Advanced Multidimensional Data Visualization via Point Placement and Dimension Reduction

Rosane Minghim
Instituto de Ciências Matemáticas e de Computação - Universidade de São Paulo - Brazil

Abstract

In this text we present an extract of the latest research results of the **VICG - Visualization, Imaging and Computer Graphics** group, working at Instituto de Ciências Matemáticas e de Computação (ICMC), University of São Paulo (USP), Brazil, related to the development of new visualization techniques and applications.

1 Introduction

The VICG research group counts with roughly 20 years of experience in visualization. In particular, the last 5 years have seen considerably important for the groupt's research in multidimensional data visualization via mappings from multidimensional spaces to 2D or 3D visual spaces, as well as in applications of these approaches to data analysis. Here we present a sample of our recent work and refer to `http:\\vicg.icmc.usp.br` for additional information as well as software and data made available. The VICG group is composed by 30 members approximately, including faculty and research students. The faculty involved in the work are Maria Cristina Ferreira de Oliveira, Rosane Minghim, João do Espírito Santo Batista Neto, Luis Gustavo Nonato, Fernando Vieira Paulovich, and Moacir Ponti Jr., some of which especialize in imaging techniques. The work reported also counts with collaboration of other members of ICMC, of other Institutes in our University and other Universities in Brazil and abroad. Due to restriction of space we focus this article on multidimensional projection techniques, visual analysis based on similarity trees, and applications.

2 Multidimensional Visualization via Multidimensional Projections

A Multidimensional Projection (MP) is a type of Multidimensional Scaling and an alternative to conventional Dimension Reduction techniques to map multidimensional data into visual spaces with two or three dimensions. They focus on reflecting a perspective of the multidimensional space that favors grouping and segregation of highly related data. The work by Paulovich et al. [8], one of our first results in the field, presents a successful projection technique called *Least Square Projection* (**LSP**), based on the application of a least squares mesh reconstruction technique to the problem of reflecting the configuration of multidimensional spaces. LSP applies a Laplacian operator in which neighbor data in the multidimensional space are projected to close positions on the visual layout. LSP performs two main steps: first, a subset os points called *control points* is chosen by a clustering technique and projected using a conventional high precision projection technique. Second, a linear system is constructed, based on the neighborhood relationship of the points in the original space, and the cartesian coordinates of the control points, in the reduced space. The solution of this linear system will determine the position of each instance on the projection space. The method is $O(n\sqrt{n})$ and presents a good compromise between precision and time, with graceful handling of very high dimensional data.

With the goal of expanding the underlying ideas and treating scalability issues as well as improving both local control and user control of the outcome, a family of new projection evolved from LSP. One of the main evolutions of this class of technique is the *Part-Linear Multidimen-*

sional Projection (**PLMP**) [6], which handles large collections of data with features defined on a Cartesian multidimensional space. Its computational time is lower than previous techniques at least by one order of magnitude. Another version of MP is the *Piecewise Laplacian-based Projection* (**PLP**) [5], that produces a layout capable of adaptation to user's input and reconfiguration from an initially sampled layout, obtaining a higher degree of interactivity over previous MPs. A further level of evolution was achieved by the *Local Affine Multidimensional Projection* (**LAMP**) method [2], which relies on a mathematical formulation derived from orthogonal mapping theory. This formulation admits a reduced subset of sample instances as input, and prevents transformation effects, a feature that preserves original distances as much as possible as well as maintaining the good features of previous MPs.

Additionally, we have extended LSP to perform projections in 3D, and, within that work, devised a strategy based on clustering and enclosing surfaces to interact with three-dimensional visual spaces [11].

3 Multidimensional Visualization via Similarity Trees

Some of the limitations of projections are intrinsic, such as reflecting similarity at global and local levels with same precision. As a partner technique to projections, we have adapted the ideas of reconstruction of phylogenies to the case of multidimensional data, giving rise to the similarity trees, such as the **NJ-tree** [1]. By positioning similar objects on close branches, this relationship is organized into levels, a natural way of interpreting degrees of similarity and allowing grouping and sub-grouping of similar objects. The first version of the tree was based the original Neighbor Joining [12], with a complexity too high to handle reasonably large data sets and the creation of too many internal nodes. The technique continued to evolve with the employment of faster NJ algorithms, and the *Promoting Neighbor Joining* (**PNJ**) tree [3] also employs a deterministic ordered graph rewriting operation to the tree in order to reduce the number of internal nodes. PNJ improves NJ both in visual and time scalability.

4 Applications

The overall goal of VICGt's research is to provide techniques to improve analysis of data sets that are large, multidimensional, and complex. Frequently, all those characteristics are present in the same application. On top of data nature there is also the complexity of the tasks involved. In the following text we exemplify but a few of the real life applications we have dealt with lately.

Visual Classification of Multidimensional Data Dimension reduction is an important task involved in data analysis in general, and in particular for data sets that reach a high number of dimensions or attributes. In this sense, we proposed a novel mapping process for visual analysis based on supervised dimension reduction [4]. It employs Partial Least Squares (PLS) [13]. The resulting process allows users to input their knowledge into the dimension reduction stage by means of a training set, and, from that, to be able to create a reusable model for improved dimension reduction and visualization of larger data sets. A similar procedure can be used to evolve a classification process from a sample set to a fully classified data set. Although PLS is a supervised approach, we also devised a strategy for PLS reduction in cases where no labeling of any part of the data set is available, based on the application of a clustering technique to produce labels and compose the training set. The performance of visual mappings after application of PLS was improved and our work illustrates its use for text and image collections, comparing favorably against other well known dimension reduction techniques, such as PCA, PivotMDS, ISOMAP and LLE, with advantage in terms of computational time and separability of classes. This work is a sequence of our work on visual data mining illustrated in previous publications such as [3].

Text, Imaging, Brain and Sensor Data Many have been the applications of interest for the techniques we work on. Previously mentioned work illustrated their use for text, imaging, music, time series, and scalar volume data. Additional applications developed by the group included visualization of web search results, time varying textual information, and water inflow in hydropower plants. Particularly interesting applications developed recently are the exploration of text collections with referenced mappings by Word clouds created on top of projections, with semantical ordering of words in documents [9], and the application of 2D and 3D projection techniques for analysis of brain fiber data sets [10]. A very successful cooperation of our group with the Physics department at USP - São Carlos generated a series of contributions of these techniques to the analysis of sensor data (see [7] for example).

Acknowledgements

The work of the VICG group has been sponsored by Brazilian research support agencies such as FAPESP, CNPq and CAPES, as well as the University of São Paulo. Collaborative work has also been supported by DAAD (Germany), NUFFIC (The Netherlands), and Fulbright (USA). We acknowledge the support of Jose Gustavo S. Paiva and Paulo Joia in writing pieces of this article.

References

[1] A. M. Cuadros, F. V. Paulovich, R. Minghim, and G. P. Telles. Point placement by phylogenetic trees and its application to visual analysis of document collections. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology - IEEE VAST*, pages 99–106. IEEE CS Press, October 2007.

[2] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local Affine Multidimensional Projection. *IEEE Transactions on Visualization and Computer Graphics*, 17:2563–2571, 2011.

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

[4] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim. Semi-Supervised Dimensionality Reduction based on Partial Least Squares for Visual Analysis of High Dimensional Data. *Computer Graphics Forum*, 31(3):1345–1354, 2012.

[5] F. Paulovich, D. Eler, J. Poco, C. Botha, R. Minghim, and L. Nonato. Piece wise laplacian-based projection for interactive data exploration and organization. *Computer Graphics Forum*, 30(3):1091–1100, 2011.

[6] F. Paulovich, C. Silva, and L. Nonato. Two-phase mapping for projecting massive data sets. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1281 –1290, nov.-dec. 2010.

[7] F. V. Paulovich, M. L. Moraes, R. M. Maki, M. Ferreira, O. N. Oliveira Jr., and M. C. F. de Oliveira. Information visualization techniques for sensing and biosensing. *Analyst*, 136:1344–1350, 2011.

[8] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least square projection: a fast high precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575, 2008.

[9] F. V. Paulovich, F. M. B. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato. Semantic Wordification of Document Collections. *Computer Graphics Forum*, 31(3):1145–1153, 2012.

[10] J. Poco, D. M. Eler, F. Paulovich, and R. Minghim. Employing 2D Projections for Fast Visual Exploration of Large Fiber Tracking Data. *Computer Graphics Forum*, 31(3):1075–1084, 2012.

[11] J. Poco, R. Etemadpour, F. Paulovich, T. Long, P. Rosenthal, M. Oliveira, L. Linsen, and R. Minghim. A framework for exploring multidimensional data with 3d projections. *Computer Graphics Forum*, 30(3):1111–1120, 2011.

[12] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[13] H. Wold. Partial Least Squares. In *Encyclopedia of Statistical Sciences*, volume 6, pages 581–591. Wiley, New York, NY, USA, 1985.