Journal of Visual Languages and Computing 48 (2018) 144–156

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

View Points

Exploring high-dimensional data through locally enhanced projections

Chufan Lai[a], Ying Zhao[b], Xiaoru Yuan[*,a,c]

[a] Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, 100871, PR China
[b] School of Information Science and Engineering, Central South University, Changsha, HB 410083, PR China
[c] Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University, Beijing, 100871, P.R. China

ARTICLE INFO

Keywords:
Dimension-reduced projection
Local data analysis
High-dimensional data
Subspace analysis

ABSTRACT

Dimension reduced projections approximate the high-dimensional distribution by accommodating data in a low-dimensional space. They generate good overviews, but can hardly meet the needs of local relational/dimensional data analyses. On the one hand, layout distortions in linear projections largely harm the perception of local data relationships. On the other hand, non-linear projections seek to preserve local neighborhoods but at the expense of losing dimensional contexts. A sole projection is hardly enough for local analyses with different focuses and tasks. In this paper, we propose an interactive exploration scheme to help users customize a linear projection based on their point of interests (POIs) and analytic tasks. First, users specify their POI data interactively. Then regarding different tasks, various projections and subspaces are recommended to enhance certain features of the POI. Furthermore, users can save and compare multiple POIs and navigate their explorations with a POI map. Via case studies with real-world datasets, we demonstrate the effectiveness of our method to support high-dimensional local data analyses.

1. Introduction

Dimension-reduced projections are widely used for high-dimensional data analysis. They approximate distributions of high-dimensional data in low-dimensional spaces. Such approximations are often made as global ones that improve the overall mapping by striking a balance among all data. Two well-known examples are Principle Component Analysis (PCA) and Multidimensional Scaling (MDS). They generate good overviews of the data, but still unable to preserve all information without any loss [1]. Local distortion is an example of such loss, where inaccurate distance mapping may lead to unfaithful interpretations of data relationships [2,3].

Users are often not aware of the existence of distortions, and hence easily get misguided [4]. Even when distortions are informed [5,6], there are seldom interactive approaches for users to control them [7]. As a result, users often find it difficult to trust the projections [4,8]. On the other hand, users may wish to observe some local POI regions more precisely, while not so concerned about the other data. It inspires us to develop a dimension reduction scheme where users are able to decide which part of the projection is more precise and can be trusted.

Over the last few decades, different kinds of non-linear dimension reduction techniques have been developed to promote local data analyses [9–12]. Recent works further allow users to control the local mapping quality during a progressive rendering process [13]. Despite their abilities to preserve local structures, non-linear projections are not designed to visualize dimensional information. Cheng et al. [14] proposed to use interpolation and iso-contours for displaying attribute values. However, iso-contours may not be as simple and intuitive as axes in conveying dimensional semantics [15,16]. They are also prone to overlapping when the dimensionality increases. In this work, we choose to stay in the linear framework, since the linear projections provide intuitive dimensional semantics, are generally simple to use and interpret, and are also computationally efficient.

For the linear projections, Choo et al. proposed to preserve local structures by including supervised dimension reduction [17,18]. Along with similar works focusing on machine learning [19] and quality metrics [20,21], these approaches require the knowledge of data classification. It makes them unsuitable for general data explorations where no prior knowledge should be assumed. On the other hand, explorational methods allow users to manually adjust dimension weights of the projection [22–24]. However, the parameter search is often blind and time-consuming. It could be difficult to find a satisfying projection for a certain POI. Yuan et al. [25] proposed a framework where users are able to create new projections for data subsets. But the approach still requires a manual search of dimensional subspaces. In comparison, our approach generates projections and subspaces based on users' POIs and

* Corresponding author.
E-mail addresses: chufan.lai@pku.edu.cn (C. Lai), zhaoying@csu.edu.cn (Y. Zhao), xiaoru.yuan@pku.edu.cn (X. Yuan).

https://doi.org/10.1016/j.jvlc.2018.08.006
Received 21 July 2018; Accepted 11 August 2018
Available online 21 August 2018
1045-926X/

tasks. Users steer the projections by choosing their interests, without the need for any manual search.

In this paper, we propose an interactive scheme to help users customize a linear projection based on their POI data and analytic tasks. Specifically, users are able to specify a focus in the projection, which could be a single datum or a group of data. Then we offer multiple ways to alter the projection to enhance different features/aspects of the focus while maintaining the other data as contexts. Based on the locally enhanced projection, we further reveal dimensional subspaces that are most likely related to the features. It helps to interpret the features in the context of dimensions. In addition, we provide various means to support the data exploration at different stages. Users are assisted to discover, analyze, modify and compare different focuses. In summary, our contributions include:

- Given the user-defined POI data, we provide linear projections with enhanced POI features to support different kinds of local analytic tasks.
- Our method supports an interactive high-dimensional data exploration, where users are assisted to discover, analyze, modify and compare interesting pieces of local data.

The remainder of this paper is structured as follows. In the next section, we briefly review the related literature. Section 3 gives an overview of the proposed method. Then we elaborate each part of the method in detail in Section 4. Section 5 presents case studies to demonstrate the effectiveness of our method. In Section 6, we discuss weaknesses and potential improvements. At last, we end this paper with the conclusions.

2. Related work

Out method facilitates local data analysis in linear projections. We adopt the strategy of feature-driven projection pursuit [26,27], as opposed to dimension-driven methods [22–24]. We will briefly introduce the related works.

2.1. Data locality analysis in projections

Data locality has been extensively studied in high-dimensional data research. There are roughly two branches focusing on different aspects.

The major branch aims to improve global projections, focusing on preserving data localities. Many non-linear projections have been proposed for this purpose, such as Laplacian Eigenmaps (LE) [9], Locally Linear Embedding (LLE) [10], Local Tangent Space Alignment (LTSA) [11] and the T-distributed Stochastic Neighbor Embedding (t-SNE) [12]. These methods are fit for data lying on a low-dimensional manifold (e.g. face images of the same person), but the semantics of dimensions are lost. Cheng et al. [14] proposed to visualize attribute values using isocontours. But given their discrete natures, contours are not as intuitive as axes in conveying dimensional information [15,16]. They are also prone to overlapping when there are multiple layers. In comparison, our method keeps all projections in the linear framework. It helps users intuitively perceive and interpret the dimensional semantics of data relationships.

Another branch aims to reveal distortions in a projection. Martins et al. [28] examined distortions in different types of projections. They used color mapping to indicate distortion levels, and searched for real neighbors using automatic algorithms. Liu et al. took a step [6] further by analyzing data structures based on distortions. But none of them provide means to correct the distorted layout. Stahnke et al. [7] proposed a simple correction by directly mapping distances to the POI. Effective as it is, the approach loses dimensional contexts and is not suitable for situations where the POI is a group of data.

2.2. Projection assisted data exploration

Dimension reduced projections are often used to explore high-dimensional data. They are intuitive overviews, but hard to be changed interactively. Jeong et al. [22] proposed to change a projection by updating dimension weights in the PCA algorithm. Nam et al. [23] further enable users to freely decide the dimension components of a projection. Beyond parameter tuning, Lehmann et al. [24] proposed a more intuitive interaction, with which users can alter the dimension axes while maintaining an orthogonal mapping. These methods are indeed effective in updating a projection, but users need to go through a trial-and-error process to learn about the unpredictable effects of parameter changes. In comparison, our method helps users choose local POIs and their enhanced features, rather than dimension weights. Users are able to directly decide and predict the outcomes.

In a projection assisted exploration, subspace clusters are often provided beforehand [23,29,30]. In other methods [29,31,32], users can further participate in the clustering process. But in either way, users don't fully understand the given clusters or subspaces. It's hard for them to modify the results, let alone discovering more hidden clusters. Yuan et al. [25] proposed a hierarchical subspace exploration, which allows users to analyze a local subset in different subspaces. The approach helps to discover hidden clusters, but it doesn't provide any guidance for subspace selection.

2.3. Feature driven projection selection

Projection pursuit [26,27] is a well-known technique for finding interesting projections. It generates a series of projections to optimize a certain index. Gleicher et al. [19] used machine learning to train compositive dimensions for classification. Choo et al. [18] made the process interactive by involving users in a semi-supervised Linear Discriminant Analysis (LDA) process. In both works, user-defined classes are imported as the pursuit index. Apart from class labels, user-defined layouts can also function as the pursuit indices [33–35]. However, these methods require prior knowledge of the data, which cannot be assumed in a data exploration.

The rank-by-feature framework [36] is a variant of projection pursuit. It ranks existing projections according to feature strengths. Various kinds of metrics [37] are defined to measure different features, including class separation [20,21], clustering/outliers [38,39], and more complex topological properties [40]. They are helpful for analyzing a large group of scatterplots [41,42]. But most of them are result-oriented and computationally expensive, and thus unsuitable to guide the generation of projections. Otherwise, the time spent to find and score a projection will be unbearable in an interactive exploration. In this work, we only consider simple metrics when pursuing projections with desired features. The simple criteria are not only more efficient, but also easier to interpret.

3. Overview

In this work, we aim to facilitate local data analysis in linear projections. We propose an interactive and exploratory scheme to help users discover, analyze, modify, and compare different POI local data. To be specific, our method supports a four-step data exploration (Fig. 1):

Step. 1: **Focus Search**: First, we present a global projection as an overview (Fig. 1(a)) of the data. Users can choose any data subset in the layout and name it as the POI, which is also called a **focus**. We define two types of focuses, i.e. the focus point and the focus group, regarding whether the POI includes multiple samples. We also make recommendations for both types. Users can simply follow our suggestions if they don't know what to choose.

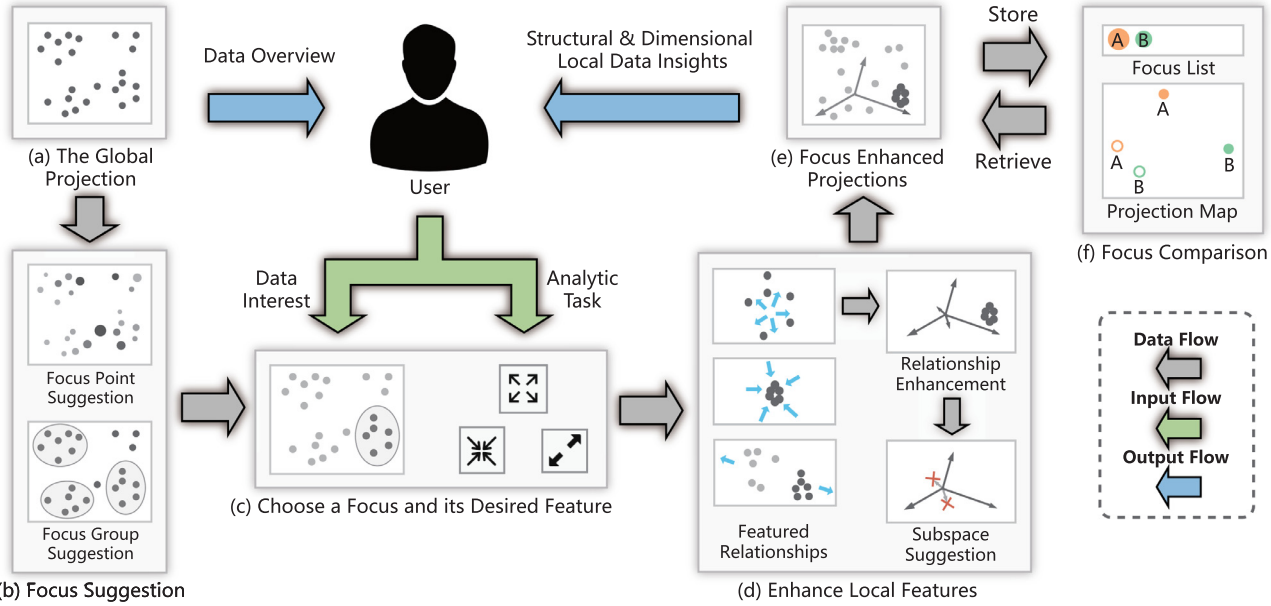


Fig. 1. We propose an interactive and exploratory scheme for local analysis in high-dimensional data. (a) We first present a global projection as the overview. (b) Then suggestions are made to help user find an interesting piece of local data. (c) User chooses some local data as the focus, and named a feature to be analyzed. (d) We make linear projections to enhance local features of the chosen data. Different features are defined for different analytic tasks. (e) A feature-related subspace is also revealed to support dimensional analysis. From the resulting projection and subspace, users can gain both structural and dimensional insights about the data focus. (f) Finally, the user can move on to a new focus, and store valuable findings in Focus List for further study. Projection Map is provided to help compare multiple focuses in the list based on their features.

Step. 2: **Enhancing Features of the Focus**: After some focus is chosen, we customize a linear projection to enhance features of the focus. By features, we refer to three kinds of featured relationships we find most informative in the local analysis (Fig. 1(d)). Different features serve different analytic tasks. Besides the enhancement, we further reveal dimensions that may relate to the feature, and thus suggest a feature-prominent subspace. It helps to interpret dimensional causes of the underlying data relationships.

Step. 3: **Modifying the Focus**: When analyses are finished in the projection, the user may need to modify the current focus or simply change his/her interest. We provide various interactions for the modification. The user can choose a new focus, or add / remove samples from the current POI without being interfered by irrelevant data. Projections will update along with the focus.

Step. 4: **Focus Comparison**: When a valuable POI is found, the user can store it in the Focus List for further analysis (Fig. 1(f)). We provide the Projection Map (Fig. 1(e)) that shows featured projections as glyphs for all focuses in the list. It helps to compare different focuses, and navigate the high-dimensional exploration.

In this four-step exploration, users can pick up any local data and feature, and get both structural and dimensional insights from the locally enhanced projections. We expect the analysis to be free from invisible and uncontrollable distortions, since the feature to be analyzed is enhanced to the greatest extent. Moreover, users are able to handle and compare multiple focuses, and retrieve the explored ones at any time.

We develop a prototype system called FocusChanger (Fig. 2) to support this exploratory pipeline. It consists of five parts: Projection View, Information Panel, Control Panel, Focus List and Projection Map. We will introduce the function of each part as we elaborate the technical details in the next section.

4. High-dimensional local data analysis in locally enhanced projections

As stated before, our method supports a four-step exploration. In this section, we'll introduce in detail how we support this POI-based exploration in each step.

4.1. Discovering interesting local focus

Shneiderman has suggested in his information seeking mantra [43]: "Overview first, then detail on demand". Following the suggestion, we first provide a PCA projection as an overview of the data. Despite the inherent distortion issues, PCA are widely used, easy to interpret, and also generates reasonably good data overviews. More importantly, it fits well into our linear framework and allows for seamless view transformations. Therefore we choose the PCA projection as a start point.

Users can brush any part of the data they feel interesting and claim it as a focus. However, it may not be easy for users who have no analytic backgrounds or prior knowledge about the data. Hence, we also recommend to users some potential focuses generated via automatic detection algorithms. Different suggestions are made for different types of POIs (point or group).

Note that, we make all recommendations based on the projected data, i.e. a distorted copy, rather than the original high-dimensional data. There are two reasons for such a practice. Firstly, no additional or prior knowledge should be assumed in a free exploration. Users choose their focuses based on what they perceive. Therefore, we should also recommend based on what is presented to users. Secondly, the locally enhanced projections are capable of revealing underlying data features. What's been misunderstood due to distortions can be corrected in the POI-enhanced layouts. By changing the focus, users can gradually clarify data structures in different local regions.

4.1.1. Focus point suggestion

Given a projection, we consider a datum interesting in the following

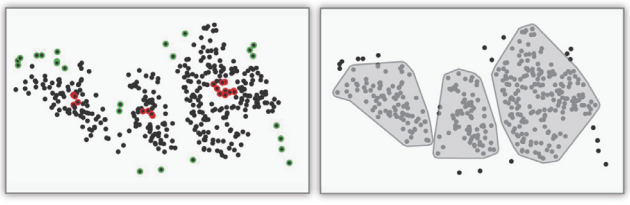
Fig. 2. An overview of our prototype system called FocusChanger. (a) **Projection View** allows users to choose the POI data, and returns a locally enhanced projection accordingly. (b) **Information Panel** shows details about the POI and its featured projection, including data size, dimension weights, and so on. (c) **Control Panel** provides different options for choosing POIs and enhancing the projection. (d) **Focus List** is where users store the POIs they found valuable for further analysis. (e) **Projection Map** shows featured projections of all POIs in the Focus List. Each projection is shown as a glyph while the in-between distance denotes dissimilarity.

cases: (1) it's a representative datum; (2) it's an outlier; (3) it's misplaced in a distorted neighborhood. The former two cases are helpful to identify popular data and anomalies, which are both widely studied in data analysis. We also consider the last case important, given that misplaced points often misguide the interpretation of data locality. Users have no need to pay full attention to these data, but knowing the distortions helps to avoid misunderstanding.

We define representative points and outliers based on clusters in the projection. To obtain clusters, we adopt a variant of DBSCAN, details of which are described in Section 4.1.2. For each cluster, we choose top 5% members that are closest to the cluster center as recommended representative points. Each of them is representative of its cluster. As for outliers, we simply recommend data that falls out of any cluster. Since density clustering is used, such data are outliers in the sense that they have sparse neighborhoods. We mark the suggested POI points with colored boundaries (Fig. 3(a)). Red denotes representative points, while green denotes outliers.

In order to quantify distortions, we assess the accumulated distance errors for each datum:

$$Error(\mathbf{x}_i') = \sum_{j=1}^{n} (Dist(\mathbf{x}_i, \mathbf{x}_j)^2 - Dist(\mathbf{x}_i', \mathbf{x}_j')^2),\ i = 1, 2, \cdots n \tag{1}$$



(a) Focus Point Suggestion (b) Focus Group Suggestion

Fig. 3. Focus Suggestions. (a) Representative points (red) and outliers (green) are recommended as potential POI points. (b) Clusters are revealed in the projection, guiding users to choose a proper POI group. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

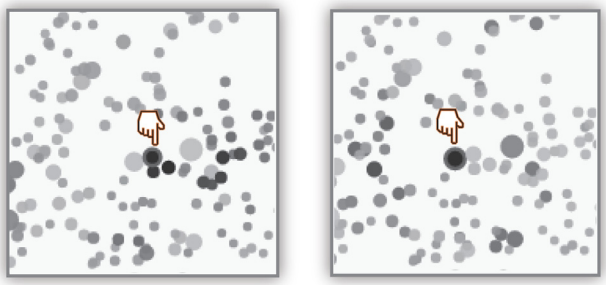


Fig. 4. Visualizing distortion with point size and color. When a datum is hovered, the saturation of other data reflects their high-dimensional distances to the hovered one. This image shows a case where two projection neighbors are probably far way in the high-dimensional space.

Here $\mathbf{x}_i$ and $\mathbf{x}_i'$ represents a high-dimensional datum and its projected counterpart. Distance is measured by the Euclidean metric. We use point size to show the distortion levels (Fig. 4). Larger points are more likely misplaced with "false neighbors", that are projected nearby but actually far away in the high-dimensional space. When users hover on a datum, the saturation of other points changes to reflect their high-dimensional distances to the hovered one. Closer data get higher saturation. This technique has also been used in [7]. It allows users to find distorted datum / neighborhoods that may benefit a lot from local enhancements.

4.1.2. *Focus group suggestion*

In a projection, we assume a group of data interesting if they appear as a cluster. Hence, we detect clusters in the projection and recommend them as potential POI groups.

Many clustering algorithms can be applied to identify projection clusters [45]. We adopt a variant of DBSCAN [44] whose parameters are adaptive to the data. We choose DBSCAN because it can efficiently identify clusters in any shape. It also reveals outliers in sparse regions. The self-adaptive parameters make it applicable to most datasets without the need for manual tuning. Refer to Fig. 3(b) for actual effects

of the algorithm. Clusters are shown as contours around data points. Users can choose any suggested cluster by clicking on the contour.

Note that, either points or groups, we only make suggestions to guide and support users' decisions, not to replace them. Users can hide the suggestions at any time, and choose his own POI by brushing the desired data. All options are provided in Control Panel (see "Focus" and "Suggestion" in Fig. 2 (c)).

4.2. Featured projections of the focus

We call a chosen datum the focus point, and call a chosen group the focus group. After some focus is chosen, we generate projections to either reduce its distortion, or enhance its local features. The locally enhanced projection is updated in Projection View (Fig. 2(a)) with highlighted POI data and grayed out contexts.

4.2.1. Focus point enhanced projection

For a focus point, we assume that the user is interested in its high-dimensional neighbors. In other words, he/she may wish to find more data similar to the POI in different aspects. It's especially useful in scenarios where users start searching from a familiar case. For example, a college student longs for a good car but cannot afford the price. He/she can search for other cars with similar qualities but at a lower price. In such a case, mapping distortions could be a major obstacle. To reveal the true neighbors, we seek a projection to reduce the overall distance distortions for the POI. Let $\mathbf{P}$ be the focus point, we aim to solve the following optimization problem:

$$\min Error(\mathbf{P}) = \min_{\mathbf{A}} \sum_{i=1}^{n} (Dist(\mathbf{P}, \mathbf{x}_i)^2 - Dist(\mathbf{PA}, \mathbf{x}_i')^2)$$
$$s.\ t.\ \ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{2}$$

Here the term $\mathbf{A}$ represents a projection matrix containing unit vectors, and $\mathbf{I}$ is the identity matrix. The projected focus is $\mathbf{P}' = \mathbf{PA}$. Via this optimization, we push aside all false neighbors that are far from POI in the high-dimensional space. The projected neighborhood is then left with real neighbors. Since the high-dimensional distances are invariant between POI and other data, the term $Dist(\mathbf{P}, \mathbf{x}_i)^2$ is simply a constant. The minimization problem can be written in the following form:

$$\max_{\mathbf{A}} \sum_{i=1}^{n} Dist(\mathbf{PA}, \mathbf{x}_i')^2,\ \ s.\ t.\ \ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{3}$$

Note that if we replace $\mathbf{P}$ with the mean data $\bar{\mathbf{x}}$, the optimization will directly lead to PCA projection in Euclidean distances. In other words, the locally enhanced projection can be regarded as PCA with a shifted center (i.e. the chosen focus). Thus the optimization is no harder than a simple PCA algorithm. Accelerating skills for PCA are also applicable to speed up the optimization.

Compared to the direct distance correction used in [7], our method preserves all benefits of a linear projection rather than mere point-wise distances. It provides rich dimensional contexts, and help maintain a consistent mental model of the data space. In Section 4.2.3, we will further integrate this projection into a larger framework.

4.2.2. Featured local relationships

For a focus group, we first examine what kinds of features are of interest in the local analysis. By features, we refer to featured relationships (e.g. clustering, sub-grouping, outliers, etc.) that involve the focus group. Since relationships are defined based on distances, we can look into the distance matrix for an answer. Given the POI group, the whole distance matrix is divided into three parts (Fig. 5(a)). The first part describes distances between group members. The second part is about distances between the group and the other data. The last part describes distances among the context data. Since the last part has nothing to do with the focus, we simply ignore it. For the remaining parts, we consider the data to be either "similar" or "dissimilar" (Fig. 5(b)).

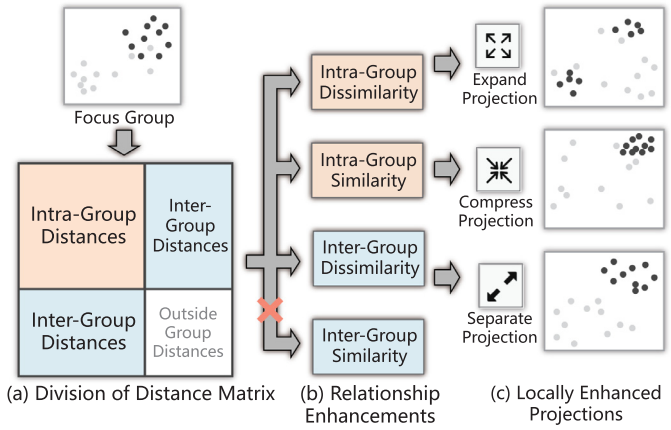


Fig. 5. Enhancing featured local relationships. Given a focus group, the distance matrix is divided into three parts (a). We enhance data similarity/dissimilarity based on each part (b). Three types of projections are generated via the local enhancements (c).

By revealing similarities among group members, we can show users in which aspects the data are most similar. It helps to comprehend why these data gather into a cluster in the projection. Enhancing dissimilarities, on the other hand, tells about the major differences among group members. If there are sub-clusters within the group, their differences will be more prominent when enhanced. It helps to reveal hidden local structures.

Currently, we did not consider similarities between the group and the others. It's based on the fact that, we will have to distinguish between two parts before talking about their similarities. If we ignore the differences, the similarity between two complementary parts is always equal to similarities among all data, i.e. the inter-group similarity of the whole dataset. In that case, there is no need for such a setting. In contrast, by enhancing the dissimilarities, we can show why the focus group is different from the others. The idea resembles that of Linear Discriminant Analysis (LDA), except that we did not regard the other data as the same class.

In summary, three types of relationships are found informative in the local analysis. They are named intra-group similarity, intra-group dissimilarity and inter-group dissimilarity respectively (Fig. 5(b)).

4.2.3. Focus group enhanced projections

With the three types of relationships, we first translate them in the language of data distances. Then we adopt projection pursuit to find linear projections for the enhancement.

Enhancing similarities or dissimilarities, equals to decreasing or enlarging data distances in the projection. For a focus group G, we enhance the intra-group dissimilarities by:

$$\max \sum_{\mathbf{x}_i', \mathbf{x}_j' \in G} Dist(\mathbf{x}_i', \mathbf{x}_j')^2 = \max_{\mathbf{A}} \sum_{\mathbf{x}_i, \mathbf{x}_j \in G} Dist(\mathbf{x}_i\mathbf{A}, \mathbf{x}_j\mathbf{A})^2$$
$$s.\ t.\ \ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{4}$$

For simplicity, we call this optimization the **Expand** metric, since the focus group will be expanded in the resulting projection. It actually leads to a local PCA projection (see Appendix). Likewise, we enhance the similarities by minimizing the same metric:

$$\min_{\mathbf{A}} \sum_{\mathbf{x}_i, \mathbf{x}_j \in G} Dist(\mathbf{x}_i\mathbf{A}, \mathbf{x}_j\mathbf{A})^2,\ \ s.\ t.\ \ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{5}$$

It is named the **Compress** metric, as the opposite of Expand. At last, we enhance the inter-group dissimilarities by enlarging distances between the group and the other data:

$$\max_{\mathbf{A}} \sum_{\mathbf{x}_i \in G} \sum_{\mathbf{x}_j \in \overline{G}} Dist(\mathbf{x}_i\mathbf{A}, \mathbf{x}_j\mathbf{A})^2, \ \ s.\ t.\ \ \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{6}$$

It is named the **Separate** metric. We use glyphs to denote different feature enhancements throughout the whole system. Fig. 5(c) illustrates those glyphs, as well as projections generated from different metrics. The user can change the enhanced feature in the Control Panel (see "Features" in Fig. 2 (c)), regarding the analytic task at hand.

In fact, a focus point can be regarded as a group containing only one datum. There will not be Compress or Expand projections, but the Separate metric simply degrades to the center-shifted PCA (compare Eq. (3) and (6)). This enables us to combine all featured projections into the same framework.

At last, solutions to all optimization problems can be approximated via eigen-decompositions in $\mathbf{O}(D^3)$ time (see Appendix), with D being the number of dimensions. The complexity can be further reduced by numerical techniques. It makes our method well scalable to larger datasets with higher dimensions.

4.2.4. Subspace suggestion

When pursuing the feature enhanced projections, we take into account all dimensions. However, only a few of them truly contribute to the features. Redundant dimensions will obscure the patterns and interfere with dimensional analysis. Hence, we need to reveal a feature-related subspace to promote subsequent analysis.

In a sense, projection pursuit itself is a process to identify the most featured dimensions. Based on its results, we can make reliable judgements on dimension contributions. To be specific, we first generate an enhanced projection with all dimensions considered. Then we examine dimension weights in that projection, and rank all dimensions by their weights: $W(d_1^*) > W(d_2^*) > \cdots > W(d_m^*)$. All weights are normalized and sum to 1. At last, we pick out dimensions with large weights, until their sum exceeds a certain threshold:

$$\begin{aligned} &Subspace = \{d_i^* | i = 1, 2, ...L\}, \\ &s.\ t.\ \ \textstyle\sum_{j=1}^{L} W(d_j^*) \le R \text{ and } \sum_{j=1}^{L+1} W(d_j^*) > R \end{aligned} \tag{7}$$

Threshold R is set as 0.75 by default, cutting down at least 25% redundant dimensions. The result would be a subspace that is most related to the enhanced feature. The sum of weights is named the **Subspace Score**. It indicates how strong the chosen subspace is related to the current feature. Dimension weights are always displayed as a bar chart in the Information Panel (Fig. 2 (b)). Users can include / exclude dimensions by brushing in this view.

After gaining a subspace, we run optimizations again in that subspace to get the final result. The refined projection will be easier to interpret with only the most related dimensions. Features will also be more prominent.

4.3. Modifying the focus

For a focus point, a distortion reduced projection is the final step. But for a focus group, it may still need to be modified. The featured projections support this task by revealing local insights.

The Expand projection shows minor relationships hidden in the group. It reveals sub-clusters and outliers, and helps to trim the POI into a more consistent group. The Compress projection enhances similar aspects within the group. If there are other data that resemble the focus in such aspects, they will also be drawn closer to the group, claiming to be potential members. It helps to regain the missing members. The Separate projection highlights differences between group members and the others. Boundary points will stand out in this case, which helps to clarify cluster boundaries. We support the modification by providing focus-aware brushing techniques (see "Selection" in Fig. 2(c)), in order to avoid interference from the context. When the user needs to add more members, he/she can use the "Increase" brush. Whatever chosen by the brush will be added into the current focus group. When there is a need for decreasing members, the "Decrease" brush can be used. The intersection of the current group and the brushed data will be chosen as the new focus.

Once the focus group is changed, the projection will also be updated. Smooth transitions are applied to avoid swift changes (see the supplementary video). We keep an orthogonal mapping in each frame of the transition [46], so as to maintain an intact mental model of the high-dimensional data structure.

4.4. Storing and comparing multiple focuses

During the exploration, there will be times when users need to store the results. For example, when sub-groups are found, the current group shall be stored before focusing on a smaller group. Besides, there are needs to compare different focuses regarding their features. We provide the **Focus List** and the **Projection Map** to support such tasks.

In the Focus List (Fig. 2(d)), users can store the current focus or retrieve a previous one. Each focus is represented as a node. Its size denotes the data size, while colors differentiate different focuses. When hovering on a node, users can name the focus or change its color. Specially, there is a fixed node called "All Data". Enhancing its Expand feature will lead to a global PCA projection.

For every focus in the list, its enhanced projections for all three features are shown as glyphs in the Projection Map (Fig. 2(e)). Different glyphs denote different features (Fig. 5(c)). Dimension weights are displayed as small histograms on top of each glyph to help compare the projections. Clicking on a glyph can retrieve the focus and corresponding projection.

To measure the similarity between projections, we refer to the manifold learning domain. It has been proved that any two 2D projections lie on the same Grassmann manifold. Their dissimilarity can thus be measured by their geodesic distance on this manifold [47]. We choose this metric since it reflects dimensional diversity between projections, which is important in feature comparison. After gaining the distances, we construct the final map using MDS algorithm. Due to the mapping technique, there may be occlusions between glyphs. Nevertheless, users can always clarify the clutter by hovering on a focus node, which can highlight related glyphs in the map.

5. Case study

In this section, we demonstrate the effectiveness of our method with two real-world datasets.

5.1. Cars data

For the first case, we present a usage scenario in the Cars dataset [48], where accurate neighborhood analysis is required. The data contains 392 samples with 8 numerical/categorical attributes: displacement, MPG, cylinder number, horsepower, weight, acceleration time, year and origin. Names of the cars are also given as a textual attribute.

Tom plans to buy a new car, but he doesn't know much about the automobile market. So he decides to look into the cars data, and try to find some suitable targets. After a quick overview of the projection, he is attracted by some representative points in the right-side cluster (Fig. 6(a)), which seem to be popular cars. When he hovers on one of them, a tooltip pops up to show the details. It is Celica GT from Toyota, a popular Japanese car famous for its excellent performance. Tom likes what he just found but is worried about its displacement (144 CL). He wonders if there are other cars with similar performance but are more environment friendly.

He first notices a recommended point next to Celica called Plymouth Cricket. However, when the point is hovered, all its neighbors fade out with low saturation (the left part in Fig. 6(a)). Tom knows immediately

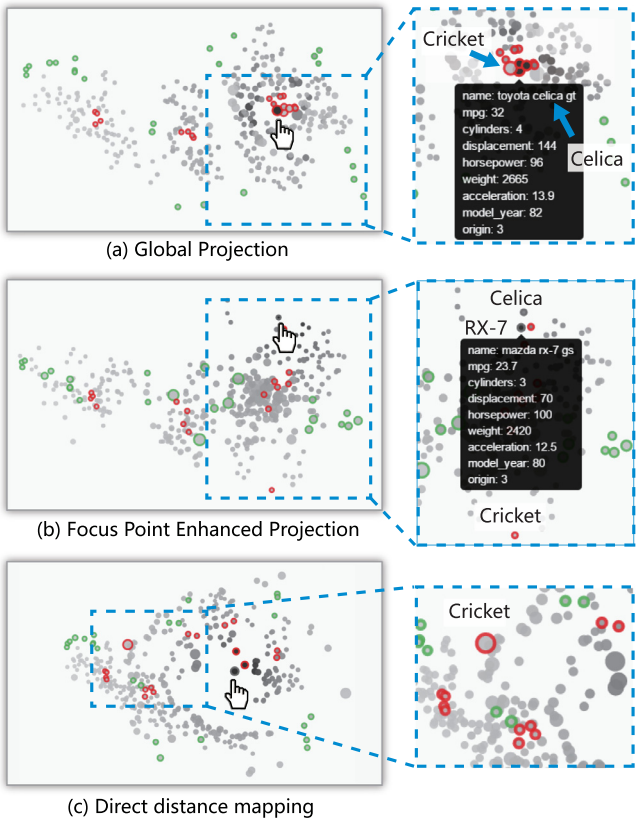


Fig. 6. Cars Data: (a) In the global projection, the user finds a good car named Celica among the suggested POIs. He also notices that a neighbor named Cricket may be misplaced, given its inconsistent saturation. (b) He chooses Celica as the focus, and enhances the projection. A similar car is found near Celica in the new layout, while Cricket is placed far wary. (c) Directly mapping distances to the layout [7] may create false data patterns. Colored points are kept across all projections to keep track of the data.

that it could be a misplaced datum. In order to reveal the real neighbors, he chooses Celica as the focus, and enhances the projection. In the resulting layout, he finds that Cricket is placed far away from Celica (Fig. 6(b)). In addition, he discovers some new neighbors that have not been noticed in the original projection. Most of these neighbors have lower displacement than Celica, especially a close one named Mazda RX-7 GS, which is also a very popular Japanese car. RX-7's performance is close to Celia, but its displacement is only half of the latter at 70 CL (right part in Fig. 6(b)). Tom is now satisfied with his new options.

To quantify the effectiveness of the algorithm, we further assess the neighborhood of Celica in different spaces. Specifically, we compare the k-nearest neighbors in a global/locally-enhanced projection to the "real neighbors" in high-dimensional space. We varied k from 10 to 40, and found that PCA preserves at most 37.5% neighbors, while the locally enhanced layout preserves at least 85% neighbors (see Appendix for more details). In the Celica case, our mapping indeed provides a more authentic neighborhood for the POI.

We also compare our method with that in [7] (Fig. 6(c)). The latter approach authentically restores distances from the neighbors to the focus, which may create misleading patterns. As an example, there are three major clusters shown in the overview, which have vanished in the direct mapping but are still maintained in our locally enhanced projection. Compared with direct distance mapping, our method keeps a better balance between the local neighborhood and the background structure.

5.2. USDA Food data

The second case is from the USDA Food Composition Data (http://www.ars.usda.gov/). The dataset describes nutrients of a collection of raw or processed foods, with 722 records and 18 dimensions after preprocessed. The recorded nutrients include fibers, carbohydrates, water, vitamins (e.g. VitARAE, VitD, VitB12), microelements (e.g. iron, calcium, sodium) and so forth. This data has been used in some previous works [25,29] for case studies. However, those methods focus on subspace mining, while we concern more about local data analyses.

Jean is a nutritionist whose daily work is to examine various foods, and analyze their nutrients. At the first sight of the projection, she can roughly identify three clusters (Fig. 8(a)). But there are no clear boundaries or subtle structures. She brushes a cluster and chooses the Expand feature, in order to discover hidden subtle structures.

In the enhanced projection (Fig. 8(b)), there seems to be some unclear pattern. Jean turns on the subspace suggestion, gaining the result in Fig. 8(c). Now she clearly sees a large cluster with several outliers. The outliers even include two tiny clusters dominating two dimensions (vitamin D and Sodium). A closer examination reveals that these are salty foods and cakes, which are rich in sodium and vitamin D respectively.

Jean chooses the major cluster as a new focus, using the "Decrease" brush to avoid including contextual data. She wants to study both similarities and dissimilarities among cluster members, so she applies Compress and Expand projections respectively (Fig. 9(b),(c)). The Compress view reveals that this cluster contains more water and energy than the other foods, but is poor in carbohydrate. On the other hand, data in the Expand view are separated into three sub-clusters along four dimensions, all of which are vitamins.

She applies Compress to two of the sub-clusters, gaining results in Fig. 9(g) and (h). Sub-cluster 1 exhibits a very strong feature (all data gathering around the origin) that, none of its members contains any fiber or beta-carotene. Jean suspects these are animal-based or processed foods, which is validated by a later examination. Sub-cluster 2 does not show much new information, but it seems to be the most similar to the contextual data (Fig. 9). She stores the two sub-clusters in Focus List for further study, before going deeper into the last sub-cluster.

After expanding the last group, she quickly identifies a large cluster with a few outliers. She picks out the cluster and applies the enhancements. There is again a strong pattern that no cluster member contains any fiber or vitamin C. These should also be animal-based or processed foods. A later examination proves that most of them contain beef, chicken, cream or fruit juice. But the Expand view shows two subgroups, which again surprises her. The two sub-group are mainly diverse in vitamin B6 and Sodium. She also stores them in the list.

Jean finally gets to enjoy all the trophies she has collected in the exploration. She names the stored clusters by their hierarchies, and then compares them in the Projection Map (Fig. 7). The map shows a very close relationship between cluster 3-1 and 3-2, in all three featured projections (Fig. 7(e), (f), (g)). She also finds the strong Compress features she met in the exploration (Fig. 7(c), (d)). Hovering on each cluster allows her to highlight them in the global projection (Fig. 9(i), (j) and Fig. 10(g), (h)). Different clusters occupy different layers in the overview, which seems reasonable.

6. Discussion

In this work, we propose to customize linear projections to facilitate analyses regarding the user-specified local POI data. In Section 5, we prove the effectiveness of our method in revealing hidden relationships and reducing local distortions. However, there are still flaws in the

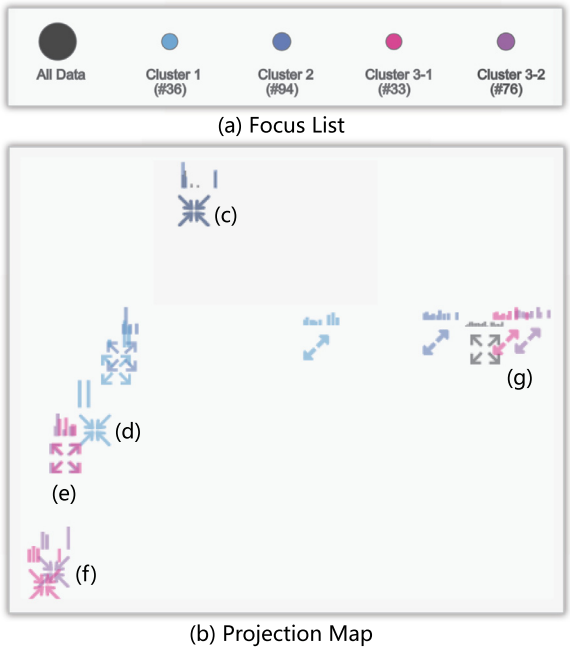


Fig. 7. USDA Food Data: Four clusters found in the exploration are stored in Focus List (a). Projection Map helps to compare them based on the featured projections. It can be seen from dimensional weights that, Compress projections of Cluster 1 and 2 are highly featured (c, d). Cluster 3-1 and 3-2 are very similar in all their projections (e, f, g).

proposed method, which we would like to discuss in this section. After that, we illustrate the relationships and differences between our method and some state of art machine learning techniques. At last, we show that our method can be extended into a more general methodology.

The first flaw we'd like to mention, is the lack of effective navigation in the exploration. Dimension weights are like the steering wheel. They provide good controls to fine-tune a projection, though the tuning is often aimless. In our method, the steering wheel is the focus with its features. The exploration inevitably jumps between diverse projections when the focus or metric is updated. Even supplied with animations, users may still get confused when faced with the abrupt changes. Some intermediate results could be lost if they are missed by mistake. A proper remediation is to provide controls over animations, dimensions, and even the navigation history.

The second flaw is about the dimensional analysis. At present, we interpret the projected data by perceiving their distribution along the projected axes. But such interpretation is not precise, since the projected directions often interfere with each other. As the dimensionality increases, it becomes more and more difficult for individual axes to stand out. In higher-dimensional cases, the dimensional analysis requires to handle a large amount of dimensions at each time (see Appendix). The interference problem will then be alleviated.

In the machine learning field, the commonly studied data use to have huge amounts of dimensions, while each dimension alone is not so meaningful. Such data are assumed to be samples on a low-dimensional manifold, where neighborhood relationships are more important than dimensional semantics. In this work, we also seek to preserve data locality, but we decide to maintain the dimensional context by using linear projections. That's because, in a more general case, it's of interest what factors account for the data differences. The explanatory power of dimensions directly differentiates our method from the local-preserving mappings. Besides, as shown in the supplemented cases (see Appendix), our method is also helpful in interpreting a very-high-dimensional dataset.

In the case studies, we only demonstrate the method in some small datasets. Nevertheless, there is no substantial difficulty in applying the method to higher-dimensional datasets, such as the image data (please refer to the Appendix). Given the $O(D^3)$ complexity, the computation time of our method seems tolerable in practice (around 2 s for a 1000-D dataset). But a performance test is still needed to validate its efficiency. We also plan to conduct a user study in real-world scenarios to further examine the usability of our approach.

Finally, we would like to discuss about possible extensions of this work. Our method is built upon linear projections. Nevertheless, the focus-based framework, either point-based distortion reduction or group-based relationship enhancement, is also applicable to non-linear projections. Currently, we only define three kinds of featured projections based on a direct division of the distance matrix. But the framework is totally compatible with more complex metrics, as long as they can be described in the form of data distances.

7. Conclusion

In this paper, we propose an interactive method that helps customize a linear projection to facilitate the analysis of user-defined local data. By incorporating users' POI with dimension reduced projection, we make more efficient use of the display space, and more effectively meets users' analytic needs. We also provide various kinds of techniques to support a fluent high-dimensional data exploration. Users are assisted to discover, analyze, modify and compare multiple pieces of local

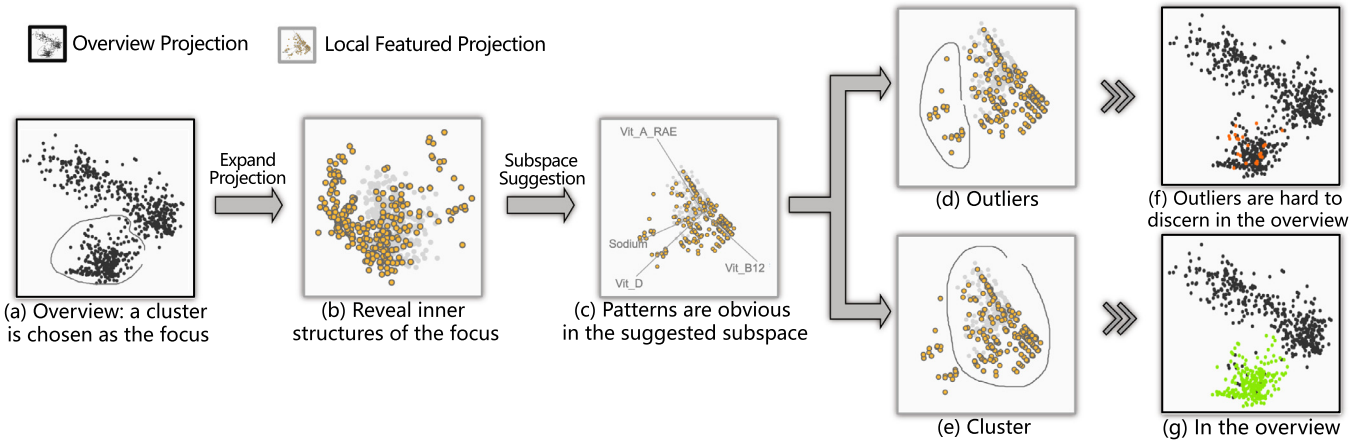


Fig. 8. USDA Food Data: User chooses a global cluster as the focus (a), and enhances the projection to explore its features (b, c). Outliers can be found in the suggested subspace (d), which are hard to discern in the global projection (f).

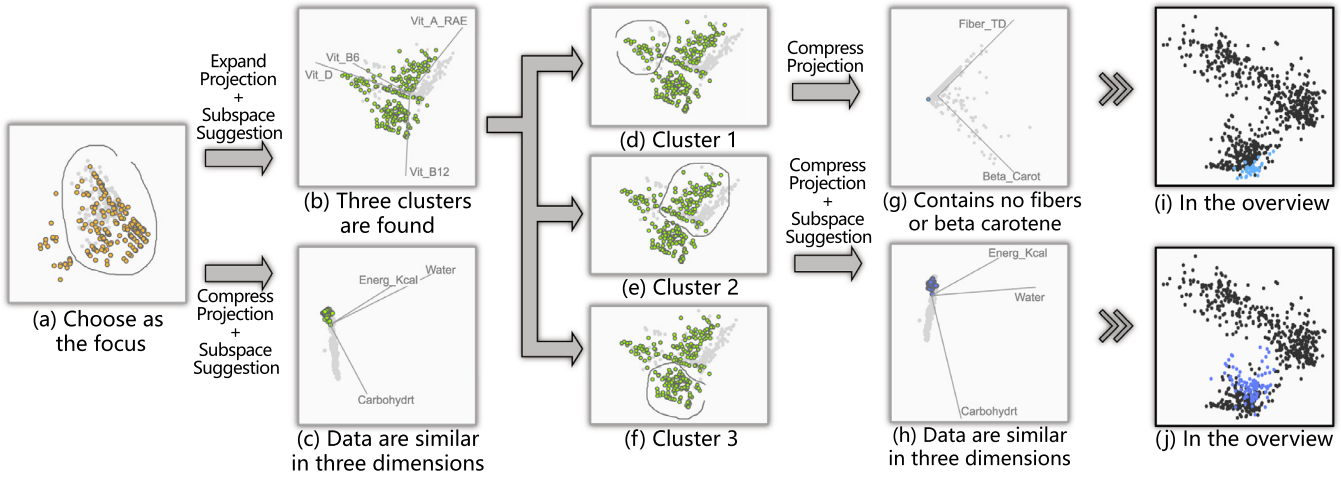


Fig. 9. The user focuses on the cluster found in a previous step. Similarities (b) and diversities (c) are revealed among the cluster members. Three sub-clusters are found. One of them exhibits a very strong feature that it contains no fibers of beta-carotene (g). It could be a group of animal-based foods.

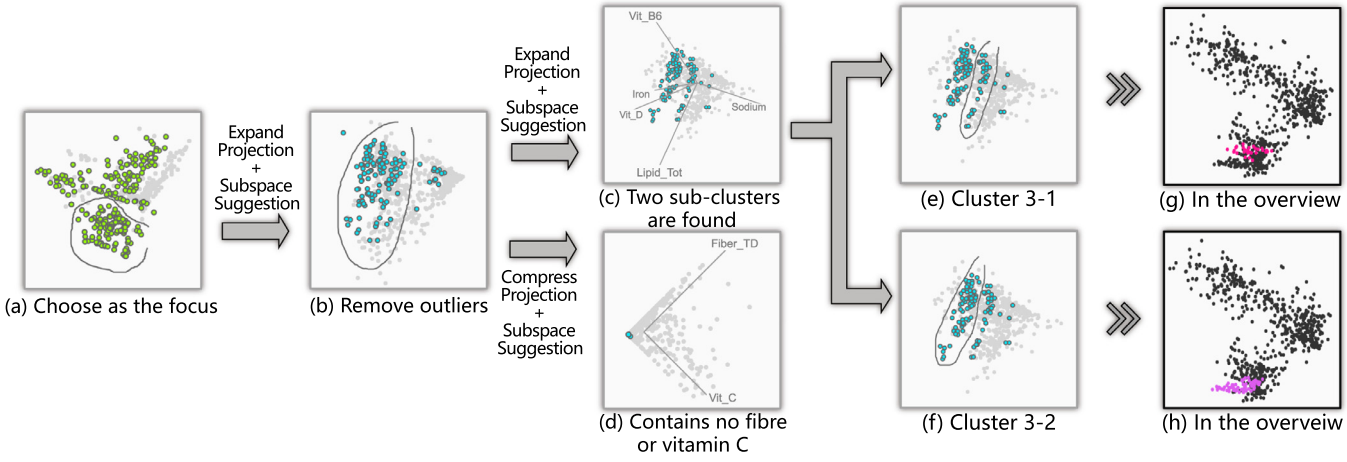


Fig. 10. The user focuses on a sub-cluster found in Fig. 9(f). She gets the major group by removing a few outliers (b). This group shows a strong Compress feature (d), as well as an interesting inner structures (c). Again, two sub-clusters (e, f) are found within the focus. They seem to be small neighboring clusters in the overview (g, f).

data. At present, our method only provides several kinds of local improvements, including local distortion reduction and local relationship enhancements. However, our method is totally compatible with more complex enhancements. It's also applicable to more general kinds of projections. The integration of user-defined focuses will largely benefit dimension reduction techniques. We look forward to extending our method to a broader field in the future.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2016QY02D0304), National Basic Research Program of China (973) (2015CB352503), and National Natural Science Foundation of China (61672055). This work is also supported by PKU-Qihoo Joint Data Visual Analytics Research Center.

Appendix A. Solving the optimization problems

We approximate solutions to the three optimization problems in $O(D^3)$ time, with D being the dimensionality of data.

The **Expand** optimization:

$$\max_{\mathbf{A}} \sum_{\mathbf{x}_i,\mathbf{x}_j \in G} Dist(\mathbf{x}_i\mathbf{A}, \mathbf{x}_j\mathbf{A})^2, \quad s.t. \quad \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{A.1}$$

We approximate the solution by solving:

$$\max_{\mathbf{A}} \sum_{\mathbf{x}_j \in G} Dist(\bar{\mathbf{x}}\mathbf{A}, \mathbf{x}_j\mathbf{A})^2, \quad s.t. \quad \mathbf{A}^T\mathbf{A} = \mathbf{I} \tag{A.2}$$

, which leads to a local PCA projection.

The **Compress** optimization:

$$\min_{\mathbf{A}} \sum_{\mathbf{x}_i,\mathbf{x}_j \in G} Dist(\mathbf{x}_i\mathbf{A}, \mathbf{x}_j\mathbf{A})^2, \ \ s.\ t.\ \ \mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{I} \tag{A.3}$$

Similarly, we approximate the solution by solving:

$$\min_{\mathbf{A}} \sum_{\mathbf{x}_j \in G} Dist(\overline{\mathbf{x}}\mathbf{A}, \mathbf{x}_j\mathbf{A})^2, \ \ s.\ t.\ \ \mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{I} \tag{A.4}$$

The **Separate** optimization:

$$\max_{\mathbf{A}} \sum_{\mathbf{x}_i \in G} \sum_{\mathbf{x}_j \in \overline{G}} Dist(\mathbf{x}_i\mathbf{A}, \mathbf{x}_j\mathbf{A})^2, \ \ s.\ t.\ \ \mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{I} \tag{A.5}$$

We approximate the solution by solving:

$$\max_{\mathbf{A}} \sum_{\mathbf{x}_j \in \overline{G}} Dist(\overline{\mathbf{x}}_G\mathbf{A}, \mathbf{x}_j\mathbf{A})^2, \ \ s.\ t.\ \ \mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{I} \tag{A.6}$$

In the approximation, we replace the many-to-many distances with many-to-one distances, allowing the mean data to represent the whole group. All the approximations can be summed up as:

$$\sum_{\mathbf{x}_i} Dist(\mathbf{y}\mathbf{A}, \mathbf{x}_i\mathbf{A})^2, \ \ s.\ t.\ \ \mathbf{A}^{\mathbf{T}}\mathbf{A} = \mathbf{I} \tag{A.7}$$

, with *Dist* being Euclidean distance. The problem can be solved by eigen-decomposing the matrix $\mathbf{P}^{\mathbf{T}}\mathbf{P}$, where each row of $\mathbf{P}$ is $\mathbf{P_i} = \mathbf{y} - \mathbf{x_i}$.

Since the $\mathbf{P}^{\mathbf{T}}\mathbf{P}$ is a D*D matrix, the eigen decomposition can be finished in $O(D^3)$ time. For the maximization problem, we choose eigenvectors with the largest eigenvalues to construct the projection matrix **A**. Similarly, the smallest eigenvalues are used to solve the minimization problem.

Appendix B. Neighborhood preservation rate

We examine the *K*-nearest neighbors of a POI datum named Celica in the Cars dataset. Specifically, each datum has a unique integer ID. Let N_H, N_G, and N_L be the **Neighborhood Sets** of Celica in the high-dimensional space, the global projection, and the locally enhanced projection respectively.

We measure the **Neighborhood Preserving Rate** by $Rate = \frac{N_H \bigcap N}{N_H}$. We vary *K* from 10 to 40 and get the following results:

K	10	20	30	40
$N_H \bigcap N_G$	2	5	8	15
$Rate_g$	20%	25%	27%	37.5%
$N_H \bigcap N_L$	9	15	25	34
$Rate_l$	90%	75%	83%	85%

The locally enhance projection only misses 6 in the 40 neighbors, and restores 9 out of 10 nearest neighbors. It preserves at least 75% neighbors, which is two times the best performance of the original projection. This proves the effectiveness of our method in preserving a high-dimensional local neighborhood.

Appendix C. A Supplemental case study

To demonstrate the scalability of our method, we provide a case study on the image data. Due to space limitations, we only include it here in the appendix. The dataset is called Yale Faces [49], shared by the Department of Computer Science, Yale University. It has been widely used over decades for computer vision research. The dataset contains 165 grayscale facial images of 15 individuals (see Fig. C.1). There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. All images are in the 32 × 32 resolution. In other words, it is a 1024-D dataset.

C1. The effect of lighting

In the global projection, the data is divided into three parts (see Fig. C.2(a)). We take the dimension weights and have them rendered in a 32 × 32

Fig. C1. Yale Faces: the 15 subjects with their 'normal' faces.

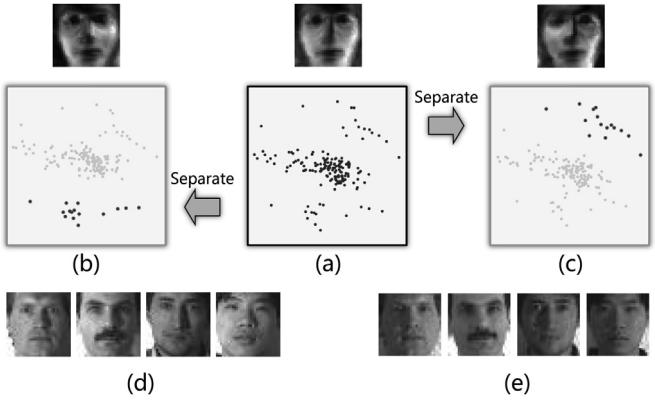


Fig. C2. In the global projection (a), the lower/upper cluster contains images taken in the left-light (d)/right-light (e) condition. Dimension weights of the Separate projection are rendered as a feature bitmap (top row of (b) and (c)). It nicely captures the lighting direction as the feature of these images.

bitmap, as shown in the top row of Fig. C.2. Brighter areas indicate higher weights. In bitmap (a), we can see that the background and the facial features (i.e. the eyes, nose, and mouth) are dark, meaning that these are the common parts shared by most images. The main differences lie in other parts like the cheeks and the shape of a face.

Then we select the lower cluster and apply the Separate projection to see why it's different from others (see Fig. C.2(b)). The locally enhanced projection looks similar to the global one, but the corresponding bitmap (b) is quite different from before. The lower-left and upper-right parts of the face are brighter, implying that these are probably the major featured areas. We further examine the data points, and find that these are exactly the 15 images taken in the left-light condition (see Fig. C.2(d) for some examples). The lower-left part of the face is brighter while the upper-right part is darker. The feature bitmap nicely captures the lighting direction to explain why this cluster stands out.

Likewise, we analyze the upper cluster and find that these are exactly the 15 images taken in the right-light condition (see Fig. C.2(c) and (e)). Once again, the feature bitmap (c) nicely captures the lighting direction. Now we understand that lighting dominates the distribution of these facial images.

C2. What makes her different

We expand the middle cluster, and immediately see some highly notable outliers (see Fig. C.3(b)). These images all belong to subject No. 11. It is the only female subject in the 15 individuals. But what makes her so different from others?

We continue to find out by narrowing down the selection, and applying the Separate projection (see Fig. C.3(c)). The feature bitmap (c) highlights two specific areas: the long hair and the prominent chin. One spot is especially bright in the forehead. It turns out this lady is the only one with hanging bangs, making her forehead area darker than anyone else.

C3. Beard or mustache

Apart from free explorations, we can also demonstrate the effectiveness of our approach by verifying a known fact in the data. As shown in Fig. C.1, only 4 subjects have facial hairs: No. 2, No. 7, No. 9 and No. 13. Among them, only No. 13 has a full beard while the other three have mustaches. The beard should be enough to distinguish No. 13 from all the others. But given the complexity of the data, is our method able to correctly separate this subject and identify his feature?

We highlight all images of No. 13 in the global projection (see Fig. C.4(a)). It turns out he is indeed quite distinguishable, since all related data are distributed at the edge of the projection. We choose the corresponding part in the middle cluster to avoid lighting effects, and apply the Separate projection. An outlier pops out in the new projection (see Fig. C.4(b)). It is a center-light image of subject No. 13, who has a mustache and dark skin. This proves the ability of our approach to avoid perceptual pitfalls in the global projection. After removing the outlier, we successfully obtain all 9 center-light images of subject No. 13. The corresponding bitmap nicely captures his full beard feature (see Fig. C.4(c)).

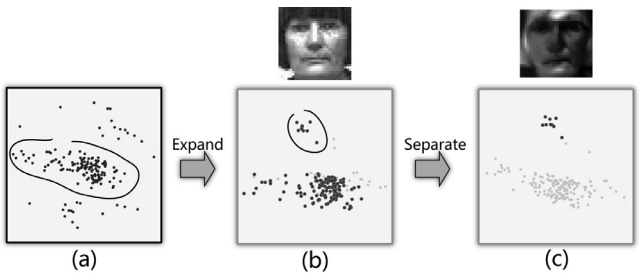


Fig. C3. Some outliers are found in the middle cluster (b), all belonging to subject No. 11, the only female subject. After the Separate projection is applied, the feature bitmap reveals the reason for this lady to stand out (c): she is the only one with bangs.

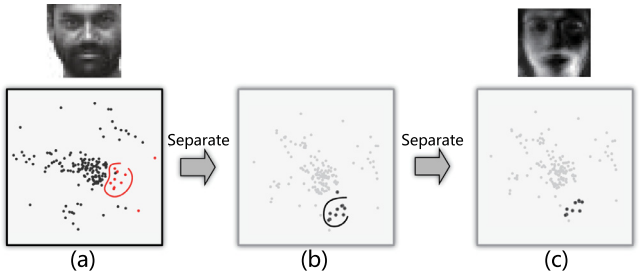


Fig. C4. We choose the center-light images of subject No. 13 in the global projection (a). An accidentally included outlier is revealed in the Separate projection (b). After removing the outlier, the feature bitmap successfully captures the full beard feature of the subject.

C4. Performance in practice

In our current implementation, the computation is supported by numerical algorithms based on C++. Parallel computing is not included. Users are able to get instant feedbacks when Interacting with small datasets (up to 18-D). For the 1024-D Yale Faces dataset, there is a delay of around 2 s for the projection to update. Note that this rough assessment includes not only the time of computation but other parts including data transmission, graphic rendering, animation and so on. A formal performance study is still needed in the future to assess the scalability and feasibility of this approach.

References

[1] B.C. Geiger, G. Kubin, Relative information loss in the PCA, 2012 IEEE Information Theory Workshop, Lausanne, Switzerland, September 3–7, 2012, (2012), pp. 562–566.
[2] S. Lespinats, M. Aupetit, Checkviz: sanity check and topological clues for linear and non-linear mappings, Comput. Graph. Forum 30 (1) (2011) 113–125.
[3] N. Heulot, J. Fekete, M. Aupetit, Visualizing dimensionality reduction artifacts: an evaluation, CoRR (2017), arXiv:abs/1705.05283.
[4] J. Chuang, D. Ramage, C.D. Manning, J. Heer, Interpretation and trust: designing model-driven visualizations for text analysis, CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012, (2012), pp. 443–452.
[5] M. Aupetit, Visualizing distortions and recovering topology in continuous projection techniques, Neurocomputing 70 (7–9) (2007) 1304–1330.
[6] S. Liu, B. Wang, P. Bremer, V. Pascucci, Distortion-guided structure-driven interactive exploration of high-dimensional data, Comput. Graph. Forum 33 (3) (2014) 101–110.
[7] J. Stahnke, M. Dörk, B. Müller, A. Thom, Probing projections: interaction techniques for interpreting arrangements and errors of dimensionality reductions, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 629–638.
[8] D. Sacha, H. Senaratne, B.C. Kwon, G.P. Ellis, D.A. Keim, The role of uncertainty, awareness, and trust in visual analytics, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 240–249.
[9] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (6) (2003) 1373–1396.
[10] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[11] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, CoRR (2002). cs.LG/0212008
[12] L. van der Maaten, G.E. Hinton, Visualizing high-dimensional data using t-sne, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[13] N. Pezzotti, B.P.F. Lelieveldt, L. van der Maaten, T. Höllt, E. Eisemann, A. Vilanova, Approximated and user steerable tsne for progressive visual analytics, IEEE Trans. Vis. Comput. Graph. 23 (7) (2017) 1739–1752.
[14] S. Cheng, K. Mueller, The data context map: fusing data and attributes into a unified display, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 121–130.
[15] H. Kim, J. Choo, H. Park, A. Endert, Interaxis: steering scatterplot axes via observation-level interaction, IEEE Trans. Vis. Comput. Graph. 22 (1) (2016) 131–140.
[16] B.C. Kwon, H. Kim, E. Wall, J. Choo, H. Park, A. Endert, Axisketcher: interactive nonlinear axis mapping of visualizations through user drawings, IEEE Trans. Vis. Comput. Graph. 23 (1) (2017) 221–230.
[17] J. Choo, S. Bohn, H. Park, Two-stage framework for visualization of clustered high dimensional data, Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11–16 October 2009, part of VisWeek 2009, (2009), pp. 67–74.
[18] J. Choo, H. Lee, J. Kihm, H. Park, ivisclassifier: an interactive visual analytics system for classification based on supervised dimension reduction, Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24–29 October 2010, part of VisWeek 2010, (2010), pp. 27–34.
[19] M. Gleicher, Explainers: expert explorations with crafted projections, IEEE Trans. Vis. Comput. Graph. 19 (12) (2013) 2042–2051.
[20] M. Sips, B. Neubert, J.P. Lewis, P. Hanrahan, Selecting good views of high-dimensional data using class consistency, Comput. Graph. Forum 28 (3) (2009) 831–838.
[21] M. Sedlmair, A. Tatu, T. Munzner, M. Tory, A taxonomy of visual cluster separation factors, Comput. Graph. Forum 31 (3) (2012) 1335–1344.
[22] D.H. Jeong, C. Ziemkiewicz, B.D. Fisher, W. Ribarsky, R. Chang, Ipca: an interactive system for pca-based visual analytics, Comput. Graph. Forum 28 (3) (2009) 767–774.
[23] J.E. Nam, K. Mueller, Tripadvisor^{n-d}: a tourism-inspired high-dimensional space exploration framework with overview and detail, IEEE Trans. Vis. Comput. Graph. 19 (2) (2013) 291–305.
[24] D.J. Lehmann, H. Theisel, Orthographic star coordinates, IEEE Trans. Vis. Comput. Graph. 19 (12) (2013) 2615–2624.
[25] X. Yuan, D. Ren, Z. Wang, C. Guo, Dimension projection matrix/tree: interactive subspace visual exploration and analysis of high dimensional data, IEEE Trans. Vis. Comput. Graph. 19 (12) (2013) 2625–2633.
[26] J.H. Friedman, J.W. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Trans. Comput. 23 (9) (1974) 881–890.
[27] D. Cook, A. Buja, J. Cabrera, C. Hurley, Grand tour and projection pursuit, J. Comput. Graph. Stat. 4 (3) (1995) 155–172.
[28] R.M. Martins, D.B. Coimbra, R. Minghim, A.C. Telea, Visual analysis of dimensionality reduction quality for parameterized projections, Comput. & Graph. 41 (2014) 26–42.
[29] A. Tatu, F. Maass, I. Färber, E. Bertini, T. Schreck, T. Seidl, D.A. Keim, Subspace search and visualization to make sense of alternative clusterings in high-dimensional data, 2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14–19, 2012, (2012), pp. 63–72.
[30] S. Liu, B. Wang, J.J. Thiagarajan, P. Bremer, V. Pascucci, Visual exploration of high-dimensional data through subspace analysis and dynamic projections, Comput. Graph. Forum 34 (3) (2015) 271–280.
[31] K. Chen, L. Liu, Ivibrate:interactive visualization-based framework for clustering large datasets, ACM Trans. Inf. Syst. 24 (2) (2006) 245–294.
[32] E.J. Nam, Y. Han, K. Mueller, A. Zelenyuk, D. Imre, Clustersculptor: a visual analytics tool for high-dimensional data, Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2007, Sacramento, California, USA, October 30–November 1, 2007, (2007), pp. 75–82.
[33] P. Joia, D. Coimbra, J.A. Cuminato, F.V. Paulovich, L.G. Nonato, Local affine multidimensional projection, IEEE Trans. Vis. Comput. Graph. 17 (12) (2011) 2563–2571.
[34] E.T. Brown, J. Liu, C.E. Brodley, R. Chang, Dis-function: learning distance functions interactively, 2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14–19, 2012, (2012), pp. 83–92.
[35] X. Hu, L. Bradel, D. Maiti, L. House, C. North, S. Leman, Semantics of directly manipulating spatializations, IEEE Trans. Vis. Comput. Graph. 19 (12) (2013) 2052–2059.
[36] J. Seo, B. Shneiderman, A rank-by-feature framework for interactive exploration of multidimensional data, Inf. Vis. 4 (2) (2005) 96–113.
[37] G. Albuquerque, M. Eisemann, M.A. Magnor, Perception-based visual quality measures, 2011 IEEE Conference on Visual Analytics Science and Technology, VAST 2011, Providence, Rhode Island, USA, October 23–28, 2011, (2011), pp. 13–20.
[38] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M.A. Magnor, D.A. Keim, Combining automated analysis and visualization techniques for effective exploration of high-dimensional data, Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11–16 October 2009, part of VisWeek 2009, (2009), pp. 59–66.
[39] S. Johansson, J. Johansson, Interactive dimensionality reduction through user-defined combinations of quality metrics, IEEE Trans. Vis. Comput. Graph. 15 (6) (2009) 993–1000.
[40] L. Wilkinson, A. Anand, R.L. Grossman, Graph-theoretic scagnostics, IEEE Symposium on Information Visualization (InfoVis 2005), 23–25 October 2005, Minneapolis, MN, USA, (2005), p. 21.
[41] D.T. Nhon, L. Wilkinson, Scagexplorer: exploring scatterplots by their scagnostics, IEEE Pacific Visualization Symposium, PacificVis 2014, Yokohama, Japan, March 4–7, 2014, (2014), pp. 73–80.

[42] A. Anand, L. Wilkinson, D.T. Nhon, Visual pattern discovery using random projections, 2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14–19, 2012, (2012), pp. 43–52.
[43] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3–6, 1996, (1996), pp. 336–343.
[44] H. Zhou, P. Wang, H. Li, Research on adaptive parameters determination in dbscan algorithm, J. Inf. Comput. Sci. 9 (7) (2012) 1967–1973.
[45] E. Kandogan, Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations, 2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14–19, 2012, (2012), pp. 73–82.
[46] A. Buja, D. Cook, D. Asimov, C. Hurley, Computational Methods for High-dimensional Rotations in Data Visualization, in: C. Rao, E. Wegman, J. Solka (Eds.), Data Mining and Data Visualization, Handbook of Statistics, 24 Elsevier, 2005, pp. 391–413.
[47] P.-A. Absil, R. Mahony, R. Sepulchre, Riemannian geometry of grassmann manifolds with a view on algorithmic computation, Acta Applic. Math. 80 (2) (2004) 199–220.
[48] M. Lichman, UCI machine learning repository, 2013.
[49] A. Georghiades, P. Belhumeur, D. Kriegman, Yale Face Database, 2 Center for computational Vision and Control at Yale University, 1997, p. 6. http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html