instance dimension reduction techniques (self-organizing maps [30], multidimensional scaling [32], principal component analysis), dimension subsetting techniques (scatterplot matrix) or dimension embedding techniques (worlds within worlds [17]) handle large number of dimensions. This is done either by reducing the number of dimensions by computations in the data space and by sophisticated graphical layout in the screen space.

2.3 Large number of samples

The main focus of this project is to handle the increasing number of multidimensional data records processed by visualization. Nowadays databases store millions of items and every day this number increases. The contemporary data processing technology more or less scales to this high number of records, at the least because it actually is the technology that produces it. However visualization techniques are not that scalable and therefore a significant improvement is necessary to allow for effective large data visualization.

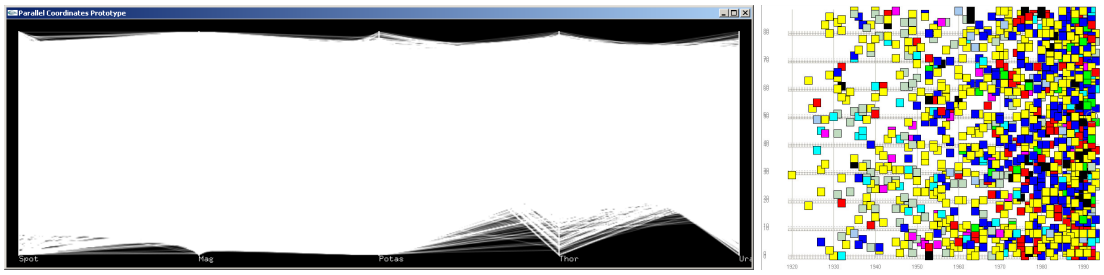

Figure 2.2: Large data causes a significant clutter in information visualization displays. For instance the parallel coordinates view does not show any apparent structures (left) and in the scatterplot many of samples are hidden beneath the others (right).

The limitations that obstruct the visualization techniques from displaying a large number of samples reside on both sides of the user interface. Computers, even with their powerful hardware and graphics acceleration, soon reach their performance limits and are not able to transform a large number of data values to their geometrical projections in a reasonable time. On the other hand, the human perception has also its limitations that stem from the properties of the human visual system. This project considers both of these aspects and tries to develop solutions to improve the overall situation.

2.4 Definition of the problem

Increasing the volume of the data records in visualization has several undesired effects. The relevance of either of them depends on the actual kind of visualization technique but all of them negatively affect the way the user perceives the graphical form of information. An incorrect or misleading visualization can lead to an incorrect judgment or decision about the observed model. Moreover many interesting features of the data might remain hidden to the observer and the value of the visualization is hence damaged [37].

Aggregation

One of the main visual judgments a user makes when observing a data visualization is the relative density of items or the relative population of certain subareas of the model space. This forms the notion of the data distribution and the structure of the features inside the data.

Every graphical representation of an item occupies a certain portion of the screen space and the screen space is limited in its capacity. Therefore every visualization method has a capacity limit when it is no longer possible to ascertain which parts of the screen contain more samples and which parts contain less. This effect is called aggregation and it affects the density information of the display (Figure 2.2)

Occlusion

Another important issue also stems from the capacity limits of a visualization. Due to the low dimensionality of a computer display it naturally happens that after a projection from the multidimensional data space certain items are given the same screen position. Depending on the type of the visualization it often happens that one item occludes the other one and therefore the occluded item remains hidden to the observer (Figure 2.2)

In a multivariate case this might create an undesired information alias as illustrated by the following example. Consider a data set describing a part of the human population with respect to their age, weight and gender. We draw a scatterplot of age against weight and we color the spots according to the gender – blue for males, red for females. If the data records are rendered without considering the drawing order it might happen that the spots representing males occlude the spots representing females resulting in elimination of many of the male samples. This incorrect visualization would then mislead the observer into a judgment that the data contains only a few (or none at all) female samples.

Speed

The last but not least important issue is the actual speed of the rendering. Infovis is not a typical performance-oriented graphics application, however the response time it takes for the display to provide the user with visual feedback is a critical parameter when it comes to interactive visualization. If the display does not operate interactively the user loses the coherence between his/her actions and the reactions of the computer. Apart from losing touch with the data, the negative effects of high response time include change blindness and attention distraction. Not speaking of prolonging the time to finish the given exploratory task.

Considering speed, the increased rendering time might seem as an issue to be solved by having more powerful computers in the future. But it is a presumption that can not be relied on, since with the growing rendering power of computer the data acquisition and data storage power of computers will rise as well. It is more likely that the size of data will grow at least as fast as the rendering performance of the hardware will. Therefore a more principal solution has to be employed even for the speed issue.

The relative importance of the individual large data problems can hardly be judged. Generally said, neither one ”wins” over the others. There are several successive steps in the infovis pipeline leading from digital data to a graphical display (see Figure 3.1) and none of these steps should be weighed a lower measure. It is clear to see that the complications mentioned in this chapter can no longer be believed to be solved by more powerful computers in the future. Either they are of perceptual nature or simply the balance between computational performance and size of data is not going to change dramatically in future. The next chapter introduces several approaches to improve large data visualization on a technical and (which is more important) also on a fundamental level.

Chapter 3

Solutions

By introducing large data to an information visualization system, every single part of the whole system becomes a potential weak spot of the design with respect to large data. Only the construction of the actual application framework influences the actual location of the bottleneck. Therefore the solutions can not focus on a certain part of the visualization pipeline (Figure 3.1) and have to treat the problem either on a global scale or at every single step. Preferably on a level that is rather technological than technical.

In general there are two basic approaches to deal with large data in information visualization. The modifications might be oriented to operate on the data-side of the process, incorporating statistical and data mining techniques to organize the data or simply reduce its size. Another approach focuses on the graphical output by creating novel visualization methods or by increasing the capacity of other methods.

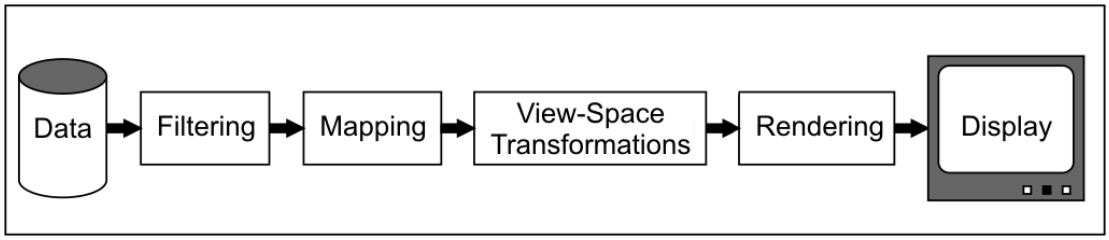

Figure 3.1: Visualization pipeline

There are many pros and cons to each of these approaches. The recent development in the field shows that an effective large data visualization can only be achieved by a reasonable combination of both approaches [46, 35, 22]. The following sections categorizes the state of the art in large data visualization. In final, these approaches are summarized and the fundamental directions of research are lined out.

3.1 Data space methods

A natural approach to solve large data visualization problems is to remove the cause of the complications – the large data themselves. The relatively old and well-based grounds of automatic data processing, statistics and data mining are the background of all data-oriented methods that strive to re-organize the data to a smaller scale. It is the common aim of the data-oriented methods to provide a less demanding data set to describe the same model.

The motivation to use less data to contain the same information is obvious – to save data storage capacities and data processing power. However, with data reduction comes the question of the balance between the size of the data and the truthfulness of it. It is clear that these two variables are bound by a, more or less, inverse proportion. The speculations about the effective tradeoff between data simplicity and data authenticity are a common issue in all research domains that incorporate numerical models of real world.

Information visualization has a special set of priorities when considering this tradeoff. It originates from the needs of the actual graphical representation of data. Data entries should be treated according to their visual importance. A well-designed data processing method in an infovis environment does not ignore the screen space. On the contrary, it exploits the specifics of the given visualization to process the data in the most effective way with respect to the resulting visualization.

3.1.1 Sampling

The most simple way to reduce the size of the data is to simply ”cut” or ”pick” a certain portion of it. Of course the consequences of an unwary data reduction are very likely to be disastrous to the final visualization. Intelligent ways to select a subset of data have to be employed to prevent from damaging the information during sampling.

Some of them include random sampling, for instance in large data visualization by Ellis and Dix [13]. The randomness in sampling preserves the relative density of individual sub-areas – an effect that is important to preserve the information about the core structures of the data. However if the sampling is applied in a uniform manner over the whole scope of the data, structures of low density or small scale might disappear. A working solution is to devise non-uniform

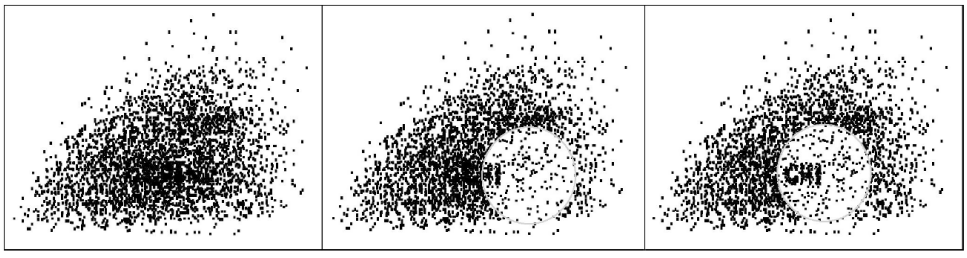

Figure 3.2: Example of sampling in a cluttered scatterplot. The subsampled area under the sampling lens reveals a hidden structure

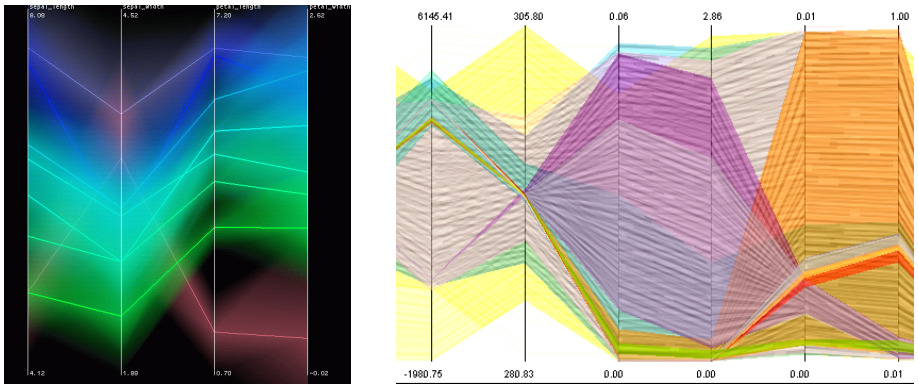

Figure 3.3: Visualization of clustered data by hierarchical parallel coordinates (left, [22]) and by visual abstraction for parallel coordinates (right, [37])

sampling [11] which considers the relative densities and choosing the sampling rate accordingly. For example the low density areas might not be sampled at all to preserve outliers and small scale features and areas of the highest density are sampled using the lowest sampling rate.

A prototype of an intelligent sampling of large data in information visualization is presented using the Sampling Lens by Ellis, Bertini and Dix [16] as illustrated in Figure 3.2

3.1.2 Data abstraction

A more radical approach, compared to filtering or sampling the original data, is to replace the original data by a derived form. If this form is less demanding and preserves the original information, the resulting data abstraction provides a promising solution to help large data visualization. The information from the original data can be abstracted using various techniques, most of them are based on statistical or data mining methods.

One of the drawbacks of the automatic data abstraction methods is that, due to their statistical nature, they might produce too crude and too abstract results without any human-given semantics. These methods were not originally designed to help human-based data exploration but to substitute for it. Automatic (or unsupervised) data mining or feature extraction are computer-oriented approaches that do not involve any particular visual feedback. Many of them do not include any human feedback at all, operating autonomously and giving results in the form of statistical values – mean, variance, clusters, purity index etc.

The nature of these supportive methods predefines them to be used in the early stages of the exploratory session. Using them on a small scale helps to reduce the data and, at the same time, create a effective data representation that works fine with the visualization. Moreover, if the visualization is aware of the data abstraction, a joint effort can be designed in the form of visual abstraction ([22], also Figure 3.3)

Data preprocessing with the help of abstraction is a successful way to reduce data and to create a simple representation for the contained information. The used methods include, among others, clustering, vector quantization or feature extraction.

Clustering

The primary presumption to introduce clustering as a pre-processing step in visualization is that the observer of the visualization is looking for groups and trends inside the data. This holds for most of the visual exploration applications and therefore clustering (together with other data abstraction methods) provide useful improvement to large data visualization.

Clustering is a statistical and data mining method that groups items of similar properties into clusters. Usually a certain form of the mutual distance of data records is used to decide the similarity. Different multidimensional metrics include Euclid, Mahalanobis, Chebychev, or k-nearest neighbors for large numbers of dimensions.

Vector quantization

Originating in signal processing and compression, vector quantization offers another way of data abstraction. The result of vector quantization is a set of code vectors that give a locally optimal approximation to the original data. The concept allows for specifying the final number of code vectors and the algorithms scale well even for large data.

If used properly, vector quantization can be used to reduce data size in visualization and thus to help large data infovis.

Feature extraction

If the data comes with a certain domain knowledge – e.g. physical simulation data or financial data – the data can be abstracted and thus reduced by the means of feature extraction. Features are groups of records that follow a certain criterion with respect to their behavior. For example econometric data produced by analyzing parameter space of a set of differential equations contain basins or attractors [40]. These features are typical to the domain of differential analysis and can be detected in the data. Extracting them from the original data creates an abstraction above the data and the originally large data are reduced to a set of features and descriptors.

3.1.3 Density-based representation

The data sampling and data abstraction methods investigate single data records and then form (often synthetic) data that approximates the original information or model. However, starting from a certain size of data, the different individual records lose their individual importance and it starts to be reasonable to consider relative density of different data areas instead.

Continuous density information is computed by fitting a density distribution function of a certain probabilistic model (or a mixture of them) to approximate the original data. This approach tends to be extremely demanding in large data cases, therefore the discrete density is usually used [9, 34].

Discrete density information is obtained by dividing the original data space into different areas and computing the population of the individual areas. The position and size of the areas is the matter of choice of the respective technique. Regular techniques divide the original data

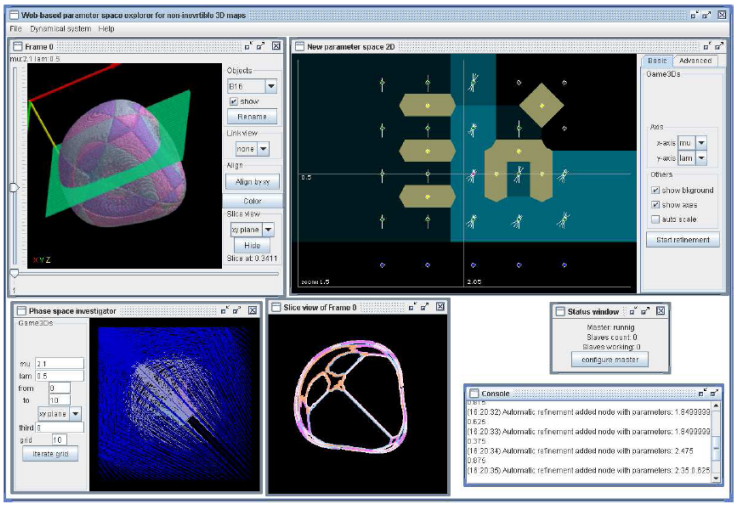

Figure 3.4: The Parameter Explorer application [41] extracts special features from the parameter space of a system of differential equations and visualizes them.

space into a set of equally-sized n-dimensional intervals – bins. This approach is well scalable and has predictable error and complexity [38].

However, real-world data often consists of different structures of different size and even different distribution models. In such cases it is more efficient to sacrifice the predictiveness of regular binning and to employ adaptive space subdivision. Similarly to quadtree and octree algorithms used in three-dimensional scene representation, the data space is repeatedly subdivided until the desired precision is reached.

The disadvantages of space subdivision emerges in data spaces of extremely high dimensionality. Memory complexity of binning grows exponentially with the number of dimensions and even for data sets with tens of dimensions it is not feasible to achieve a reasonable precision within the bound of physical memory.

3.1.4 Hierarchical data organization

As have already been explained, there are different ways of decreasing the complexity of the incoming data in information visualization. Most of them create data of smaller size and eliminate the original data by merging the individual data samples (clustering), hiding some of them (sampling) or discarding them all (density representation) . But often the original data contains important detailed information and it is undesired to obstruct the observer in investigating the individual records of the original data set.

The methods are therefore combined to create a hierarchical data representation that holds different levels of detail of the data. The lowest level contains the original data and upper levels of the hierarchy provide data representation of smaller size [22]. The construction of the upper levels may utilize any of the above mentioned data-oriented approaches. Even a combination of

them – for instance the first level above the original data might be created by intelligent sampling and the level on top of that might be clustered to create an abstract data representation [36].

An important issue with respect to hierarchical data organization is the interaction with the user. The system must provide efficient ways of navigating through the hierarchy otherwise the actual purpose of creating it diminishes. This depends on the respective visualization method and some implementations even contain a special display for exploring the data hierarchy [22].

3.2 Screen space methods

The set of methods dealing with large data in the screen space is smaller than the one of data-oriented methods. The reason is simple: the statisticians and information scientists have dealt with data sampling, data abstraction or density estimation long before computer-based visualization. In spite of the short history of large data information visualization, there already are several contributions designed to handle data of large size within the graphical part of the whole process.

3.2.1 Pixel-oriented techniques

Many of the large data visualization issues are caused by the situation when multiple data records are displayed at the same position. The probability of this situation increases with the size of the screen space occupied by the graphical representation of a single data record. The pixel-oriented (or *pixel-based*) techniques decrease the elementary screen space. Potentially to the size of a single pixel. This enables the methods to display potentially millions of data records [18] in a single screen. In combination with large screens (such as the PowerWall [2]) this creates an interesting solution to large data visualization.

Pixel-oriented techniques are being successfully implemented especially for data sets of low dimensionality or displays with low number of dimensions displayed. Algorithms like e.g. the recursive pattern [28] or the gridfit algorithm [29] are used in visualization for financial analysis. Some of them [25, 18] are also capable of displaying large hierarchies down to the level of a single record (see Figure 3.5)

By having a single record represented by such small graphical element, potentially only one pixel, not much of the screen resources is left to display additional information about the particular data record. Usually an additional property is mapped to the color of the pixel. For example stock value trend [7] or file type [18]. The handicap of the pixel-oriented methods dwells in the incapability of displaying multiple attributes at once. The additional data dimensions are therefore often displayed in the form of tooltips or in a separate linked display.

3.2.2 Transparency in visualization

Similarly to density-based representation in data space, the transparency in visualization aggregates density information in screen space. By utilizing semi-transparency the display gains an additional dimension that represents the relative population of the particular screen area. The structures and trends can then be told apart one from another by observing the different groups

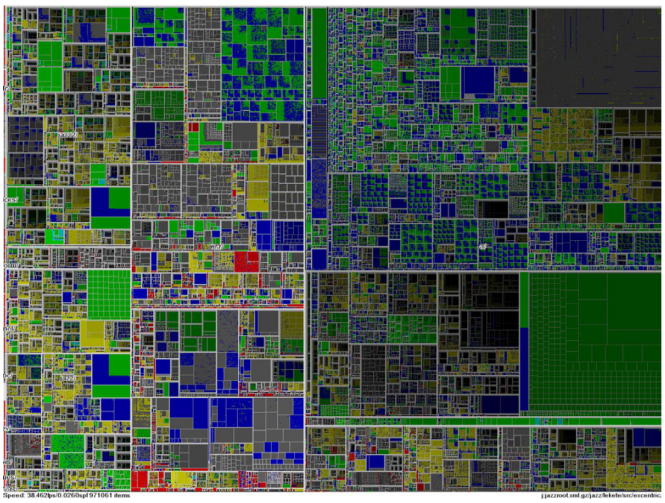

Figure 3.5: Interactive information visualization of a million items [18] using pixel-based methods. Visualization of a file system with the directory structure as nested rectangles.

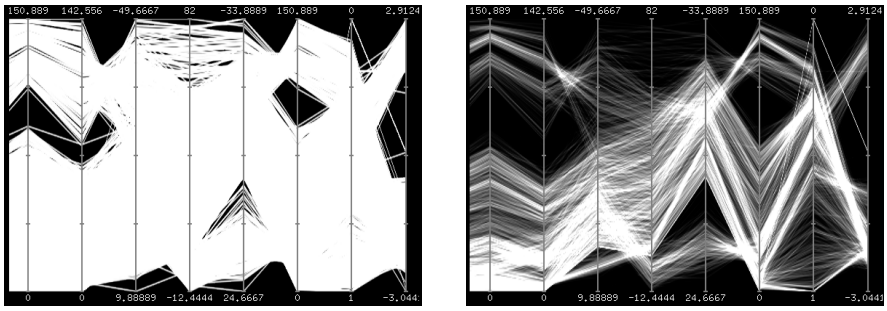

Figure 3.6: The same data visualized without transparency (left) and with transparency (right).

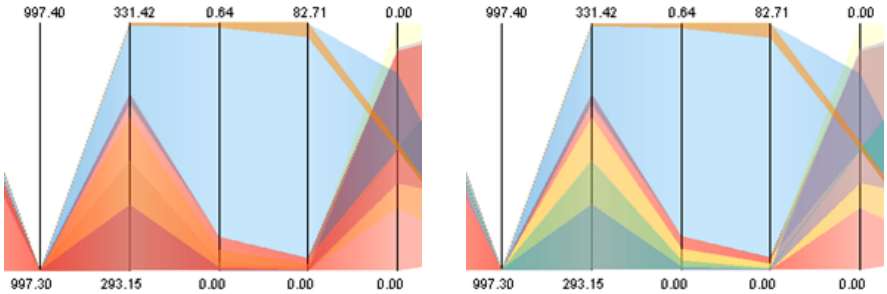

Figure 3.7: Semi-transparent clusters visualized with random rendering order (left) and with rendering order depending on the size of the cluster (right).

of adjacent visual elements [45]. The screen-space density is usually implemented and represented as the alpha channel in the color space of the display, where high alpha (or opacity) values represent high density and vice versa.

Areas that are sparsely populated only receive the basic opacity while heavily populated areas aggregate the transparency values and are highly opaque. This helps to perceive the relative population of different areas and thus improves the situation with aggregation problem in large data visualization (Figure 3.6) The second benefit of semi-transparency is that the graphical elements that are rendered later do not occlude the earlier elements. This fixes many cases of occlusion (see Figure 3.7)

Additionally, the alpha values can be stored in a high-precision texture and then the texture can be transformed using different transfer functions [24]. This allows for instance for observing structures in low-population areas independently from high-population areas. Without the transfer function properly applied, the result will either be clamped by the highest alpha value, leaving no details in the high opacity area. Or it will be scaled down to the highest alpha value, diminishing the importance of the low-population areas.

Semi-transparency and alpha blending bring up the question of correct blending mode and, depending on the blending mode, the appropriate rendering order (see Figure 3.7) The different applications choose specific approaches to this, mainly because different combinations provide different visual clues. For instance additive blending is suitable to emphasize the population of areas. To achieve the highest resolution with respect to opacity, the elementary transparency value for an individual graphical element has to be rather low otherwise the opacity limit would soon be reached (especially for large data). But the low elementary transparency discards low population areas or outliers because they are barely visible. This might lead to information loss in particular applications.

3.2.3 Output-sensitive rendering

While pixel-based and density-oriented methods focus on improving occlusion and aggregation in information visualization of large data, the speed issue is given little attention. One way to improve the speed of the rendering within the screen space is to use the rendering performance wisely. If the actual technique is aware of the rendering specifics and also of its interaction capabilities an intelligent scheme can be designed in order to render only those records or screen portions that really contribute to the final rendition. This approach is called output-sensitive rendering.

In the first step, the data records that would not affect the final image are left out. The capacity limitations or the focus of the display determine criteria to detect such data records. For example if the display utilizes additive transparency, the overlapping graphical elements soon sum up to the full resolution of the alpha channel. The data records that would to be placed in the areas of full opacity can be omitted from the rendering process.

Another application of output-sensitive rendering concept is to quantize the data in a manner that the visual output does not change. If the quantization is designed according to the visualization, for instance if bin size is set to correspond to one pixel, the information value of the display

does not change. In many large data cases, optimizations of this kind save significant amount of processing and rendering time.

The rendering can also take advantage of visual coherence before and after interaction. Not all portions of the screen have to be updated and much of the already produced imagery can be saved. By organizing the display into layers and/or segments, only the particular portion needs to be updated with keeping the rest intact [38].

3.2.4 Hardware acceleration

One part of the whole visualization process is often neglected in many visualization prototypes and tools. It is the technical implementation of the graphical end and the rendering part of the projection. This is mostly due to the implementation demands that favor either novelty (in academic research) or compatibility, portability and data-oriented robustness and reliability (in commercial applications.)

It often happens that a popular visualization tool suffers from low interaction feedback or low rendering performance because of not considering the technical improvements that are possible using the advanced features of today's GPUs.

It has been possible for a long time now to store many layers of rendering output in textures or to perform geometric transformation on the GPU. The recent development of the GPUs (even though powered mostly by the entertainment industry) brings many new features that can help large data visualization either by decreasing the feedback time or by enhancing the display with advanced rendering capabilities.

The GPUs nowadays provide high definition floating point textures that can contain density information of an infovis display. Rendering of large number of graphical elements can be performed faster with vertex programs and novel features like instancing. All these and many forthcoming features should be considered when designing an information visualization tool that is supposed to handle large data [19].

3.3 Summary

As mentioned in the beginning of this chapter, effective information visualization of large data should exploit the advantages of both approaches – working in data space and in screen space as well. At the very least because many screen-space methods need to derive information from the original data and, on the other hand, many data-space methods need appropriate visualization for their special data structures. The particular combinations of data- and screen-based methods plus the balance between them is a question of the respective application and exploratory task. What is common to all of them, is the fact that when large data have to be visualized using information visualization tools, analysis of the data has to be performed first.

The experience with large data visualization was summarized by Daniel Keim into the Visual Analytics mantra, a modification to the original visual information seeking mantra of Ben Shneiderman – *"Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand"*. The Visual Analytics agenda [43] realizes the importance of large data visualization

in many critical fields – national security, computer networks, financial transactions etc. Following this motivation, the Visual Analytics combines the information processing performance of machines and humans through visualization. This approach appears to be the future orientation of all information visualization since large data is quickly becoming a prevalent issue in all research domains.

The best example for the effective combination of data-oriented and screen-oriented methods is the Focus+Context paradigm [12]. The F+C techniques divide the data according to user interaction into two groups – the focus, which is the area of interest and is displayed in high details, and the context, which represents the rest of the data and uses low graphical resources. The benefit of using this two-fold visualization is that the user sees the area of interest in detail and does not lose the relation of the focus to the context within the whole data set.

In practice, data-oriented methods are used to build an abstract representation of the data records that lie inside the context and also to separate the focus from the context. Screen-oriented methods handle effective visualization of focus, graphical representation of the abstract context level and the combination of these two. The F+C concept is a popular approach to visual exploration and it is one of the core topics of this project.

Chapter 4

Preliminary results

The results achieved during the course of this project and the respective research have already been presented in various forms. This chapter tries to summarize the effort and achievement. Also the preliminary results have to be put to the overall domain context to clearly determine the contribution and weigh the importance of the effort.

4.1 Similarity Brushing

Similarity brushing deals with large number of dimensions in a low-dimensional display. Large number of dimensions is not the primary focus of this project, but similarity brushing is an effective presentation of the abstraction concept in visualization. As well as it is an example of benefits of joining the power of human and computer to achieve effective data exploration via visualization [39].

The multidimensional information stored in original data space is abstracted to sets of relations between individual data samples. This relations represent the mutual similarity of two samples. If this kind of derived information is present, it provides an abstract way of representing multidimensional structures in a low-dimensional space.

Traditional brushing is usually performed in the screen space as a selection of constraints on points of interest – e.g. a rectangular brush. Then it is decomposed to several one-dimensional interval queries and the eventual selection is evaluated as a composition of the queries. This determines the nature of the brush to be a low-dimensional slice across the full dimensionality of the original data space. Only the dimensions depicted in the screen space are taken into account in traditional brushing. The final selection is a cartesian product of the one-dimensional selections and therefore the n-dimensional brush performed using traditional technique is always an axis-aligned n-dimensional box.

On the contrary, the similarity brushing uses the abstracted similarity information to extend the brush into its full dimensionality. It is not inevitably box-shaped and axis aligned, which provides better selection of real world features which rarely come in the shape of axis-aligned boxes (see Figure 4.1)

The advantage of similarity brushing lies in the combination of human and computer actions.

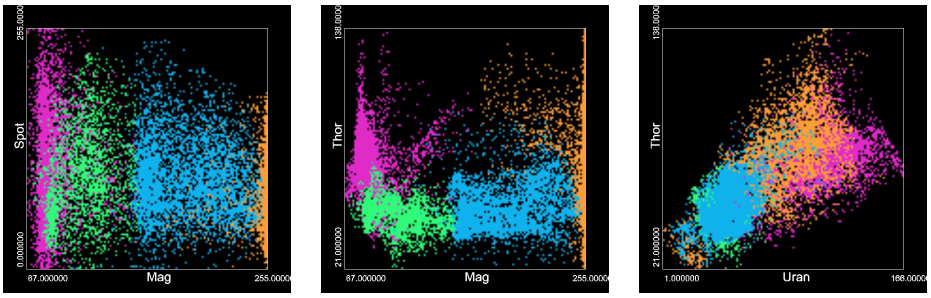

Figure 4.1: Segmentation of data using similarity brushing – complex, unsharp and overlapping segments are seldom feasible when conventional brushing techniques are applied. With similarity brush the segmentation process took less than a minute and required only simple interaction.

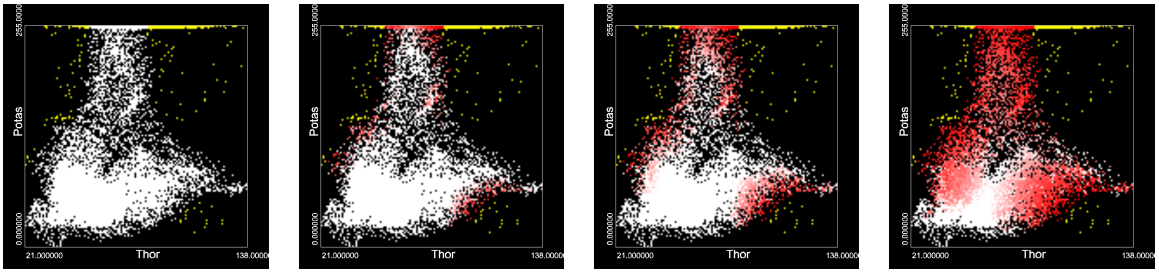

Figure 4.2: Extending the screen-based brush into the data space. The traditional screen-based brush (painted in yellow) is extended in the data space by decreasing the similarity threshold.

The humans are not able to perceive real multidimensional information. Our perception capabilities are quite low with respect to dimensions. However the computer can process multiple dimensions and derive the abstract similarity information out of it. This similarity information (which is a scalar value) is presented visually to the user and can be used to select structures of shape and behavior that exceeds the dimensionality of the screen. Clearly the most obvious benefits come in cases of low-dimensional visualizations, such as the scatterplot. The actual performance of the similarity brush does not necessarily depend on the dimensionality of the original data space.

The similarity brush starts as a traditional screen-based brush but then it is extended to the data space (through the similarity information) an it can be repeatedly refined in the screen space (see Figure 4.2) This tool can be successfully used to steer exploration and decision making in low-dimensional visualizations. The similarity brush is a true multidimensional brush that does not revert to axis-aligned bounding boxes brushes like the previous approaches to data-driven brushes [33].

4.2 Binning and output-sensitive rendering

The density-based representation approach is utilized in the advanced implementation of parallel coordinates [38]. The resulting visualization is output-sensitive and is capable of visualizing datasets of even millions of data records. At first the data is processed from the aspects of two-dimensional subspaces, each corresponding to two adjacent axes of the parallel coordinates display. Each of these subspaces is binned to $m \times m$ bins. A bin in this representation is rendered as a parallelogram connecting its boundary values on the respective axes. The finest precision corresponds to the highest m. At the precision of $m = 256$ the binned visualization based on the density-based representation is visually as precise as the original original parallel coordinates. But compared to the original approach it communicates more information (see Figure 4.3) and it also no longer depends on the size of the original data; it renders and reacts smoothly.

The most important benefit of binning lies not in the rendering speed. By aggregating the data into larger bins, the display becomes clearer and the different structures are distinguishable thanks to the scalable transparency mapping of the bins [36]. This would be almost impossible

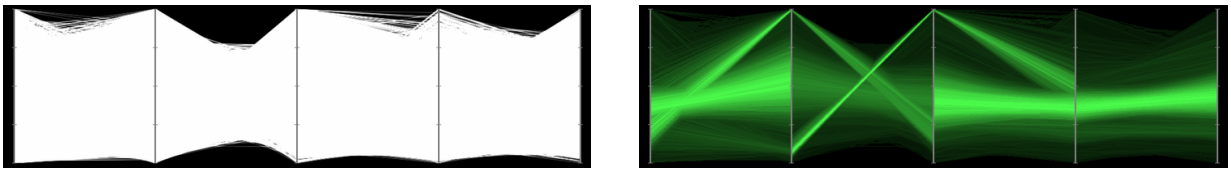

Figure 4.3: Remote sense data [1](interpolated to 100.000 samples) rendered using conventional parallel coordinates (left) and after binning to 128×128 bins (right). Not only the binned representation is precise enough to preserve the details (note the width of a bin) it also clarifies the visualization thanks to the density-based representation of data.

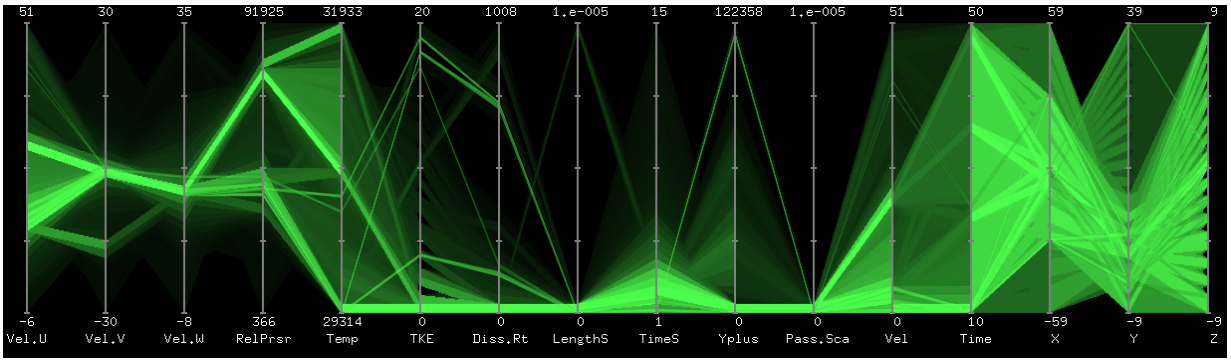

Figure 4.4: Output-sensitive, outlier-preserving focus+context visualization allows to even render more than three million data items (a CFD simulation) into a parallel coordinates plot. Not only is the binned representation spared from being cluttered by a million of data items, it also renders interactively.

in original parallel coordinates if large data would have to be visualized.

To avoid losing details in low density areas, the outliers are separated before the actual bin visualization and are drawn on top of the bins. Using this combined approach, similar to Focus+Context techniques, the trends and the outliers can be visualized in parallel without having to switch to different modes of focus.

The perception issues and output-sensitiveness were also considered in this sub-project. For example the bins do not use transparency but are sorted according to their population and drawn in the ascending order one on top of the others. This avoids enormous numbers of unnecessary fragments and visual attractors that would be present if transparency had to be incorporated.

The screen is divided into layers, which means that if the focus changes, only the focus layer is updated leaving the context layer intact. Also the display is divided into segments. One segment holds the graphical information between a pair of adjacent axes. This allows for per-axis interaction where only the adjacent segments (usually two) have to be updated and not the whole display. By dividing the screen space into a three-dimensional space of $segments \times layers$, the area of the screen that actually has to be updated is heavily decreased resulting in an output-sensitive rendering.

4.3 Visual abstraction

In extension to the previous effort dedicated to binning, the resulting density-based representation can be used to perform fast data abstraction such as clustering. The reasons for using the density information instead of the original data are two-fold. The density information uses much smaller data, therefore the heavy data processing that is performed during data abstraction takes less time and can even be done in real-time. The second reason stems from the fact that many data-abstraction methods, e.g. clustering or multidimensional scaling, only provide locally optimal approximations. By having the option to run a single iteration of an algorithm in a very short time (thanks to the simple density information), we can afford to run more iterations of the particular algorithm or even process several runs with different starting conditions and then combine the results to achieve a better approximation (Figure 4.5.)

In the parallel coordinates project, this is done using the two-dimensional binning in the two-dimensional subspace defined by a pair of data dimensions that are mapped to adjacent. Binning in two dimensions is very similar to creating a two-dimensional histogram of the particular subspace. This histogram is treated as a height map with height of a particular vertex being the population of the according bin. This creates a notion of a terrain or a relief, which can be divided into clusters by decreasing the height threshold.

This idea is novel in at least two ways. First of all it does no longer tamper the original data and therefore its complexity does no depend on the size of the original input data. The source data is uses – the density information – is a by-product of another process – the binning – and therefore it can be considered a zero cost to the whole algorithm complexity. Second the clustering based on the two-dimensional histogram is an original idea that combines the advantages of statistical density estimation methods with the benefits of fast discrete data representation [38].

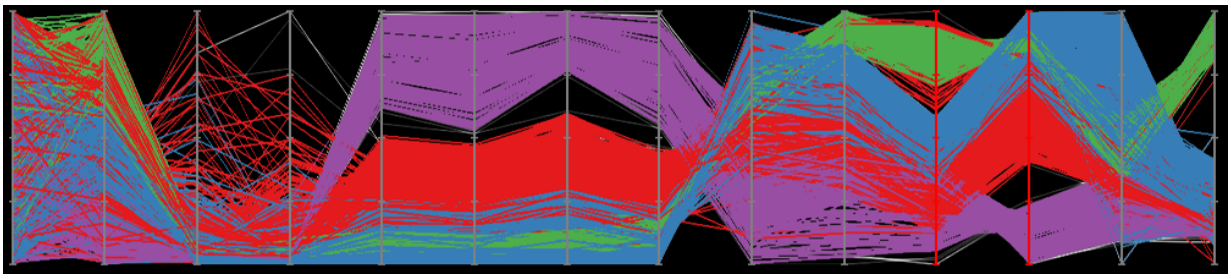

Figure 4.5: Clustering of the binned data performed between the 11th and the 12th axis (red axes). The respective two-dimensional subspace was binned to 64×64 bins and clustered by the occupancy values of the bins. The distinct colors show four clusters extended to all the dimensions.

4.4 ffVis – Hardware acceleration of parallel coordinates

The technical background of rendering for information visualization is also considered in this project. The ffVis application is a prototype implementation of the state of the art GPU features to improve information visualization of large data. The parallel coordinates is a very demanding graphical application consisting of many overlapping lines. An effective large data visualization using this popular and useful technique would be impossible without using proper hardware acceleration [19].

Some of the features the program implements and successfully tests are:

- Vertex arrays and vertex programs – the line geometry is stored in vertex arrays and the GPU is used to compute its screen coordinates a and color. By exploiting the native parallel processing pipelines of the GPU, this improves the performance greatly compared to the original CPU-based computation.

- Stencil test – rendering semi-transparent lines clears up the display but it also introduces a serious performance hit. For large data cases with many overlapping lines, the rendering speed is decreased by 70%. By using stencil test this damage is lowered and the overall penalty in using transparency decreases speed only by 25%.

- Frame buffer objects and advanced blending – the actual rendition is stored in a high definition texture. This stores the screen density to future re-use and the density mapping function can be changed without having to re-render the view. Changing the density mapping function, or the transfer function (as some call it), changes the focus between areas of high density and areas of low density. The exploration of trends and outliers and differences between them is thus feasible even for large data cases.

Chapter 5

Future Work

The documented results together with the related work give many ideas for future research. The large data in information visualization (and in visualization generally) is an exciting phenomenon that motivates combination of different approaches and stimulates research in many directions. The future of the project is oriented mostly on improving the combination of data-oriented and screen-oriented methods.

One of the practical issues is an intelligent outlier detection and special handling of the outliers. Pre-processing in the form of binning proved to be a good starting point for outlier detection, mainly because it eliminates the large data. The importance of outliers differs between eventual visualization application, but in general they should not be thrown away by the data abstraction or data reduction methods. They often contain valuable information and therefore they need to be treated separately both in the data space and in the screen space.

Another demand extends the popular Focus+Context concept. Certain applications dealing with large data need more levels than the two present in F+C. Levels such as *SuperContext*, *Mid-Context* or *Superfocus* need to be devised and carefully visualized.

The binning is a good starting point for other advanced data-processing techniques. Using the binned data representation, techniques such as Focus+Context or direct manipulation can be introduced to displays that were not capable of handling those because of arduous effort spend on dealing with large data.

These and others are the future directions of the research with respect to the documented project. The importance of large data visualization is becoming hot topic in more and more research domains and therefore new improvements and contributions are necessary. The project is taking a promising direction in introducing new and effective solutions to the problems caused by large data in information visualization.

Bibliography

[1] http://davis.wpi.edu/xmdv/datasets.html.

[2] http://dbvis.fmi.uni-konstanz.de/pics.php?type=application.

[3] http://www.census.gov/prod/cen2000/doc/pums.pdf.

[4] http://www.missionforvisionusa.org.

[5] http://www.mvrf.org.

[6] http://www.simvis.at.

[7] http://www.smartmoney.com/marketmap/.

[8] http://www.turbosquid.com/fullpreview/index.cfm/id/255161.

[9] Almir Olivette Artero, Maria Cristina Ferreira de Oliveira, and Haim Levkowitz. Uncovering clusters in crowded parallel coordinates visualizations. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'04)*, pages 81–88, Washington, DC, USA, 2004. IEEE Computer Society.

[10] Michelle Q. Wang Baldonado, Allison Woodruff, and Allan Kuchinsky. Guidelines for using multiple views in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 110–119. ACM Press, 2000.

[11] Enrico Bertini and Giuseppe Santucci. By chance is not enough: Preserving relative density through non uniform sampling. In *IV '04: Proceedings of the Information Visualisation, Eighth International Conference on (IV'04)*, pages 622–629, Washington, DC, USA, 2004. IEEE Computer Society.

[12] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman, editors. *Readings in information visualization: Using vision to think*. Morgan Kaufmann Publishers, San Francisco, 1999.

[13] A Dix and G P. Ellis. By chance: enhancing interaction with large data sets through statistical sampling. In *Proc. AVI'02*, pages 167–176. ACM Press, 2002.

[14] Helmut Doleisch, Martin Gasser, and Helwig Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *VISSYM '03: Proceedings of the symposium on Data visualisation 2003*, pages 239–248, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[15] S. Eick and G. Wills. High interaction graphics, 1995.

[16] Geoffrey Ellis, Enrico Bertini, and Alan Dix. The sampling lens: making sense of saturated visualisations. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1351–1354, New York, NY, USA, 2005. ACM Press.

[17] S. Feiner and C. Beshers. Worlds within worlds: metaphors for exploring n-dimensional virtual worlds. In ACM, editor, *Third Annual Symposium on User Interface Software and Technology UIST*, pages 76–83, New York, NY 10036, USA, October 1990. ACM Press.

[18] Jean-Daniel Fekete and Catherine Plaisant. Interactive information visualization of a million items. In *INFOVIS*, pages 117–124, 2002.

[19] Martin Florek. Interactive information visualization using graphics hardware. Master's thesis, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, 2006.

[20] J.H. Friedman and J.W Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers*, C-23(9):881–889, 1974.

[21] Michael Friendly. *Visualizing Categorical Data*. SAS Publishing, 2001.

[22] Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In David Ebert, Markus Gross, and Bernd Hamann, editors, *IEEE Visualization '99*, pages 43–50, San Francisco, 1999. IEEE.

[23] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multidimensional geometry. In *IEEE Visualization '90 Proceedings*, pages 361–378. IEEE Computer Society, October 1990.

[24] Jimmy Johansson, Patric Ljung, Mikael Jern, and Matthew Cooper. Revealing structure within clustered parallel coordinates displays. In *IEEE Symposium on Information Visualization (INFOVIS)*, 2005.

[25] Brian Johnson and Ben Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *VIS '91: Proceedings of the 2nd conference on Visualization '91*, pages 284–291, Los Alamitos, CA, USA, 1991. IEEE Computer Society Press.

[26] Daniel A. Keim. Visual exploration of large data sets. *Communications of the ACM (CACM)*, 44(8):38–44, 2001.

[27] Daniel A. Keim. Scaling visual analytics to very large data sets. In *Workshop on Visual Analytics*, 2005.

[28] Daniel A. Keim, Mihael Ankerst, and Hans-Peter Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, page 279, Washington, DC, USA, 1995. IEEE Computer Society.

[29] Daniel A. Keim and Annemarie Herrmann. The gridfit algorithm: An efficient and effective approach to visualizing large amounts of spatial data. *vis*, 00:181, 1998.

[30] Teuvo Kohonen. *Self organizing maps*. Springer, New York, 2000.

[31] Robert Kosara, Helwig Hauser, and Donna Gresh. An interaction view on information visualization. In *EUROGRAPHICS 2003*, 2003.

[32] J. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.

[33] Allen R. Martin and Matthew O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, page 271, Washington, DC, USA, 1995. IEEE Computer Society.

[34] John J. Miller and Edward J. Wegman. Construction of line densities for parallel coordinate plots. pages 107–123, 1991.

[35] T.D. Nguyen, T.B. Ho, and H. Shimodaira. A visualization tool for interactive learning of large decision trees. *ictai*, 2000.

[36] Matej Novotný. Visual abstraction for information visualization of large data. Master's thesis, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, 2004.

[37] Matej Novotný. Visually effective information visualization of large data. In *Proceedings of the 8th Central European Seminar on Computer Graphics*, 2004.

[38] Matej Novotný. Outlier-preserving focus+context visualization of large data. Technical report, VRVis Research Center, Vienna, 2006.

[39] Matej Novotný and Helwig Hauser. Similarity brushing for exploring multidimensional relations. *Journal of WSCG*, 14, 2006.

[40] Vlado Roth. Web-based parameter space explorer for non-invertible 3d maps. In *Proceedings of the 9th Central European Seminar on Computer Graphics*, 2005.

[41] Vlado Roth. Computation and visualization of bifurcation diagrams for non-invertible 3d maps, rigorous thesis. Master's thesis, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, 2006.

[42] Robert Spence. *Information visualization*. Addison-Wesley, 2000.

[43] James J. Thomas and Kristin A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

[44] Edward R. Tufte. *Envisioning Information*. Graphics Press, 1990.

[45] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[46] E. Wegman and Q. Luo. High dimensional clustering using parallel coordinates and the grand tour. Technical report, Center for Computational Statistics, George Mason University., 1996.

[47] G. J. Wills. Selection: 524,288 ways to say ”this is interesting”. In *INFOVIS ’96: Proceedings of the 1996 IEEE Symposium on Information Visualization (INFOVIS ’96)*, page 54, Washington, DC, USA, 1996. IEEE Computer Society.

[48] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *VISSYM ’03: Proceedings of the symposium on Data visualisation 2003*, pages 19–28, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

Publications and International Presentations

- M. Novotný. *Visually Effective Information Visualization of Large Data*
  In CESCG 2000 – 2005 Best Paper Selection, ISBN 3-85403-204-8, 2006
- M. Novotný, H.Hauser. *Similarity Brushing for Exploring Multidimensional Relations*
  In Journal of WSCG, Vol.14, No.1-3, ISSN 1213-6972, ISBN 80-86943-09-7, 2006
- M. Novotnỳ, H.Hauser. *Outlier Preserving Focus+Context Visualization of Large Data*
  Technical report, VRVis Research Center Vienna, 2006
- M. Novotný. *Interactive Visual Exploration of Large Multidimensional Data (or Cleaning The Traffic Jam on The Way From Screen To Brain*
  Presentation at the University of Konstanz, Germany, 2005
- M. Novotný. *Visual Abstraction for Information Visualization of Large Data*.
  Master thesis at the Comenius University, Bratislava, 2004.
- M. Novotný. *Visually Effective Information Visualization of Large Data*
  Student Scientific Conference, Brno, 2004.
- M. Novotný. *Visually Effective Information Visualization of Large Data*
  Proceedings of the 8th Central European Seminar on Computer Graphics (CESCG 2004).
- M. Novotný, R.Kosara. *Visually Effective Information Visualization of Large Data*
  Technical report TR-VRVis-2004-006, VRVis Research Center Vienna, 2004
- M. Novotný. *Visually Effective Information Visualization of Large Data*
  Presentation at VisForum #86, VRVis Research Center Vienna 2004