A Framework for Exploring Multidimensional Data with 3D Projections

J. Poco\textsuperscript{1,2}, R. Etemadpour\textsuperscript{3}, F.V. Paulovich\textsuperscript{1}, T.V. Long\textsuperscript{3,4},
P. Rosenthal\textsuperscript{3,5}, M.C.F. Oliveira\textsuperscript{1}, L. Linsen\textsuperscript{3} and R. Minghim\textsuperscript{1}

1University of São Paulo, São Carlos, Brazil
2 University of Utah, USA
3 Jacobs University, Bremen, Germany
4 University of Transport and Communication Hanoi, Vietnam
5 Chemnitz University of Technology, Germany
Introduction

• MDS for visualization

• Intuitively 3D is better at group separation

• But difficult to interact with

• Framework to create and interact with 3D data projections
Outline

• Motivation

• Projections Algorithms
  – Quantitative Analysis

• 3D Cluster Visualization

• The Framework

• User Studies
  – 3D vs. 2D LSP
  – Surfaces Evaluation

• Conclusions
Motivation

Datasets

- Document Collection
- Image Collection
- Multi Scalar Field

Questions:

- Are there well-defined groups of similar objects?
- How are different groups related?
- What about instances within a group?
- Which data features determine the grouping?
Least Square Projection (LSP)

Points in $\mathbb{R}^m$ → Choose Control Points → Project Control Points → Determine the neighborhood for control point → Solve a Sparse Linear System

\[
\begin{bmatrix}
L \\
C
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix} = \begin{bmatrix}
b
\end{bmatrix}
\]
Extending LSP to 3D

- This system is solved for each dimension in visual space.
- In 3D case, we need to solve 3 linear systems.
LSP 3D: Quantitative Evaluation

Neighborhood preservation - neighborhood after projection

Neighborhood Preservation

Precision

Number Neighbors

-all_2D_cosine  - all_3D_cosine
LSP 3D: Quantitative Evaluation

Neighborhood Hit – neighbors’ classes after projection
3D Cluster Visualization

**Color coded:**

- **✓** Fast and robust.
- **✗** Loss in perception of depth.
3D Cluster Visualization

Convex Hull:

- Fast, robust and simple meshes.
- Effective in convex clusters.
- Misinterpretation in non-convex clusters.
3D Cluster Visualization

Enclosing Surface iso-distance to cluster points:

✓ Works for non-convex clusters.
✓ Visually pleasing rendering.
✗ Adaptive adjustment of the radius of influence when clusters come close to each other.
3D Cluster Visualization

Non-Convex Hull:

- Doesn’t need radius adjustment.
- A distance field is created computing Voronoi diagram in GPU
- Works for non-convex cluster.
3D Cluster Visualization

Enclosing surfaces iso-distance to non-convex Hull:

- Doesn’t need radius adjustment.
- A distance field is created computing Voronoi diagrams in GPU
- Works for non-convex cluster.
The System
Coordination of 2D and 3D views.

- From 3D to 2D: selecting enclosing surface.
- From 2D to 3D: selecting well “resolved” groups that are preserved in both spaces.
- From 2D to 3D: selecting hidden groups
Hierarchy Clustering

- Root
  - (music, audio, proc, signal, int)[50.55]
  - (logic, program, induct)[30.51]
    - (inform, retriev)[55.58]
    - (case-bas, reason, learn)[17.91]
    - (learn, algorithm, comput, queri, statist)[48.70]
    - (logic, program, learn, induct, muggleton)[22.94]
    - (logic, program)[30.77]
  - (inform, retriev)[68.22]
  - (reason, case-bas)[23.74]
  - (case-bas, reason)[25.70] (network, rout, wireless,
Exploring image sets
Number of Instances in Selection: 0

Number of Instances in Selection: 73 (7.3%)

Mar 3, 2011 6:25:54 [EXECUTED]: Identity Coordination – execution time 0.0s
Mar 3, 2011 6:25:54 [EXECUTED]: Points matrix reader – execution time 0.8s
Evaluation user study set up

• 12 participants
• Experience with computers
• Short training session
• 2 datasets
  – Document data (681 objects, 2993 dimensions)
  – Medical imaging data (1000 objects, 150 dimensions)
Quality Measures

- Reliability
- Efficiency
- Learn ability
- User-friendliness

- Correctness
- Response time
- User satisfaction
- User preferences
Set Up

- Set of tasks
- Record time
- Participants confidence using Likert scale
- User preference
Statistical Significance

• Shapiro-Wilk test against normal distribution

• 2 projections
  – Wilcoxon Matched-Pairs Signed-ranks test
  – T-test

• 5 cluster visualizations
  – Friedman's $\chi^2$-test or test for repeated measures
  – ANOVA test
2D vs. 3D Hypotheses

• Data were clustered before projection

• Two hypotheses were formulated
  – 2D projection is optimal, hard to find in 3D.
  – 3D projection has better separation leading to more correct answers.
Tasks

1) Count the clusters
2) Order the clusters by their density
3) List all overlaps of clusters
4) Detect an object within a cluster
5) Find closest cluster to a specific point
6) Repeat Task 5 with a different point
Correctness 2D vs. 3D

The user’s perception on detecting the similarities among clusters.

<table>
<thead>
<tr>
<th>Q3</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01562</td>
<td>0.01953</td>
<td></td>
</tr>
</tbody>
</table>
Response Times 2D vs. 3D

### Table

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q5</th>
<th>Q6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymp. Sig.</td>
<td>0.01855</td>
<td>0.8501</td>
<td>0.5186</td>
<td>0.375</td>
<td>0.4922</td>
</tr>
<tr>
<td>Mean</td>
<td>58.58</td>
<td>22.75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The two-tailed P value equals 0.0189
User Satisfaction

<table>
<thead>
<tr>
<th>Mean Value</th>
<th>User satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Asymp. Sig. 0.007812

3D vs. 2D
User Study Evaluation
Cluster Visualization

- Convex hull
- Enclosing surfaces isodistant to points
- Non-convex hull
- Enclosing surfaces isodistant to non-convex hull
- Color-coded point cloud.
Tasks

1) Count the clusters
2) List cluster overlaps
3) Identify most distant clusters
Set up

- Different techniques coupled with different datasets
- Half of the participants complementary combinations to the other half

Document data, 5 classes
Medical data, 12 classes
Correctness

No significant difference
Correctness

<table>
<thead>
<tr>
<th>Different Techniques</th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friedman's Chi-Square</td>
<td>11.24</td>
<td>6.8</td>
</tr>
<tr>
<td>Asymp. Sig.</td>
<td>0.025</td>
<td>0.149</td>
</tr>
</tbody>
</table>

Document Dataset
Correctness

![Bar chart showing Correctness percentages for different interfaces with Friedman's Chi-Square results]

<table>
<thead>
<tr>
<th></th>
<th>Q2</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friedman's Chi-Square</td>
<td>9</td>
<td>2.5</td>
</tr>
<tr>
<td>Asymp. Sig.</td>
<td>0.063</td>
<td>0.645</td>
</tr>
</tbody>
</table>

Medical Image Dataset
Response Time

Anova Probability = 0.022
Less Accurate
Longer Time
User Satisfaction

Highest rate of satisfaction

No significant difference
Conclusion

- Framework for multidimensional data visualization
- 3D projection
- 3D cluster visualization
- Quantitative evaluation for 2D vs. 3D projection
- Higher precision in 3D
- User study for 2D vs. 3D
- Higher correctness & user satisfaction in 3D
- User study different cluster visualization
Acknowledgements

• FAPESP, CNPq, CAPES
  – Brasil

• DAAD, DFG
  – Germany

• Available at
  http://infoserver.lcad.icmc.usp.br/infovis2/Tools