Visually Effective Information Visualization of Large Data

Matej Novotny*

VRVis Research Center for Virtual Reality and Visualization
Vienna / Austria
http://www.VRVis.at/vis/

Abstract

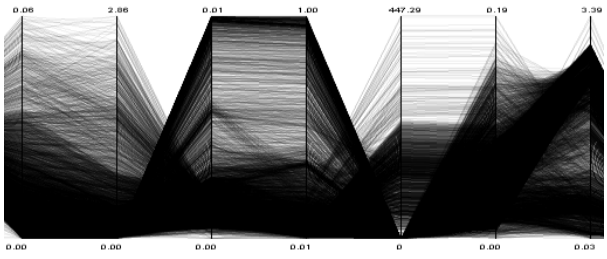
Recent technology developments produce an increasingly large volume of information. Therefore visualization of these data requires sophisticated and efficient methods that take the amount of data into account. The information often gets lost or hidden in displays of traditional information visualization techniques. A significant improvement can be achieved using clustering and visual abstraction. The synergetic approach introduced here combines visual and computer data mining. Its effect is demonstrated on a popular information visualization method – the parallel coordinates.

Keywords: Large Data, Information Visualization, Parallel Coordinates, Visual Abstraction, Clustering

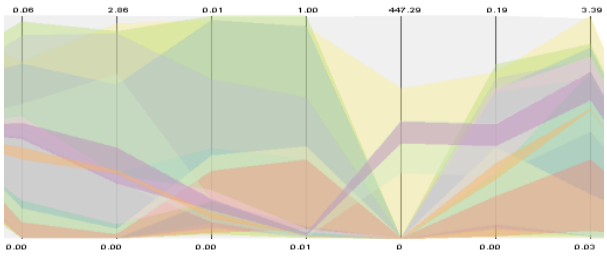
1 Introduction

Visualization plays an important role in the communication with computers. The results of any simulation, scan or survey are easily hidden or lost when an improper visualization is used. The data are often multivariate, structured or hierarchical and it is hard to communicate this information to a human through the limited resources of a computer display [22]. An example of such a display, a parallel coordinates view of ten thousand samples can be seen in Figure 1, top. This and similar views get easily cluttered by large data sets and visual exploration in that cases becomes a difficult task. Many graphical and human processing resources can be saved by abstracting into a simplified representation of the data, Figure 1, bottom. Displaying less information while keeping the visualization relevant to the original meaning of the data is an useful means for large data visualization. Although the human visual system is capable of perceiving a rich volume of a wide variety of impulses, it might get overloaded by the large number of visual cues from an information visualization display. It is therefore necessary to use the communication channels between the human and the computer in the most efficient way. The goal of this work is to introduce a new approach to information visualization, the visual abstraction.

*8novotny@st.fmph.uniba.sk



a)



b)

Figure 1: The same data is visualized by traditional parallel coordinates (top) and using visual abstraction (bottom).

1.1 The Data of Interest

The information to be visualized comes from various sources and in various forms. For example, physical simulations, economical surveys, weather observations etc. Whether the data origin is a simulation or a real world observation, its dimensionality is usually greater than the mere three that we are used to. Moreover, information visualization (InfoVis) suffers from another problem compared to scientific, medical or flow visualization: it can rarely assume any a priori knowledge about the nature of the data.

1.2 InfoVis Techniques

There are many approaches to handle these problems and it is hard to define a working taxonomy that will cover all of them. But many of the information visualization techniques share some common properties depending on how they deal with the multivariate data and how they exploit the display [17].

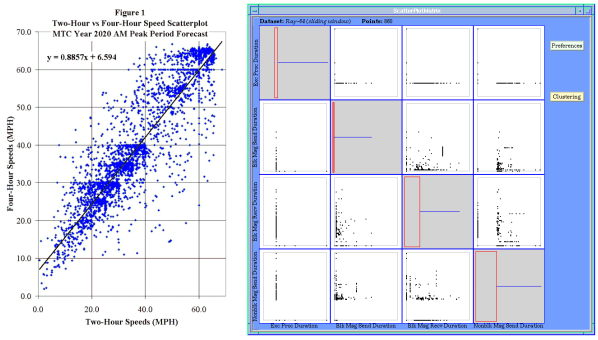


Figure 2: Example of a scatterplot and a scatterplot matrix. Image courtesy of Charles L. Purvis, Metropolitan Transportation Commision.

Dimensional subsetting is a simple approach, where a slice at a given position is extracted from the data. The slice usually has two or three dimensions and it can be visualized more easily. Different slices can be combined to cover all the dimensions. For example scatterplots [2, 5] or scatterplot matrices, (Figure 2).

Dimensional embedding also works with slices of lower dimensionality. The dimensions are divided into those that are in the slice and those that create the wrapping space where these slices are then embedded at their respective position. This method is used in dimensional stacking [14] or worlds within worlds [7]

Axis reconfiguration involves a projection into another coordinate system. Chernoff faces [4] map several dimensions onto facial features. Parallel coordinates [10] create a planar representation of an n-dimensional space by mapping points to polylines, (Figure 3). This method is further described later in this paper.

Pixel oriented techniques are particularly useful for displaying very large amount of data of relatively low dimensionality, since they try to represent each data sample by a pixel-based area of certain properties such as color or shape, [8, 11].

Dimension reduction methods try to reduce the dimensionality of the data into a number that is easier to display. The concept of these methods aims to represent the relations inside the data mostly in a two dimensional space. For example multidimensional scaling or self organizing maps [12, 13].

1.3 Parallel Coordinates

This popular method for visual exploration utilizes the axis reconfiguration approach to information visualization. Every n-dimensional point is represented by a polygonal line according to its position in the original space, (Figure 3).

N copies of the real line are placed equidistant and parallel to each other. They are the axes of the parallel coordinate system for R^N. A point C with coordinates $(c_1, c_2, \ldots, c_N)$ is represented by a polygonal line connecting the positions of c_i on their respective axes [10].

This projection provides a 2-dimensional display of the whole data set and is capable of displaying up to tens of different dimensions. An unpleasant drawback is the cluttered display when trying to render a large number of samples. Interaction and even mere understanding of such a display is complicated (Figure 4). Numerous modifications to parallel coordinates [18, 21] and efforts towards efficient displaying of large data in parallel coordinates [9] make this method a promising working ground for visual abstraction and large data visualization. This paper addresses the situation and uses parallel coordinates as the showcase for visually effective information visualization.

2 Large Data Problems

Large data sets can mean a serious problem when it comes to their visualization. Information visualization treats data from different sources and most of them are strongly influenced by the ever growing processing power of computers. They produce increasingly large volumes of data and traditional methods start to fail in efficient visualization of this amount of data. The consequential problems can be categorized into three main groups.

2.1 Loss of Speed and Interaction

Visual exploration requires a user-friendly interface for the most convenient way to focus on areas of interest and zoom in on desired details. It can be a problem to guarantee a sufficiently short response time if the application handles millions of multivariate observations.

The speed issue can be improved by using faster computers and more sophisticated graphical hardware. Yet the visual efficiency still strongly depends on minimizing the influence of the other remaining problems connected to large data visualization. A promising means to decrease

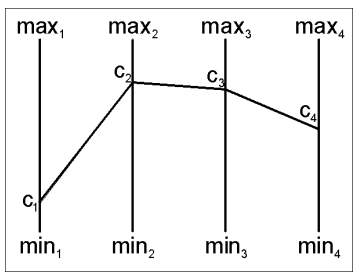


Figure 3: Parallel Coordinates. Point $C(c_1, c_2, c_3, c_4)$ is represented by a polygonal line.

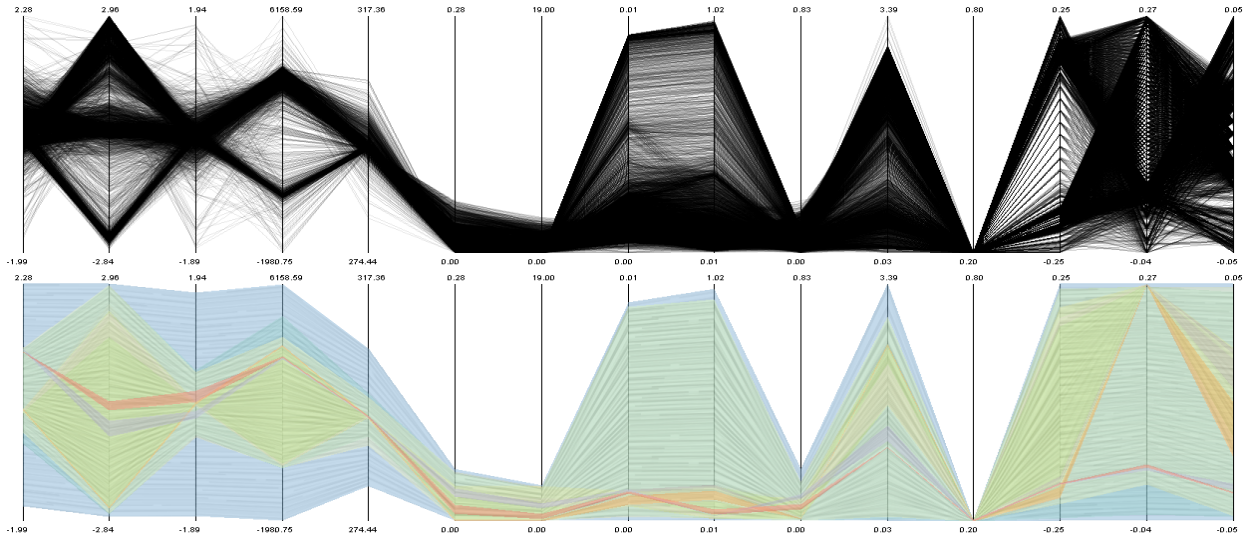


Figure 4: Large data visualization in traditional parallel coordinates. Virtually no patters or groups can be observed in the cluttered areas (top). It is also hard to determine the number of samples in different regions. Visually effective information visualization using parallel coordinates (bottom). 10 thousand samples, 15 dimensions.

the amount of displayed data while preserving the information contained in it is described in Section 3.

2.2 Occlusion

This problem is not strictly connected to large data itself. It is a common problem of any visualization and many efficient ways to deal with it emerged in the past. Occlusion happens when two or more visual elements overlap each other in a way, that obstructs in perception of some of them (Figure 4), top. It can lead to bad interpretation of the view or to missing some important details or relations. Occlusion is a significant concern when processing large data since the probability of overlapping of two or more items rapidly grows with the amount of items in the display.

2.3 Aggregation

Similarly to occlusion, aggregation is a frequent phenomenon in large data visualization. When objects are drawn over each other it is hard to determine the number of them or to compare the volume of samples in aggregated areas (Figure 4), top. Misinterpretation of the data can happen again. This problem is very common in scatterplots or parallel coordinates, but can also be found in many other visualization techniques.

Both occlusion and aggregation problems can be partially solved using semitransparent lines [23] or by coloring the lines. But these modifications to parallel coordinates still fail when really large data sets have to be visualized. Also additional processing needs to be done in order to preserve outliers in the data. The visual abstraction approach presented in this paper doesn't suffer from these problems.

3 Visual Abstraction

During the process of visual data exploration it is usually not necessary to show all the details at once. Many visual and computing resources can be saved by displaying a less complicated image with respect to the viewer's needs. This concept is familiar to computer graphics and was incorporated into many techniques for level-of-detail or non-photorealistic rendering together with information visualization itself. Abstraction always discards some information. The purpose of such behavior is to reduce the volume of information while preserving its meaning. It is therefore essential to know which details carry the most information. Some applications can exploit the known nature of the data they treat and can estimate the importance of various details more easily, (Figure 5). But most of the information visualization methods try to be as general as possible in order to be compatible with different kinds of data. One of the ways to obtain the importance information in such cases is further described in Section 4.

The ratio between simplicity and truthfulness is always a subject of discussion. The solution presented in this paper offers a possibility to observe the data at different levels of abstraction and even to change the levels across different areas in the same display.

4 Clustering

Data that describe the real world usually contain groups of samples sharing similar properties. Most of the sam-

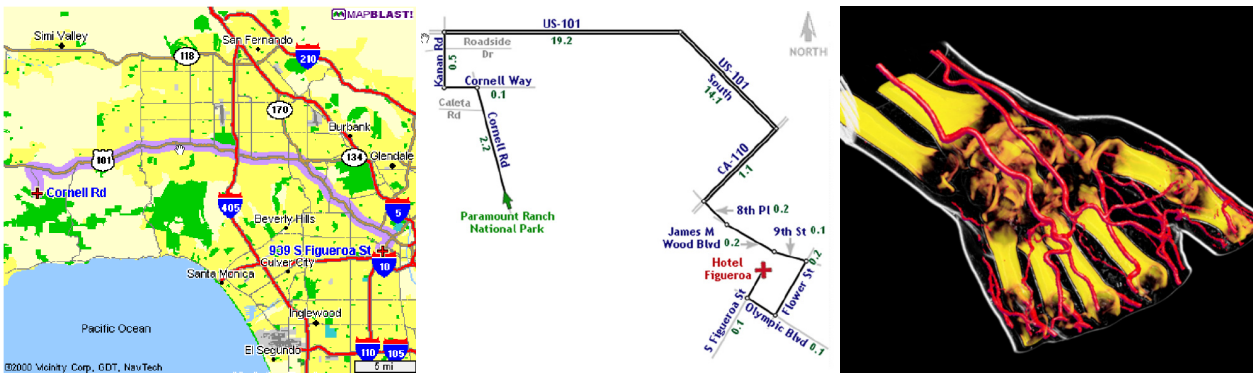


Figure 5: Examples of visual abstraction: Route map visualization in traditional software (left) and using visual abstraction (middle) [1]. Non-photorealism in scientific visualization: the skin is rendered as a contour to enhance the view (right) [16]. Images courtesy of Maneesh Agrawala and VRVis Research Center.

ples inside such group do not bring any new information and thus they can be filtered out without any significant damage to the meaning of the data. An abstract representation would have all these samples discarded and the group replaced by a larger and simpler element. Such a representation is a good starting point for visual abstraction and a promising solution for large data information visualization.

Clustering helps to obtain an estimation of presence and position of these groups. The process assigns similar samples to the same cluster and forms a partitioning of the space into these clusters. Such clusters are an approximation of the real world groups. The precision of the approximation depends on the clustering algorithm and on the data itself.

There are two types of clustering algorithms depending on the direction of their progress [19]. Top-to-bottom algorithms group the samples by partitioning the whole set into decreasingly smaller areas until a certain level of precision is reached. They are less time consuming, but might produce wrong results by accidentally drawing the partitioning border between similar units. In contrast, the bottom-to-top algorithms start on the lowest level of abstraction, grouping pairs of similar units and clusters together. This approach produces safer results for the price of higher time complexity, (Figure 6).

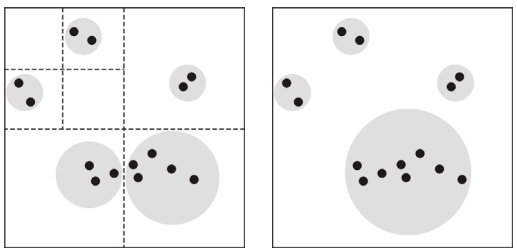


Figure 6: Clustering approaches: The top-to-bottom approach (left) accidentally divides the largest group into two clusters in contrast to the bottom-to-top approach (right).

For the purposes of visually effective large data visualization the bottom-to-top approach seems to be a better choice. The top-to-bottom algorithms consume an exponential amount of memory depending on the number of dimensions, which can be a large number when dealing with multivariate data for information visualization. Furthermore, the process of bottom-to-top clustering follows the process of building a level-of-abstraction hierarchy. The different abstract layers correspond to clusters obtained at different similarity thresholds.

4.1 2D Binning

Though not precisely a clustering, the 2D binning offers a helpful insight on the data. The binning occurs within a chosen combination of 2 dimensions and in that case it is often convenient to implement a bottom-to-top algorithm. The 2D binning repeatedly divides a two dimensional subspace of the whole data set and forms clusters at each level of precision. An abstract structure built by this algorithm significantly reduces clutter in scatterplots and is also useful to explore data in parallel coordinates when the data are investigated according to their behavior between two of the axes. It is also useful for viewport linking between the scatterplots and the parallel coordinates, see Figure 7.

4.2 K-means

A simple bottom-to-top algorithm called k-means was chosen as the basic clustering algorithm for this work [15]. K-means begins with a selection of k starting samples. Every sample is then merged with the nearest sample or cluster centroid and the position of the centroid is recalculated until all the samples are assigned to a cluster. It is obvious that the key factor in such clustering is the selection of the starting points.

A popular criterion on the starting points is their maximal mutual distance. This choice produces nice results,

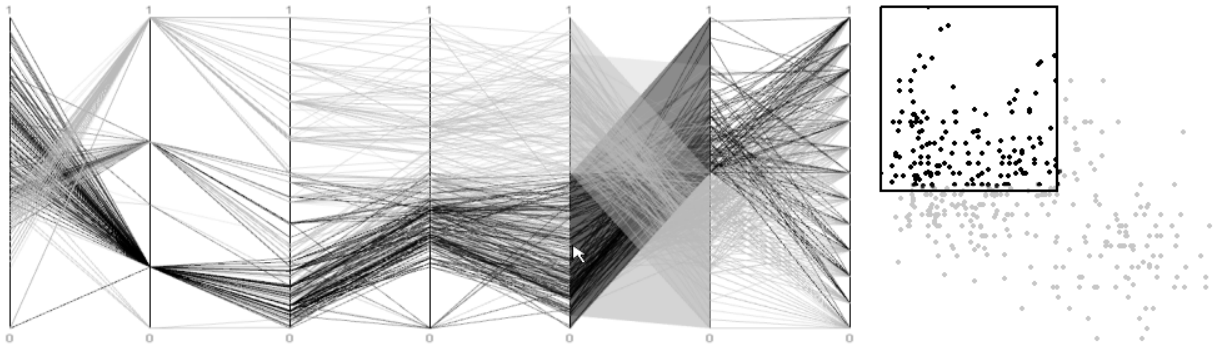


Figure 7: 2D binning in parallel coordinates and scatterplots: Data are highlighted according to the occupied portion of the two dimensional subspace between the fifth and the sixth channel.

but it is virtually impossible to compute all distances within a large multivariate data set in a reasonable time. Therefore it is convenient to have a similar, but less time-demanding algorithm.

4.3 Modified k-means

To avoid extensive computing of a high number of distances, the following modification was implemented [15]. The cluster assignment is chosen randomly for every sample and the centroid of every cluster is computed. Every sample is then reassigned to the cluster with the nearest centroid. The former and the new centroids are recomputed and another sample is reassigned until a stable solution is reached. The selection of starting points does not play such an important role, but the order in which the samples are reassigned does. This order can be randomized to avoid most critical cases.

4.4 Implementation details

The k-means as a non-hierarchical algorithm would seem against the plans to form a structure with different levels of abstraction. But this algorithm is used only to form the lowest layer of the cluster tree. Further abstraction can then use different similarity measures for clusters than for samples. For example clustering can use Euclidean distance to detect similar samples. The clusters on higher abstract levels can be compared according the similarity of their size, variance or other properties.

Nevertheless, k-means and its modifications sometimes fail to copy the real nature of the data. Other approaches like clustering through fractionation / refractionation [20] or the vector quantization [6] seem to be a promising solution towards data reduction and abstraction in large data sets.

5 Contribution

The solution presented in this paper modifies traditional approaches on both sides of the visualization process, in data space and in visualization space. New visual elements and new data structures described below have been designed to maximize the effectiveness of visualization with respect to large data. The projection of the data space onto the screen has been extended into a longer pipeline that offers better visual results for the price of a simple preprocessing.

5.1 Visual Elements

Introducing visual abstraction to the parallel coordinates requires new visual representation for the new kind of hierarchical data. Some previous work on this topic was described in [9] and this paper presents another approach to depicting clusters in the parallel coordinates.

A cluster is visualized by a polygonal area representing the bounding box of the cluster in the N-dimensional space. Different clusters are colored using different colors to distinguish them, (Figure 4, bottom). The colors are chosen to be as different as possible while having the same intensity [3]. This prevents false emphasis on clusters of colors that the human visual system is more sensitive to.

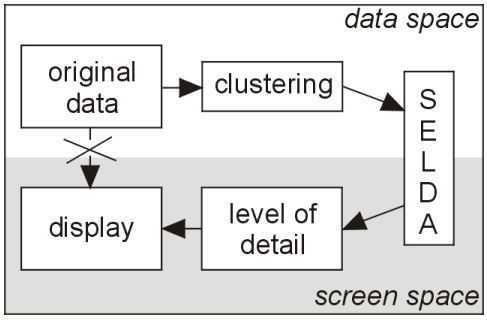


Figure 8: The shortest route is not always the fastest one.

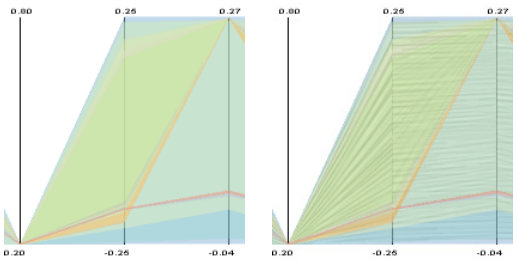


Figure 9: Using texture to distinct overlapping areas.

To avoid occlusion, opacity mapping was implemented and different modes of the mapping are available:

- uniform mapping – all clusters have the same opacity value, usually $1/k$, where k is the number of them
- population mapping – clusters with higher data population are emphasized. The opacity value is the ratio of the cluster population to the number of samples in the whole data set.
- density mapping – the opacity value is the ratio of the cluster population to the size of the cluster. This mapping provides the best results, because the less populated but dense clusters are emphasized.

These methods can be used to observe various properties of the clusters and are helpful in finding the areas of interest. To improve the results of the semitransparent overprinting, the clusters are sorted according to their sizes before the rendering. This prevents from overdrawing of the smaller clusters by the larger ones.

The data behavior in overlapping areas can be enhanced by applying a stripe or hatch texture to the polygons. This texture helps distinguish clusters that occupy the same area between two axes, but have different behavior within the two dimensions, (Figure 9).

5.2 SELDA

During the process of visualization, an efficient structure to store the abstract information is necessary. Especially with large data. This novel structure called SELDA (Structure for Efficient Large Data Abstraction) is similar to a space partition tree but is adapted to certain needs. The original data are left intact, memory consumption is minimized with a slight tradeoff towards faster level-of-abstraction transitions and convenient bidirectional assignment is available between samples and clusters. The whole structure consists of a list of levels of abstraction (usually no more than three or four levels of abstraction are used), a tree of the clusters and a data layer, where two tables are stored – cluster index table (CIT) and data index table (DIT). These tables store the mutual assignment between the clusters and the data samples, (Figure 10). Every node in the cluster tree stores information about the number of

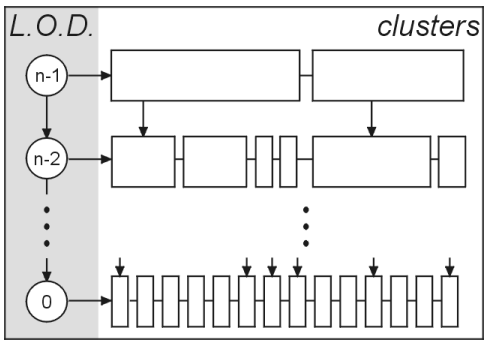


Figure 10: Diagram of the SELDA structure. Notice that only one pointer is used to address the children of a node.

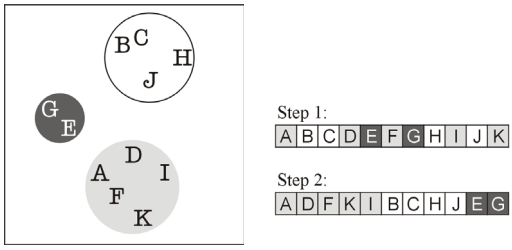


Figure 11: Data Index Table reordering. In the first step, clusters are assigned to data. In the second step, the indices are ordered according to their cluster membership.

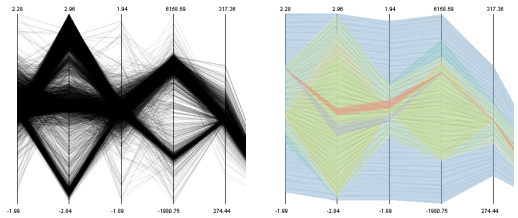
samples contained in it, their mean, variance and boundary values for all dimensions. Unlike a regular tree structure, pointers to neighbors on the same level are stored to facilitate easy access to the nodes of the same level of abstraction.

Building of SELDA is a three-step process. First a clustering has to be performed and the cluster membership of every sample is written into CIT. Then the DIT table is sorted and basic clusters are formed. The new order of the indices in DIT enables that only two numbers are necessary to store the information about the children of a node: the pointer to the first one and the total number of them, (Figure 11).This assures predictable and fixed memory demands. In the third step the higher levels of the tree are abstracted from the basic clusters.

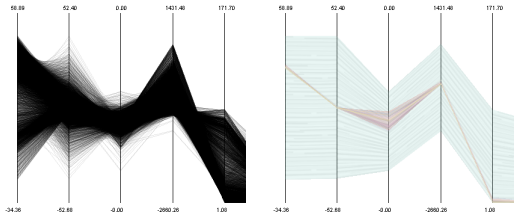
6 Results

Using the modification of traditional methods mentioned in the previous sections, significant improvement in large data visualization was achieved. For the price of relatively cheap preprocessing (tens of thousands of samples, 12 to 20 dimensions each, organized into clusters in less then a minute on a consumer PC) a much clearer display and easily perceivable information is the result of this approach. The interaction with such a display is many times

faster than with traditional information visualization and occlusion together with aggregation effects are minimized, see Figure 12. Along with that, the details of the data are easily accessible via decreasing the level of abstraction in the areas of interest.



(a)



(b)

Figure 12: Minimizing the problems with large data visualization.
(a) Occlusion: What appeared as a homogenous region in the middle part of the traditional parallel coordinates (left) reveals two different groups in abstract view (right).
(b) Minimizing the aggregation: It is now clear to see (right) where the most of the data lies and that the rest is less populated.

7 Conclusion and Future Work

We have proposed a promising solution and a relatively novel approach to information visualization – the visual abstraction. Together with special treatment of the data and the graphical resources this approach offers an interesting improvement towards visually effective information visualization of large data. The results of these methods are encouraging and several areas for future work are revealed. Namely experiments with other algorithms for clustering, extensions to other information visualization displays and special improvements when data of particular properties are provided.

8 Acknowledgments

This work was done as a part of the basic research on information visualization at the VRVis Research Center in Vienna, which is funded by an Austrian research program called K plus. The author would like to thank Robert Kosara, Helwig Hauser, Martin Gasser and all the VRVis staff for cooperation. Special thanks go to Andrej Ferko.

References

[1] Maneesh Agrawala and Chris Stolte. Rendering effective route maps: Improving usability through generalization. In *(SIGGRAPH) 2001, Computer Graphics Proceedings*, pages 241–250. ACM Press / ACM SIGGRAPH, 2001.

[2] Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[3] Cynthia A. Brewer. The color brewer: http://www.personal.psu.edu/faculty/c/a/cab38/colorbrewerbeta2.html.

[4] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the Americal Statistical Association*, (342):361–368, 1973.

[5] William C. Cleveland and Marylyn E. McGill. *Dynamic Graphics for Statistics*. CRC Press, Inc., 1988.

[6] Juan A. Corral, Miguel Guerrero, and Pedro J. Zufiria. Image compression via optimal vector quantization: A comparison between SOM, LBQ and K-means algorithms. In *Proc. ICNN'94, International Conference on Neural Networks*, pages 4113–4118, Piscataway, NJ, 1994. IEEE Service Center.

[7] S. K. Feiner and Clifford Beshers. Worlds within worlds: metaphors for exploring n-dimensional virtual worlds. In *Proceedings of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology*, pages 76–83. ACM Press, 1990.

[8] Jean-Daniel Fekete and Catherine Plaisant. Interactive information visualization of a million items. In *Proceedings of IEEE Symposium on Information Visualization 2002 (InfoVis 2002)*, October 2002.

[9] 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.

[10] 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.

[11] Daniel A. Keim, Ming C Hao, Umesh Dayal, and Meichun Hsu. Pixel bar charts: a visualization technique for very large multi-attribute data sets. *Information Visualization*, 1(1):20–34, 2002.

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

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

[14] Jeffrey LeBlanc, Matthew O. Ward, and Norman Wittels. Exploring n-dimensional databases. In *Proceedings of the 1st conference on Visualization '90*, pages 230–237. IEEE Computer Society Press, 1990.

[15] Clustan Ltd. K-means cluster analysis: http://www.clustan.com.

[16] Helwig Hauser Markus Hadwiger, Christoph Berger. High-quality two-level volume rendering of segmented data sets on consumer graphics hardware. In *Proceedings of IEEE Visualization 2003*.

[17] Donna L. Gresh Robert Kosara, Helwig Hauser. An interaction view on information visualization. In *Proceedings of EUROGRAPHICS 2003, (EG 2003)*, pages 123–137, 2003.

[18] Harri Siirtola. Direct manipulation of parallel coordinates. In *Proc. of IEEE Conf. on Info. Vis.'00*, 2000.

[19] Serengul Smith. Machine learning techniques: http://www.cs.mdx.ac.uk/staffpages/serengul/clustering.htm.

[20] Jeremy Tantrum, Alejandro Murua, and Werner Stuetzle. Hierarchical model-based clustering of large datasets through fractionation and refractionation. In David Hand, Daniel Keim, and Raymond Ng, editors, *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-02)*, pages 183–190, New York, July 23–26 2002. ACM Press.

[21] Holger Theisel. Higher order parallel coordinates. In *Proc. Vision, Modeling and Visualization 2000*, pages 119–125, 2000.

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

[23] Edward J. Wegman and Qiang Luo. High dimensional clustering using parallel coordinates and the grand tour. Technical Report 124, Fairfax, Virginia 22030, U.S.A., 1996.