This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

Perception-based Evaluation of Projection Methods for Multidimensional Data Visualization

Ronak Etemadpour, Robson Motta, Jose Gustavo de Souza Paiva, Rosane Minghim, Maria Cristina Ferreira de Oliveira, *Member, IEEE,* and Lars Linsen, *Member, IEEE*

Abstract—Similarity-based layouts generated by multidimensional projections or other dimension reduction techniques are commonly used to visualize high-dimensional data. Many projection techniques have been recently proposed addressing different objectives and application domains. Nonetheless, very little is known about the effectiveness of the generated layouts from a user's perspective, how distinct layouts from the same data compare regarding the typical visualization tasks they support, or how domain-specific issues affect the outcome of the techniques. Learning more about projection usage is an important step towards both consolidating their role in high-dimensional data analysis and taking informed decisions when choosing techniques. This work provides a contribution towards this goal. We describe the results of an investigation on the performance of layouts generated by projection techniques as perceived by their users. We conducted a controlled user study to test against the following hypotheses: (1) projection performance is task-dependent; (2) certain projections perform better on certain types of tasks; (3) projection performance depends on the nature of the data; and (4) subjects prefer projections with good segregation capability. We generated layouts of high-dimensional data with five techniques representative of different projection approaches. As application domains we investigated image and document data. We identified eight typical tasks, three of them related to segregation capability of the projection, three related to projection precision, and two related to incurred visual cluttering. Answers to questions were compared for correctness against 'ground truth' computed directly from the data. We also looked at subject confidence and task completion times. Statistical analysis of the collected data resulted in Hypotheses 1 and 3 being confirmed, Hypothesis 2 being confirmed partially and Hypotheses 4 could not be confirmed. We discuss our findings in comparison with some numerical measures of projection layout quality. Our results offer interesting insight on the use of projection layouts in data visualization tasks and provide a departing point for further systematic investigations.

Index Terms—Projections, dimension reduction, multidimensional data, perception-based evaluation

1 Introduction

Visual exploration methods rely on 2D or 3D visual encodings. In handling high-dimensional data, dimension reduction techniques such as projections from the input space to a 2D/3D visual space are widely applied. A large variety of such techniques exist, targeting distinct design goals or applications, which typically output similarity-based layouts often displayed as scatter plots. Multidimensional data sets may include hundreds to thousands of objects described by tens to hundreds of attributes. Data characteristics regarding the distribution in the multidimensional space vary for different application domains, e.g., consider document data and image data: text usually produces sparse spaces and image produces dense spaces.

Commonly desired properties of projected layouts are similarity or distance preservation (with a distance metric properly defined in the given multidimensional attribute space), as well as cluster or class preservation and separation (or segregation). Several measurements have been introduced to assess projection methods with respect to such properties. However, user perception has not played a significant role in quality investigations, and very little is known about how subjects perceive projection layouts. This work takes a step in this direction by systematically evaluating the performance of projection methods in various categories from a user perspective. We designed and implemented a controlled user study to test human performance on multiple tasks, using image and textual data sets.

We chose representatives from certain interesting groups of techniques to investigate (see Sections 2 and 3.1). We decided to focus on two data domains with distinct characteristics, namely image data, where each object is an image described by a number of derived features, and document data, where each object is a document described by the frequencies of certain terms (see Section 3.2). We collected common questions arising in both application domains, which led to the formulation of eight abstract tasks for the experimental study. These can be grouped into tasks that favor good performance on cluster segregation, distance preservation, and visual clutter avoidance, as explained in Section 3.3. Based on the given tasks, data, and projections, we designed a user study to test against four hypotheses. They are concerned with the performance of the layouts with respect to different (groups of) tasks and different data properties, as well as subject preference. The set-up of the study and the hypotheses are specified in Section 3.4.

We derive the ground truth for each task from the multidimensional data sets, compute errors with respect to that ground truth, and perform a statistical analysis of the data collected in the user study in Section 4. The results regarding the formulated hypotheses are presented and discussed in Section 5. We summarize overall findings

• *R. Etemadpour and L. Linsen are with Jacobs University, Bremen, Germany. J.G.S. Paiva is with Federal University of Uberlândia, Brazil. R. Motta, R. Minghim and M.C.F. Oliveira are with University of São Paulo, São Carlos, Brazil.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

and derived guidelines in Section 6, and we also discuss our findings relative to existing quality measurements for projection layouts in Section 7.

We were able to investigate how the choice of the projection technique affects human perception and performance when executing typical tasks on projected layouts of multidimensional data. Some hypotheses on the influence of certain parameters on the performance of projections regarding those tasks could be confirmed, while others had to be rejected. These findings will help visualization experts and domain scientists to choose suitable projections for a given task and a given data set. They also may trigger further investigations in various directions such as interactive tasks or use of supporting visual artifacts.

2 Related Work

Many techniques are currently available to generate 2D similarity-based layouts from high-dimensional data, here called "projection techniques", or simply "projections", often displayed as 2D scatter plots. It is possible to identify several distinct approaches to reduce data dimensionality, and a full discussion is beyond the scope of this paper. We briefly refer to a few classic and recent contributions from both the data mining and data visualization fields, shown to be effective in particular scenarios. A detailed discussion of several nonlinear dimensionality reduction techniques can be found in [1][2].

Classical dimension reduction algorithms, such as PCA - *Principal Component Analysis* [3] - are often employed to generate similarity layouts by reducing data to two or three dimensions, despite the fact that they have not been developed for this purpose.

Multidimensional Scaling (MDS) [4] refers to a broad range of techniques that transform points defined in a higher-dimensional input space into points represented in a lower-dimensional visual space such that relative distances are preserved. Alternative strategies to achieve distance preservation result in different MDS techniques. *Classical Scaling* [4][5] is a metric MDS that applies a spectral decomposition on the distance matrix in input space to obtain a spatial embedding of the points while preserving their known distances.

Force-directed Placement approaches [6] rely on iterative algorithms that model the data points as a system of particles attached to each other by springs. The length of the spring connecting two particles is given by the distance between their corresponding data points. A spatial embedding is obtained with an iterative simulation of the spring forces acting on this hypothetical physical system.

A distinct category of 2D mappings employs tree layouts to convey similarity levels contained in a distance matrix. The algorithms to generate similarity layouts [7][8] are inspired on the well-known Neighbor-Joining (NJ) heuristic originally proposed to reconstruct phylogenetic trees. NJ builds unrooted trees, for which the leaf nodes represent the data points and edge lengths indicate dissimilarity. The heuristic produces a tree in which highly similar data points are placed at the ends of branches, progressing towards the top with those points less similar to each other.

Many measures have been introduced to estimate, numerically or graphically, the quality of layouts produced by projection methods. Estimates such as the silhouette coefficient [9] and neighborhood hit [10] evaluate clustering capability, while the correlation coefficient [11] and distance plots evaluate distance. Some criteria involve ranks of sorted distances and analyze K-ary neighborhoods computed in both the high- and low-dimensional spaces [12], [13], for a varying value of K, which yield curves that must be scrutinized on several scales. Lee and Verleysen [14] suggested a measurement based on the summarization of these scale-dependent measures into a single scalar value, enabling simple and direct comparison of dimensionality reduction methods. Venna et al. [15] introduced a quality measurement for an information-retrieval task based on minimizing the cost of a query, the amount of missed instances, and the amount of those erroneously retrieved. Aupetit [16] propose to visualize quality measures associated to a set of projected instances by coloring the corresponding Voronoi cell in the projection space according to local distortion. Lespinats and Aupetit [17] use stress functions, calculated over the entire layout, to characterize each projection data point as a false neighborhood or a tear. Their visualization shows where structures are projected reliably or have been distorted.

Although these measures are useful to compare multiple layouts regarding their faithfulness to the original data embedding, they do not consider user perception. We compare our findings based on human perception to some of these estimates in Section 7. For a survey on quality metrics for high-dimensional data visualization the reader is referred to Bertini et al. [18].

Recently, Tatu et al. [19] investigated quality measures computed from projections from a human perception perspective. They conducted a user study, where the subjects were asked to select and rank the five most useful scatter plots for the task of best separating three given classes encoded by color, considering a single data set. The scatter plots were selected views from a scatter plot matrix, and the focus was not on evaluating the performance of projection methods, but to compare the best views detected against the computed Class Consistency Measure (CCM) [20] and Class Density Measure (CDM) [21]. In contrast to their work, in this paper we investigate non-orthogonal projections and users' performance under multiple projection methods, multiple tasks and multiple data sets.

Lewis et al. [22] presented a study to investigate human agreement on layout quality, as well as what types of layout structures they find appealing. They concluded that expert users are reasonably consistent judges of layout quality, in contrast with novices, which are very inconsistent. Also, humans do not appear to have strong layout structure preferences. The idea of this study differs from the one presented here in the sense that it evaluates the overall usefulness of the layouts according to several criteria, such as variance, skewness, etc., without associating them to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

any specific task, whereas we evaluate the dependency of layouts on tasks and data characteristics.

Recent studies [23][24] investigated the accuracy of Clustering Quality Measures (CQMs) of cluster separation capability in scatter plots depicting multidimensional projection layouts, and whether certain CQMs correlate better with human judgments than others. They found that quality assessment of cluster separation by these measures was highly discrepant with human assessments – obtained from systematic inspection by two researchers – with the measures showing a high number of failure cases. They show that a natural mathematical formalization does not suffice to guarantee that the evaluations of clusterings produced using the CQM seem natural to the users. Sedlmair et al. [23] present a detailed taxonomy of factors that affect the human perception of cluster separation. By comparing the consistency of expert and non-expert users, Lewis et al. [24] state that the general population have a natural clustering evaluation skill, which does not require a specific training.

Albuquerque et al. [25] have attempted to find a perception-based quality measure for scatter plots. First, users were asked to identify similarity between scatter plots, which was used to train a MDS embedding. Then, users were asked to rank scatter plots according to their appropriateness to a given task leading to a quality metric for assessing the test data sets. Scatter plots were ranked according to the quality measure on two tasks, namely, correlation of two attributes and separation of two color-coded clusters. The derived quality measure is task-dependent and requires a proper training set. In contrast to their work, we look into evaluating users' performance on layouts generated by different projection techniques. Similar to us, they handled different tasks separately, although considering rather simple tasks and data sets.

Rensink and Baldridge [26][27] have investigated the perception of correlation in scatter plots purely from a psychological perspective, not considering real-world data sets or tasks motivated by the visual analysis process of certain applications. They have tested whether users could discriminate pairs in a set of generated scatter plots with points distributed within a certain range from the diagonal, and concluded that the perception of correlation in a scatter plot is completely specified by two easily-measured parameters. In a follow-up study, Rensink [28] showed that the perception is rapid.

3 Design of User Study

3.1 Projections

We have selected four techniques as representatives of three distinct strategies for embedding data in two dimensions, namely statistical dimension reduction, MDS, and Force-directed Placement. We have also included a technique based on Similarity Trees [7], which is a different type of point placement and had not been previously used as a projection. The techniques picked are PCA [3], Isomap [29], LSP [10], Glimmer [30], and NJ tree [8]. Our choice covers modern and classic techniques that have been introduced aiming at capturing different data behaviors.

PCA - *Principal Component Analysis* [3] has been included in the study because it is a classical dimension reduction strategy often employed to generate visual embeddings of data. It applies an orthogonal transformation to compute linear combinations of the original data attributes, outputting a reduced number of descriptive dimensions that best capture data variance in the input space. 2D layouts are obtained considering the two first principal components, at the risk of disregarding other potentially relevant components.

Isomap - *Isometric Feature Mapping* [29] is a variant of Classical Scaling MDS. It replaces the original distances by geodesic distances computed on a graph to obtain a globally optimal solution to the distance preservation problem. A weighted nearest-neighbor graph is built from the data, with pairwise point distances as edge weights. The shortest path in this graph gives the distance between two points. Isomap is effective on data that present non-linear relationships, that both PCA and Classical Scaling typically fail to detect.

LSP - *Least Square Projection* [10] first samples a reduced sub-set of points representative of the data distribution in the input space and projects them to the target space with a precise MDS, force placement or dimension reduction technique. It then builds a linear system from information given by the projected points and their neighborhoods, which is solved to obtain a 2D embedding of the remaining data points. A Laplacian operator ensures that data points in a particular neighborhood remain proximate in the target space. The choice of representatives affects precision of the resulting layout, with good results achieved with sampling by clustering. LSP is a modern techique that is both cost-effective and highly precise according to objective quality measurements.

Glimmer [30] is a recent technique representative of force-directed placement MDS. In Glimmer the iterative point placement procedure is highly optimized by usage of GPU hardware combined with a multilevel strategy that operates on a hierarchical model of the underlying particle-spring system. It is also fast and generates good quality layouts as evaluated by stress preservation measures.

Finally, we had evidence that similarity tree layouts favor good performance on tasks that require visual segregation of clusters, and wanted to check whether their good grouping and distance properties would be perceived by subjects in the same way as the projections if the edges are removed from the layouts. We picked the *Neighbor-Joining* (NJ) tree layout computed by the algorithm recently introduced by Paiva et al. [8], which is faster than the original NJ-tree layout algorithm [7] and generates more precise layouts.

3.2 Data

Accounting for different data types and characteristics is important when investigating projection methods applied to multidimensional data. Thus, we wanted to conduct the study on data with different characteristics, and preferably

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

on real data. Moreover, we have a particular interest in text analytics. These factors motivated the choice of document and image collections as the target domains. Text modeled as vector spaces have very peculiar characteristics, quickly reaching high dimensionalities. Most prominently, higher dimensionality usually imposes higher data sparseness. Image data sets, on the other hand, are usually of much lower dimensionality, albeit very sensitive to the choice of the feature space. We thus selected two real document and two real image data sets to further investigate the hypothesis that techniques are sensitive to data characteristics. Document collections are taken as representative of sparse data typically embedded in very high-dimensional feature spaces, whereas image collections are representative of lower-dimensional and less sparse feature-spaces. The different characteristics are also reflected by the choice of distance metrics. Cosine distance is the usual choice for text data and has been taken as default. For the image data the choice of the distance function was made after comparing the layouts produced with the specific technique considering both Cosine and Euclidean distances, and picking the layout displaying the best point segregation on a visual assessment.

The chosen document data sets are referred to as CBR and KDViz. CBR comprises 680 documents in four different topics, with the number of documents unbalanced between labels[1]. A bag of words representation has been created with 1,423 terms, or dimensions. Similarly, the KDViz documents have been collected from an Internet repository[1], again addressing four topics, with 1,624 objects, 520 dimensions and four highly unbalanced labels. The image data sets are referred to as Corel and Medical. The Corel data[2] includes 1,000 photographs on ten different themes, described by 150 dimensions (SIFT descriptors). The Medical data is of magnetic resonance (MRI) images[3] and has 540 objects and 28 dimensions (Fourier descriptors and energies derived from histograms, plus mean intensity and standard deviation).

For reference, Table 1 shows the projected layouts of all four data sets obtained with each of the five projections identified in Section 3.1, with class labels mapped to colors.

3.3 Tasks

To define representative user tasks we identified typical questions raised when visually analyzing document and image data and abstracted them from the underlying application. Following the task framework defined by Andrienko and Andrienko [31], we are looking into synoptic tasks including the whole reference set or subsets thereof, as "elementary tasks play a marginal role in exploratory data analysis" [31]. Andrienko and Andrienko grouped synoptic tasks into pattern identification, behavior (pattern) comparison, and relation-seeking.

First, we have been looking into pattern identification tasks, where the targets were finding groups of similar objects (clusters) or finding outliers, while the constraints were considering the whole data set or a subset thereof. We formulated the tasks:

#Clu Estimate the number of observed clusters.

#SClu Estimate the number of observed subclusters of a given cluster.

#Out Estimate the number of outliers.

Second, we looked into comparison tasks with respect to a given reference set, which can be a cluster or an individual object. We formulated the tasks:

fCluClu Identify the closest cluster to a given cluster.

fCluObj Identify the closest cluster to a given object.

rKnn Identify and rank the k nearest objects to a given object.

Third, we looked into a relation-seeking task between different reference sets, where the reference sets were clusters, and formulated the task:

rDens Rank given clusters by density.

Finally, another task about cluster properties was formulated for a single reference set as:

#Obj Estimate the number of objects in a selection.

Figure 1 shows one example stimulus for each task.

3.4 Set-up and Hypotheses

We applied each of the five projection techniques to each of the four data sets, leading to the 20 scatter plots shown in Table 1. The parameters in generating the scatter plots were the default ones adopted in the implementations employed and are detailed in the Supplementary Material. We generated for each scatterplot one stimulus for each of the eight tasks, leading to 160 stimuli like the ones shown in Figure 1. Given the large number of scatter plots, the body of subjects was divided into two groups. The first group of 31 subjects was assigned the tasks #Clu, #SClu, and #Obj; the second group of 30 subjects was assigned the tasks #Out, fCluClu, fCluObj , rKnn, and rDens. Subjects assigned the same task set executed them in the same (random) sequence and saw the same images. All subjects fulfilled their tasks in two sessions with a short break in between.

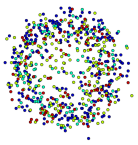
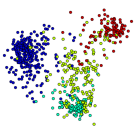
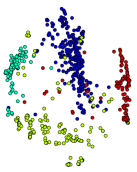
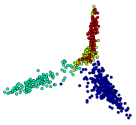
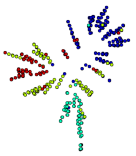
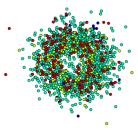
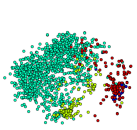
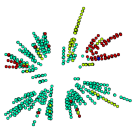
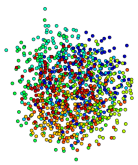
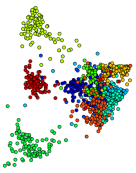
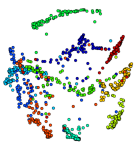
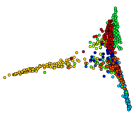
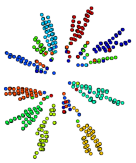
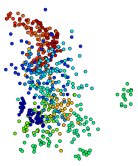
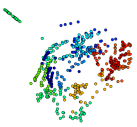
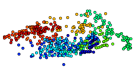
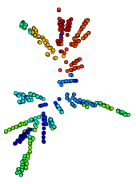
The body of subjects for the study consisted of 61 students at an undergraduate or graduate level in the fields of applied mathematics and computer science. They had not been engaged with projections in depth, although they possibly had different levels of knowledge about projections. They were provided with a 20-minute introduction on projections, scatter plots, and the set-up of the user study. It was not necessary to confront them with the applications behind the data (document and image data).

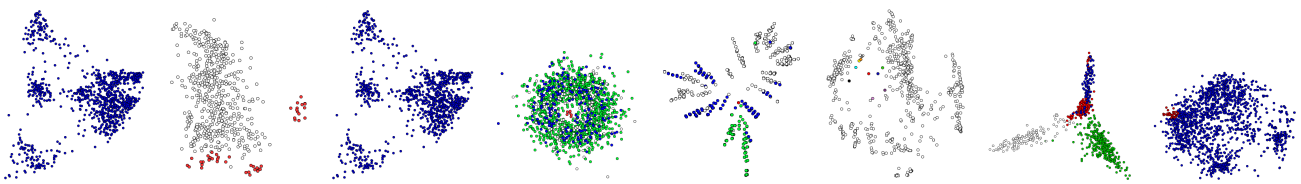
For some tasks some points were highlighted by color. For Tasks #Clu and #Out, all data points were shown in a single color; for Tasks #SClu and #Obj points from a target cluster/selection were highlighted with a different color. For Tasks fCluClu and fCluObj the target cluster/object was shown in one color (red) and two other clusters in further colors (green and blue), from which the one closer to the target cluster/object should be identified. For Task

1. http://vicg.icmc.usp.br/infovis2/DataSets
2. UCI KDD Archive, http://kdd.ics.uci.edu
3. made available by a collaborator

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

TABLE 1. The layouts obtained with the five tested projections on the four data sets investigated. Circle color indicates instance class label.

	Glimmer	Isomap	LSP	PCA	Tree
CBR					
KDViz					
Corel					
Medical					



(a) #Clu (b) #SClu (c) #Out (d) fCluClu (e) fCluObj (f) rKnn (g) rDens (h) #Obj

Fig. 1. Instances of task stimuli: (a) Estimate number of clusters, (b) estimate number of subclusters of red group, (c) estimate number of outliers, (d) determine whether green or blue cluster is closer to red object, (e) determine whether green or blue cluster is closer to red cluster, (f) find five closest objects to red object, (g) rank red, green, and blue clusters by density, and (h) estimate number of objects in red group.

rDens, three clusters were shown in distinct colors and subjects should rank them based on density. For Task rKnn, ten points were highlighted using ten different colors and the users were asked to identify and rank the five points most similar to the purple one. To avoid bias due to the choice of colors we randomly assigned colors to each subject individually. The PCA scatter plot of KDViz was too cluttered to allow for distinguishing the colored points on Tasks fCluClu and rKnn, we thus removed those from the collection and assumed maximum error in the analysis. The complete collection of tasks and respective images is provided as supplementary material to this paper.

The system always first presented the task to the subjects. Once they felt comfortable about having understood the task, they were confronted with a sequence of still images showing the respective stimuli. For each image they were asked to answer the question as soon as they knew the answer. To force participants to act as quickly as possible, we introduced a time limit. In a pilot study with eight participants we observed that it took them on average 7.7 seconds to fulfill the tasks and the average maximum time was 24.75 seconds. Therefore, in the actual study we gave the participants 30 seconds to complete the tasks, after which the stimulus disappeared. The question would remain until answered. After each task, the subjects were asked to rank their confidence in the given answer (on a five-step Likert scale). We also recorded the subject answering times.

A more detailed inspection reveals that the tasks require different properties of the projection for best performance. Tasks #Clu, #SClu, and #Out require good spatial segregation of clusters (and outliers), Tasks fCluClu, fCluObj and rKnn require good precision to preserve distances between objects and/or clusters, and Tasks rDens and #Obj require the projections to avoid object clutter. Based on these

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

observations, we formulate the following hypotheses:

H1 Different projections perform better on different tasks.

H2 Performance of projections are similar within the three groups of Tasks (#Clu, #SClu, and #Out), (fCluClu, fCluObj, and rKnn), and (rDens and #Obj).

As different types of data (document vs. image) have different characteristics,another hypothesis is formulated to investigate the impact of such differences:

H3 Performance of projections depends on data characteristics.

We also ask about subject confidence when performing the tasks. Here, we formulate the hypothesis

H4 User confidence for projections is governed by good segregation.

4 Statistical Analysis

4.1 Estimation of Data Properties

Tasks #Clu and #SClu require comparing the number of clusters and sub-clusters as perceived by subjects on the projected layouts with the 'real' cluster structure in the data. Generally speaking, there is no ideal cluster structure, as distinct solutions may be acceptable. Still, cluster-related tasks are indisputably highly relevant to projection usage [23], [32], e.g. "the fundamental reason that people look at scatterplots of dimensionally reduced data is to see if the implicit groups match their mental model of the dataset." [23]. The tasks included in the study would hopefully give evidence for the projections' capabilities of visually grouping similar content, as well as in the differences in user perception of 'visual' groups. The sub-cluster Task #SClu, in particular, was motivated by the observation that projection layouts often depict subgroups within larger groups.

As we need a baseline for comparison, for Task #Clu we chose to assume that the given class structure gives a good approximation of the cluster structure that the projections should help to retrieve, and considered the number of labeled classes in each data set as the ground truth for its number of clusters. Similarly to other authors [23][33] [34] we acknowledge that taking classes as clusters is arguable as a general assumption, but in some situations analysts do take the class structure as reference, motivating its adoption as a valid baseline for comparing the projections.

This reasoning does not apply to Task #SClu, as no sub-class structure is given. We thus chose to run a clustering algorithm on the target class and take its output as an approximation to the expected number of sub-groups. Again, although no solution can be interpreted as 'the correct one' and projections are not meant to reflect a particular clustering strategy, a specific solution provides a valid baseline for comparison as long as it is a reasonable one: if a cluster structure exists, a good projection should be able to recover it, to some extent. We favored the *X-Means* clustering [35], which extends the well known K-Means algorithm to automatically find the best choice for the number of clusters k. In a posterior step we run two other clustering algorithms on the target classes, namely the divisive hierarchical *Bisecting K-Means* and the *Agglomerative Hierarchical Single Link* [36] for a number of clusters in the range (2-20) and identified as the 'optimal' number that of the solution with the highest *Dunn Index*. We observed that the three clustering algorithms only agreed completely for one data set (KDViz), while resulting numbers of clusters deviated for the other three data sets. Hence, we restricted our analysis of Task #SClu to KDViz only.

To detect the outliers (Task #Out), i.e., objects that differ significantly from all others [37], one may look into distances [38], classifier models [39], clusters [40], or densities [41]. We favored a density-based method because distance- and cluster-based methods tend to perform poorly on clusters of varying densities, and classifier-based methods are sensitive to parameter choice. The Local Outlier Factor (LOF) method [41] can identify outliers in different densities and requires a single parameter, the number of nearest neighbors. The LOF algorithm first computes a reachability distance measure between objects, then creates a local reachability density for each object by considering its nearest neighbors, and finally compares the object's local density with that of its neighbors. A LOF value close to one indicates that the local density at an object is comparable to that of its neighbors, a LOF value below 1 indicates a denser region, and a LOF value significantly above 1 indicates an outlier. The LOF algorithm estimated four outliers for CBR, three outliers for KDViz, one outlier for Corel, and ten outliers for Medical.

To identify the closest cluster to a given cluster (Task fCluClu), we compute pairwise distances between all objects of the target cluster and those of the remaining clusters in the multidimensional space and identify the smallest distance. Analogously, we identify the closest cluster (Task fCluObj) and the k nearest objects (Task rKnn) to a given object.

To estimate cluster density (Task rDens), known non-parametric density estimator methods are kernel estimator [42] and local likelihood [43]. Due to high computational costs and slow convergence, these estimators only work well for data sets with up to six dimensions [44]. We considered a simple distance-based approach because we must estimate densities in high-dimensional data, but it suffices to do so comparatively. A minimum spanning tree is created for each cluster and its density is defined as the inverse of the average edge length in the minimum spanning tree, as it has short edges in dense regions and long edges in sparse regions. Because it considers only distances, this solution scales well to high dimensions. Moreover, it is not biased towards any shape and insensitive to density changes.

Counting the number of points in a selection (Task #Obj) is trivial.

4.2 Computation of Errors

Given the ground truth, we can compute the errors in the answers of the subjects for each task. For the tasks that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

required the subjects to estimate a number (Tasks #Clu, #SClu, #Out, and #Obj), the error is computed by

$$e = \frac{|n_{true} - n_{answer}|}{n_{true}},$$

where n_{true} is the estimated ground truth and n_{answer} is the reported answer. For Task #Out, there was a large spread in the answers n_{answer} and we normalized the estimated errors to the interval $[0,1]$ by dividing by the maximum error reported. For the tasks that required a cluster to be identified (Tasks fCluClu and fCluObj), the error is either zero or one. For the ranking tasks (Tasks rKnn and rDens) we estimated the number of swaps required to get from the reported answer to the ground truth. For example, if (s_1, s_2, s_3) is the correct ranking and (s_3, s_1, s_2) the reported answer, one needs to first swap s_3 with s_1 and then with s_2 to get from the reported answer to the correct one. Hence, the number of swaps is two. The error is computed by the number of necessary swaps for the reported answer divided by the number of necessary swaps for the worst answer. For the given example, the number of swaps for the worst answer (s_3, s_2, s_1) would be 3 and the error would be $\frac{2}{3}$.

4.3 Investigations and Statistical Methods

Several aspects were considered for the statistical analysis of the results of the experimental study. First, we compared the five projection methods for each of the eight tasks by looking into the mean errors over all subjects and all data sets. Second, we did the same comparisons considering document data and image data separately. Third, we compared the mean errors of document vs. image data over all projections and analogously compared the two image data sets against each other, as well as the two document data sets. In addition to the mean error, we also evaluated the confidence ratings the subjects reported and the time it took them to fulfill the tasks.

For all analyses, we computed means and standard deviation of the errors. To test for statistical significance of the individual results, we first tested the distribution of the error values against normality using the Kolmogorov-Smirnova and the Shapiro-Wilk tests. In case of non-normal distribution, we applied the Wilcoxon test on non-parametric two related samples when comparing two groups and the Friedman test on K related samples when comparing more than two groups. Since the Friedman test is an omnibus test and only computes overall differences but does not report which groups particularly differ from each other, we also perform pairwise comparisons of the groups using a Wilcoxon test and Holm's sequential Bonferroni adjustment on the results from the Wilcoxon tests at the 0.05 level. In case of normal distribution, we used t-test when comparing two groups and ANOVA test when comparing more than two groups. For pairwise comparisons in case of more than two groups we run a series of Tukey's post-hoc tests.

5 Results

Figure 2 summarizes the comparative analysis of the five projections for each of the eight tasks. Note that Task #SClu is restricted to only use results from KDViz, as explained above. The bar charts show the mean error values and the standard error from the mean. The omnibus tests for statistical significance showed that there is statistical significance in the mean errors for all tasks. The outcome of the pairwise significance test is indicated by the red horizontal lines color coded on a scale from red to white. More precisely, groups of projections with no pairwise significant difference among their mean error have lines of the same color. However, please note that different colors do not necessarily indicate pairwise significant difference. For example for Task #Clu, Friedman test showed significant difference ($\chi^2(4,31) = 68.982, p < 0.05$) among five projections and Bonferroni test across pairwise Wilcoxon comparisons showed significant differences between all pairwise comparisons except for LSP vs. Isomap ($Z = -0.901, p = 0.367$) and Tree vs. Glimmer ($Z = -0.078, p = 0.938$). Hence, Isomap and LSP form the winner group indicated by the dark red line, Tree and Glimmer form the loser group indicated by no (or white) line, while PCA is in-between as indicated by the light red line. Similarly we analyze all the other tasks. The Friedman test delivers statistical significance ($p < 0.05$) for all tasks and for Tasks rKnn and rDens even strong statistical significance ($p < 0.01$). The p-values for all pairwise Bonferroni tests are provided in the supplementary material.

Although it can be observed that Isomap is doing very well on five of the eight tasks (ranked first on #Clu, #SClu, fCluClu, fCluObj, ranked second on rDens), it is not performing so well on the other three tasks (#Out, rKnn, #Obj). Similarly, Glimmer is performing very well on four tasks (rDens, #Obj, fCluClu, fCluObj), but very poorly on three other tasks. LSP, in general, did well (except for Task fCluObj), as mostly there is no significant difference between LSP and the method ranked first. PCA and Tree did well for one task each only (PCA on Task #SClu, Tree on Task rKnn, ranked second in both cases), but performed very badly several times (PCA was the ranked last on Tasks fCluClu, rDens, and #Obj, and second-to-last on Task fCluObj; Tree was the ranked last on Tasks #Out and #SClu and second-to-last on Tasks fCluClu and rDens). Hence, we can conclude that some methods did generally better than others, but that there is no method that performed best for all tasks. Thus, Hypothesis H1 is confirmed.

To determine which projections did generally better than others, we counted how often a projection belongs to the winner group and how often it belongs to the loser group. The winner group is generated by traversing the ranking top-down and adding all methods that do not exhibit a significant difference to the top-ranked method. Analogously, we determine the loser group for each task. There was one exception, as PCA for Task rKnn did not show significant difference neither to the best nor to the worst group and was added to none. We obtained that both Isomap and LSP were five times among the winners and two times among the losers such that these methods can be considered as the overall winners of the study. Glimmer had mixed results, as it did well on some tasks and bad on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

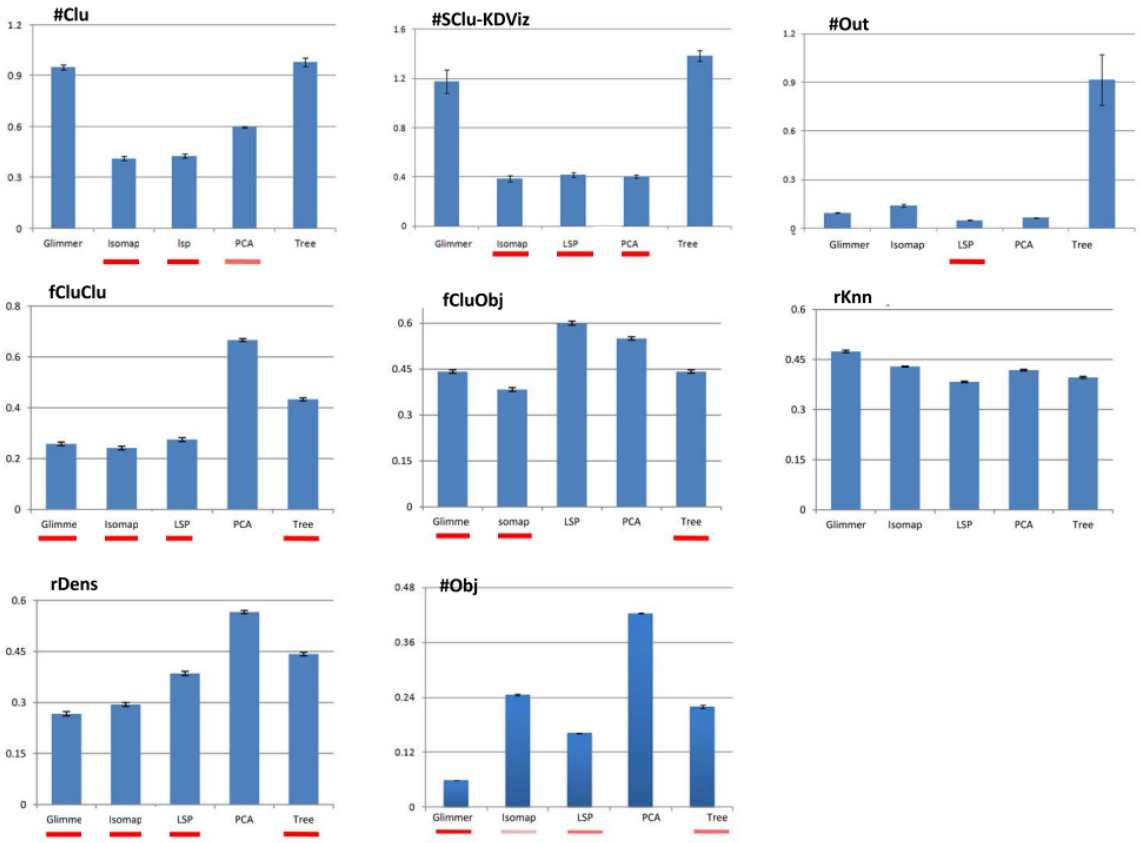


Fig. 2. Correctness (bar charts show mean error and standard error from the mean): results of comparing the projection methods on the tasks considered. There is statistical significance for all tasks. The horizontal lines encode rankings of pairwise statistical significance using a red-to-white color transition, i.e., if there are pairwise significant differences the winners are underlined with dark red color, the losers in white (or no) color, and the ones in-between in light red color.

others. Consequently, the outcome is balanced for Glimmer belonging four times to the winner group and four times to the loser group. Tree and PCA both ended up being only once among the winners, while Tree was three times among the losers and PCA was five times among the losers. Hence, PCA can be considered as the method that overall performed worst.

Next, we looked into the groups of tasks mentioned in Hypothesis H2. When considering segregation Tasks #Clu, #SClu, and #Out, there is some consistency in that Glimmer and Tree tend to do worse than the other three methods. Tasks #Clu and #SClu delivered very consistent results, while Task #Out somewhat weakens the observation. When considering the precision of Tasks fCluClu, fCluObj , and rKnn, results are somewhat consistent among Tasks fCluClu and fCluObj, but not quite so for Task rKnn. The reason may be that Task rKnn only considered individual objects, while Tasks fCluClu and fCluObj considered clusters, which also involves segregation. When looking into the clutter avoidance Tasks rDens and #Obj, one observes a clear pattern with Glimmer being ranked first and PCA being ranked last. Hence, we can conclude that Hypothesis H2 can be confirmed partially.

Figure 3 shows a similar comparison as Figure 2, but considering the results for document data and image data separately. Task #SClu is excluded here, as we only had reliable ground truth for KDViz, see above. Tests for statistical significance showed that there are significant differences among the five projections for all tasks and both types of data with one single exception, which is Task #Out on document data. It can be observed that the results for document and image data are somewhat consistent for some tasks when considering winner and loser groups, but differ substantially for other tasks. For example, for Task #Clu Tree was ranked first for image data but did significantly worse than any other method for document data. Similarly, for Task fCluObj Glimmer is ranked last for image data and first for document data, and for Task rKnn Isomap was ranked last for image data and first for document data. Given those differences, we analyzed for each of the eight tasks and each of the five projections whether there is a significant difference in the results for image and document data. Statistical significance was reported in 30 out of these 40 cases. More precisely, the results for document data were significantly better in 22 cases and the ones for image data in eight cases. Hence, Hypotheses H3 also holds, and we can conclude that the performance of the projections was affected by the data characteristics.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

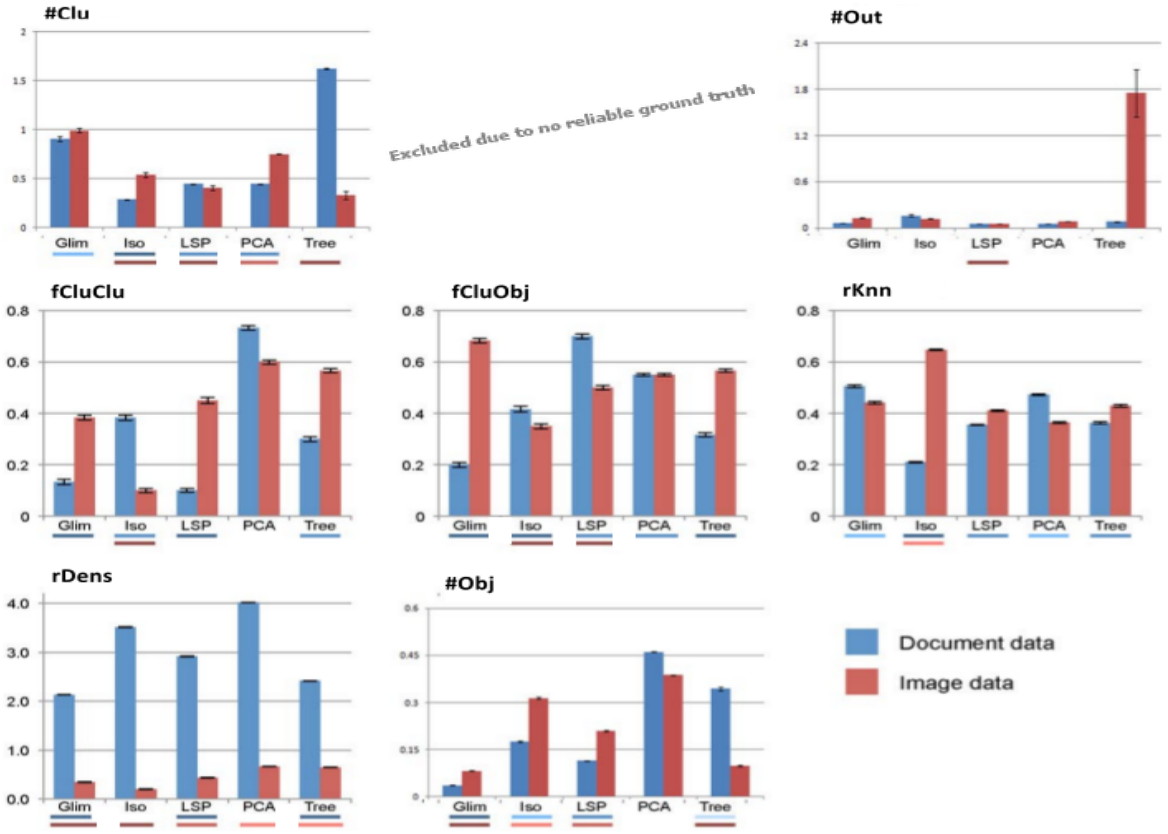


Fig. 3. Correctness (bar charts show mean error and standard error from the mean): results of comparing the five projection methods on the eight tasks for image (red bars) and document data (blue bars). The horizontal lines encode rankings of pairwise statistical significance using a red-to-white color transition for image data and a blue-to-white color transition for document data.

This conclusion was also stated by Sedlmair [45], in the sense that no single dimensionality reduction technique can be considered superior to all others, and some techniques may not reveal certain structures in real world data sets. Moreover, we observe that the projections tend to perform better on document data.

We also analyzed whether there are significant differences in the results within document and image data, respectively. For document data, there was a significant difference in 24 out of 40 cases, where CBR had better results in 11 cases and KDViz in 13 cases. For image data, there was a significant difference in 26 out of 40 cases, where Medical had better results in 16 cases and Corel in 10 cases. So, there are also differences within the groups of data we have been investigating. While the situation was balanced for document data, for image data Medical was overall producing slightly better results.

We now investigate the subjects' confidence when performing the tasks. When looking into the average confidence values for all tasks, all data sets, and all subjects, no statistically significant difference was observed in the confidence levels when comparing the five projection methods. However, there are significant differences when looking at each task individually. In fact, all comparisons show a significance difference for each of the eight tasks when averaging the confidence estimates over all data sets and all subjects. Figure 4 shows a summary of the findings for each of the eight tasks. We report the mean confidence for each task and projection method. As before, we use a red-to-white color transition to indicate statistical significance, i.e., groups of projections with no pairwise significant difference among their mean confidence are shown in the same color. Consequently, dark red fields indicate the winners, white indicate the losers, and light red indicates the ones in between. From Figure 4, we observe that Glimmer is actually five out of eight times in the top-ranked group (more often than any other projection), while PCA is five out of eight times in the bottom ranked group (again, more often than any other). This result is somewhat consistent with the correctness analysis. In fact, when comparing Figure 4 to Figure 2, there seems to be the expected overall correlation between correctness and confidence. However, there are some deviations. Most remarkably, Isomap had low confidence values on several questions with high correctness. As Isomap was one of the winners for segregation tasks, but does not have highest confidence values here, Hypothesis H4 was not confirmed.

Finally, we look into the subjects' task fulfilling times. Findings are summarized in Figure 5. Analogously to Figure 4, it shows the mean completion times for each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

	#Clu	#SClu	#Out	fCluClu	fCluObj	rKnn	rDens	#Obj
Glimmer	3.67	2.80	4.31	3.13	3.65	3.17	3.30	4.31
Isomap	3.58	3.19	3.72	3.10	3.23	3.53	3.57	3.72
LSP	3.12	3.23	3.57	3.4	3.58	3.45	3.78	3.57
PCA	3.61	3.63	3.10	2.77	3.12	2.81	3.75	3.10
Tree	2.91	3.31	3.74	3.26	3.84	3.17	3.61	3.74

Fig. 4. Confidence: Comparing mean confidence values for completing tasks with different projection methods. Colors indicate groups of no significant pairwise differences in form of winners shown in dark red, loser in white, and the ones in between in light red.

task and projection method and groups of no significant pairwise differences by color. Again, all comparisons show statistical significance. One expects the time needed to complete a task to be correlated with the confidence (unless the subject is overwhelmed by the task and quickly gives up). When comparing Figure 5 to Figure 4, there is indeed an overall correlation with two remarkable exceptions. First, tasks were always answered quickly when using PCA, while the confidence was low. Indeed, when looking at the correctness, the answers given are often among the worst. Second, tasks for Glimmer frequently took longer times, while confidence and correctness were still quite high. Hence, confidence showed a higher correlation to correctness than to time on these examples.

	#Clu	#SClu	#Out	fCluClu	fCluObj	rKnn	rDens	#Obj
Glimmer	9.01	8.58	14.78	5.84	7.47	32.58	9.63	8.48
Isomap	10.77	6.64	16.05	6.53	7.94	27.77	7.07	10.80
LSP	12.59	6.60	14.49	4.65	5.70	27.19	6.85	10.95
PCA	8.63	4.56	10.50	5.08	4.21	26.69	6.11	8.99
Tree	18.12	7.02	13.24	5.47	4.74	28.77	7.5	15.85

Fig. 5. Times: Comparing mean times for completing tasks with different projection methods. Colors indicate groups of no significant pairwise differences in form of winners shown in dark red, loser in white, and the ones in between in light red.

For a complete presentation of all numbers (for correctness, confidence, and timings) and all results from statistical tests (omnibus and pairwise), we refer to the extensive supplementary material.

The outcome of the cluster-related task deserves further discussion. Table 2 shows histograms of the answers for Task #Clu. Note that the histogram distributions confirm that Glimmer did not favor cluster detection, since subjects most often identified a single cluster. They did identify more clusters (mostly two or three) in the Medical data set, but numbers differ highly from the reference. The scatter plot display of the Tree layout apparently favors the perception of a high number of clusters. For the other projections one observes better grouping, with Isomap doing particularly well when few clusters are to be identified. The same holds for PCA, except that it rarely supports identification of more than three clusters. More subjects reported numbers closer to the reference with the LSP layouts, with most getting it right for both text data sets. The same observations hold for Task #SClu.

6 Discussion

In this section, we attempt to interpret the results, draw conclusions about the findings, and formulate guidelines. A first broad guideline we can infer is that, when restricted to using a single projection method for several tasks and different data types, one shall consider using Isomap or LSP. Glimmer has some problems with revealing clusters, PCA has problems with clutter and cluster separation when the data sets' main characteristics do not align with the two principal directions, and using Tree as a point placement strategy without rendering the tree structure only works well if the placement matches the investigated clusters.

In general, however, the study indicates that it is worth to investigate the data characteristics first and examine the required properties for the given tasks, as already stated by Lewis et al. [22]. With that information at hand, one can make a better design choice. Glimmer is not suitable in tasks that require cluster segregation. Also, Tree (in the scatter plot set-up adopted) is often not a good choice, although it can be an excellent choice in specific cases. The other three methods performed better. LSP performed best on representing the outliers, and although Lewis et al. [22] suggest the use of Isomap if the data lie on a convex manifold, this technique also performed well on the cluster-related tasks.

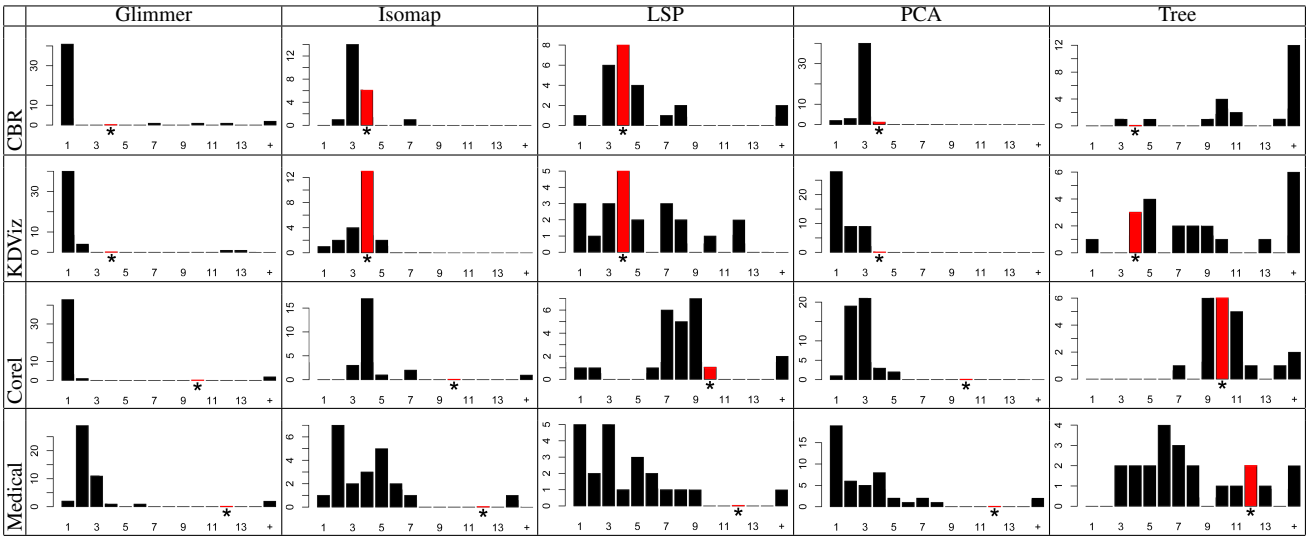
If distance preservation and interpretation are most important for the given task, LSP and PCA are reasonable choices, as they use linear projections, while Glimmer, Isomap, and Tree perturb distances. Still, Tree produced good results for distance interpretations among points, which is probably due to the one-dimensional interpretation of a branch structure that maps to a sorting. On the other hand, PCA did worse than expected, mainly due to clutter.

If clutter avoidance is most important for the given task, Glimmer excels, as the spring model tends to separate points. PCA, on the other hand, exhibits severe cluttering artifacts, as using only the first two principal components does not guarantee separation of data points. The other three techniques are in between and perform equally well. It also became evident that clutter affects the confidence levels and response times. High clutter led to fast answers with low confidence, while a widely spread-out distribution led to higher response times but increased confidence.

Concerning the data types, we observed that, in general, the projection methods worked better on document data, which may be caused by the choice of the features. When comparing the projection methods, there is no general trend

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics

TABLE 2. Histograms of answers to Task #Clu; red bar indicates number of given classes (also shown with '*' sign), rightmost bar accumulates answers ≥ 15.



that a particular method favors a particular type of data over others. However, within the individual tasks there were some remarkable performance differences across the two data types, as presented in the previous section. For PCA the distance distribution of document data (with larger distances) produced even more cluttered results, which negatively affected the distance interpretation results. For Isomap the more sparse document data sets led to layouts spread out more evenly, which improved distance interpretation. For Glimmer, the large distances in document data led to a very regular distribution of points in the layout, which somehow helped to interpret distances to clusters. For the Tree layout the identified clusters in the image data matched better the tree structure, which improved performance on cluster separation tasks.

7 Comparison to Numerical Measures

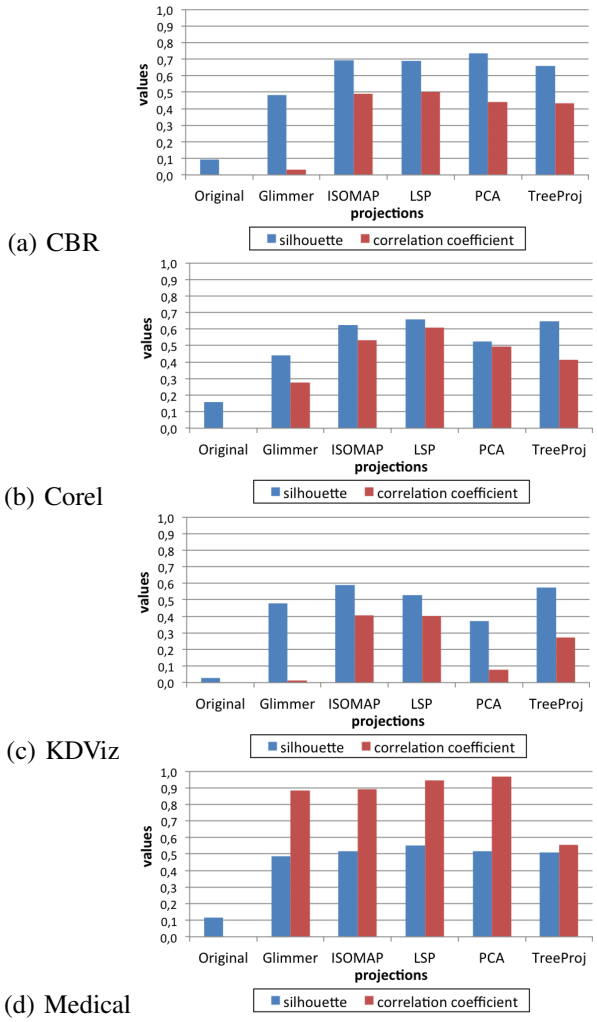
As discussed in Section 2, several numerical estimates have been proposed to compare projected layouts and assess their quality. We chose two numerical and one graphical measure to compare with the perceptual results. Both types of measures assess two important properties of a projection layout relative to the original space, namely the ability to reflect the configuration of distances (or similarities) and the ability to reflect segregation of groups. These properties, referred to as distance preservation and group segregation, are sometimes conflicting: some projections are meant to favor segregation while others are meant to preserve distances or neighborhoods. The best-known *stress* measure [4] has been shown unsuitable to assess grouping and separation [10].

We consider one numerical measure aimed at evaluating group segregation and another one targeted at evaluating distance preservation, namely the *Silhouette Coefficient* (SC) [9], which measures the cohesion and separation between groups of instances on the layout, and the *Correlation Coefficient* (CC) [11], which computes the correlation between all pairwise point distances in the original and in the reduced spaces. SC is in the range $[-1,1]$ and positive values closer to 1 indicate better cohesion and separability; CC is in the range $[0,1]$ and greater values indicate better distance preservation. Both measures rely on distance calculations, which have been computed in the original data space with the same distance functions adopted to compute the projections, and in the projected spaces with the Euclidean distance.

Figure 6 shows both measurements for each data set. One notices that the highest silhouette values (blue bars) were obtained by projections of CBR and Corel, with CBR having the highest. This is an indication of reasonable data separability in the original space, which is confirmed by its original space silhouette (the leftmost bar in each plot). However, for CBR and KDViz, all projections but Glimmer actually improved the separability coded by the original space features. Their corresponding silhouette values also show that, although it has the best value for CBR, PCA did not perform well in any of the other data sets. PCA performed worse for the unbalanced classes of KDVis, but can reasonably discriminate the somewhat more homogeneous four classes in CBR. We were expecting that it would be difficult to achieve good class separability in the Medical data set. LSP did the best job in this case, matching the silhouette of the original space.

In all cases Glimmer's silhouette values were lower than those of the original spaces, unsurprisingly, as it is not formulated to favor discriminability. Nevertheless, on the KDViz data set it did better than PCA in that regard. The distance preserving capability of Glimmer reflects on its good correlation coefficient values (red bars) on the image data sets. A distinguishing feature of the image and the document data spaces is that the first are less sparse. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2014.2330617, IEEE Transactions on Visualization and Computer Graphics



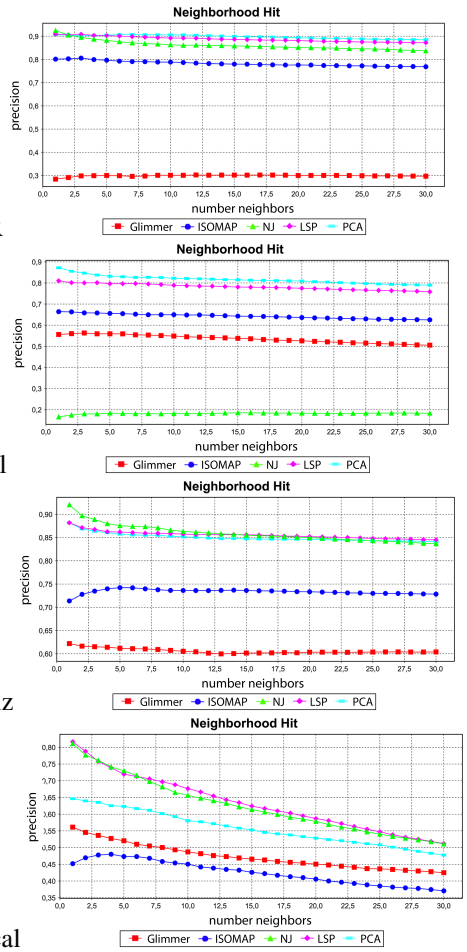
(a) CBR

(b) Corel

(c) KDViz

(d) Medical

Fig. 6. Bar charts showing Silhouette (normalized in the interval [0,1]) and Correlation Coefficient computed for the original data representation and for the five projected layouts.



(a) CBR

(b) Corel

(c) KDViz

(d) Medical

Fig. 7. Neighborhood Hit curves of the five layouts.

correlation coefficient values of the other projections were also higher on image than on document data sets.

We also considered the graphical measure neighborhood hit (NH) shown in Figure 7 aimed at evaluating group separation and distance preservation, respectively. NH curves show global neighborhood preservation by the layout, computed over a range of neighborhood values, with a value 1 indicating 100% preservation [10]. The NH curves relative to the projections of the four data sets exhibit a pattern. LSP and Tree display top precision in all plots, indicating that they reconstitute class neighborhood consistently. In the case of CBR there is no statistical difference with the performance of PCA and LSP. For Corel there is no statistical difference between Isomap, Tree, and LSP. The Medical data set has the lowest average NH curve, which confirms its lower separability. Glimmer's distance plots are as good as others for the image data (Medical and Corel), but it has worse NH measurements in all cases.

There is mostly agreement between the numerical measurements just discussed and the findings of the perceptual evaluation, but a few discrepancies deserve noting. First, one easily notices that the Tree layouts did not perform as well perceptually as in the numerical and graphical evaluations. That is not surprising, since the tree branches have been removed from the layouts shown in the study (we wanted to test the tree's ability as a point placement strategy). Nonetheless, many similarity interpretation tasks on an NJ-tree, such as assessing group formation and group similarity, depend on the branches rather than to spatial proximity on the plane only: points spatially close but placed in distinct branches may actually be quite dissimilar. The tree still performed well in cluster proximity tasks for documents and cluster density tasks for images. It also championed estimation of number of clusters for image data and cluster proximity for document data.

Also worth mentioning is that, while Glimmer did not perform well numerically in terms of segregation, and did mostly well in terms of distance, it championed many questions related to clutter density. In fact, Glimmer has the best clutter ratio, favoring tasks that required recognizing densities, a particular aspect of the projections not reflected in the measurements employed.

PCA performed mostly as expected, doing badly on segregation tasks due to the difficulties of group separation on a reduced space of only two dimensions. Nonetheless, it still

performed well in tasks related with detecting subclusters and detecting outliers. The CBR data influenced that result, since it produces the best PCA segregation, but, even if the clusters are badly estimated, the recognized clusters can still be well analyzed regarding outliers and proximity of groups. PCA also produced the fastest responses, though more often than not leading to wrong answers.

LSP and Isomap resulted in the best overall performances, both numerically and perceptually. Perceptually, LSP and Isomap were quite often among the winners (five tasks) and rarely among the losers (two tasks) when averaging over all data sets. Isomap was in the middle to top range in most numerical evaluations and LSP close to the top in most. When dismissing Tree, they were best both perceptually and numerically.

8 Conclusion

We have conducted a controlled study to evaluate how subjects perceive multidimensional data projection layouts. We compared layouts obtained with five projection methods on data sets with distinct characteristics in terms of sparseness and distance distribution, considering tasks of multiple natures. In particular, we considered tasks related with specific properties of projection layouts, namely group segregation and separability, distance preservation and outlier detection, and clutter avoidance.

Our results confirm the intuition that no projection technique is capable to perform equally well on the different types of tasks. Moreover they indicate that performance is dependent on data characteristics, particularly in tasks that require distance interpretation. Considering the set of tasks globally, the best overall subject performance was obtained on Isomap and LSP layouts, but still other techniques did better than these two on some tasks. Glimmer layouts resulted in poor performance in group segregation tasks, but very good performance on clutter avoidance. PCA was particularly poor on segregation and clutter avoidance, but good in distance preservation in certain situations, e.g., when clutter was low. Regarding the experiment of using the NJ tree as a projection layout, it becomes evident that the branches are actually required for correct layout interpretation in some tasks.

As reported in the literature [23], we are convinced that density and surrounding information affects the perception of clusters and that perception is an important aspect that is not captured in numerical estimates of projection quality. Still, we compared our findings with common numerical estimates of layout segregation and distance preservation capabilities. The analysis confirmed that better segregation and distance preservation are more easily achieved on less sparse data spaces. It also confirmed the observation that Glimmer and PCA do not favor group segregation. However, numerical estimates of distance preservation are good for Glimmer, although it did not perform so well on the corresponding tasks in the study.

Our findings and derived respective guidelines can be useful to analysts selecting from projection techniques to perform specific visualization tasks on data sets of a particular type. Moreover, we believe this work can provide a starting point for further research on the role of perception in multidimensional data projections, including investigations of visual encodings. Future studies may extend this initial step by investigating more data sets, data with other characteristics, additional projection methods, additional tasks, or tasks for specific applications. Moreover, it is an interesting challenge to explore which perception rules (e.g., Gestalt laws) apply and influence the outcome and which cognitive processes are involved. Visual attention captured by eye trackers can be a valuable source of information for such analyses.

Acknowledgments

This work was supported by VisComX (Jacobs University), CNPq and FAPESP (Brazil), and CAPES/DAAD (34/10).

References

[1] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*, 1st ed. Springer Publishing Company, Incorporated, 2007.
[2] L. J. P. V. der Maaten, E. O. Postma, and H. J. V. D. Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, pp. 1–41, 2009.
[3] I. T. Jolliffe, *Pincipal Component Analysis*. Springer-Verlag, 1986.
[4] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling Theory and Applications*, 2nd ed., ser. Springer Series in Statistics. Springer, 2010.
[5] W. Torgerson, "Multidimensional scaling: I. theory and method." *Psychometrika.*, vol. 17, no. 4, pp. 401–419., 1952.
[6] P. A. Eades, "A heuristic for graph drawing," in *Congressus Numerantium*, vol. 42, 1984, pp. 149–160.
[7] 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 *Proc. IEEE Symposium on Visual Analytics Science and Technology*. IEEE Computer Society, 2007, pp. 99–106.
[8] J. G. S. Paiva, L. F. C., H. Pedrini, G. P. Telles, and R. Minghim, "Improved similarity trees and their application to visual data classification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2459–2468, December 2011.
[9] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston, MA, USA: Addison-Wesley Longman, 2005.
[10] 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*, vol. 14, no. 3, pp. 564–575, 2008.
[11] X. Geng, D. C. Zhan, and Z. H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35 Issue:6, pp. 1098 – 1107, 2005.
[12] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7, pp. 1431–1443, 2009.
[13] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castrén, "Trustworthiness and metrics in visualizing similarity of gene expression," *BMC bioinformatics*, vol. 4, no. 1, p. 48, 2003.
[14] J. A. Lee and M. Verleysen, "Scale-independent quality criteria for dimensionality reduction," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2248–2257, 2010.
[15] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *The Journal of Machine Learning Research*, vol. 11, pp. 451–490, 2010.
[16] M. Aupetit, "Visualizing distortions and recovering topology in continuous projection techniques," *Neurocomputing*, vol. 70, no. 7, pp. 1304–1330, 2007.
[17] S. Lespinats and M. Aupetit, "Checkviz: Sanity check and topological clues for linear and non-linear mappings," in *Computer Graphics Forum*, vol. 30, no. 1. Wiley Online Library, 2011, pp. 113–125.

[18] E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: An overview and systematization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2203 –2212, dec. 2011.

[19] A. Tatu, P. Bak, E. Bertini, D. A. Keim, and J. Schneidewind, "Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data," in *Proc. Working Conference on Advanced Visual Interfaces (AVI '10)*, 2010, pp. 49–56.

[20] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan, "Selecting good views of high-dimensional data using class consistency," *Computer Graphics Forum (Proc. EuroVis)*, vol. 28, no. 3, pp. 831–838, 2009.

[21] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. A. Keim, "Combining automated analysis and visualization techniques for effective exploration of high-dimensional data," in *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST '09)*, 2009, pp. 59–66.

[22] J. M. Lewis, L. van der Maaten, and V. R. de Sa, "A behavioral investigation of dimensionality reduction," *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pp. 671–676, 2012.

[23] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," *Computer Graphics Forum (Proc. EuroVis)*, vol. 31, no. 3, pp. 1335–1344, 2012.

[24] J. M. Lewis, M. Ackerman, and V. de Sa, "Human cluster evaluation and formal quality measures: A comparative study," in *Proc. 34th Annual Conference of the Cognitive Science Society*, . R. P. C. N. Miyake, D. Peebles, Ed., 2012, pp. 1870–1875.

[25] G. Albuquerque, M. Eisemann, and M. Magnor, "Perception-based visual quality measures," in *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST)*, Oct. 2011, pp. 13–20.

[26] R. Rensink and G. Baldridge, "The perception of correlation in scatterplots," *Computer Graphics Forum (Proc. EuroVis 2010)*, vol. 29, pp. 1203–1210, 2010.

[27] ——, "The visual perception of correlation in scatterplots," *Journal of Vision*, vol. 10, no. 7, p. 975, 2010.

[28] R. Rensink, "The rapid perception of correlation in scatterplots," *Journal of Vision*, vol. 11, no. 11, p. 1085, 2011.

[29] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric faramework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[30] S. Ingram, T. Munzner, and M. Olano, "Glimmer: Multilevel mds on the gpu," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 2, pp. 249–261, 2009.

[31] N. Andrienko and G. Andrienko, *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[32] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner, "Dimensionality reduction in the wild: Gaps and guidance - ubc computer science technical report tr-2012-03," The University of British Columbia, Tech. Rep., 2012.

[33] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," in *Proc. 19th Int. Conference on Machine Learning*, ser. ICML'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 307–314.

[34] K. Wagstaff and C. Cardie, "Clustering with instance-level constraints," in *Proc. 17th Int. Conference on Machine Learning*, 2000, pp. 1103–1110.

[35] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Proc. 17th. Int. Conference on Machine Learning*, ser. ICML'00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 727–734.

[36] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Addison-Wesley, 2006.

[37] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, pp. 1–21, 1969.

[38] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. ACM SIGMOD Int. Conference on Management of Data*, ser. SIGMOD'00, New York, NY, USA, 2000, pp. 427–438.

[39] H. Lukashevich, S. Nowak, and P. Dunker, "Using one-class svm outliers detection for verification of collaboratively tagged image training sets," in *Proc. 2009 IEEE Int. Conference on Multimedia and Expo*, ser. ICME'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 682–685.

[40] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, 1996, pp. 226–231.

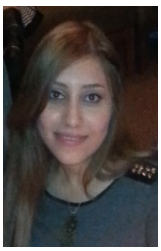
[41] M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proc. ACM SIGMOD Int. Conference on Management of Data*, 2000, pp. 93–104.

[42] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, Sep. 1956.

[43] C. R. Loader, "Local likelihood density estimation," *The Annals of Statistics*, vol. 24, no. 4, pp. 1602–1618, 1996.

[44] D. W. Scott and S. R. Sain, *Multi-Dimensional Density Estimation*. Amsterdam: Elsevier, 2004, pp. 229–263.

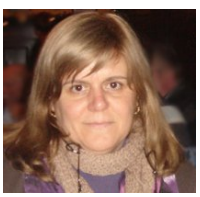
[45] M. Sedlmair, T. Munzner, and M. Tory, "Empirical guidance on scatterplot and dimension reduction technique choices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2634–2643, 2013.



Ronak Etemadpour is a postdoctoral researcher at University of Arizona, US. She received her PhD from Jacobs University Bremen, Germany. She is interested in perceptual aspects of visualization, and data analysis tasks. Visual effectiveness, efficiency and user satisfactions have been considered in her studies. She also has conducted controlled user studies that assess the user's visual attention.

Robson Motta is a PhD candidate and received the MSc degree from the Universidade de São Paulo in São Carlos, Brazil. He is interested in information visualization, data mining, and big data analytics.

Jose Gustavo de Souza Paiva is an assistant professor at the Federal University of Uberlandia, Uberlandia, Brazil. His research interests include information visualization and visual data classification, analyzing the role of the produced layouts on the perception and comprehension of the classification process.



Rosane Minghim is an Associate Professor at Universidade de São Paulo in São Carlos, Brazil. She is interested in all aspects of Visualization, Information Visualization, Visual Analytics and a wide variety of applications.

Maria Cristina Ferreira de Oliveira is currently a Professor at the Computer Science Department of the Instituto de Ciencias Matematicas e de Computacao, at the University of São Paulo, Brazil. Her research interests are in visual analytics and visual data mining techniques and applications.

Lars Linsen is a Full Professor of Computational Science and Computer Science at the Jacobs University, Bremen, Germany. He received his academic degrees in Computer Science from the Universität Karlsruhe (TH), Germany, was a postdoctoral researcher and lecturer at the University of California, Davis, U.S.A., and an Assistant Professor at the Ernst-Moritz-Arndt-Universität Greifswald, Germany, before joining Jacobs University. His research interests are in Data Visualization.