Spectral-Based Contractible Parallel Coordinates

Koto Nohno*, Hsiang-Yun Wu†, Kazuho Watanabe‡, Shigeo Takahashi† and Issei Fujishiro§
* *Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan*
† *Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 133-8565, Japan*
‡ *Dept. of Computer Science and Engineering, Toyohashi University of Technology, Aichi 441-8580, Japan*
§ *Dept. of Information and Computer Science, Keio University, Kanagawa 223-8522, Japan*

***Abstract*—Parallel coordinates is well-known as a popular tool for visualizing the underlying relationships among variables in high-dimensional datasets. However, this representation still suffers from visual clutter arising from intersections among polyline plots especially when the number of data samples and their associated dimension become high. This paper presents a method of alleviating such visual clutter by contracting multiple axes through the analysis of correlation between every pair of variables. In this method, we first construct a graph by connecting axis nodes with an edge weighted by data correlation between the corresponding pair of dimensions, and then reorder the multiple axes by projecting the nodes onto the primary axis obtained through the spectral graph analysis. This allows us to compose a dendrogram tree by recursively merging a pair of the closest axes one by one. Our visualization platform helps the visual interpretation of such axis contraction by plotting the principal component of each data sample along the composite axis. Smooth animation of the associated axis contraction and expansion has also been implemented to enhance the visual readability of behavior inherent in the given high-dimensional datasets.**

***Keywords*-parallel coordinates; axis contraction; spectral graph theory; dendrograms**

I. Introduction

Parallel coordinates provides us with an effective means of visualizing high-dimensional data by plotting each sample in terms of multiple vertical axes as a polyline. This visualization technique is useful in a sense that we can visually understand the degree of correlation between data samples in terms of two adjacent axes. Nonetheless, this representation still suffers from distracting visual clutter especially when we have to analyze a larger amount of datasets of higher dimensions, because the associated polyline samples intricately overlap with each other within the limited screen space.

Conventional techniques alleviate this technical problem by improving the readability of the parallel coordinates plots themselves, for example, by employing edge bundling or illustrative rendering techniques. This implies that previous studies rather focused on merging polyline samples to enhance the visual representation, and considered little about combining multiple axes in the parallel coordinates. However, it is also true that, if some pair of dimensions is highly correlated with each other, we can easily guess the arrangement of data samples along one dimension by referring to the other dimension. This inspires us to merge a set of multiple vertical axes into a new composite axis to systematically reduce the visual clutter arising from high dimensionality of the data.

This paper introduces a concept of *contractible parallel coordinates*, which enables the contraction of parallel coordinates axes to illuminate the global trends of the given data samples in terms of a reduced set of dimensions. This is accomplished by reordering the parallel axes through the spectral analysis of a weighted graph composed by referring to data correlation among multiple dimensions, and then merging the most correlated pair of axes to contract the overall parallel coordinates plots. The history of such axis contraction is represented by a dendrogram tree together with a reduced set of parallel axes, on which the data samples are plotted by projecting them calculating their principal components. Animated preview of the axis contraction has been also implemented to clearly illustrate how the high-dimensional data plot can be transformed into a simpler one. Figure 1 provides snapshots of the proposed contractible parallel coordinates, where data samples are classified into three groups using k-means clustering.

This paper is structured as follows. Section II provides a brief survey on existing parallel coordinates representations and relevant researches. Section III describes details of the proposed approach for contracting parallel coordinates to reveal hidden global trends of the given datasets. Section IV presents implementation details and the associated experimental results. Section V concludes this paper and refers to possible future extensions.

II. Related Work

Parallel coordinates has been popular for visualizing multivariate data since Inselberg conducted a pioneering work [2]. This representation helps us plot the samples of each variable along a vertically parallel axis and further observe the degree of correlation between the samples along two neighboring axes. As described earlier, even with parallel coordinates plots, we still suffer from visual clutter arising from a large amount of high-dimensional data due to their complicated overlaps within the limited screen space. This visual clutter problem has aggressively been tackled by devising rendering styles of the polyline samples to enhance

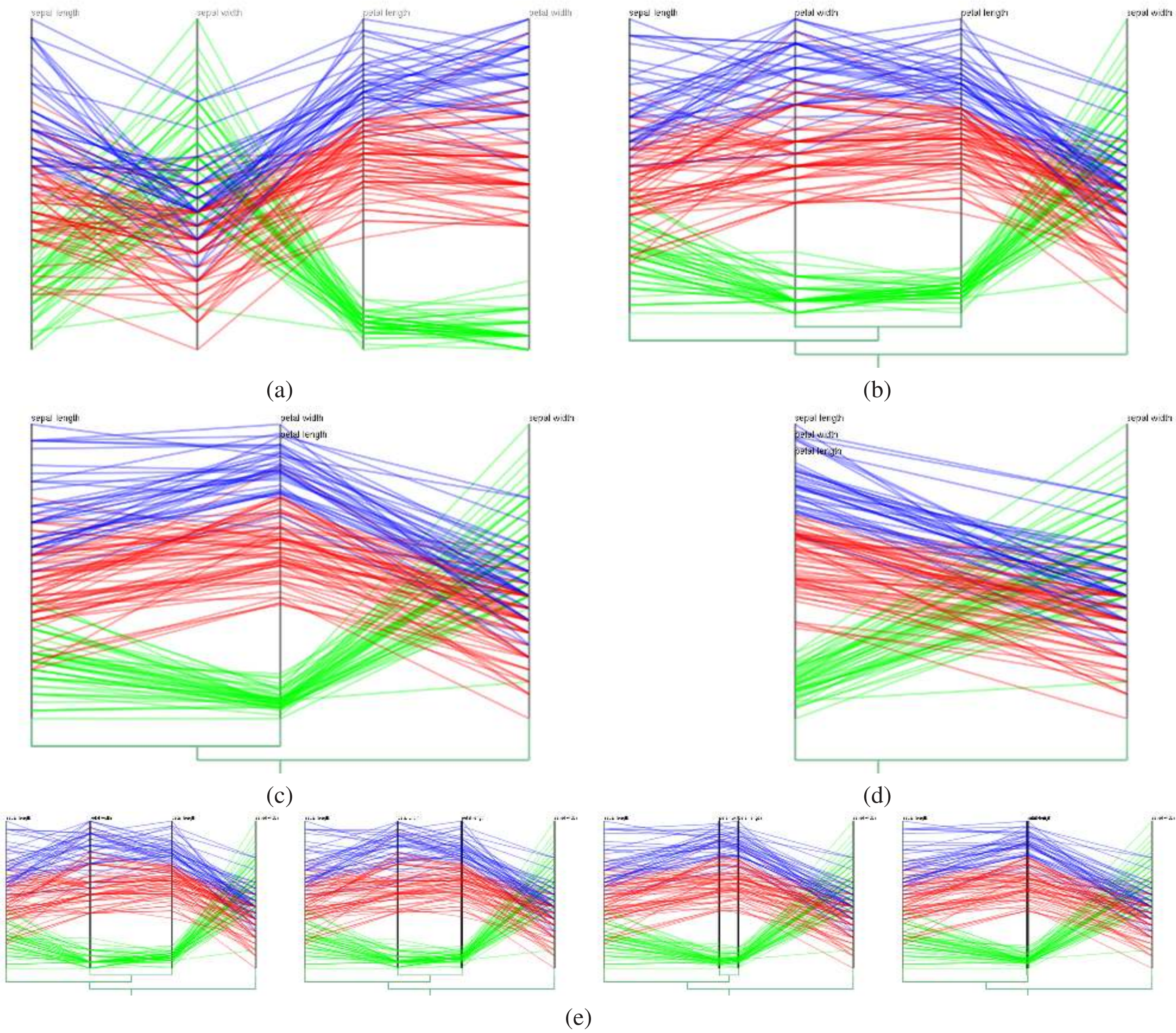
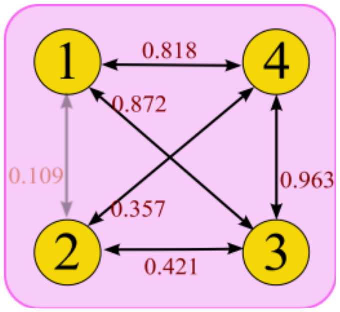


Figure 1. Contractible parallel coordinates. (a) Ordinary parallel coordinates plots of the "Iris" dataset [1]. (b) Vertical axes are reordered by referring to data correlation among dimensions. (c) Middle two axes are contracted to reduce the number of dimensions. (d) Left two axes are further combined. (e) Animation frames during the axis contraction from (b) to (c).

the readability of the parallel coordinates plots. For example, Zhou et al. [3] invented an algorithm for bundling polyline samples of high correlation, and McDonnell and Mueller [4] employed illustrative visualization by applying translucent rendering styles to a set of highly-correlated data samples.

On the other hand, the readability of the data samples is also improved by rearranging the layout of vertical parallel axes by referring to data correlation between every pair of dimensions. Yang et al. [5] proposed an interactive dimension reordering and spacing based on dimension hierarchies in parallel coordinates, which has been followed by improved reordering of parallel axes by Peng et al. [6], where they systematically reduced visual clutter by maximizing data correlation between pairs of adjacent axes.

For more sophisticated reordering of dimensions, a network structure called *pairwise correlation graph* [7] is often constructed where the nodes of the graph correspond to the dimensions while the edges represent some relationships such as data correlation between the corresponding pairs of dimensions. Hurley and Oldford [8] ameliorated parallel coordinates representation by extracting *Eulerian tours* and *Hamiltonian decompositions* of the pairwise correlation graph. Zhang et al. [9] employed the pairwise correlation graph as an interface so that users can travel along a proper order of parallel coordinates axes and further select a specific subset of the axes to conduct more detailed analysis of the data. Heinrich et al. [7] constructed *parallel coordinates matrix* representation by covering the pairwise correlation



(a)

$$
\begin{aligned}
\boldsymbol{L} &= \boldsymbol{D} - \boldsymbol{A} \\
&= \begin{pmatrix} 1.690 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.778 & 0.000 & 0.000 \\ 0.000 & 0.000 & 2.256 & 0.000 \\ 0.000 & 0.000 & 0.000 & 2.138 \end{pmatrix} - \begin{pmatrix} 0.000 & 0.000 & 0.872 & 0.818 \\ 0.000 & 0.000 & 0.421 & 0.357 \\ 0.872 & 0.421 & 0.000 & 0.963 \\ 0.818 & 0.357 & 0.963 & 0.000 \end{pmatrix} \\
&= \begin{pmatrix} 1.690 & 0.000 & -0.872 & -0.818 \\ 0.000 & 0.778 & -0.421 & -0.357 \\ -0.872 & -0.421 & 2.256 & -0.963 \\ -0.818 & -0.357 & -0.963 & 2.138 \end{pmatrix}
\end{aligned}
$$

(b)

$$
\begin{array}{llllll}
\text{Eigenvalues:} & \lambda_1 = 0.000, & \lambda_2 = 0.958, & \lambda_3 = 2.731, & \lambda_4 = 3.174 \\
\text{Eigenvectors:} & \boldsymbol{e}_1 = \begin{pmatrix} 0.500 \\ 0.500 \\ 0.500 \\ 0.500 \end{pmatrix}, & \boldsymbol{e}_2 = \begin{pmatrix} -0.453 \\ 0.847 \\ -0.185 \\ -0.209 \end{pmatrix}, & \boldsymbol{e}_3 = \begin{pmatrix} -0.728 \\ -0.176 \\ 0.331 \\ 0.573 \end{pmatrix}, & \boldsymbol{e}_4 = \begin{pmatrix} 0.119 \\ 0.045 \\ -0.779 \\ 0.615 \end{pmatrix}
\end{array}
$$

(c)

Figure 2. Reordering parallel coordinates axes by referring to the pairwise correlation among dimensions. (a) A graph representing similarity among dimensions. (b) Computation of the Laplacian matrix $\boldsymbol{L}$ of the graph from the diagonal matrix $\boldsymbol{D}$ and adjacency matrix $\boldsymbol{A}$. (c) The eigenvalues and normalized eigenvectors of the Laplacian matrix $\boldsymbol{L}$.

graph with a minimum number of *Hamiltonian paths*. Our approach also employs a variant of the pairwise correlation graph for reordering the parallel coordinates axes while we take advantage of dimensionality reduction of the graph for better grouping correlated axes. Readers can refer to a recent survey [10] for more details on the state of the art in parallel coordinates. Furthermore, our framework can also be considered as a technique for identifying feature subspaces spanned by appropriate combinations of dimensions and data samples in the original data space. Refer to several papers [11], [12] for recent technical advancements on this topic.

III. Contracting Axes in Parallel Coordinates

This section describes an algorithm for reordering vertical axes of the parallel coordinates by referring to data correlation between each pair of dimensions first and then merging two most correlated axes step by step to illuminate the global behavior of the given data samples. Our algorithm consists of four steps as follows:

1) Calculating correlation among dimensions
2) Encoding pairwise correlation as a graph
3) Reordering Axes via spectral graph analysis
4) Axis contraction based on dendrogram trees

In the remainder of this section, we describe these respective steps in the following subsections.

A. Calculating correlation among dimensions

As for the pairwise relationship between a pair of dimensions, we employed the correlation coefficient developed by Pearson [13]. Suppose that $a_i\,(i = 1, 2, \ldots, n)$ and $b_i\,(i = 1, 2, \ldots, n)$ are data values of the a- and b-th dimensions. The correlation coefficient can be given by:

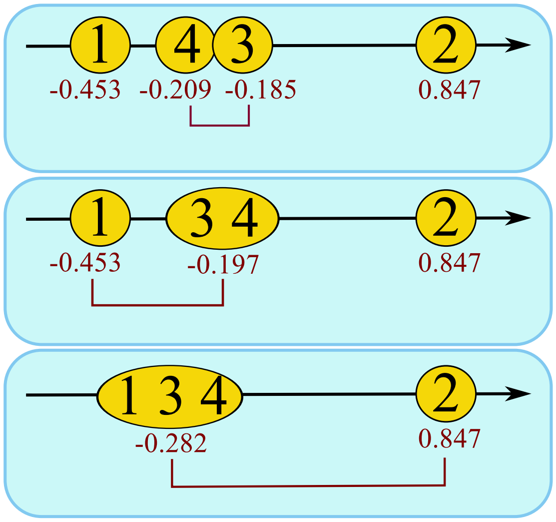
$$\frac{\displaystyle\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\displaystyle\sum_{i=1}^{n}(a_i - \overline{a})^2}\sqrt{\displaystyle\sum_{i=1}^{n}(b_i - \overline{b})^2}} \tag{1}$$

where $\overline{a}$ and $\overline{b}$ are the averages of $\{a_i\}$ and $\{b_i\}$, and thus can be written as $\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$ and $\overline{b} = \frac{1}{n}\sum_{i=1}^{n} b_i$. Note that the correlation coefficient ranges from -1 to 1 according to Eq. (1). Moreover, it approaches to 1 if a_i is likely to be large (small) as b_i increases (decreases), while -1 if a_i is likely to be large (small) as b_i decreases (increases). If there exists little correlation between the two variables, the corresponding correlation coefficient almost vanishes.

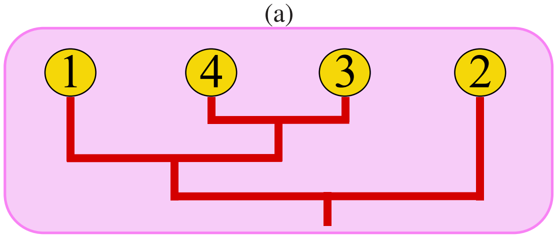
In our approach, we define the similarity between a pair of dimensions as the absolute value of the correlation coefficient. This is because, if the two dimensions have high correlation we can easily guess the behavior of one variable from that of another. On the other hand, we can know little about the relationships between the two variables if the corresponding correlation coefficient is quite small. In practice, we reorder the sequence of dimensional axes by referring to the correlation coefficients first and then merge highly-correlated ones to systematically reduce the number of axes to better illuminate the entire trends of the data.

B. Encoding pairwise correlation as a graph

Our next step is to represent the data correlation between each pair of dimensions as a graph. In practice, we construct



(a)



(b)

Figure 3. Dendrogram-based clustering of dimensions. (a) Step-by-step axis contraction. (b) A dendrogram that represents this history of axis contraction process.

a variant of the pairwise correlation graph introduced in [7], where a node represents some dimension and an edge corresponds to the relationship between a pair of dimensions. In our approach, we use the precomputed similarity values, i.e., the absolute correlation coefficients, as the respective edge weights of the graph.

Nonetheless, the definition of our graph slightly differs from that of the original pairwise correlation graph [7] in that we do not insert an edge to the graph if the corresponding absolute correlation coefficient is less than some threshold. This is because we would like to make the graph as sparse as possible so that we can localize the similarity relationships among dimensions for later use. As for the choice of the threshold, we tested a set of possible threshold values and observed to what extent we can properly discriminate between the similarity between every pair of coordinate axes for each value. In the end, we learned that 0.15 is empirically the maximum threshold that still retains the similarity relationships among coordinate axes. Thus, we set the threshold to 0.15 in our approach. Figure 2(a) shows a graph we constructed from the similarity values among dimensions of the "IRIS" dataset [1] in Figure 1(a).

C. Reordering axes with spectral graph analysis

Now we are ready to reorder the parallel coordinates axis by referring to the pairwise similarity among dimensions. For this purpose, we employ the spectral graph analysis [14], which is known as a tool for decomposing the graph into clusters by cutting edges of minimal weights [15], and projecting the graph onto low-dimensional space while keeping similar nodes maximally close to each other [16]. We use this tool to find an optimal ordering of the vertical axes in the sense that we can appropriately merge a set of dimensions of high similarity.

The spectral analysis begins with calculating the Laplacian matrix $\boldsymbol{L}$ from the node connectivity of the graph. As shown in Figure 2(b), the Laplacian matrix $\boldsymbol{L}$ can be defined as the difference from the diagonal matrix $\boldsymbol{D}$ and the adjacency matrix $\boldsymbol{A}$. Here, the adjacency matrix $\boldsymbol{A}$ can be composed in such a way that the (a, b)-th entry corresponds to the similarity value between the a- and b-th dimensions, as shown in Figure 2(b). The diagonal matrix $\boldsymbol{D}$ is constructed from $\boldsymbol{A}$ so that the a-th diagonal value is equal to the sum of the entries in the a-th row (or column) of $\boldsymbol{A}$. These two matrices allows us to obtain the Laplacian matrix $\boldsymbol{L}$ by calculating $\boldsymbol{D} - \boldsymbol{A}$ as shown in Figure 2(b).

The parallel coordinates axes are finally reordered by computing the eigenvalues and normalized eigenvectors of the Laplacian matrix $\boldsymbol{L}$. Note that the smallest eigenvalue λ_1 always becomes zero since the Laplacian matrix L is not full rank according to the definition. Thus, we refer to the entries of the eigenvector $\boldsymbol{e}_2$, which correspond to the second smallest eigenvalue λ_2. As shown in Figure 2(c), for example, we can project each dimension onto the number line by referring to the entries of $\boldsymbol{e}_2$ to seek an optimal ordering of the parallel coordinates axes, as illustrated at the top of Figure 3(a). In this way, the spectral analysis of the similarity weighted graph permits us to find an appropriate ordering of the axes in the sense that we can naturally merge a pair of axes that are similar to each other.

D. Axis contraction based on dendrogram trees

Our last task is to reduce the number of dimensions by merging axes that are adjacent to each other in the sequence we have arranged so far. In our approach, this has been accomplished by successively merging a pair of adjacent axes that are the most similar to each other. This consecutive combination of two dimensions results in dendrogram-based clustering of parallel coordinates axes.

Suppose that all the dimensions are aligned along the number line by referring to their corresponding projected coordinates as depicted in Figure 3(a), for example. First of all, we can merge Dimensions 3 and 4 as illustrated at the top of Figure 3(a), since the difference in the coordinate is the smallest among all pairs of adjacent dimensions. When we merge the two dimensions by contracting the corresponding vertical axes, we update the coordinate of the combined dimension along the number line as the average of the two coordinates of the original dimensions. In this case, the average coordinate is set to be $((-0.209)+(-0.185))/2 = -0.197$. We again try to merge the axes of Dimensions 1 and 3&4 to compose a new axis since these two axes are the closest among all the pairs of adjacent dimensions, as shown in the middle of Figure 3(a). This time, the coordinate of the merged axis is computed as a weighted average of the previous two axes since Dimension 3&4 originally consists of the two dimensions. This implies that the new coordinate of the composite axis becomes $((-0.453) \times 1 + (-0.197) \times 2)/3 = -0.282$, which is equivalent to the average coordinate of Dimensions 1, 3, and 4. This axis contraction process is repeated until all the dimensions are merged into one, as revealed at the bottom of Figure 3(a).

This step-by-step combination of parallel coordinates axes can be represented as a dendrogram tree of all the dimensions, as shown in Figure 3(b). Actually, in our visualization system, the history of dimension combinations is visualized as a dendrogram tree, which is attached to the bottom of the contracted parallel coordinates as shown in Figure 1. This indeed helps us visually identify how we have merged the parallel coordinates axes so far during the analysis of the given high-dimensional data.

Note that, after having combined multiple dimensions, we still have to plot the data samples along the composite axis so that we can visually inspect the distribution of the data samples in that composite space. For this purpose, we project the data samples of the associated set of combined dimensions to the primary axis through principal component analysis (PCA), and plot the samples along the composite axis by referring to their principal components. This successfully transforms the overall spatial distribution of the data samples within the set of merged dimensions into the contracted parallel coordinates representation.

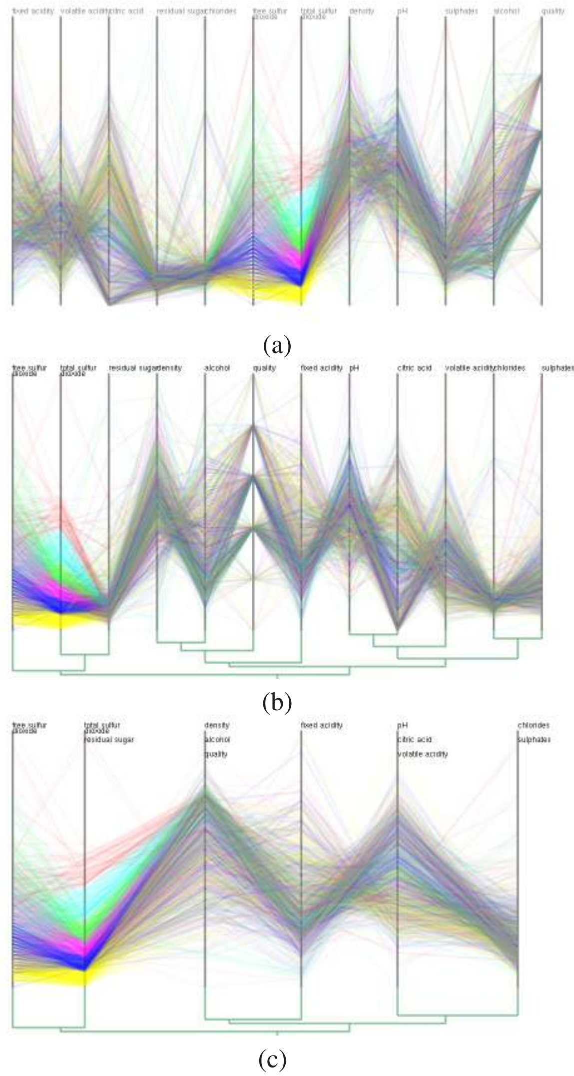


Figure 4. Visualizing the "Wine Quality" dataset [1] using the contractible parallel coordinates. (a) Ordinary parallel coordinates plots. (b) Axis reordering still cannot fully reduce visual clutter. (c) Axis contraction illuminates the global trends of the data.

IV. Implementation Details and Results

Our prototype system was implemented on a Laptop PC with an Intel Core i7 (2GHz) and 8GB RAM, and the source code has been written in C++ using OpenGL and GNU Scientific Library for matrix computation. We equipped our system with an interface for contracting and expanding parallel coordinates axes, where the associated transformation will be animated with ease-in and ease-out effects to provide more temporally coherent previews of the parallel coordinates plots.

Figure 1 shows a simple example where the "IRIS" dataset [1] has been visualized using our contractible parallel coordinates. Our approach transforms the original parallel coordinates plots in Figure 1(a) into those in Figure 1(b) by reordering the axes through the spectral analysis of a similarity-weighted graph among the dimension nodes. The number of dimensions is reduced to 3 by contracting the middle two axes as shown in Figure 1(c), and then 2 by merging left two axes as shown in Figure 1(d). Our system can also animate the axis contraction process in order to provide users with temporally coherent previews of the dimensionality reduction process, as demonstrated in Figure 1(e). Another dataset "Wine Quality" [1] has been also visualized as shown in Figure 4. In this case, the visual quality of the parallel coordinates plots of the original 12D dataset is degraded due to the visual clutter problem (Figure 4(a)). However, our contractible parallel coordinates representation significantly alleviates this problem by reordering the axes (Figure 4(b)) first and then reducing the number of dimensions to 6 while keeping the overall trends of the data (Figure 4(c)).

Moreover, we also conducted a small user study to investigate how the correlations among coordinate axes are better visualized using our approach, where we collected 8 participants (5 females and 3 males). We asked the participants to see how we can contract the parallel coordinates plots by combining and decomposing the vertical axes first, and then answer the underlying principle for such contractible representation of the multivariate data samples. In this first case study, 4 participants correctly pointed out that our approach tries to merge coordinates axes that are highly correlated with each other while leaving other uncorrelated axes untouched. In the second case study, we formally explained our mechanism for contracting parallel coordinates axes to the participants and asked them to interact with our visualization system in practice. After this stage, all the participants excluding one admitted the fact that our approach tries to merge pairs of coordinate axes in the order of correlation.

These experimental results demonstrated the capability of the proposed approach for elucidating important features inherent in the high-dimensional data.

V. Concluding Remarks

This paper has presented an approach to contracting the parallel coordinates representation to alleviate visual clutter problems arising from a large amount of high-dimensional datasets. For this purpose, we compose a graph that represents pairwise correlations among dimensions first, and then rearrange the parallel coordinates axes through the spectral analysis of that graph. Dendrogram-based clustering of dimensions are performed to successively contract the parallel coordinates axes to better visualize the global trends of the given high-dimensional data.

Our future extension includes the sophistication of the interface for interactively identifying meaningful subspaces in a set of high-dimensional data samples. Automated clustering of both data samples and dimensions will definitely augment the capability of our data analysis framework and remains to be tackled.

Acknowledgments

This work has been partially supported by MEXT KAKENHI under Grants-in-Aid for Scientific Research on Innovative Areas No. 25120014 and JSPS KAKENHI under Grants-in-Aid for Scientific Research (B) No. 24300033.

References

[1] K. Bache and M. Lichman, "UCI machine learning repository," 2014. [Online]. Available: http://archive.ics.uci.edu/ml

[2] A. Inselberg, "The plane with parallel coordinates," *The Visual Computer*, vol. 1, no. 2, pp. 69–91, 1985.

[3] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen, "Visual clustering in parallel coordinates," *Computer Graphics Forum*, vol. 27, no. 3, pp. 1047–1054, 2008.

[4] M. T. McDonnell and K. Mueller, "Illustrative parallel coordinates," *Computer Graphics Forum*, vol. 27, no. 3a, pp. 1031–1038, 2008.

[5] 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 *Proceedings of the 9th Annual IEEE Conference on Information Visualization (InfoVis 2003)*, 2003, pp. 105–112.

[6] W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," in *Proceedings of the 10th IEEE Conference on Information Visualization (InfoVis 2004)*, 2004, pp. 89–96.

[7] J. Heinrich, J. Stasko, and D. Weiskopf, "The parallel coordinates matrix," in *Proceedings of EuroVis 2012 Short Papers*, 2012, pp. 37–41.

[8] C. B. Hurley and R. W. Oldford, "Pairwise display of high-dimensional information via eulerian tours and hamiltonian decompositions," *Journal of Computational and Graphical Statistics*, vol. 19, no. 4, pp. 861–886, 2010.

[9] Z. Zhang, K. T. McDonnell, and K. Mueller, "A network-based interface for the exploration of high-dimensional data spaces," in *Proceedings of the 5th Pacific Visualization Symposium (PacificVis 2012)*, 2012, pp. 17–24.

[10] J. Heinrich and D. Weiskopf, "State of the art of parallel coordinates," in *Eurographics 2013 - State of the Art Reports*, 2013, pp. 95–116.

[11] A. Tatu, F. Maaß, I. Farber, E. Bertini, T. Schreck, T. Seidl, and D. Keim, "Subspace search and visualization to make sense of alternative clusterings in high-dimensional data," in *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012, pp. 63–72.

[12] X. Yuan, D. Ren, Z. Wang, and C. Guo, "Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2625–2633, 2013.

[13] K. Pearson, "Contributions to the mathematical thoery of evalution. II. Skew variation in homogeneous material," *Philosophical Transactions of the Royal Society of London*, vol. 186, pp. 343–414, 1895.

[14] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Maching Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[16] Y. Koren, "On spectral graph drawing," in *Computing and Combinatorics (Proceedings of the 9th Annual International Conference COCOON 2003*, ser. Lecture Notes in Computer Science, vol. 2697. Springer-Verlag, 2003, pp. 496–508.