

Interaction Design in Multidimensional Visualization

*Techniques for multidimensional data visualization,
exploration and visual analytics*

Tze-Haw Huang

A thesis submitted for the degree of
Doctor of Philosophy in Computing Sciences
at the
University of Technology, Sydney

Faculty of Information Technology
University of Technology, Sydney
Sydney, Australia

2015

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

Date	June, 2015
Author	Tze-Haw Huang
Title	Interaction Design in Multidimensional Visualization
Degree	Doctor of Philosophy in Computing Sciences

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate:

Date:

Acknowledgement

During the pursuit of higher education, I have taken on a number of roles, these being a part time Ph.D. candidate, a full time software engineer, a father and a husband. Without the support of many people it would have been very difficult for me to complete this dissertation.

First of all, a special thanks to my supervisor Associate Professor Dr. Mao Lin Huang, who contributed invaluable suggestions and knowledge to my work. I would never have been able to finish my Ph.D. without his exemplary and much appreciated guidance.

Secondly, I want to give my utmost thanks to my wife Yi Lau for her understanding, patience, care and continuous support throughout my PhD candidature. My sincere thanks is also extended to my parents for their consistent and very welcome encouragement over the duration of my doctoral studies.

Finally, I want to express my thanks to all of the people that I encountered who provided direct and indirect support while I worked on my dissertation.

Tze-Haw Huang

Sydney, September, 2014

Table of Contents

Acknowledgement	ii
Table of Contents	iii
List of Figures	vii
List of Algorithms	xiii
Abstract	xiv
Chapter 1 Introduction	1
1.1 From InfoVis to Visual Analytics	1
1.1.1 Problem Statement	4
1.2 Challenges and Goals	5
1.3 Contributions.....	6
1.4 Outline.....	6
Chapter 2 Background	8
2.1 Terminology	8
2.1.1 Curse of Dimensionality	9
2.2 Multidimensional Visualizations	11
2.2.1 Parallel Coordinates	11
2.2.2 Scatterplot Matrix.....	17
2.2.3 TableLens.....	18
2.2.4 Space Filling Curve.....	19
2.2.5 Star Coordinates	21
2.2.6 TreeMap	21
2.3 Interaction in Multidimensional Visualization	24
2.3.1 Data Retrieval.....	24
2.3.2 Interaction for View Change	26
2.3.3 Interaction for Analytical Reasoning	29
2.3.3.1 Clustering	29
2.3.3.2 Dimensionality Reduction.....	31
2.4 Discussion	34
Chapter 3 A New Framework of Visual Interaction	36

3.1	Introduction.....	36
3.2	3-Layers Framework of Visual Interaction.....	38
3.2.1	Tasks by Dynamic Selection.....	39
3.2.2	Tasks by Dynamic Viewing.....	39
3.2.3	Tasks by Dynamic Scoping.....	40
3.2.4	Discussion.....	40
Chapter 4 Hierarchical Virtual Node.....		41
4.1	Interaction or Selection?.....	41
4.2	Revisiting the Data Selection Models.....	42
4.2.1	Rectangular Selection Model.....	43
4.2.2	Value Range Model.....	48
4.2.3	Point Selection Model.....	52
4.2.4	Discussion.....	54
4.3	Implementing the HVN.....	57
4.3.1	System Overview.....	57
4.3.2	Data Classification.....	59
4.3.3	Non-parametric Partitioning by Hierarchical Clustering.....	59
4.3.4	Mapping Virtual Nodes into Visual Space.....	64
4.3.5	Building a Dendrogram.....	68
4.3.6	Constructing Parallel Coordinates.....	70
4.3.6.1	Polyline.....	70
4.3.6.2	Bezier Curve.....	72
4.3.6.3	Bezier Virtual Nodes.....	76
4.3.7	Overview Presentation by Virtual Nodes Density.....	82
4.4	Performance.....	85
4.5	Discussion.....	86
Chapter 5 Interactive Techniques for Visual Analytics.....		87
5.1	Task by Dynamic Selection.....	87
5.1.1	Interact with Data by the HVN.....	88
5.1.2	Dynamic Brushing via HVN.....	93
5.1.3	Highlighting Detail on Demand.....	97
5.1.4	Discussion.....	98

5.2	Task by Dynamic Viewing	98
5.2.1	Hierarchical Local Drill-Down	98
5.2.2	Hierarchical Global Drill-Down	101
5.2.3	Probability Density Estimation	101
5.2.4	Variable Overview of Big Dataset	106
5.2.4.1	Divide-and-Conquer Model	108
5.2.4.2	Overview by Correlation Matrix	110
5.2.5	Discussion	115
5.3	Task by Dynamic Scoping	115
5.3.1	Dimensionality Reduction by RST	115
5.3.2	Discussion	119
Chapter 6 Technical Evaluations.....		120
6.1	Visual Clutter of Overview	120
6.2	Data Selection	122
6.2.1	Continuous Neighbour Selection	123
6.2.2	Non-Continuous Selection	124
6.3	Drill-Down.....	127
Chapter 7 Case Studies.....		132
7.1	Case Study 1.....	132
7.2	Case Study 2.....	136
7.3	Case Study 3.....	139
Chapter 8 Extended Works.....		144
8.1	Flow based Scatterplot Matrix	144
8.1.1	Interaction by Point-to-Region.....	145
8.2	Space Filling Multidimensional Visualization.....	148
8.2.1	Properties and Definitions.....	149
8.2.2	SFMDVis	150
8.2.3	Color Models.....	153
8.2.3.1	RGB Color Ramping.....	153
8.2.3.2	Single-Hue Ramping	154
8.2.4	Interaction Techniques in SFMDVis	156
8.2.4.1	Zooming	156

8.2.4.2 AND and OR Operator for Data Selection.....	157
8.3 Discussion	159
Chapter 9 Conclusion	160
9.1 Summary	160
9.2 Final Conclusion	161
Appendix A Publications.....	162

List of Figures

Figure 1.1. A botanic visualization of a hard disk, practical or art?	2
Figure 1.2. A simple model of visualization.	3
Figure 1.3 Visual analytics is an integration of interdisciplinary theories.....	4
Figure 2.1. Illustration of visual and data complexities.	10
Figure 2.2. A comparison of the Google search results.	12
Figure 2.3. Mapping a data point to a vertical axis.....	13
Figure 2.4. A parallel coordinates visualization.....	15
Figure 2.5. A parallel coordinates visualization in default variable ordering.....	15
Figure 2.6. The perception of linear correlations between parallel coordinates and scatterplot matrix.....	17
Figure 2.7. A visualization of the scatterplot matrix.....	18
Figure 2.8. The TableLens visualization.....	19
Figure 2.9. Building block of Hilbert curve.....	20
Figure 2.10. Visualizations of Hilbert space filling curve.	21
Figure 2.11. Star coordinates visualization.	21
Figure 2.12. Map of the market.....	22
Figure 2.13. Hierarchical data mapping.....	23
Figure 2.14. Widget based data selection.....	25
Figure 2.15. An example of scatterplot.....	26

Figure 2.16 Direct data selection by a 2D rectangle.	26
Figure 2.17 An illustration of zooming technique.	28
Figure 2.18 Reordering of variables in parallel coordinates.	29
Figure 2.19. A K-means clustering.	30
Figure 2.20. Hierarchical clustering categories.	31
Figure 2.21 A plot of eigenvalues for the Scree test.	33
Figure 3.1 A cognitive model of gaining data insight.	37
Figure 3.2. An example of dynamic selection operation.	39
Figure 3.3. An example of dynamic viewing operation.	40
Figure 3.4. An example of dynamic scoping operation.	40
Figure 4.1 Rectangular selection model.	44
Figure 4.2 Properties of the Liang-Barsky algorithm.	45
Figure 4.3. Rectangular selection model within a crowded visualization.	48
Figure 4.4 The value range model with AND and OR operators in parallel coordinates.	50
Figure 4.5 An application of the value range model in 2D.	51
Figure 4.6. Point selection with a tolerance.	52
Figure 4.7 A web application with a node-link based navigation. The u.	53
Figure 4.8. A tile-based parallel coordinates.	53
Figure 4.9. An application of the indirect point selection in parallel coordinates.	54
Figure 4.10. Illustration of the hierarchical virtual node design.	55

Figure 4.11. System flowchart of the HVN based parallel coordinates.....	58
Figure 4.12 An illustration of the single, complete and average linkages.....	62
Figure 4.13. Logical groups partitioned by the hierarchical clustering.....	63
Figure 4.14. Virtual node depth in the hierarchy.....	64
Figure 4.15. Virtual node layout definitions.....	67
Figure 4.16. Illustration of the virtual node layout.....	68
Figure 4.17 Connection of virtual nodes.....	68
Figure 4.18 Types of virtual node.....	70
Figure 4.19. Severities of the overlapped polylines.....	71
Figure 4.20. A snapshot of polyline primitive in our system.....	71
Figure 4.21. Bezier curve with control points.....	72
Figure 4.22. Evaluation of points in a Bezier curve.....	73
Figure 4.23. Bezier curve segments.....	74
Figure 4.24. Bezier curves with various intervals of t	75
Figure 4.25 Geometric drawing of the selected data.....	76
Figure 4.26. Bezier virtual nodes style drawing.....	77
Figure 4.27 Redefinition of a data matrix.....	78
Figure 4.28 An illustration of geometric mapping of data and virtual node to parallel coordinates.....	79
Figure 4.29. A snapshot of the Bezier virtue nodes drawing in our system.....	79
Figure 4.30. Comparison of the overplot severity between geometric primitives.....	81

Figure 4.31. Overview presentation of the HVN in parallel coordinates.....	82
Figure 4.32. An illustration of ordinary histograms.....	83
Figure 4.33. A parallel coordinates with histograms embedded.....	84
Figure 4.34. Hardware environment for benchmarking.....	85
Figure 5.1. Data query in the HVN.....	89
Figure 5.2. Notations used for query the global data matrix.....	90
Figure 5.3. Direct data selection via a virtual node.....	92
Figure 5.4. Brushing task via the HVN.....	94
Figure 5.5. Alpha blending for uncovering a major pattern.....	95
Figure 5.6. Comparison of alpha blending with various alpha values.....	96
Figure 5.7. An application of detail on demand.....	97
Figure 5.8. Hierarchical local-drill-down.....	99
Figure 5.9. Remapping maximal and minimal values in local drill-down.....	100
Figure 5.10. Comparison of different bandwidth selection in KDE.....	103
Figure 5.11. Gaussian kernel with various bandwidths.....	106
Figure 5.12. Visualization of NYTS 2009 dataset in parallel coordinates.....	107
Figure 5.13. Circle segments visualization.....	108
Figure 5.14. An interactive divide-and-conquer model.....	109
Figure 5.15. Multivariate correlation matrix view of a car dataset.....	111
Figure 5.16. Correlation matrix view for the NYTS 2009 dataset.....	112

Figure 5.17. Performance of building the multivariate correlation matrix.	113
Figure 5.18. Application of MDS map for variable overview.	115
Figure 5.19. Applications of RST.	118
Figure 6.1. Subband entropy measure of visual clutter.	121
Figure 6.2. Feature congestion measure of visual clutter.	122
Figure 6.3. 2D rectangular data selection in GGobi.	123
Figure 6.4. 2D rectangular data selection in Mondrian.	124
Figure 6.5. Non-continuous data selection in the HVN.	125
Figure 6.6. Coerce data selection in GGobi.	125
Figure 6.7. An application of parallel coordinates visualization in d3.js.	127
Figure 6.8. Evaluation of drill down feature in GGobi.	128
Figure 6.9. Convolutd result of data alignment in Mondrian.	129
Figure 6.10. Evaluation of the drill down feature in Mondrian.	129
Figure 6.11. Evaluation of the local drill down feature in our HVN.	130
Figure 7.1. Local drill-down scenario 1.	133
Figure 7.2. Local drill-down scenario 2.	134
Figure 7.3. Local drill-down scenario 3.	134
Figure 7.4. Local drill-down scenario 4.	135
Figure 7.5. Local drill-down scenario 5.	135
Figure 7.6. Global drill-down scenario 1.	136

Figure 7.7. Global drill-down scenario 2.	137
Figure 7.8. Global drill-down scenario 3.	138
Figure 7.9. Global drill-down scenario 4.	138
Figure 7.10. Initial view of the multivariate correlation matrix.....	140
Figure 7.11. Case study step 1.	141
Figure 7.12. Case study step 2.	142
Figure 7.13. Case study step 3.	143
Figure 8.1 An example of the cross product.	146
Figure 8.2 An example of point-to-region interaction.	147
Figure 8.3 From scatterplot to flow based scatterplot.....	147
Figure 8.4 A visualization of the flow based scatterplot matrix.	148
Figure 8.5. The properties of SFMDVis.	149
Figure 8.6. A visualization of SFMDVis.	153
Figure 8.7. An illustration of using RGB color remapping to denote the value magnitude.	154
Figure 8.8. An example of single-hue progression in the purple color.....	155
Figure 8.9. Single-hue color ramping in blue and green colors.	155
Figure 8.10. Zooming in SFMDVis.	157
Figure 8.11. Interactive AND and OR data selection in SFMDVis.....	159

List of Algorithms

Algorithm 2.1 An implementation of the parallel coordinates visualization.....	14
Algorithm 4.1 An implementation of the Liang-Barsky.....	47
Algorithm 4.2 An implementation of the hierarchical clustering using average linkage.....	63
Algorithm 4.3 An algorithm that computes the depth of a virtual node in the hierarchy.....	65
Algorithm 4.4. Algorithms of mapping a virtual node to the screen coordinate.....	67
Algorithm 4.5. An implementation of drawing a dendrogram.....	69
Algorithm 5.1. An algorithm for hit test.....	90
Algorithm 5.2. Local drill-down algorithm.....	100
Algorithm 5.3. Implementation of KDE.....	106
Algorithm 5.4. An implementation of the multivariate correlation matrix.....	111
Algorithm 8.1. The core algorithm of SFMDVis.....	152

Abstract

Interaction is an overloaded term in information visualization. Basically, every software tool is interactive but mostly through the manipulation of a widget. Broadly speaking, a visualization is just a software application. What makes the interactive component of a visualization really distinctive is how well it supports an arbitrary selection of data directly in the interface in order to facilitate subsequent analytic tasks. This is challenging due to over-plotting and visual clutter in the multidimensional space and such phenomenon is commonly known as the *curse of dimensionality*.

Data selection is a frontier of a visualization and too many multidimensional visualizations claiming to be interactive mostly address the change of view without explicitly specifying the core technique of how to materialize such operations. Perhaps, the interactive component is achieved through the traditional widget.

To overcome the complexity of truly interacting with multidimensional data for effective visual analytics, we first propose an interactive framework for better understanding of the problem domains. Dynamic data selection is materialized by a novel and sophisticated technique called the *Hierarchical Virtual Node* which opens an application to interact with data directly in parallel coordinates that would otherwise have been impossible or difficult to achieve by existing methods. It works well even under the circumstance of the curse of dimensionality and offers several advantages over others. For example, the use case only requires a mouse click to select a set of data item(s). To achieve an efficient visual analytics, a set of analytic tasks are also developed in each layer of the proposed framework.