Hifocon: Object and Dimensional Coherence and Correlation in Multidimensional Visualization

Soon Tee Teoh[1] and Kwan-Liu Ma[2]

[1] Department of Computer Science, San Jose State University
[2] Department of Computer Science, University of California, Davis

Abstract. In any multidimensional visualization, some information has to be compromised when projecting multidimensional data to two- or three-dimensional space. We introduce the concepts of dimensional and object coherence and correlation to analyze and classify multidimensional visualization techniques. These concepts are used as principles for our design of Hifocon, a new multidimensional data visualization system.

1 Introduction

Multidimensional visualization is challenging because humans live in a three-dimensional world and have no intuition of higher dimensional space. Therefore, any attempt to visualize multidimensional data must find a projection from the high-dimensional space to a two- or three-dimensional visual space that is intelligible to humans. As a result, different multidimensional visualization ideas have been proposed.

To classify multidimensional visualization methods in terms of their emphases and tradeoffs, we introduce the principles of dimensional and object coherence and correlation. We discuss the strengths and limitations of existing multidimensional visualization techniques, and analyze and classify them using these principles.

We then introduce Hifocon (High-Dimensional Focus+Context), a multidimensional data visualization system designed to provide strong coherence and correlation. We show how Hifocon helps users to gain useful and interesting information. A particular strength of Hifocon is that it can be used to find hierarchical clusters, and clusters which are outliers in other dimensions.

2 Object and Dimension Coherence and Correlation

In high-dimensional visualization, not only are there too many objects to visualize at once, but there are also too many dimensions to visualize at once.

To discuss multidimensional visualization, we introduce a new concept: *object coherence* and *dimension coherence*. In a visual representation with object coherence, each object is represented as a single and coherent visual entity, such as a point. Lack of object coherence happens when an object is visually represented as separate visual entities such as several points. In such a visualization,

G. Bebis et al. (Eds.): ISVC 2005, LNCS 3804, pp. 235–242, 2005.

the user cannot see clearly the properties of the object all at once; therefore we say that the visualization of the object is "not coherent".

Dimension coherence is satisfied when the distribution of objects' attribute values in each dimension is clear. As we show in Section 3, many multidimensional visualization methods do not satisfy object and dimension coherence.

In general, object coherence is desirable when the user is interested in knowing the object's attribute values in many different dimensions and how they relate to one another. Dimension coherence is desirable when the user wants a clear picture of how objects are distributed in this dimension; for example, whether there are clusters present.

Correlation is another important aspect of understanding data. We define *object correlation* to be the property of a visualization that allows a user to tell whether two objects are similar in their attribute values, and to visually group similar objects. We say that a visualization has *dimension correlation* among a certain number of dimensions when the user is able to easily tell whether these dimensions are correlated according to the attribute values of the objects in the dataset.

Object and dimension correlation are very desirable and useful properties of a visualization, but because of the difficulty of visualizing high-dimensional data, they are often not achieved.

3 Related Work

Some existing techniques serve as examples to illustrate the concept of object and dimension coherence.

In parallel coordinates [6], each dimension is represented as a vertical line. Each object is mapped to one point on each line according to its attribute value in that dimension. A poly-line is then drawn to connect all the points. In parallel coordinates, there is good dimension coherence because for each dimension, the distribution of all the objects' attribute values for that dimension is clear. Furthermore, the correlation between adjacent dimensions is also visible. However, object coherence is not achieved in parallel coordinates because from the visualization, one cannot tell all the attribute values of any single oject. Similarly, object correlation is bad in parallel coordinates. In this respect, the primary focus of parallel coordinates is the dimensions, not the objects, and furthermore, when the user looks at a certain dimension, the focus is on that dimension and its adjacent dimensions because the relationship between those dimensions are obvious while all the other dimensions are still visible in context. Yang et al. [21] presented one way to enhance the perception of dimensional correlation in parallel coordinate is to order the dimensions such that similar dimensions are placed adjacent to each other.

Another popular multidimensional visualization method is the scatterplot matrix. In each position *(i,j)* in the matrix, a scatterplot is drawn with dimension i as the x-axis and dimension j as the y-axis. In this visualization, there is no object coherence because each object is shown as multiple points and the

user cannot tell all the object's attribute values for any single object. There is dimension coherence because the user can tell the distribution of all the objects' attribute values for any dimension that the user is interested in. Furthermore, the correlation between any two dimensions i and j is clear from looking at the scatterplot in position *(i,j)*. However, the user cannot simultaneously observe the correlation between more than two dimensions.

Dimension-reduction techniques are also commonly used in visualization. For example, Principal Component Analysis (PCA) [7] can be conducted on the data, and a scatterplot is shown with the first two principal components as the x- and y-axis. In such a visualization, there is no dimension coherence because each principal component is a combination of different original dimensions and therefore, from the visualization, one cannot tell an object's attribute value in any of the original dimensions. There is very good object coherence because each object is simple shown as a point, and the screen position of each object in relation to other objects can be easily observed. Similarly, there is also good object correlation because PCA tends to place objects similar in high-dimensional space close together in 2-D display. Another dimension-reduction method, MDS [19], is designed especially to preserve in 2-D the inter-object distances in higher dimensions. Projection Pursuit [5] methods provide more general projections of high-dimensional to low-dimensional space.

There are other existing multidimensional visualization techniques. Some are variants of the above-discussed methods, some combine different methods, and some, such as the Grand Tour methods [3], use animation and interaction techniques to enhance the visualization, linking multiple views. All these different methods can be analyzed based on their choices of what to show in focus and what to show in context, the smoothness of their transitions between focus and context, and their trade-offs between object and dimension coherence. For example, in animated visualizations, the correlation among the objects/dimensions shown between two adjacent frames is more obvious than between two frames separated by a long period of time.

Several multidimensional visualization systems have been built and are publicly available for download and use. The XGobi [17] package includes many built-in visualization tools such as scatterplots and parallel coordinates, and has the ability to link different scatterplots. Xmdv [20] is similar, and also includes dimensional stacking [12] and star glyphs [8]. VisDB [9] includes pixel-oriented techniques, and is used for visually exploring large databases These systems allow the user to conveniently choose different visualization display methods to explore the high-dimensional data.

4 Hifocon

Hifocon is the multidimensional visualization system we designed for improved coherence and correlation. In Hifocon, there are two display areas, called "primary" and "secondary". A scatterplot is shown in each. The user is allowed to choose which dimensions to use on the 4 axes. For example, the user may choose

to use the first principal component for the x-axis and the second principal component as the y-axis in the primary display, and the original dimensions 5 and 8 as the x- and y-axes respectively in the secondary display.

Fisheye [4] magnification with star-glyph [8] are used to enhance coherence and correlation. A typical use of fisheye magnification is as follows. The user has chosen to display a scatterplot using MDS to layout the points. Then the user places a fisheye magnification lens on the display. This focuses the user's attention on the magnified objects, and these objects can be shown in more detail. Each magnified object is no longer shown as a single point, but as a star-shaped glyph. This star has n sticks radiating from the center, where n is the number of original dimensions the user has selected to focus on. The length of each stick is determined by the object's attributed value in the the stick's represented dimension; the larger the value, the longer the stick.

In this way, the user is allowed to select a subset of objects and a subset of dimensions to focus on simultaneously. Fisheye magnification with star-glyphs give good object coherence. However, there is not much dimension coherence because if MDS is used as the layout, the attribute values of the objects in any of the original dimensions cannot be discerned from the visualization. Furthermore, in star-glyphs, the focus dimensions are shown as disparate sticks on each object, so the distribution of the all the objects' values in any dimension cannot be clearly seen.

For better object coherence and dimension correlation, we designed another visualization metaphor: Arcs. In the two-scatterplot Hifocon display, each object is shown as two points, one in the primary and one in the secondary scatterplot. For better object coherence, a curved line is drawn between these two points. Now, an object is no longer two points, but one arc. Using arcs rather than straight lines to connect points give better perception of their endpoints. The arc representation has previously been used successfully in Thread Arcs [10].

Dimension correlation can also be enhanced in arcs. For example, if four different original dimensions were chosen as the four axes in the two scatterplot displays, dimension coherence is achieved for each of the four dimensions in focus. Dimension correlation is satisfied between the two dimensions shown in each scatterplot but not across the two scatterplots. Now, if arcs are drawn to connect objects in the two displays, then the dimension correlation between dimensions in the primary and secondary scatterplots becomes more obvious.

When there are too many objects in the display, the arcs can make it visually cluttered. When that happens, the user is allowed to move a focus lens (like the fisheye focus mentioned in the previous section). Only the objects covered by the lens would have arcs drawn in full color. Other objects are either displayed just as points, or have arcs drawn in less saturated color. These objects and arcs provide good context to the objects in focus.

Sometimes, a certain choice of axes provides a scatterplot visualization of the entire dataset that shows clearly the overall distribution of entire dataset. However, the visualization of parts of the data may not be clear. In such cases, Hifocon allows the user to *paint* a region of the display which the user is not

satisfied with. A new pair of scatterplots will be shown and all objects falling on the painted region in the previous scatterplot will be re-displayed in this new scatterplot-pair. Different axes can be selected for this new display that would give a better visualization.

For example, a cluster in two dimensions *(a,b)* may become two clusters in two other dimensions *(c,d)*. In this case, the first two dimensions *(a,b)* can be chosen for the parent scatterplot, and a region is painted over the cluster, and a new scatterplot is plotted for the cluster, using dimensions *(c,d)*. In this way, hierarchical clustering can be observed.

PCA and MDS can also be performed only for this subset of data to more accurately show the statistical distribution and covariance of the subset.

Painting is performed simply by clicking on the mouse and dragging over the desired region of the display. The creation of regions and new scatterplots results in a heirarchy of scatterplots. Hifocon allows the simultaneous display of a scatterplot pair in focus together with its parent and children. Arcs can also be drawn to connect points representing the same object in the different scatterplots. This enhances object coherence and dimension correlation, so that the context shown by the parent scatterplot is more intuitive.

5 Examples

The Segment dataset from the Statlog [13] database is used to evaluate Hifocon. Each object in this dataset represents an image. Each image is of one of seven types: brickface, sky, foliage, cement, window, path or grass. These seven types are thus the classes an object can belong to. Each object is defined in 19-dimensional space. An example of a dimension is the average red value over the region. Another dimension is the contrast between vertically adjacent pixels, used to detect horizontal lines.

We use some examples from the visual exploration of the Segment dataset with Hifocon to illustrate how the visualization features of Hifocon can uncover important knowledge in high-dimensional data.

The left picture in Figure 1 shows a scatterplot pair with its parent scatterplot. In the secondary scatterplot, two clusters of red points are clearly distiguishable. However, in the primary scatterplot, there is only one cluster of red points. This shows that in the x-axis of the secondary scatterplot (which the user has chosen to be the original dimension *exred-mean*), there are two distinct clusters, however, in the other three axes (which the user has chosen to be original dimensions *region-centroid-row* , *wedge-mean* , and *exgreen-mean*), there is only one cluster.

Looking at the distribution of the red points on the secondary scatterplot along the y-axis, it is also clear that the left cluster has a higher value in the y-axis (which is *exgreen-mean*). This shows that even though the two clusters are not separately clustered in *exgreen-mean*, they are still separable. Now, the user is interested in finding out if there is any such correlation with the two dimensions used as axes in the primary scatterplot. This is done by drawing

arcs and placing a focus point on one cluster and then the other, as shown in Figures 1. The results show that these two clusters are also separable in the two dimensions of the primary scatterplot, even though no clustering occurs in these two dimensions. This shows that there are two different types of class *brickface* surface type in this dataset, and these dimensions can be used to distinguish between the two types.

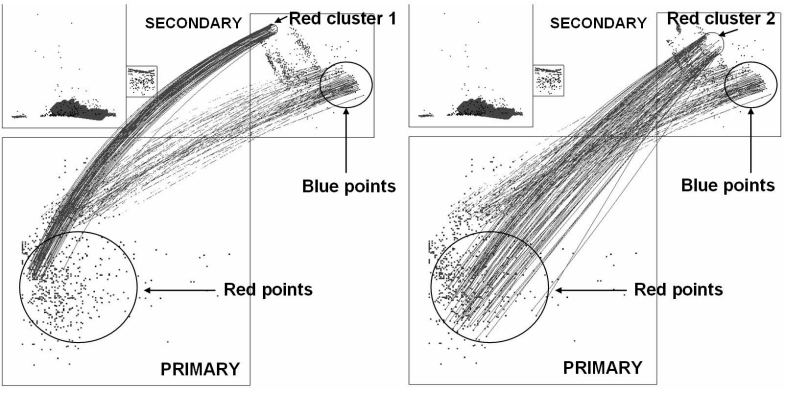

Fig. 1. Annotated screenshots. No clustering in the primary scatterplot, but the two clusters of the secondary scatterplot are separable in the primary scatterplot.

Another interesting discovery made in Hifocon visualization of the Segment dataset is that clusters in one dimension can be outliers in another dimension. This is shown when objects from one cluster in one plot has lines connected to objects which are outliers in another plot. This shows the interesting phenomenon that points in a cluster in one dimension can become outliers in other dimensions.

Figure 2 visualizes the objects belonging to the *sky* class. In the parent scatterplot (with *rawblue-mean* as the x-axis and *exred-mean* as the y-axis), there is a cluster with slightly larger value in *exred-mean* than *rawblue-mean*. By connecting the points in that cluster to the primary scatterplot (with *rawred-mean* as the x-axis and *intensity-mean* as the y-axis) with arcs, the user observes that the cluster also has slightly larger value in *rawred-mean* and smaller value in *intensity-mean*. Connecting lines to the child scatterplot shows that this cluster does not deviate in the two dimensions used for the child scatterplot. This shows that *sky* images contains a cluster that is slightly more red than other *sky* images.

6 Conclusions

We have defined the concepts of object coherence, object correlation, dimensional coherence, and dimensional correlation to help discuss and analyze multi-

dimensional visualization. We find that object and dimension correlation are not satisfied in many existing multidimensional visualization methods. Using these concepts, we have provided an analysis of well-known existing multidimensional visualization methods.

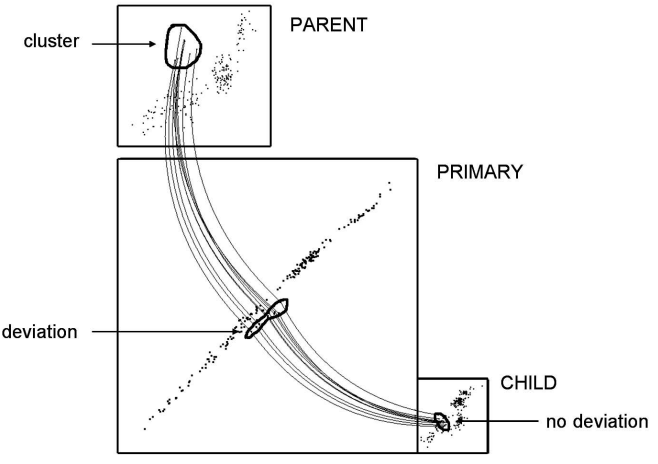

Fig. 2. Annotated screenshot. A cluster is observed in the parent scatterplot. This cluster is linked to the primary scatterplot and the child scatterplot. Deviation of the cluster is observed in the primary scatterplot but not in the child scatterplot. Such deviation is very hard to detect in parallel coordinates.

We then introduced Hifocon, a multidimensional visualization system we designed for better coherence and correlation. We incorporated dimension-reduction techniques like PCA and MDS to place objects on scatterplot displays, and used fish-eye magnification to show focus objects in detail with star-glyphs. We also use an arc to link two points representing the same object in two different scatterplots. This allows the relationship between four dimensions to be observed. Arcs are drawn for all points within a focus area specified by the user, while other points are shown as context. Coherence and correlation for both objects and dimensions are improved with arcs. With arcs, many interesting observations have been made. For example, we have shown an example of hierarchical clusters and an example of a cluster which becomes outliers in another dimension. Arcs are well-suited to discover such relationships because arcs link multiple dimensions together.

The ability to plot a new scatterplot for a subset of the data is also provided in Hifocon. This is important because axes can be custom-chosen to best reveal patterns, clusters and outliers in the subset. Arcs can be drawn back to the parent scatterplot for better object coherence and dimension correlation, so that the context shown by the parent scatterplot is more intuitive.

References

1. K. Alsabti, S. Ranka, and V. Singh. Clouds: A decision tree classifier for large datasets. In *Proc. 4th Intl. Conf. on Knowledge Discovery and Data Mining (KDD '98)*, pages 2–8, 1998.
2. M. Ankerst, M. Ester, and H.-P. Kriegel. Towards an effective cooperation of the user and the computer for classification. In *Proc. 6th Intl. Conf. on Knowledge Discovery and Data Mining (KDD '00)*, 2000.
3. D. Asimov. The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143, January 1985.
4. G. Furnas. Generalized fisheye views. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'86)*, pages 16–23, 1986.
5. P.J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.
6. A. Inselberg. The plane with parallel coordinates. *Special Issue on Computational Geometry: The Visual Computer*, 1:69–91, 1985.
7. I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.
8. E. Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proc. ACM SIGKDD '01*, pages 107–116, 2001.
9. D.A. Keim and H.-P. Kriegel. Visdb: Database exploration using multidimensional visualization. *IEEE Computer Graphics and Applications*, 14(5):40–49, 1994.
10. B. Kerr. Thread arcs: An email thread visualization. In *Proc. IEEE Symposium on Information Visualization*, 2003.
11. J. Lamping, R. Rao, and P. Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems (CHI'95)*, pages 401–408, 1995.
12. J. LeBlanc, M.O. Ward, and N. Wittels. Exploring n-dimensional databases. In *Proc. IEEE Visualization '90*, 1990.
13. D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.
14. R. Parekh, J. Yang, and V. Honavar. Constructive neural-network learning algorithms for pattern classification. *IEEE Trans. on Neural Networks*, 11(2), 2000.
15. R.M. Pickett and G.G. Grinstein. Iconographics displays for visualizing multidimensional data. In *Proc. IEEE Conference on Systems, Man, and Cybernetics*, pages 514–519, 1998.
16. J.C. Shafer, R. Agrawal, and M. Mehta. Sprint: A scalable parallel classifier for data mining. In *Proc. 22nd Intl. Conf. on Very Large Databases (VLDB '96)*, pages 544–555, 1996.
17. D.F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998.
18. S.T. Teoh and K.-L. Ma. Paintingclass: Interactive construction, visualization and exploration of decision trees. In *Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining (KDD '03)*, 2003.
19. W.S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.
20. M.O. Ward. Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proc. IEEE Visualization '94*, pages 326–336, 1994.
21. J. Yang, W. Peng, M.O. Ward, and E.A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. IEEE Symposium on Information Visualization*, 2003.