Eurographics / IEEE Symposium on Visualization 2011 (EuroVis 2011)
H. Hauser, H. Pfister, and J. J. van Wijk
(Guest Editors)

Volume 30 (2011), Number 3

A Framework for Exploring Multidimensional Data with 3D Projections

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

[1]University 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

Abstract

Visualization of high-dimensional data requires a mapping to a visual space. Whenever the goal is to preserve similarity relations a frequent strategy is to use 2D projections, which afford intuitive interactive exploration, e.g., by users locating and selecting groups and gradually drilling down to individual objects. In this paper, we propose a framework for projecting high-dimensional data to 3D visual spaces, based on a generalization of the Least-Square Projection (LSP). We compare projections to 2D and 3D visual spaces both quantitatively and through a user study considering certain exploration tasks. The quantitative analysis confirms that 3D projections outperform 2D projections in terms of precision. The user study indicates that certain tasks can be more reliably and confidently answered with 3D projections. Nonetheless, as 3D projections are displayed on 2D screens, interaction is more difficult. Therefore, we incorporate suitable interaction functionalities into a framework that supports 3D transformations, predefined optimal 2D views, coordinated 2D and 3D views, and hierarchical 3D cluster definition and exploration. For visually encoding data clusters in a 3D setup, we employ color coding of projected data points as well as four types of surface renderings. A second user study evaluates the suitability of these visual encodings. Several examples illustrate the framework's applicability for both visual exploration of multidimensional abstract (non-spatial) data as well as the feature space of multi-variate spatial data.

1. Introduction

Document and image collections, time series, or multiple scalar fields related to a single phenomenon are just a few examples of high-dimensional data. Creating visual representations that provide insight into the global behavior of such data is challenging. Typical questions are: Are there well-defined groups of similar objects? How are different groups related? What about instances within a group? Which data features determine the groupings? Seeking answers to those questions requires intuitive visual representations and effective user interaction. Multidimensional projection techniques offer a unifying framework in this scenario, by mapping data to a low-dimensional visual space suitable for user interaction, i.e., 2D or 3D. While 2D maps afford easy interaction, 3D projections decrease information loss allowing for better group discrimination. However, interacting in 3D in everyday applications is more difficult.

In this paper, we introduce a framework for interactive visual exploration of multidimensional data using 3D projections. A 3D projection obtained by generalizing the *Least Square Projection* technique (LSP) [PNML08] from a 2D to a 3D scheme builds the core of our framework. It is presented in Section 3. We investigate its effectiveness using quantitative measures and a user study. For quantitative evaluation we apply the similarity metrics of neighborhood hits and neighborhood preservations, which confirm the intuition that 3D projections outperform 2D in terms of precision, as discussed in Section 3. The user study compares the suitability of 2D and 3D projections to perform analysis tasks, see Section 4. We found that better separation in 3D projections allowed for a more precise analysis result.

We also address the problem of interacting with 3D projections, when depth perception becomes an issue. We compare standard 3D scatter plots with color-coded clustering results to visualizations that use cluster hulls and embedding

surfaces.We consider several cluster enclosing surfaces: a convex hull surface, a surface isodistant to the cluster points, a non-convex hull computed from that surface, and a surface isodistant to the non-convex hull, see Section 5. Again, we conducted a user study to compare the effectiveness of these alternative strategies, as detailed in Section 6.

Finally, we describe how our framework is implemented into a publicly available system that supports a range of interaction mechanisms, allowing for predefined optimal 2D views, coordinated 2D and 3D views using a brushing-and-linking technique, hierarchical cluster exploration, and cluster modifications, see Section 7. Also in Section 7 we illustrate how this approach is applicable to visual exploration of real-world data, considering both multidimensional abstract (non-spatial) data and the multidimensional feature space of multi-variate spatial data, following the ideas presented by Linsen et al. [LLRR08, RLL08].

Our main contributions can be summarized as follows:

- A framework for interactive visual exploration of multidimensional data with 3D projections, based on generalizing LSP to project into 3D visual spaces.
- A comparative study of 2D vs. 3D projections.
- Deploying multiple surface representations for 3D cluster visualization and user interaction, and comparing them through a user study.
- Implementation of the framework in a dataflow system and illustration of its applicability to visual exploration of abstract and spatial data.

2. Related Work

2.1. Projection-based Information Visualization

Classical Information Visualization relies heavily on visual representations embedded in 2D space. Multiple data dimensions may be anchored on 2D axes layouts, arranged either in parallel [ID90] or radially [Kan00]. *Multidimensional Scaling* (MDS) approaches, also known as multidimensional projections, also derive 2D layouts, but establish no direct association between the original data dimensions and 2D axes. A set of points $S = \{p_1, \ldots, p_n\}$ in $\mathbb{R}^m$ is mapped into a space $\mathbb{R}^d, d \leq m$ based on some criterion, e.g., attempting to preserve the original neighborhood (similarity) relationships. The projected points afford visual representations that reveal groups of similar/dissimilar elements, such as point clouds, 3D surfaces or graphs. Several classical and novel techniques handle diverse high-dimensional data, e.g., Sammon's Mapping [Sam69], FastMap [FL95], the Nearest-Neighbor Projection (NNP) [TMN03] and the Least Square Projection (LSP) [PNML08].

LSP has been conceived to handle large data sources characterized by sparse data distributions in high-dimensional spaces. The projection process comprises two steps. First, an *MDS* method is employed to project a number of 'control points' into the low-dimensional space $\mathbb{R}^d$. Based on these projected points and on neighborhood relationships among the m-dimensional data points, a linear system is built and then solved to obtain the projected coordinates of the remaining points.

The control points comprise a sample of S, carefully chosen to reflect its distribution in $\mathbb{R}^m$. The data points are clustered and the cluster medoids chosen as the control points. The clusters also define a neighborhood relationship, i.e., a list of neighboring points $V_i \subset S$ for each point $p_i \in S$. A point p_i is placed in the convex hull of V_i by generating a final layout based on local relations in $\mathbb{R}^m$. For each cluster, a nearest neighbor search of its medoid defines its k nearest clusters. When the nearest neighbors of p_i are sought, only the point's own cluster and its nearest clusters are examined, an approximation that yields good results at a reasonable cost.

An advantage of LSP (and other multidimensional projections) is that dimensionality of the projected space is just an input parameter. Therefore, mappings to 2D or 3D visual spaces are equally possible. Nonetheless, since the standard display is a 'point cloud', the lack of a more compact geometry tends to impair user orientation and interaction in 3D. Unlike Scientific Visualization, Information Visualization solutions in general seem to favor 2D interaction over 3D. Some exceptions are briefly reviewed in the following.

2.2. Information Visualization in 3D

A few classical Information Visualization techniques have been extended to create data representations embedded in 3D. *Viz3D* [AdO04], for example, extends RadViz by adding a third dimension represented by an axis orthogonal to the radial axes – the added spatial dimension improves group differentiation. In projection pursuit and grand tours the goal is to select and display a sequence of interesting 2D projections of pairs of data attributes. A 3D projection pursuit [Nas95] and a 3D Grand Tour [Yan99] have been introduced that generate cluster-guided 3D data projections and render the resulting visualizations in a CAVE virtual environment. Scatterdice [EDF08] is a system that employs 3D animated transitions for user navigation in scatterplot matrices depicting multidimensional data.

Additionally, several contributions on the problem of clustering high-dimensional data rely on visual representations, either in 2D or 3D, to assist or improve clustering decisions [HKW99]. Likewise, visualizing high-dimensional data sometimes relies on clustering the data prior to visualization [BDY03]. Nonetheless, pre-clustering has many subtleties in this scenario and data objects may not be easily associated with unique clusters.

Multidimensional projections to 3D spaces offer an interesting alternative to creating 3D 'general-purpose' representations of abstract high-dimensional data. They favor the

perception of object similarity both across and within groups and also support visual identification of groups of highly correlated objects and user focusing on groups of interest. Furthermore, intuition is that projections on 3D space are likely to improve user perception of data groups, as compared to 2D. Since every dimension 'lost' in the projection process contributes to mixing unrelated objects, the added visual dimension should improve precision and enhance exploration capabilities.

Sprenger et al. [SBG00] advocate enclosing surfaces as a strategy to break down visual complexity when visualizing multidimensional data. Our approach is, in some aspects, akin to their H-BLOB hierarchical visual clustering. They employ a spring-embedding strategy to place multidimensional objects in a 3D visual space, creating a mass-spring graph representation of the data with spring stiffness reflecting similarity measurements. The H-BLOB algorithm comprises two stages: computing a cluster hierarchy from the graph with an algorithm based on edge collapsing, and then visualizing resulting clusters with nested implicit surfaces. A drawback of their approach is that the clustering based on graphs does not scale well. Also, the derived hierarchical cluster tree is static. The surface renderings employed are similar to our enclosing surfaces isodistant to cluster points, which we compare against other alternatives.

Tory et al. [TKAM06] investigated when 2D views, 3D views, or their combination are most effective. They employed several scenarios, but it is hard to draw conclusions from their results for separability in projected cluster visualizations. Consequently, we decided to perform our own experiments, as described below.

3. 3D Least Square Projection and Quantitative Evaluation

The original LSP has been successfully applied to various types of multidimensional data, including document and image collections [PNML08, EN*09]. Exploration is conducted on a 2D display by locating, selecting and examining visible groups. Despite its effectiveness, larger data sets incur in severe visual clutter, causing different groups to mix and impairing identification of subgroups.

The technique as proposed originally solves a linear system for each coordinate of the projected space. Extension to 3D (or any higher-dimensional space) is trivial as long as one can obtain the 3D coordinates of the control points. Choosing the control points and solving the resulting systems is analogous to the 2D case, only an additional system must be solved for the third coordinate. We modified the original algorithm to generate 3D projections as follows: A k-means clustering is applied over the vector space, the cluster centroids (or medoids) are chosen as control points and the cluster organization is discarded. The 3D positions of the control points are obtained with the Force Scheme [TMN03] extended to handle 3D coordinates. Neighborhoods for each control point are defined with a k-nearest neighbor search. Details relative to each of these steps may be found in the original LSP paper [PNML08], as the 3D implementation handles them similarly.

Projecting in 3D improves both neighborhood preservation and identification of highly related data points when compared to 2D. We consider the case of visualizing text collections, a typical example of abstract high-dimensional data. Visualization usually requires obtaining a vector space representation of the collection, a process that requires stopword elimination, stemming (reducing words to their radicals), and term counting and weighting.

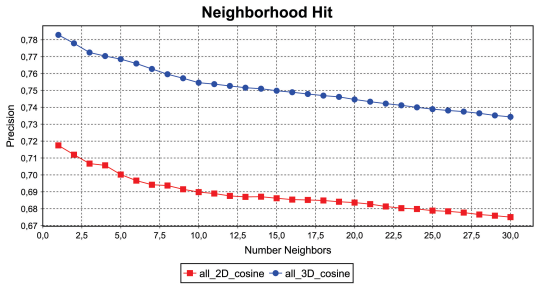
We consider initially a collection of 2,841 scientific papers (preserving title, abstract, authors, and affiliations) in 8 areas of knowledge, each represented by a different number of documents. The areas are: case-based reasoning, inductive logic programming, information retrieval, sonification, bibliographic coupling, co-citation analysis, milgrams, and information visualization. The latter has been extracted from the data made available for the IEEE InfoVis 2004 contest (10 years of InfoVis). The remaining articles were obtained from internet repositories and library searches. After preprocessing, 1,269 terms define the resulting vector space, document similarity is estimated with the cosine measure over the vector representation.

We compared 2D and 3D LSP visual maps of this collection employing two quality metrics known as *neighborhood hit* and *neighborhood preservation* [PNML08, PM08]. Both metrics consider the capability of a projection to preserve the data neighborhoods found in the original space, considering each data point. Neighborhood hit computes the percentage of a point's neighbors that have been human-assigned to its own class and averages the values. The neighborhood preservation metric computes the percentage of a point's neighbors, in the projected space, that belong to the same neighborhood in the original space, and averages the values. Results are shown in Figure 1: the blue curves confirm that precision is superior in the 3D version for any number of neighbors considered.

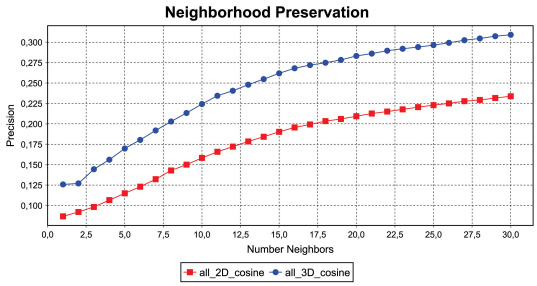
Although, as expected, the 3D projections do improve group separation as compared to 2D, interaction becomes critical, as discussed in Section 5. In the following we describe an empirical user study to evaluate the usability of 3D vs. 2D projections obtained with LSP.

4. User Study Evaluation of 3D LSP

The user study involved 12 participants with different background including visualization experts and non-experts, but all with significant experience in working with computers. They got a brief training about the system and, afterwards, were asked to answer six questions for 2D and 3D LSP projections. All participants answered all the six questions for

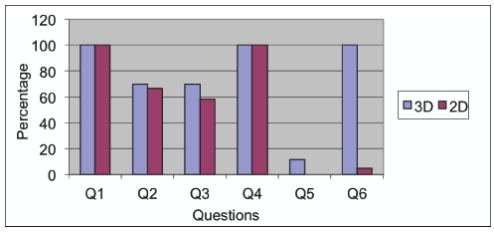


(a) Neighborhood hit metric for 2D and 3D LSP.

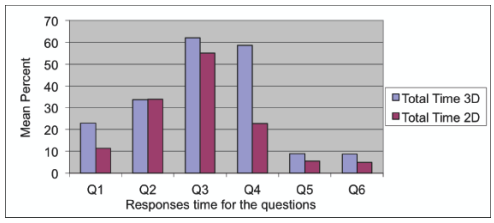


(b) Neighborhood preservation metric for 2D and 3D LSP.

Figure 1: *Increased precision of 3D LSP as compared to 2D LSP, measured by Neighborhood Hit and Neighborhood Preservation, for a corpus of scientific papers.*



(a) Correctness



(b) Response Times

Figure 2: *Fulfilling of the individual tasks with 2D (red) vs. 3D projections (blue).*

both the 2D and 3D system. A document data set with 681 objects and 2,993 dimensions was used. The data set was clustered before being projected and the clusters were color-coded as shown in Figure 6. Two hypotheses were formulated: Hypothesis 1 assumes that the 2D view is optimal and that users would fail to find that optimal view in the interactive 3D system and, thus, would deliver incorrect answers. Perceptional issues of the 3D views when looked at with a 2D monitor may enhance this effect. Hypothesis 2 is based on the fact that the 3D projection allows for better separation and assumes that the users would find proper views to use this advantage and answer the questions more correctly.

The following tasks were asked: 1) count the clusters; 2) order the clusters by their density; 3) list all (pairwise) overlaps of clusters; 4) detect an object within a cluster (labels are shown at mouse-over events); 5) find closest cluster to a specific point (excluding the point's own cluster); and 6) repeat Task 5 with a different point. For each task the time has been recorded by the examiner and the participants had to state their confidence about the conducted tasks using a Likert scale (1 to 5). Moreover, we asked all participants whether they preferred the 2D or the 3D system.

We first computed the correctness of the given answers. For the tasks that required users to list clusters, we counted the percentage of correct answers. Our findings were that the correctness averaged over all participants and all questions was at 74.4% for the 3D system and at 64.3% for the 2D system. The results for the individual tasks are shown in Figure 2a. To analyze the results, we performed statistical tests. Correctness values for Questions 1 and 4 were 100% for both 2D and 3D. For the other questions, the Shapiro-Wilk test was used to check against a normal distribution. As not all of them were normally distributed, we applied the non-parametric Wilcoxon matched-pairs signed-ranks test to check for statistical significance. Only the findings for Questions 5 and 6 were statistically significant (p-values of 0.0156 and 0.0195, respectively). Both questions exhibit a significantly higher correctness for the 3D system.

Then, we computed the average time spent by users to fulfill the tasks. The average time over all participants and tasks were 59.3*s* for the 3D system and 40.7*s* for the 2D system. The times for the individual tasks are shown in Figure 2b. We performed the same statistical tests as for the correctness. Question 4 was the only one with normal distribution such that we applied a paired t-test, for all the others again the Wilcoxon test. Questions 1 and 4 were the ones found to be statistically significant (p-value of 0.0189 for both). Both questions exhibit higher response times for the 3D system.

Evaluating the confidence measures, we compute that the user satisfaction / confidence was somewhat higher when using the 3D system (3.9 as opposed to 3.6). A Wilcoxon matched-pairs signed-ranks test showed that this difference is actually statistically significant (p-value 0.0078). Moreover, 100% of the users answered that they preferred operating with the 3D system. We also checked whether there is a difference between expert and non-expert users, but it was not statistically significant.

Our tests indicate that Hypothesis 2 was confirmed. The interactive 3D system really allowed users to produce better results, which dismisses Hypothesis 1. However, this higher correctness comes at the expense of spending more time. Still, users prefer to use the 3D system.

5. 3D Cluster Visualization

3D interaction is critical in visualization, particularly in displays depicting sparse point glyph representations, as depth perception becomes an issue. A strategy whereby groups of points are enclosed by selectable surfaces may assist refined data exploration. The projected data points are clustered in viewing space and each data cluster is represented by a surface, as illustrated in Figure 3.

We investigated several approaches to generate enclosing surfaces for clusters and discuss their applicability. The simplest solution is to generate the convex hull of the point clusters, for which existing solutions are fast, robust, and produce simple meshes. Figure 3(c) shows the result of applying this approach to the color-coded data point clusters shown in Figure 3(b). Such surfaces are effective to reflect the shape of convex clusters, however, convex hull visualizations of non-convex clusters may lead to misinterpretations.

Alternatively, one may generate enclosing surfaces in which the data points lie in the surface interior. A common approach for non-convex point clusters is the blobs method by Blinn [Bli82], who associates each point with a radial basis function and defines some parameters to control surface "blobbiness". Along this line we have implemented a kernel-based approach to generate enclosing surfaces. We chose the kernel $K(p) = (1 - ||p||^2)^2$ for $||p|| \leq 1$ and $K(p) = 0$ otherwise. Given the set of projected cluster points $\{p_i = (x_i, y_i, z_i)\}$, we compute the function

$$f_h(p) = \sum_{i=1}^{m} K\left(\frac{p - p_i}{h}\right), p \in \mathbb{R}^3$$

with h being the minimum distance of two projected cluster points in a minimal spanning tree. We resample the function over a regular grid and apply marching cubes [LC87] to extract and render an isosurface that assures cluster connectedness. Figure 3(d) shows the result of computing this enclosing surface to the clusters in Figure 3(b).

The previous solution creates nice renderings, but requires an adaptive adjustment of the radius of influence when clusters come close to each other. The third and fourth surface representations chosen to overcome this problem go back to an approach by Rosenthal and Linsen [RL09]. It is based on the GPU computation of 3D discrete Voronoi diagrams. When computing a 3D discrete Voronoi diagram from the points of the target cluster, a discrete distance field is computed that describes the distance to the points of the point cluster. Extracting isosurfaces from the distance field produces an enclosing surface similar to the one in Figure 3(d). Moreover, one can exploit the natural neighborhood property induced by the Voronoi tesselation and construct a hull from those cluster points whose Voronoi cells were intersected by the described enclosing surface. This hull is non-convex for non-convex clusters. Results are visually similar to the well-known α-ball approach [EM94], but our solution is significantly faster and optimal in the sense that it includes the minimal volume. Figure 3(e) shows the result when applied to the clusters in Figure 3(b).

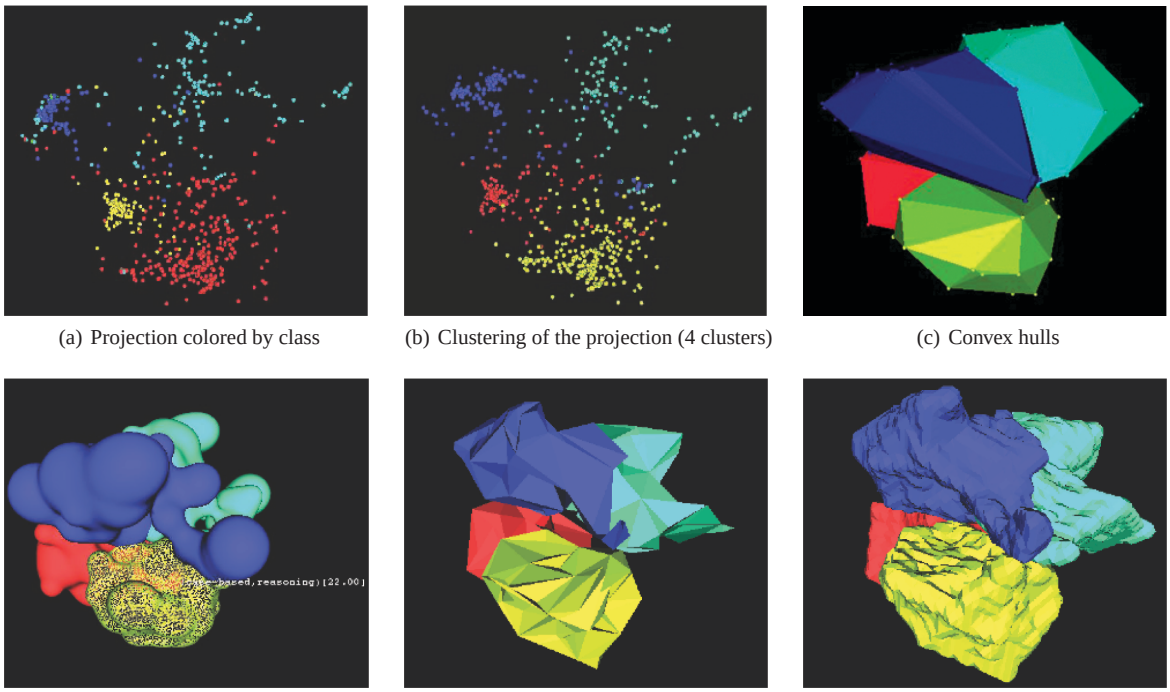
The fourth approach computes an enclosing surface equidistant to the non-convex hull in Figure 3(e), which can be obtained by extending the discrete Voronoi diagram computation from point clusters to a discrete Voronoi diagram computation from polygonal models. Consequently, enclosing Voronoi-based surfaces can be generated with any distance to the hull by extracting isosurfaces from the new distance field. The resulting enclosing surfaces stick close to the point cluster. Figure 3(f) shows the surface obtained from the clusters in Figure 3(b).

6. User Study Evaluation of Cluster Visualization

We conducted another user study with the same format and participants as in Section 4. In addition to the document data set, we used a medical image data set with 540 objects and 28 dimensions. For this experiment, the data sets have been clustered after projection (into five clusters). The first investigation was to compare surface-based cluster visualizations to visualizations with color-coded point clouds. The second investigation was to compare the four surface-based cluster visualizations introduced in Section 5. The questions to the participants were reduced to three, namely: 1) count the clusters; 2) list cluster overlaps; and 3) identify clusters most separate from each other. All participants had to fulfill all three tasks for all five cluster visualization techniques. Different techniques were coupled with the two different data sets leading to two disjoint sets of combinations. Half of the participants used the combinations complementary to those used by the other half.

Figure 4 shows correctness, times, and user confidence for fulfilling the tasks with the different visualization techniques in the order 1) convex hull, 2) enclosing surfaces isodistant to cluster points, 3) non-convex hull, 4) enclosing surfaces isodistant to non-convex hull, and 5) color-coded point cloud. We evaluate the correctness results for each data set individually. As the distribution is non-normal, we use Friedman's χ^2 test. Question 1 was excluded, as all answers for all interfaces were 100% correct. The Wilcoxon matched-pairs signed-ranks test delivers that only Question 2 is statistically significant (p-value of 0.025) and only for the document data set. For Question 2 on the document data set, the enclosing surface isodistant to non-convex hull showed the best results, whereas the convex hull ended up last. The other three approaches ended up in between with comparable numbers.

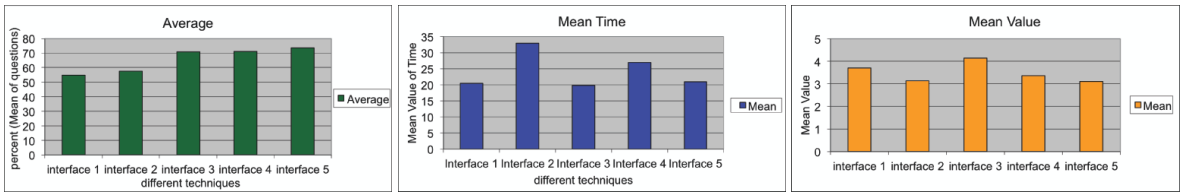
For the evaluation of the response times, the Shapiro-Wilk



(a) Projection colored by class (b) Clustering of the projection (4 clusters) (c) Convex hulls

(d) Enclosing surfaces isodistant to cluster points. (e) Non-convex hulls computed from (d) (f) Enclosing surfaces isodistant to non-convex hulls in (e)

Figure 3: *Projection, clustering, and surface generation in visual space of a 3D LSP on a data set with papers in four areas of knowledge.*



(a) Correctness (b) Times (c) Confidence rates

Figure 4: *Tasks with the four different surface visualizations (Interface 1-4) and color-coded point clouds (Interface 5).*

test delivered a normal distribution: Consequently, we applied the one-way parametric ANOVA test. Again, only results for Question 2 are statistically significant (p-value of 0.022). Here, the two hull approaches had best results, while the two enclosing surfaces ended up last.

Finally, we evaluated the user confidence results. The Shapiro-Wilk test delivered a normal distribution, but the ANOVA test indicated no statistically significant difference among the approaches. However, we also asked the users to rate their preference. 50% of the users answered that they liked the non-convex hull best, followed by point clouds (25%) and enclosing surfaces isodistant to non-convex hull (18%).

The results show that point clouds are still an alternative to surface rendering, as the lower confidence rates they got are not statistically significant. Among the surface representations those with smaller volume were preferred. The convex hull and the enclosing surfaces isodistant to cluster points create too large volumes to keep clusters separated. In terms of efficiency, the hull approaches allowed for lower response times. Overall, the non-convex hull was the one that performed well in all aspects and, thus, may be regarded as the winner of the study.

7. Interactive Exploration Results

The functionalities for multidimensional exploration described in this paper were gathered in a Dataflow type system. Data transformation functionalities may be accessed just by plugging the proper module chosen from a module directory, and interaction capabilities are added to the viewer modules.

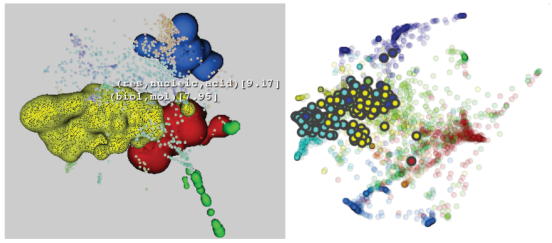
The *3Dproj* system and demonstration video are made available at: http://infoserver.lcad.icmc.usp.br/infovis2/Tools. System features include:

- Selecting from several existing 2D projection techniques to generate 2D views from the original data.
- Performing coordinated interaction of 2D and 3D LSP projection views.
- Projecting from 3D to 2D. Orthogonal projection is available as well as several dimension reduction and multidimensional projection techniques.
- Generating a 2D frame from a current 3D viewing position for further exploration in 2D.
- Coordinating visualization windows so that selection of groups affect all visualizations. This also includes object space views, i.e., spatial data visualizations.
- Providing point- and various surface-based cluster visualizations.
- Combining groups of individual items and groups of clusters to generate larger groups whose elements may be saved for further manipulation.
- Dynamic post-clustering of projected results.
- Hierarchical clustering, including interactive cluster selections to generate subclusters.

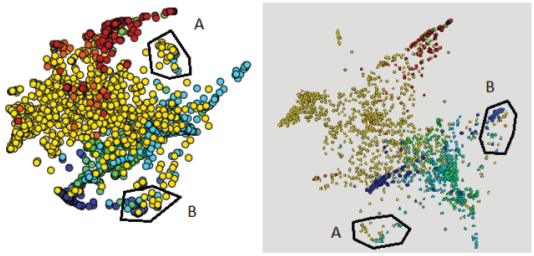
Coordination of 2D and 3D views. The ability to coordinate multiple 2D and 3D projections also encourages novel exploration strategies, as illustrated by the examples depicted in Figure 5. Visualizations depict the corpus of 2,814 scientific papers employed in the quantitative analysis in Section 3.

Figure 5(a) illustrates the coordination from 3D to 2D. The user selects an enclosing surface (selection indicated by the wire-frame view); selected points (displayed with full opacity and bold circumference) are highlighted in the 2D view. Notice that user selection in 3D triggers the display of the main topics addressed by the papers in the group (nucleic acid and molecular biology). Other types of data could trigger other summarization strategies.

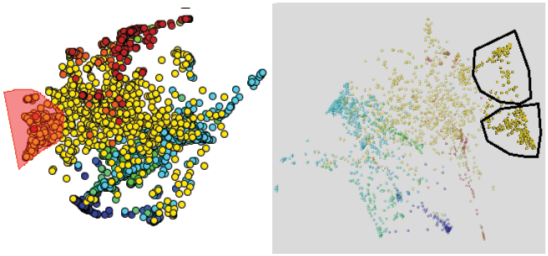
Figure 5 also shows how coordination allows a user to brush the 2D view to highlight groups of possible interest in 3D. Users can locate well 'resolved' groups that are preserved in both spaces and are not strongly affected by the dimensionality reduction to 2D, such as the one in Figure 5(b). They may also identify hidden sub-groups. For instance, Figure 5(c) illustrates that a group of points in the 2D view, which looks also grouped in the original 3D view, actually separates in two sub-groups when the 3D view is rotated. This prompts the user to analyze the group for sub-groups of interest within.



(a) Selecting a document group in 3D and examining the corresponding points in the coordinated 2D view.



(b) Selecting two well formed groups in 2D (A and B) and highlighting corresponding documents in the coordinated 3D view.

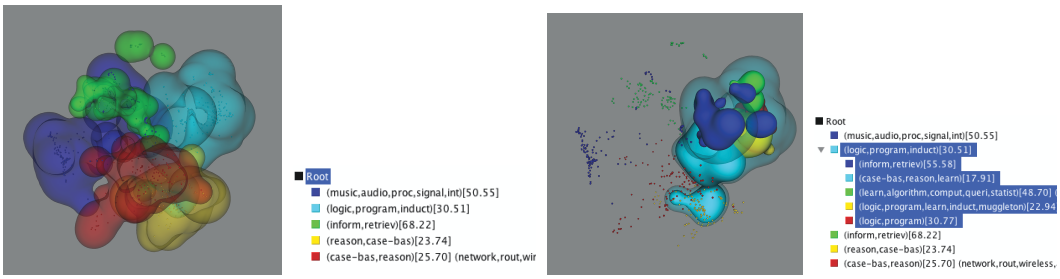


(c) Selecting a group in 2D that separates in 2 groups in 3D.

Figure 5: *Coordinating 2D and 3D views for data exploration.*

Such coordinated multi-space environment supports multiple ways of drilling down in the data. By selecting groups of papers, users may see their topic and content and focus on that in a separate window. In further refinements a group may be split into sub-groups, so that users gradually find their way to the documents that require further examination.

Hierarchical clustering. We have implemented a simple yet useful mechanism to help navigation in the 3D point representation, based on a user-guided clustering process carried out hierarchically. First-level clustering may be seen in Figure 6(a) and its cluster tree is shown in Figure 6(b). The tree both guides the hierarchical clustering process and assists user navigation through the cluster hierarchy. At each level, users may select a cluster and refine it further by applying another clustering process. Figure 6(c) shows the sec-



(a) Clustering after projection. (b) Cluster tree. (c) Subclusters of selected cluster. (d) Hierarchical cluster tree.

Figure 6: *Hierarchical clustering.*

ond level of the tree after a cluster has been refined, and Figure 6(d) shows the updated cluster tree with two levels. At each step the user may choose between bisecting k-means [SKK00, Mac67] or k-medoids [KR90] as clustering method. We refer to the accompanying video for further illustration.

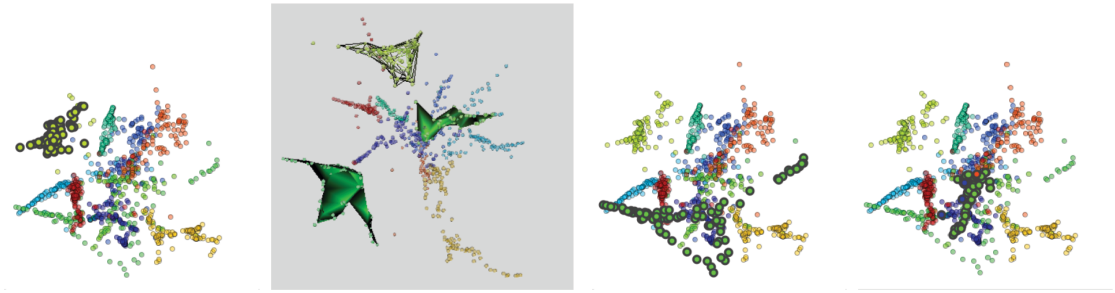
Exploring image sets through their feature spaces. A particularly difficult type of data to explore is that of image collections. While automatic classification solutions are progressing, there is the need for image set exploration so that the user has final control over their interpretation, and judgement of the adequacy of feature sets for their representation. This example aims at illustrating the value of the third dimension in the exploration of this type of highly sensitive, difficult to describe, data sets.

The data set is a segment of 1,000 photos from the Corel photo and picture data set, composed by 10 different classes, with 100 pictures in each class. 2D and 3D visualizations were generated from the same set of 150 sift features [LW03] extracted for each image. Figure 7 illustrates the exploration of the data. It shows that 2D projections (see Figure 7(a)) were capable of separating some groups of photos, while mixing various other groups. We have created the top level 3D cluster tree with 10 clusters, employing the k-medoids technique and Euclidean distance as a measure of dissimilarity. Three of those clusters are shown in Figure 7(b). That view is coordinated with the 2D projection. The first cluster, on top of Figure 7(b) and highlighted in Figure 7(a), grouped well in both 2D and 3D, and corresponds to 86 of the 100 photos in the flower class, with no element from other classes.

From Figures 7(c) and 7(d) it can be easily noticed that the two other well-separated 3D clusters were very mixed with other points in the 2D projection. The bottom cluster contains 97 of the 100 dinosaur drawings without any items from other classes (Figure 7(e)). The middle cluster contains 55 photos, 50 of which belong to the elephant class (Figure 7(f)). This is a very difficult class to separate due to other photos similar in information content, mainly in the classes of horses and landscapes. This example illustrates both the capability and the need for the extra projection dimension, particularly in challenging applications. Figure 7(g) shows the relative precisions of the 2D and 3D projections, by means of neighborhood hit calculations, also reflecting the better separability of 3D. In our system, images corresponding to groups can be loaded under users' selection (by drawing a region in 2D or by a single click on the surface in 3D).

Exploring the feature space of multi-dimensional spatial data. The same interaction framework applies to multi-field scalar data, as illustrated in Figure 8(a), which depicts visualizations of particle data. This is a time-varying spatial data set with a multidimensional feature space [WN08], output by the simulation of the propagation of an ionization front instability. Although the spatial data are sampled on a volumetric grid, data points are unstructured in multidimensional feature space. The data set includes multiple attributes, namely density, temperature, mass abundances of eight chemical species, and velocity. It spans 200 time steps, each with $37 \cdot 10^6$ points, resulting in 1.7 GB of data per time step. For performance purposes, data has been spatially sampled. For an unbiased sampling, random sampling is desirable, resulting in a set of points randomly distributed in the volumetric space, where each point carries multiple properties.

The left frame in Figure 8(a) depicts a 3D Star Coordinates view of the data feature space, as in [LLRR08], for a single simulation time step. Our visualization, on the right side, has been created by pre-clustering followed by projection in 3D using LSP. Our approach conveys the same global shape of the data space, but it stresses local cluster formation and separation. Users may narrow down the examination to particular coherent spatial regions by identifying and delimiting clusters interacting with their enclosing surfaces. This is exemplified in Figure 8(b), left and right frames, which display enclosing density surfaces generated for some 3D clusters, in Star Coordinates and LSP views, respectively. Users interact to define how many clusters and corresponding surfaces they want to extract.



(a) 2D LSP. Highlighted points correspond to top cluster in 7(b).

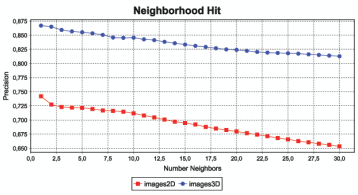
(b) 3D projection, with three enclosed clusters.

(c) 2D LSP. Highlighted points correspond to bottom cluster in 7(b).

(d) 2D LSP. Highlighted points correspond to middle cluster in 7(b).

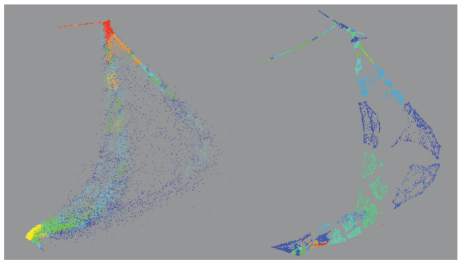
(e) Some images in the bottom cluster.

(f) Some images in the middle cluster.

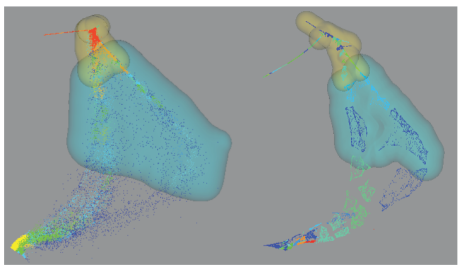


(g) 2D vs. 3D precision.

Figure 7: *Photo features projections. Color is target class.*



(a) Placement of particles with original grouping by property



(b) Surface view of particle groups, after clustering in visual space to obtain 30 groups, and zooming in

Figure 8: *Visualizations of the feature space of the particle simulation data. Star Coordinates views on the left, LSP views on the right.*

8. Conclusions

We discussed how projections in 3D space offer a unifying framework for handling both abstract and spatial high-dimensional data. A similarity-based multidimensional projection technique was adapted to generate projections in 3D space. We illustrated how the added dimension reduces information loss and brings enhanced group differentiation capability to users when compared to 2D displays by means of quantitative metrics, practical usage scenarios, and a user study. However, interacting with 3D projections requires suitable strategies. We proposed interaction facilities that combine information on post-clustering of the projected data with geometrical information derived from the clusters to obtain selectable enclosing surfaces. Several alternative approaches for extracting surfaces that shape up the partition of the 3D space, while retaining interactive rendering rates, were considered and compared. Facilities are integrated into a system that incorporates the projection, clustering and visualization techniques of the proposed framework. As additional interaction facilities 2D-3D coordinated views and hierarchical cluster generation are included, as well as user-driven hierarchical clustering.

Acknowledgments

The authors acknowledge the financial support of CAPES/DAAD PROBRAL (344/10 and 415-br-probral), CNPq (305079/2009-3 and 301295/2008-5), DFG (LI 1530/6-1), and the VisComX Center at Jacobs University.

References

[AdO04] ARTERO A. O., DE OLIVEIRA M. C. F.: Viz3d: Effective exploratory visualization of large multidimensional data sets. In *SIBGRAPI '04: Proceedings of the Computer Graphics and Image Processing, XVII Brazilian Symposium* (Washington, DC, USA, 2004), IEEE Computer Society, pp. 340–347.

[BDY03] BORYCZKO K., DZWINEL W., YUEN D. A.: Clustering revealed in high-resolution simulations and visualization of multi-resolution features in fluid-particle models. *Concurrency and Computation: Practice and Experience 15*, 2 (2003), 101–116.

[Bli82] BLINN J. F.: A generalization of algebraic surface drawing. *ACM Trans. Graph. 1*, 3 (1982), 235–256.

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J. D.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics 14* (2008), 1141–1148.

[EM94] EDELSBRUNNER H., MÜCKE E. P.: Three-dimensional alpha shapes. *ACM Trans. Graph. 13*, 1 (1994), 43–72.

[EN*09] ELER D. M., NAKAZAKI M. Y., , PAULOVICH F. V., SANTOS D. P., ANDERY G. F., OLIVEIRA M. C. F., NETO J. B., MINGHIM R.: Visual analysis of image collections. *The Visual Computer 25*, 10 (October 2009), 923–937.

[FL95] FALOUTSOS C., LIN K.-I.: Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data* (New York, NY, USA, 1995), ACM Press, pp. 163–174.

[HKW99] HINNEBURG A., KEIM D. A., WAWRYNIUK M.: Hdeye: Visual mining of high-dimensional data. *IEEE Comput. Graph. Appl. 19*, 5 (1999), 22–31.

[ID90] INSELBERG A., DIMSDALE B.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. pp. 361–378.

[Kan00] KANDOGAN E.: Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. In *In Proceedings of the IEEE Information Visualization Symposium, Late Breaking Hot Topics* (2000), pp. 9–12.

[KR90] KAUFMAN L., ROUSSEEUW P.: *Finding Groups in Data An Introduction to Cluster Analysis*. Wiley Interscience, New York, 1990.

[LC87] LORENSEN W. E., CLINE H. E.: Marching cubes: A high resolution 3d surface construction algorithm. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques - SIGGRAPH 1987* (1987), ACM Press, pp. 163–169.

[LLRR08] LINSEN L., LONG T. V., ROSENTHAL P., ROSSWOG S.: Surface extraction from multi-field particle volume data using multi-dimensional cluster visualization. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (Nov.-Dec. 2008), 1483–1490.

[LW03] LI J., WANG J. Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*, 9 (2003), 1075–1088.

[Mac67] MACQUEEN J. B.: Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (1967), Cam L. M. L., Neyman J., (Eds.), vol. 1, University of California Press, pp. 281–297.

[Nas95] NASON G.: Three-dimensional projection pursuit. *J. Royal Statistical Society, Series C 44* (1995), 411–430.

[PM08] PAULOVICH F. V., MINGHIM R.: HiPP: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (2008), 1229–1236.

[PNML08] PAULOVICH F., NONATO L., MINGHIM R., LEVKOWITZ H.: 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 (May-June 2008), 564–575.

[RL09] ROSENTHAL P., LINSEN L.: Enclosing surfaces for point clusters using 3d discrete voronoi diagrams. *Computer Graphics Forum, EuroVis 09 Proceedings 28*, 3 (2009), 999–1006.

[RLL08] ROSENTHAL P., LONG T. V., LINSEN L.: Shadow clustering: Surface extraction from non-equidistantly sampled multifield 3d scalar data using multi-dimensional cluster visualization. In *Winner of 2008 IEEE Visualization Design Contest, VisWeek 08 Conference Compendium* (2008), IEEE Computer Society, pp. 64–65.

[Sam69] SAMMON J. W.: A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers 18*, 5 (1969), 401–409.

[SBG00] SPRENGER T. C., BRUNELLA R., GROSS M. H.: H-blob: a hierarchical visual clustering method using implicit surfaces. In *Proceedings of IEEE Visualization '00* (Los Alamitos, CA, USA, 2000), VIS '00, IEEE Computer Society Press, pp. 61–68.

[SKK00] STEINBACH M., KARYPIS G., KUMAR V.: *A Comparison of Document Clustering Techniques*. Technical Report 00-034, University of Minnesota, 2000.

[TKAM06] TORY M., KIRKPATRICK A. E., ATKINS M. S., MÖLLER T.: Visualization task performance with 2d, 3d, and comnination displays. *IEEE Transactions on Visualization and Computer Graphics 12*, 1 (2006), 2–13.

[TMN03] TEJADA E., MINGHIM R., NONATO L. G.: On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization 2*, 4 (2003), 218–231.

[WN08] WHALEN D., NORMAN M. L.: Competition data set and description. In *2008 IEEE Visualization Design Contest* (2008), http://vis.computer.org/VisWeek2008/vis/contests.html.

[Yan99] YANG L.: 3d grand tour for multidimensional data and clusters. In *IDA '99: Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis* (London, UK, 1999), Springer-Verlag, pp. 173–186.