Evaluating Density-based Motion for Big Data Visual Analytics

Ronak Etemadpour
School of Information
University of Arizona
retemadpour@email.arizona.edu

Paul Murray
Dept. of Computer Science
University of Illinois at Chicago
pmurra5@gmail.uic.edu

Angus Graeme Forbes
Dept. of Computer Science
University of Illinois at Chicago
aforbes@uic.edu

***Abstract*—A common strategy for encoding multidimensional data for visual analysis is to use dimensionality reduction techniques that project data with a very large number of objects and dimensions from higher dimensions onto a lower-dimensional space. In visual analytics tasks, the *density* of the multidimensional clusters can strongly affect how these clusters are perceived. However, this feature can be lost when that dataset is projected into a 2D space, adversely affecting the effectiveness of visual analytics tasks. Thus, it makes sense to preserve, as far as possible, information about the density during the dimensionality reduction. This paper is a study of motion-enhanced cluster perception where the clusters are shown in 2D scatterplots and cluster density is mapped to the motion of the individual constituent points. We consider different types of *density-based motion*, where the magnitude of the motion is directly related to the density of the clusters. We conducted a series of user studies with large datasets to investigate how motion is a powerful perceptual cue well-suited for grouping or segmenting types during perceptual tasks. We found that the use of motion enabled users to be easily able to distinguish between clusters with different densities. The amount of visual change per unit time was different for the different motions, and we describe the ranges and thresholds for each of them. Specifically, we looked at two projection techniques that output 2D scatterplots for a range of data analysis tasks. We focus on high-dimensional, real-world datasets that might require analyses involving cluster identification, similarity seeking, and cluster ranking tasks. Our results indicate that incorporating density-based motion into visualization analytics systems effectively enables the exploration and analysis of multidimensional datasets.**

***Keywords*-Big Data, multidimensional data analysis, high-dimensional data, projection methods, density-based motion, user evaluation.**

I. Introduction

To facilitate data analysis tasks, multidimensional reduction techniques map high-dimensional data onto a lower-dimensional visual space in form of 2D or 3D scatterplots. Projections using different algorithms generate scatterplots with particular point placements as the most common visual encoding. Typically, two-dimensional visual encodings displayed as scatter plots use similarity-based or distance preservation layouts, utilizing a properly defined distance metric in the given multidimensional attribute space. Data analysis tasks are primarily concerned with the detection of structures, patterns, groups, and similarities with the data. Within a multidimensional dataset, data points can be grouped manually into classes or automatically into clusters. However, multidimensional projection mappings are especially prone to distortion because projection methods may not necessarily preserve the spatial relations of the data. Several measurements have been introduced to assess projection methods with respect to properties such as cluster preservation and separation (or segregation) [38]. Projecting the higher dimensions onto 2D spaces introduces some loss of information that causes difficulties for some data analysis tasks such as pattern identification, similarity seeking, and cluster rankings [34]. Healey argues that the "strongest" feature should be used to encode the relationships that are most relevant to a user's task [13]. As documented by Etemadpour et al. [7] and Sedlmair et al. [37], density strongly affects the perception of clusters. Since density is an important feature of the multidimensional data, it makes sense to preserve, as far as possible, this information during the dimensionality reduction.

While "Big Data" commonly refers to datasets that contain a very large number of data points, the term also often implies the use of datasets with a very high amount of dimensionality. In this paper, we explore how augmenting visualization analytics techniques with motion makes it easier to make sense of high-dimensional data. Specifically, we explore various methods to encode density using *density-based motion*. We demonstrate how tasks involving clusters, such as pattern identification, similarity seeking, and ranking, can be enhanced by motion. Usually, in a static depiction, the proximity of data points is used for clustering. Adding motion to the visualization can emphasize these clusters, making them easier to distinguish. We also believe that motion could be effective for uncovering data points that become visually cluttered when projected into a lower dimensional space. As noted by by Bartram, et al., since motion does not seem to interfere with existing color and form coding [5], using motion to represent density would allow visualization designers to communicate extra information using different modalities to represent other aspects of the data. On the other hand, the human visual system groups elements that are physically separated but that are similar to each other. Thus, *common-fate* grouping can be used as a process of grouping. Levinthal and Franconeri [22] showed that the visual system is effective at searching for motion-linked

groups among non-linked objects. For these reasons, we believe that motion could be a useful modality for assisting in visualization tasks related to projected multidimensional clusters. However, detailed perceptual guidelines on the use of motion in high-dimensional data projections have not yet, to our knowledge, been documented.

We describe and analyze two experiments that study the effects of moving clusters on human perception over multidimensional data projections; in each of the experiments we investigate different types of motion ("circular," "wiggle," and "pulse"). In the first set of experiments we analyze a user's perception when he or she is given typical analysis tasks for 2D scatterplots that have been generated synthetically. In this user study, we evaluated the ability of users to perceive clusters and their relationships via density-based motions. In a second user study, using real-world multidimensional datasets from two different domains (an image collection and a collection of documents) projected into 2D visual spaces, we evaluate the ability of users to discriminate between different magnitudes of motion. Users are asked to perform three main analysis tasks on the resulting 2D scatterplots. To decide on the projection methods to investigate, we chose Isomap [42] as the representatives of MDS approach, and PCA [17] a classical dimension reduction strategy. PCA projection methods are usually more prone to distort relations within and between clusters. We also show that users perform well if clusters are mapped to density-based motions even when examining projected data that distorts their spatial relations. As Ware has stated [45], understanding the perceptual processing of users can provide design guidelines for visualization systems. We draw conclusions on how the different density-based motions influence visual interpretation and how this supports or hinders effective task completion.

We provide a systematic user-centered examination of visual tasks related to projected multidimensional data. Our results show that users were able to perform visualization tasks more effectively using density-based motion, and, more specifically, that circular motions are especially effective for most tasks. Our results show improvements on visualization tasks related to the analysis of multidimensional data, including relation seeking and pattern identification tasks between or within clusters. Projection methods sometimes obscure particular patterns by grouping points close together. By giving users the ability to control the amount of motion in a visualization we augment these existing projection methods to make these points easier to perceive, enabling a more effective visual analysis of high-dimensional datasets.

II. Related Work

Ware et al. [47] shows that animation is a strong perceptive attention draw that consequently may distract people from their primary task. However, other studies have found that motion is a useful modality for encoding or augmenting data for information visualization tasks. It has been shown that some characteristics of animation may facilitate information-centric tasks and can be effectively used to show large amounts of information in a small space [24], [32]. Many other researchers have also examined motion in visual search tasks [10], [12], [20], [23], [33]. Motion can indicate a global movement of a single entity through, for example, animation of particles or glyphs that represent magnitude and orientation. Bartram et al. [5] describes an empirical investigation of use of variations in color, shape, and motion in information-dense displays to see how dynamic information is communicated from the system to the user. Their results showed that when motion was applied to a static glyph, even small linear oscillations were significantly easier to recognize than a change to the glyph's color or shape. Motion is perceptually efficient for visualizations incorporating multiple groups of data objects, and, in particular, circular motion is more easily perceivable but demands more attention that other types of motion [4].

Ware and Bobrow [46] suggest that the rapid visual querying of nodes is possible when using highlighting methods with interactive diagrams. In their investigation, evaluations were carried out with networks on moderately large node-link diagrams containing up to a few thousand nodes. These previous studies show that motion is a powerful perceptual cue that is effective for for a variety of perceptual tasks. Nonetheless it remains a relatively under-explored visual modality for practical applications, especially in relation to tasks important for the analysis of Big Data. Therefore, we suggest using motion for visual analytics tasks on high-dimensional data.

Many projection methods exist to generate 2D similarity-based layouts from a higher dimensional space, such as principal component analysis (PCA) and multidimensional data scaling (MDS). PCAs generate similarity layouts by reducing data to lower dimensional visual spaces [17]. MDS 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 while maintaining pairwise distances between points [6]. Some projection methods, such as isometric feature mapping (Isomap), favor maintaining distances between clusters instead. It replaces the original distances by geodesic distances computed on a graph to obtain a globally optimal solution to the distance preservation problem [42]. A number of studies have explored numerical methods to evaluate the quality of layouts [38], [39], [27], [11]. However, Etemadpour et al. [8] shows that no projection technique is capable of performing equally well on every type of task; performance is also dependent on specific data characteristics. Considering the set of tasks globally, the best overall subject performance was obtained on Isomap layouts, and PCA has problems with cluster segregation and led to mis-interpretations in a projected data. Thus, the correlations of data points

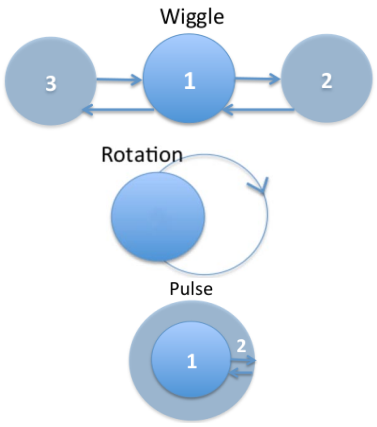


Figure 1. Instances of motion stimuli: *wiggle*, a combination of linear transitions; *rotation*, or circular motion; *pulse*, an in place expansion/contraction motion.

or clusters are not always known after they have been mapped from a higher dimensional data space to 2D display space. How well groups of points can be distinguished by users, or visual class separability, is investigated in different studies [1], [36], [41]. Sedlmair et al. [35] investigates the accuracy of class density measures in multidimensional projection 2D layouts. Etemadpour et al. [9] investigates the role of visual attention and guidance of attention for 2D projection layouts from the user's perspective. Rensink and Baldridge [30] investigate the perception of correlation in scatterplots from a psychological perspective. They explore the use of simple properties such as brightness to generate a set of scatterplots and they found that perception of correlations in a scatterplot is rapid, and that in order to limit visual attention to specific information it is more effective to group features. Robertson et al. [31], in examining animated graphs, found that using motion as way to display trends over time did not lead to effective visual analysis amongst users. However, other researchers, including Kehoe et al., found the exact opposite [19]. Our work also indicates that motion, judiciously used, can be helpful for analysis tasks.

We created series of user studies to look at three different motions that are similar to motions that have been investigated by other researchers and, as discussed above, that represent simple motions readily perceivable by humans. We call these three motions: *wiggle*, *pulse*, and *rotation*. As shown in Figure 1, wiggle indicates a back-and-forth translation along the horizontal axis, rotation indicates the circular movement of a point while retaining its orientation, and pulse indicates a repeating expansion and contraction in scale. The magnitude of each of these movements is in proportion to the density of the cluster to which a point belongs. Here, velocity as one of the factors that change the magnitude is considered. Specifically, the velocity is correlated to the *inverse* of the cluster density. Thus, all points belonging to the same cluster have the same rate and magnitude, and the denser the cluster the less movement there is. Specifically, we investigated point cloud scatterplots without connectivity containing up to one thousand nodes generated from high-dimensional datasets. We investigated the use of each of these different motions to the static scatterplots as well contrasting the use of motion with static scatterplots. The static scatterplots already use proximity for clustering, thus the motion condition is used to augment this proximity feature. Bartram et al. [5] have stated that, unlike hue or shape discrimination, motion is well-suited to extracting information from "noisy" environments across the entire visual field. Thus, motion may be especially important in visualizations that are cluttered and difficult to extract information from, such as PCA projections of high-dimensional data.

III. Density-based motion and synthetic data

In the following sections, we introduce terms and tasks relevant to all of our experiments. Thereafter, we discuss specific user studies in more detail.

A. Definition of Cluster Density

The *density* of a cluster is defined via a minimum spanning tree algorithm that is applied to that cluster. This gives us a optimal set of edges each with the shortest possible length. Etemadpour et al. [8] define the density as the number of points in the cluster divided by the sum of lengths of the edges in the spanning tree:

$$density = n_p / \sum_{i=1}^{n_e} length(e_i)$$

where n_p is the number of points in the cluster; n_e is the edges created by the minimum spanning tree algorithm; and e_i indicates the ith edge in the spanning tree. A minimum spanning tree is an especially good solution for multidimensional datasets because it considers distances in high-dimensional space, scales particularly well when data contains many dimensions, and is not sensitive to differences in the shape of clusters.

B. Detection of Discrimination

Huber and Healey [16] explored the perceptual properties of flicker, finding lower bounds of differentiability in terms of frequency, direction, and velocity of flicker motions. They showed that that minimum visual differences are needed for flicker, direction, and velocity. But target flicker must be coherent with the background. They have studied the cycle length as the duration of the target element's cycle in milliseconds that investigate the viewer's ability to distinguish the presence or absence of a small group of target elements that flicker at a rate different from the background elements. Therefore, in the first study, we wanted to investigate the ability of users to detect and to discriminate motion information that is based on structural characteristics related to the

density of clusters. We investigated the minimum difference in magnitude (velocity) that enabled users to distinguish between similar clusters.

An Apple iMac with a 21.5" screen was used to display the scatterplots via interactive web pages (served through a locally running server) that also collected the user responses. The system started immediately with the task and its proper image after a short demographic question. The users were presented with a sequence of either still or moving images displaying the respective scatterplots. For each image they were asked to answer the question as soon as they knew the answer and to act as quickly as possible, although we did not limit the time. Specifically, this experiment studied a viewer's ability to distinguish a small group of similar points within a cluster (target elements) that move at a rate different from the other points. Nine clusters were shown overlaid onto a 3×3 grid. Each cluster was roughly centered over one cell in the grid. When participants moved the mouse over one of the clusters, that cluster was highlighted, as shown in Figure 2, where the bottom left cell is selected, highlighting one of the clusters. The magnitude of motion was the same for all but one of the clusters. The participants were asked to click on the cluster of points that appeared to have a magnitude of motion that was different from the background clusters as soon as they could identify it. They have been asked to choose the cluster that moves in a different rate once they detect it visually without using the highlighting method. By asking this, we tried to minimize the effects of highlighting that confound results.

As stated above the velocity of motion for each element in a cluster was proportional to the inverse of that cluster's density; two clusters with quite similar densities reveal similar motion information. Therefore, this section summarizes some of our main findings about an individual's ability to discriminate stimuli that involve a similar kind of movement, but with different magnitudes. In our formulated algorithms below, the target elements and background elements complete a whole cycle in a same time. Thereby, viewers can perceive a difference between the target and the background motion rates if the target elements complete a cycle with higher velocity. This velocity is teasing apart a δ value that changes the density of clusters. Direction, path curvature is similar for the target elements and background elements in order to keep them coherent.

The mapping functions for 4 different types of motion were used in the different synthetic scatterplots. The absolute difference in the magnitude of motion (velocity here) between the points in the target cluster and the points in the background clusters is defined for the different motions like so:

$$\textit{wiggle}: \begin{cases} x_t = \sin(t) \times \delta/d_n \\ y_t = \sin(t) \times \delta/d_n \\ \delta \in \lfloor d \rfloor \times \{2, 3, 4, 6, 7, 8, 8.5, 9, 9.8, 10\} \end{cases}$$

$$\textit{pulse}: \begin{cases} r_t = \sin(t) \times \delta/d_n \\ \delta \in \{3, 3.6, 4, 5, 6, 7, 8, 9, 10, 11\} \end{cases}$$

$$\textit{rotation}: \begin{cases} x_t = \sin(t) \times \delta/d_n \\ y_t = \cos(t) \times \delta/d_n \\ \delta \in \lfloor d \rfloor \times \{1, \lfloor d \rfloor/2, \lfloor d \rfloor, 3\lfloor d \rfloor/2, 5\lfloor d \rfloor/2, \\ \qquad 3\lfloor d \rfloor, 2\lfloor d \rfloor, 7\lfloor d \rfloor/2, 4\lfloor d \rfloor\} \end{cases}$$

where d is the density of the cluster and d_n is the density normalized within the interval $(0, 1]$. 10 different δ values have been examined for each motion leading to a total of 112 stimuli, similar to the one shown in Figure 2. 35 students participated in this study and a between comparison strategy was followed.

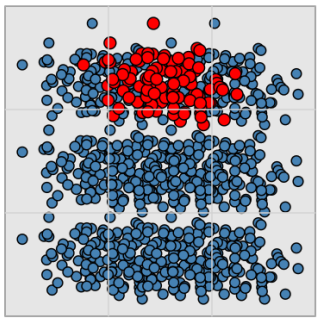


Figure 2. Detections of discrimination: Clicking on the one cluster with a different magnitude of motion.

1) Results: We originally chose our δ values based on an initial estimation of motion rate change. Because these estimations are arbitrary and not quantitatively based, we fit them to a psychometric function for detection of differential item functioning in order to determine the mean success results [43]. Moreover, we were interested in finding the smallest δ value while maintaining high accuracy. We anticipated that identification would be faster and more accurate when the target cluster was moving faster (i.e., when we used a larger δ value). Thus, the Weibull function [25] using a "maximum likelihood" procedure as a continuous probability distribution is used. The coherence level that predicts 80% correct performance as an acceptable success rate is picked to determine the values for each motion's success rate. Figure 3 summarizes the results using the fitting function. For *wiggle*, $\delta = 5.4978$ predicts 80% correctness that the parameters maximize the log likelihood. Similarly, $\delta = \lfloor d \rfloor \times 3.54$ for *wiggle* can be considered as the lower bound for an effective distinction between magnitudes. For *rotation*, $\delta = \lfloor d \rfloor \times 0.608\lfloor d \rfloor$ has the higher performance in terms of success rate.

These results provide information regarding which densities can be easily perceived within a scatterplot that encodes density using motion. A viewer's ability to distinguish the presence of a small group of target elements that move at a velocity different from background elements increased with higher values.

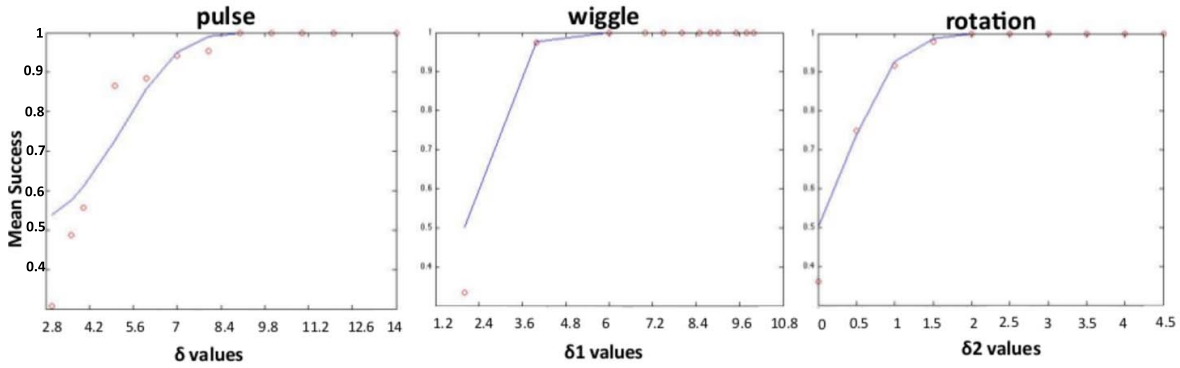


Figure 3. Fitting with a parametric function (the Weibull function): for *pulse* $\delta = 5.4978$ predicts 80% correctness; for *wiggle* $\delta 1 = 3.54$ predicts the 80% correctness and $\delta = \lfloor d \rfloor \times 3.54$ is desired; for *rotation* $\delta 2 = 0.608$ predicts the 80% correctness and $\delta = \lfloor d \rfloor \times 0.608 \lfloor d \rfloor$ is our desired significant value.

IV. Density-based motion and multidimensional datasets

In addition to investigating motion on scatterplots created from synthetic data, we also examined the use of motion in real-world datasets.

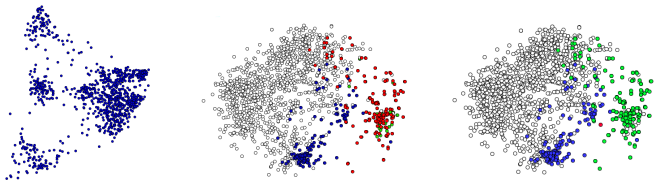
A. Datasets and Projections

We used a document collection and an image collection. Textual datasets generally have a high dimensionality even when the data is relatively sparse. We chose the KDViz[1] dataset as representative of document datasets with high dimensionality. It contains documents collected from an Internet repository related to four different topics with 1,624 unique documents, 520 different dimensions, and 4 highly unbalanced labels.

Image datasets generally have a lower dimensionality and are sensitive to the choice of the feature space. We use the Corel dataset[2] as representative of image datasets. The Corel dataset includes 1,000 photographs related to ten different themes, and each photograph is described by 150 dimensions (i.e., their SIFT descriptors). We chose these datasets because of their high dimensionality, an aspect of "Big Data" that sometimes is ignored.

We have selected two techniques as representatives of two distinct strategies for embedding data in two dimensions, namely statistical dimension reduction (PCA [17]), and MDS (Isomap [42]). PCA is a classical dimension reduction strategy often employed to generate visual embeddings of data. 2D layouts are obtained by considering the two first principal components (at the risk of disregarding other potentially relevant components). Isomap is a variant of MDS that builds a weighted nearest-neighbor graph from the data, with pairwise point distances as edge weights. The distance between two points in this graph creates the shortest path. Cosine distance is the usual choice for text data and we use this in our study for both projection techniques when examining KDViz. For the Corel dataset, the choice of the distance function was made based on the best point segregation on 2D projections, consequently Cosine distances were chosen for PCAs and Euclidean distances were chosen for the Isomap projections.

[1]http://vicg.icmc.usp.br/infovis2/DataSets
[2]UCI KDD Archive, http://kdd.ics.uci.edu



(a) Count Cluster (b) Rank Density (c)Association

Figure 4. Instances of task stimuli: (a) Estimate number of clusters, (b) rank red, green, and blue clusters by density, (c) determine whether green or blue cluster is similar to red object.

B. Analysis Tasks

We identified typical questions raised when visually analyzing multidimensional data in order to define representative user tasks. The visual analytics tasks on high-dimensional data can be usefully categorized as pattern identification, behavior comparison, and relation seeking [2], [8]. First, we have been looking into *pattern identification* tasks, where the targets were the detection of clusters within a given point distribution in a scatterplot. The *behavior comparison* task asks users to rank point densities with clusters, in effect comparing characteristics of the clusters. The *relation seeking* task asks users to determine which cluster a particular point or object is more similar to. Three different tasks as representative of each of these categorizations are used as:

Count Cluster Estimate the number of observed clusters.
Rank Density Rank the clusters by density.

Association Identify which cluster a given point is most similar to.

For task **Count Cluster**, the static layouts are either color-coded based on cluster labels or shown with a single color because we did not want colors to distract a user's attention from the main task of counting clusters. For task **Rank Density**, the colors were assigned randomly to the given clusters in order to decrease the chance of inadvertent associations with colors. Three clusters were shown in three different main colors (red, green, blue). For task **Association**, each of the two clusters were color-coded either green or blue, while the reference point is given a red color. Figure 4 shows an example stimulus for each task. For both **Association** and **Rank Density**, to avoid bias, we randomly assigned colors to each cluster.

C. Hypotheses

We formulate the following hypotheses:

H1 Density-based motion will improve user performance of **Count Cluster**, **Association**, and **Rank Density** on real-world projections.

Since both motion and color are handled by a dedicated visual processing mechanism [3], we anticipated that motion, as a preattentive visual feature, would perform as well as color.

H2a Pairwise comparisons between detection accuracy using color-coded clustering and density-based motion will not deliver any significant difference for **Count Cluster**.

H2b Pairwise comparisons between response times using color-coded clustering and density-based motion will not deliver any significant difference for **Count Cluster**.

D. Computation of Errors

Given the ground-truth, we compute the errors in the answers given by the subjects for each task. For **Count Cluster**, which required the subjects to estimate a number, the error percentage is computed by

$$e = \frac{|n_{true} - n_{answer}|}{n_{true}} \times 100,$$

where n_{true} is the estimated ground truth and n_{answer} is the reported answer. For **Association**, which required a cluster to be identified, the error is either zero or one. For **Rank Density**, which required the user to rank clusters, we first calculated the number of changes required to get from the user's reported answer to the ground truth. Each cost of transformation was then calculated as the absolute value of the difference between the densities of the clusters involved in the transformation. This was then divided by the sum of the densities of all three clusters, in order to normalize the value relative to the "worst case" answer, in which all three rankings must be swapped. For example, if the correct ranking was (c_1, c_2, c_3), and the user reported (c_2, c_1, c_3), one transformation is needed to get from the user's response to the correct answer, namely, exchanging c_2 and c_1. If the cluster densities for (c_1, c_2, c_3) were $(20, 2, 12)$, respectively, the user's error would be calculated as:

$$e = \frac{|density_{c_2} - density_{c_1}|}{\sum(density_{c_1}, density_{c_2}, density_{c_3})} = \frac{|20 - 2|}{20 + 2 + 12}$$

for a total error of 0.53. At the end of this calculation we further multiply by 100 in order to make it easier to read.

E. Investigations and Statistical Methods

For the statistical analysis of the results of the user study, we compared three types of motion and static scatterplots with no motion for each task individually. By looking into the mean errors (as computed using the methods described in Section IV-D) over all subjects and all datasets, we tested the distribution of the error values against normality using the Kolmogorov-Smirnova and the Shapiro-Wilk tests. In the case of non-normal distribution, we used the Friedman test on K related samples when comparing more than two groups. We also performed pairwise comparisons of the groups using a Wilcoxon test on the results at the 0.05 level to be able to report which pairs of groups differ from each other significantly. The Kruskal-Wallis is used as a non-parametric method for comparing more than two samples that are independent. In case of normal distribution, we used a t-test when comparing two groups and an ANOVA test when comparing more than two groups. For pairwise comparisons, in cases where there were more than two groups we ran a series of Tukey's post-hoc tests. In addition to the mean error, we also evaluated each participant's confidence ratings (on a five-step Likert scale) as well as the time it took for each participant to fulfill the tasks.

F. Set-up for User Study

We have created an interactive multidimensional data projection tool for our experimental study that is based on the concept of density-based motions. Our tool allows us to create a series of views, each of which uses either the KDViz or Corel dataset, the PCA or Isomap projection, and one of the three motions we are exploring. For instance, Figure **??** (top) shows a view using an Isomap projection applied to the Corel dataset using rotation. By changing a slider at the bottom of the tool, the user is able to interactively increase or decrease the magnitude of the motion. This slider updates a magnitude factor F which alters the motions described in Section III-B. For *wiggle* and *rotation*, F alters the range of the x and y coordinates of the points. For *pulse*, F alters the maximum radius of the point size.

We conducted a controlled user study involving 12 subjects who were students or researchers in computer science

or medicine. The primary task area was approximately 6"×6", with a margin approximately 3" from the left side of the stimulus window and 1" from the top. The stimulus subtended a field of view of approximately 24° to 27° of visual angle from the center of the stimulus window. We considered multidimensional data analysis tasks similar to the ones described for the synthetic data user study in Section IV-B. Each subject was presented with a series of 28 different scatterplots of projected multidimensional data. For each of the scatterplots, we asked the participant to complete one of these three tasks. The presented images include both animated and static projections. For counterbalancing, a random function was used to shuffle the order of the presented images. Again, as defined in Section III-A, higher density clusters move less; more movement indicates a sparser cluster. For **Count Cluster**, three motions (*pulse*, *rotation*, and *wiggle*) in addition to static layouts were considered. The static layouts are either color-coded based on cluster labels or shown with a single color. For **Association**, the given object was shown in red and the two other clusters were colored green and blue. For **Rank Density**, three clusters were shown in three different main colors (red, green, blue). For both **Association** and **Rank Density**, to avoid bias, we randomly assigned colors to each cluster. Although 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 four different clustering techniques to determine cluster assignments for **Count Cluster**. We used the adjusted Rand index [29] to compare the similarity of different cluster assignments to the class labels given in the data. As other authors have stated, taking class labels can be used to generate pairwise constraints but it is arguable whether or not this is the best solution for understanding relations within multidimensional data [37], [21]. The Rand index measures the similarity of two different partitions of data (e.g. clusters or classes). Given that each partition assigns each element into one of many subsets, the Rand index calculates the amount of agreement between the two partitions. Four clustering techniques were considered: K-means, X-means [28], hierarchical agglomerative [26], and hierarchical divisive [18]. Each clustering technique was tested across a range of values of K (with the exception of X-means, which determines the optimal number of clusters on its own). For the Corel dataset, which contains ten classes, K values from four to twenty were used; For the the KDViz dataset, which contains four classes, K values ranging from two to ten were used. For each clustering technique, and for each value of K, the adjusted Rand index was calculated, and the cluster assignment with the highest adjusted Rand index was used to assign elements to clusters in our study. For instance, a Hierarchical Agglomerative clustering with $k = 12$ yielded the best Rand index (0.69) for the Corel dataset.

1) Correctness: Figure 5 summarizes the results for real-world data and each projection separately. The omnibus tests for statistical significance showed that there is statistical significance in the mean error for all tasks. The outcome of the pairwise significance test is indicated by the red horizontal lines. More precisely, groups of motions with no pairwise significant difference among their mean error have lines of the same color. Hence, we can conclude that density-based motion outperforms static projections. Thus, Hypothesis H1 is confirmed. We considered one numerical measure that measures the cohesion and separation between groups of instances on the layout. The Silhouette [40] of a projection is obtained by averaging the Silhouette coefficients of its n instances. Resulting values vary in the range $[-1, 1]$, with a values of 1 indicating that the groups are perfectly separated. Figure 6 shows Silhouette measurements for each dataset. The highest Silhouette values (red bars) were obtained by Isomap on Corel and KDViz. Corresponding Silhouette values show that PCA did not perform well on the KDViz dataset. However, for the Corel dataset, PCA improved the separability coded by the original space features. For KDViz, Isomap enhanced separability. Two principal directions were used to compute the PCA layouts, but were not capable of effectively separating either the $k = 12$ clusters in the Corel dataset, or the unbalanced classes in the KDVis dataset. Our statistical investigation showed that density-based motion has the additional advantage of adding extra information to the display that is especially helpful when the projections create cluttered clumps of points.

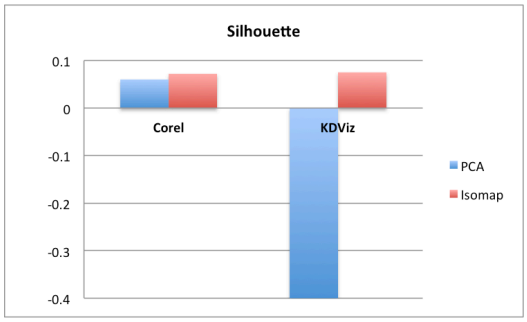


Figure 6. Silhouette Coefficient computed for the original data.

We asked users to estimate the number of clusters in color-coded projections without any motion (where the color of the projected data points were based on what cluster they belonged to). The results from a Wilcoxon Signed Ranks test did not reveal any statistical significant difference between motion and color ($Z = -0.089, p = 0.929$), Mean error in percent for motion was higher though (Mean error=13.5402) than Mean error for color (Mean error=10.9375). Therefore, we can confirm H2a. We also investigated how long it took for the subjects to complete the tasks in color-coded

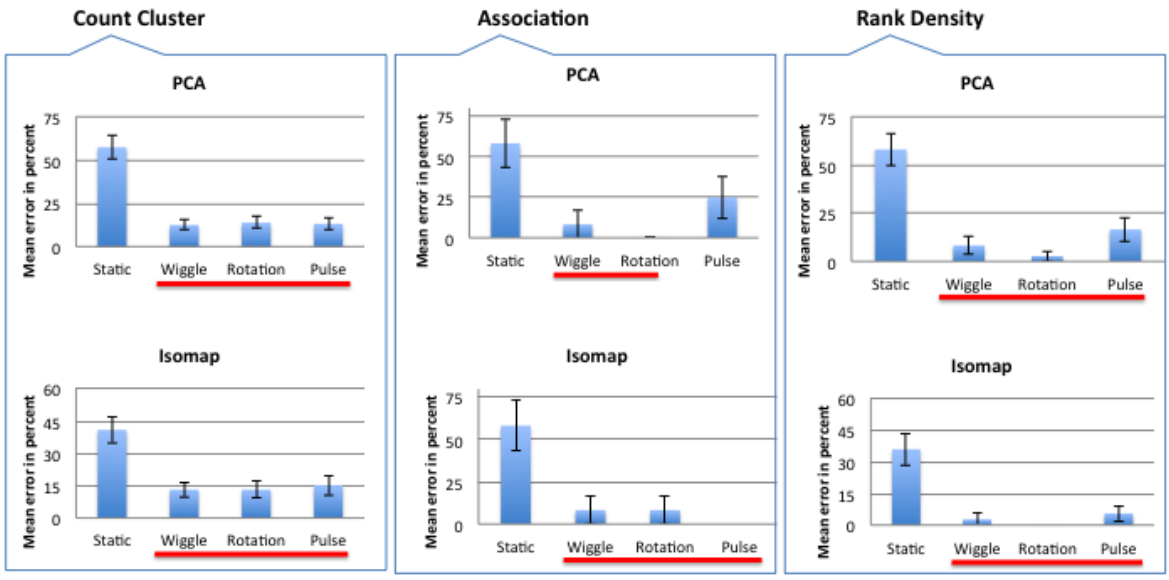


Figure 5. Here we show the results of comparing the projection methods on the tasks considered; the bar charts show mean error and standard error from the mean. There is statistical significance for all three tasks. The horizontal lines encode pairwise statistical significance using a red-to-white color transition. For example, the **Association**'s red line excludes the pulse motion because *wiggle* and *rotation* outperformed both the *pulse* motion and the *static* data.

projections and animated scatterplots. The statistical test did not show any significant different between color and motion when time is compared ($Z = -0.471, p = 0.638$). Finally, for confidence levels, comparison between color and motion did not reveal any significant differences ($Z = -0.192, p = 0.848$). Thus, H2b is also confirmed.

2) Confidence Levels: In regard to confidence levels, a one-way ANOVA test showed significant differences among all comparisons for **Count Cluster** ($F(3, 92) = 11.57, p < 0.05$). A Post-hoc Tukey showed significantly that the static projections had the lowest level of confidence ($Mean = 2.83$). For **Association**, again significant differences were seen ($F(3, 92) = 33.39, p < 0.05$) and Post-hoc Tukey confirmed that static projections had the lowest scores. For **Rank Density**, static projections also had significantly lesser confidence scores ($F(3, 92) = 11.601, p < 0.05$). These results are very consistent with the results we found in examining user accuracy of the tasks; that is, the users' confidence was warranted.

3) Time: Finally, we investigated how long it took for the subjects to complete the tasks. Findings for **Count Cluster** did not reveal any significant differences between motions and static projection ($\chi^2(3, 24) = 7.25, p = 0.064$). Similarly, for **Association** and **Rank Density**, a Friedman test also did not show any significant differences ($\chi^2(3, 24) = 1.05, p = 0.789$), ($\chi^2(3, 24) = 2.6, p = 0.457$). We can conclude that perception plays an important role in interpreting the scatterplots. For example, PCA had problems with cluster segregation and led to mis-interpretations in a static projected data. In particular, mapping of density-based motion can enhance the perception and user's performance significantly, where the cluttered layout were displayed.

We also performed a comparative analysis of motions and color-coded projection methods on two types of data, which had similar levels of accuracy. As Bartram et al. discussed, color is particularly well suited for categorization but less effective at showing other relations [5]. Our results showed that density-based motion can be an effective way to show the density and similarity relations in multidimensional data visualization. However, the cognitive costs associated with using color and motion simultaneously should be investigated because, as Healey stated, the various graphical codes may perceptually interfere with each other [14]. Nonetheless, our results indicate that motion can be used as an additional approach in order to enable users to effectively explore different aspects of data. A video demonstrating the different phases of our user study can be found on the authors' website, along with the full data collected from all 12 of the participants[3].

V. Conclusion and Future Work

We described a series of controlled user studies that evaluated how users perceive density-based motion in scatterplots. However, some papers indicated that there was no advantage to animations over static displays [44] because of their complexity. Hegarty [15] explores ways in which providing the user with interactive control of the visual representation can be a useful way to increase the effectiveness of a

[3]https://dl.dropboxusercontent.com/u/571874/MotionStudy.zip

display. Thus, in our real-world dataset analyses we gave users the ability to interact via changing a slider at the bottom of the tool to increase or decrease the magnitude of the motion. Cluster segregation, similarities, and behavior comparisons were considered. Three types of motion were chosen and the results confirm our general hypothesis that motion techniques perform well on different types of tasks. In the first experiment, motion as a low-level perceptual cue with a lower bound related to density was investigated to improve performance on similarity detection of data points and their associated clusters. This study created the best overall subject performance in enabling users to differentiate clusters. In the second part of our study, we formulated hypotheses for visual analyses of projected multidimensional data. We investigated the role of motion related to cluster characteristics (densities) in real-world data and the statistical tests confirmed those hypotheses. Multidimensional data representations are often visually very complicated. Our results showed that using a density-based motion not only is useful for representing clusters of data, but also that it can be potentially used as a means to more effortlessly inspect other interesting aspects of multidimensional data.

In the future, we plan to test the effectiveness of motions when allowing users to control the speed of movement, to specify at a certain region, to select or deselect motion methods, or to change colors. We would also like to design further user studies to explore other perceptual properties of motion, including frequency, amplitude, direction, and phase. For this paper, our evaluations focused on high-dimensional datasets, as high-dimensionality is an important and sometimes overlooked aspect of Big Data. However, we believe that density-based motion would be effective for datasets with a very large number of data points as well. Future work will explore density-based motion on projections of datasets that are made up of a very large number of samples within a very high-dimensional space.

ACKNOWLEDGMENTS

The authors would like to thank Kelland Thomas for his help in recruiting research participants for our user studies.

REFERENCES

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

[2] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer, 1 edition, Dec. 2005.

[3] L. Bartram. Perceptual and interpretative properties of motion for information visualization. In *Proceedings of the 1997 Workshop on New Paradigms in Information Visualization and Manipulation*, NPIV '97, pages 3–7, New York, NY, USA, 1997. ACM.

[4] L. Bartram, C. Ware, and T. Calvert. Filtering and integrating visual information with motion. In *In Proceedings on Information Visualization*, pages 66–79. Society Press, 2001.

[5] L. Bartram, C. Ware, and T. Calvert. Moticons: Detection, distraction and task. *Int. J. Hum.-Comput. Stud.*, 58(5):515–545, May 2003.

[6] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling Theory and Applications*. Springer Series in Statistics. Springer, 2nd. edition edition, 2010.

[7] R. Etemadpour, R. Carlos da Motta, J. G. d. S. Paiva, R. Minghim, M. C. Ferreira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. In *5th International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 107–113, Lisbon, Portugal, 2014.

[8] R. Etemadpour, R. Motta, J. de Souza Paiva, R. Minghim, M. F. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 99(PrePrints):1, 2014.

[9] R. Etemadpour, B. Olk, and L. Linsen. Eye-tracking investigation during visual analysis of projected multidimensional data with 2d scatterplots. In *5th International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 233–246, Lisbon, Portugal, 2014.

[10] A. G. Forbes, C. Jette, and A. Predoehl. Analyzing intrinsic motion textures created from naturalistic video captures. In *Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 107–113, Lisbon, Portugal, January 2014.

[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*, 35(6):1098–1107, 2005.

[12] S. Haroz and D. Whitney. Temporal thresholds for feature detection in flow visualization. In *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization*, APGV '10, pages 163–163, New York, NY, USA, 2010. ACM.

[13] C. G. Healey. Effective visualization of large multidimensional datasets. Technical report, 1996.

[14] C. G. Healey, K. S. Booth, and J. T. Enns. Harnessing preattentive processes for multivariate data visualization. In *In Proceedings graphics interface 93*, pages 107–117, 1993.

[15] M. Hegarty. The cognitive science of visual-spatial displays: Implications for design. *Topics in cognitive scinece*, 3:446–472, 2011.

[16] D. E. Huber and C. G. Healey. Visualizing data with motion. In *IEEE Visualization*, page 67. IEEE Computer Society, 2005.

[17] I. T. Jolliffe. *Pincipal Component Analysis*. Springer-Verlag, 1986.

[18] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

[19] C. Kehoe, J. Stasko, and A. Taylor. Rethinking the evaluation of algorithm animations as learning aids: An observational study. *International Journal of Human-Computer Studies*, 54:265–284, 1999.

[20] G. D. Kerlick. Moving iconic objects in scientific visualization. In *Proceedings of the 1st Conference on Visualization '90*, VIS '90, pages 124–130, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[21] 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 *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 307–314, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[22] B. R. Levinthal and S. L. Franconeri. Common fate grouping as feature selection. *Psychological Science*, 22(9):1132–1137, 2011.

[23] E. B. Lum, A. Stompel, and K. L. Ma. Kinetic visualization: A technique for illustrating 3d shape and structure. In *Proceedings of the Conference on Visualization '02*, VIS '02, pages 435–442, Washington, DC, USA, 2002. IEEE Computer Society.

[24] D. S. McCrickard, R. Catrambone, and J. T. Stasko. Evaluating animation in the periphery as a mechanism for maintaining awareness, 2001.

[25] U. Mortensen. Additive noise, Weibull functions and the approximation of psychometric functions. *Vision Research*, 42(20):2371–2393, 2002.

[26] F. Murtagh. *Multidimensional clustering algorithms*. Compstat lectures. Physica-Verlag, Vienna, 1985.

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

[28] 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*, ICML'00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[29] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[30] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Comput. Graph. Forum*, 29(3):1203–1210, 2010.

[31] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1325–1332, Nov. 2008.

[32] G. G. Robertson, S. K. Card, and J. D. Mackinlay. Information visualization using 3d interactive animation. *Commun. ACM*, 36(4):57–71, Apr. 1993.

[33] C. S. Royden and J. M. Wolfe. Visual search asymmetries in motion and optic flow fields. *Perception & Psychophysics*, pages 436–444, 2001.

[34] T. Schreck, T. von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9(3):181–193, June 2010.

[35] 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. Technical report, The University of British Columbia, 2012.

[36] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643, 2013.

[37] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Comp. Graph. Forum*, 31(3pt4):1335–1344, June 2012.

[38] 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 2009)*, 28(3):831–838, 2009.

[39] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley Longman, Boston, MA, USA, 2005.

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

[41] 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 *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '10)*, pages 49–56, 2010.

[42] J. B. Tenembaum, V. de Silva, and J. C. Langford. A global geometric faramework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[43] J. A. Teresi, M. Kleinman, and K. Ocepek-Welikson. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine*, 19(11-12):1651–1683, 2000.

[44] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: Can it facilitate? *Int. J. Hum.-Comput. Stud.*, 57(4):247–262, Oct. 2002.

[45] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[46] C. Ware and R. Bobrow. Supporting visual queries on medium-sized node-link diagrams. *Information Visualization*, 4:49–58, 2005.

[47] C. Ware, J. Bonner, R. Cater, and W. Knight. Simple Animation as a Human Interrupt. . *International Journal of Human-Computer Interaction*, 4:341–348, 1992.