Data Visualization Using Interactive Dimensionality Reduction and Improved Color-Based Interaction Model

P.D. Rosero-Montalvo[1,2(✉)], D.F. Peña-Unigarro[3], D.H. Peluffo[1,4], J.A. Castro-Silva[5], A. Umaquinga[1], and E.A. Rosero-Rosero[1]

[1] Universidad Técnica Del Norte, Ibarra, Ecuador
pdrosero@utn.edu.ec
[2] Instituto Tecnológico Superior 17 de Julio, Ibarra, Ecuador
[3] Universidad de Nariño, Pasto, Colombia
[4] Corporación Universitaria Autónoma de Nariño, Pasto, Colombia
[5] Universidad Surcolombiana, Neiva, Huila, Colombia

Abstract. This work presents an improved interactive data visualization interface based on a mixture of the outcomes of dimensionality reduction (DR) methods. Broadly, it works as follows: The user can input the mixture weighting factors through a visual and intuitive interface with a primary-light-colors-based model (Red, Green, and Blue). By design, such a mixture is a weighted sum of the color tone. Additionally, the low-dimensional representation space produced by DR methods are graphically depicted using scatter plots powered via an interactive data-driven visualization. To do so, pairwise similarities are calculated and employed to define the graph to simultaneously be drawn over the scatter plot. Our interface enables the user to interactively combine DR methods by the human perception of color, while providing information about the structure of original data. Then, it makes the selection of a DR scheme more intuitive -even for non-expert users.

Keywords: Color-based model · Data visualization · Dimensionality reduction · Pairwise similarity

1 Introduction

The advance of technology can be observed through the integration into everyday human activities in devices like sensors, mobile applications, web pages and companies integrated systems that allows the collection of user data for the purpose of finding valuable information. Subsequently, such information can be converted into useful knowledge for humans to finally make proper decisions [1]. As a consequence, there had been an increasing volume of data generating then the need for computational systems become more robust by incorporating machine learning algorithms, so that knowledge generation can be reached in an optimal way (i.e. by avoiding information redundancy and noise) mainly for unstructured and multivariate databases [2,3].

J.M. Ferrández Vicente et al. (Eds.): IWINAC 2017, Part II, LNCS 10338, pp. 289–298, 2017.
DOI: 10.1007/978-3-319-59773-7_30

The dimensionality reduction (DR) is one of the approaches to make data perceivable in a simpler and compact way, since representing a set of high dimensional data increases the complexity of user's understanding due that the information may become abstract specially, regarding the manner to describe objects being non-physical [4].

DR methods are able to simplify the description of the data set that can represent large volumes of information at optimal processing times, while keeping the same properties of the complex high-dimensional data. As a result, it favors compression, elimination of redundancy and improves the processes with the implementation of machine learning algorithms. Then, it also reduces the computational cost. In virtue of the above, the user obtain a better analysis with an effective pattern recognition and considering a smaller number of dimensions [2].

Once performed the DR stage, the interactive visualization takes place to create an interface between the human beings and the computational processes with their algorithms of machine learning. Such an interface allows to generate efficient forms of mathematical and statistical processes to the user's understanding, where he can manipulate the information until to determine the best method in each specific information type. However, presenting data in an understandable, dynamic and intuitive way with transparent mathematical processes to the user becomes a challenge [4,5]. The visualization of data only succeeds when it can encodes the information in a way that our eyes can discern and our brains can understand. To achieve this objective is more a science than an art, which can only be achieved through the study of human perception [6].

Some works [2,3,7,8] have accomplished interfaces with methods of dimensionality reduction with different approaches and ways of generating mixtures between the different DR algorithms, so that the user can intuitively select the most appropriate in a visual way. [9] also has a pairwise similarities to determinate the affinity for the DR method mixture result but all these works does not focus in the interface design and reason to applies the color inside the data visualization. The present work is an improved approach to those cited works by optimizing the user interaction with the interface by associating DR methods with colors and RGB bars that are easier to associate with processes previously learned by user with the aim of create more intuitive environments.

For experiments, we used the spherical data set in 3-D, the evaluation of the performance of the mixture was considered conventional methods of DR such as: multidimensional classical scaling (CMDS) [10], locally linear embedding (LLE) and t-Student distributed (TSNE) [11,12], in addition to provide more interactivity to the user can control color bars tone by varying their parameter, also integrates a slider to control and visualize the affinity of the points of the 2-D graphic in relation to the 3-D graphic. To perform the mixing of methods the user has the RGB bars (Red, Green, Blue) in order to modify the color tone in a container with scale from 0 to 255, for weights factors are performed by an average in relation to the tonality summation of the RGB bars, as a result the 2-D circumference is graphically observed in a friendly and interactive way [6].

The remaining of the paper is organized as follows: In Sect. 2, Data visualization via dimensionality reduction is outlined. Section 3 introduces the proposed interactive data visualization scheme. Experimental setup and results are presented in Sects. 4 and 5, respectively. Finally, Sect. 6 gathers some final remarks as conclusions and future work.

2 Data Visualization via Dimensionality Reduction

The data visualization means the interaction between the human and the system (interface) which handle thousands of complex data sets records. This allowing an in depth knowledge and pattern recognition in such a way that they become information comprehensible for the user. The 2- or 3-dimensional representation maybe can the most intuitive ways of visualizing large volumes numerical data for analyzing and find information when strong hypotheses about data are not yet available [13], besides can be readily represented using a scatter plot, giving the facility to the human eye for its interpretation, since they can see easily in two dimensions and the brain is in charge of calculating the distance between the object, giving the perception of third dimension [14]. In this way, dimensionality reduction methods are born from the need to obtain a simple representation of the complexity or relationship of big volume of data into a low dimension space, with the least loss of information possible [15]. So, when performing a DR method, a more realistic and intelligible visualization for the user is expected [11]. More technically, the goal of dimensionality reduction is to embed a high dimensional data matrix $\boldsymbol{Y} = [\boldsymbol{y}_i]_{1\leq 1\leq N}$ such that $\boldsymbol{y}_i \in \mathbb{R}^D$ into a low-dimensional, latent data matrix $\boldsymbol{X} = [\boldsymbol{x}_i]_{1\leq 1\leq N}$ being $\boldsymbol{y}_i \in \mathbb{R}^d$, where $d < D$ [11,16]. Figure 1 depicts an instance where a manifold (3-dimensional sphere) is embedded into a 2-D representation, which resembles to an unfolded version of the original manifold.

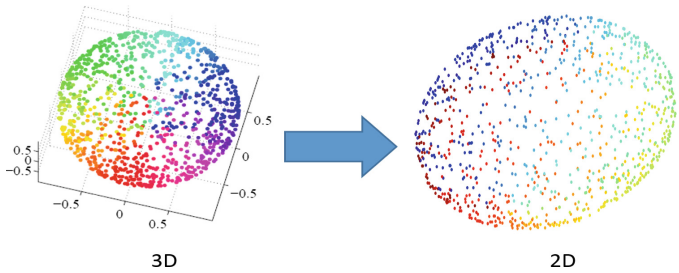

Fig. 1. Dimensionality reduction effect over an artificial (3-dimensional) spherical shell manifold. Resultant embedded (2-dimensional) data is an attempt to unfolding the original data.

3 Interactive Data Visualization Scheme

The proposed visualization improve approach, here so-called DataVisSim, involves three main stages: mixture of DR outcomes, interaction, and visualization, as depicted in the block diagram of Fig. 2. One of the most important contributions of this work is that information on the structure of the input high-dimensional space is added to the visual final representation, by using a pairwise-similarity-based scheme and the greater accuracy of the proportion of DR methods, giving the user the knowledge of their DR mixture in percentages according to the color's tonality.

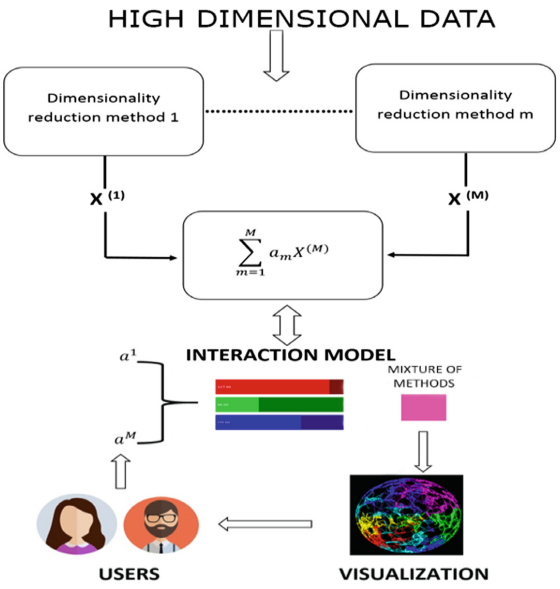

Fig. 2. Block diagram of proposed interactive data visualization using dimensionality reduction and similarity-based representations (DataVisSim). It works as follows: First the interface loads the database of high dimension and reduced dimension, in second step the user can manipulate the color bars for performs a mixture between DR methods, at third step when the user has decided the weighting factors for the aforementioned mixture we can validate his choice with a novel similarity-bases approach, and finally the embedded representation can be saved. (Color figure online)

3.1 Mixture

Let us suppose that the input matrix $\boldsymbol{Y}$ is reduced by using $\mathbf{M}$ different DR methods, yielding then a set of lower-dimensional representations: $\{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(M)}\}$. Herein, we propose to perform a weighted sum in the form:

$$\bar{\boldsymbol{X}} = \sum_{m=1}^{M} \alpha_m \boldsymbol{X}^{(m)}, \tag{1}$$

where $\{\alpha_1, \cdots, \alpha_M\}$ are the weighting factors. To make the selection of weighting factors intuitive, we use probability values so that $0 \leq \alpha_m \leq 1$ and $\sum_{m1=1}^{M} \alpha_m = 1$, and therefore all matrices $\boldsymbol{X}^{(m)}$ should be normalized to rely within a unit hypersphere.

3.2 Interaction Model

An appropriate design of an interface, allows to the user to create own mental models that help to understand the information on the screen of a computer. Through previous experiences and expectations the user shapes perceptions. The interaction between the user and the system must be a fluid dialogue in the style of the interface where the senses of vision, hearing and touch interact [17]. This work emphasizes touch and vision based on an additive synthesis model that emits light directly to the source of illumination of some kind, representing a color by mixing the 3 primary RGB light colors (Red, Green, Blue) [18]. This form of representation and creation of color is used since the human eye has photoreceptors, approximately 64% of the cones (photosensitive cells) contains photo pigments (light sensitive proteins), 32% contain green and only about 2% contains photo blue pigments [6]. Consequently the human eye has greater sensitivity RGB colors based on human perception and the combination between light, object and observer [17].

The proposed interface allows the process between luminescence, contrast, color and movement that allows a sensation of physical stimuli to the human being and can pay attention to the mixture of DR. The HSV model (Hue, Saturation and Value) represented in a computer according to Fig. 3. The user can to manipulate the values of the bars of tone of the RGB colors, the increase or decrease of their value is given according to the saturation of the bar, giving the feeling of filling or emptying it [6,18]. The interface works as follows: the user loads the sphere in third dimension, once visualized the figure has the RGB bars that can modify the percentage of tone of the same, so the user change the weight of the DR methods and they can be observed the 2-D figure about the existing blend and the resulting color of the RGB. Finally, the work can be save for later analysis of the new data set.

For the sake of interactivity, the values of every α_m -required to calculate $\bar{\boldsymbol{X}}$ according to Eq. (1)- are to be defined by the users using an color saturation-bar available in the interface. Within a friendly-user and intuitive environment, in the case than more DR methods is selected, weighting factors can be readily imputed by just select values from bars and choose the color saturation between RGB color bars are definite by fundamental counting principle, which given a set of n elements, is defined as an arrangement of n in order of k ($k <= n$) to each tuple that can be formed by taking k different elements among n given. The user can move the bars when they consider suitable.

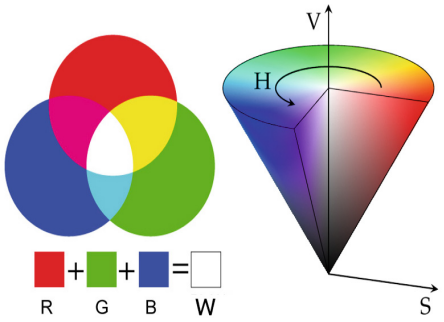

Fig. 3. The picture in the left side explain the way of the RGB color can make others colors, the right side show the saturation and hue with model HSV and the interaction with RGB color for visualize different color tone (Color figure online)

3.3 Similarity-Based Visualization

The most used method to visualize 2- or 3-dimensional data is the scatter plot. In this work, we introduce a similarity-based visualization approach with the aim to provide a visual hint about the structure of the high-dimensional input data matrix $\boldsymbol{Y}$ into the scatter plot of its representation in a lower-dimensional space To do so, we use a pairwise similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$, such that $\boldsymbol{S} = [\boldsymbol{s}_{ij}]$. In terms of graph theory, entries $\boldsymbol{s}_{ij}$ defines the similarity or affinity between the $i-th$ and $j-th$ data point from $\boldsymbol{Y}$. Doing so, we can hold the structure of original input space in a topological fashion, specifically in terms of pairwise relationships. For visualization purposes, such a similarity is used to define graphically the relationship between data points by plotting edges. In order to control the amount of edges and make an appealing visual representations, the value of $\boldsymbol{s}_{ij}$ is constrained as $\boldsymbol{s}_{ij} > \boldsymbol{s}_{max}$, being $\boldsymbol{s}_{max}$ a maximum admissible similarity value to be given by the users as well. In other words, our visualization approach consists of building a graph with constrained affinity values.

4 Experimental Setup

Database: In order to visually evaluate the performance of the DataVisSim approach, we use an artificial spherical shell (N = 1000 data points and D = 3), as depicted in Fig. 1.

Parameter Settings and Methods: In order to capture the local structure for visualization, i.e. data points being neighbors, we utilize the Gaussian similarity given by: $\boldsymbol{s}_{ij} = -exp(-0.5\|\boldsymbol{y}_{(i)} - \boldsymbol{y}_{(j)}\|^2/\sigma^2)$. The parameter is a bandwidth value set as 0.1, being the 10% of the hypersphere ratio (applicable once matrices are normalized as discussed in Sect. 3.1. To perform the dimensionality reduction we consider $M = 3$ DR methods, namely: CMDS, LLE, and t-SNE. All of them are intended to obtain spaces in dimension $d = 2$.

Performance Measure: To quantify the performance of studied methods, the scaled version of the average agreement rate $R_{NX}(K)$ introduced in [19] is used, which is ranged within the interval [0, 1]. Since $R_{NX}(K)$ is calculated at each perplexity value from 2 to $N-1$, a numerical indicator of the overall performance can be obtained by calculating its area under the curve (AUC). The AUC assesses the dimension reduction quality at all scales, with the most appropriate weights.

5 Results

Figure 4 shows the scatter plots for the resultant low-dimensional spaces obtained by the considered dimensionality reduction methods for the interface. These DR methods has been insert doing relationship with eye perception in front of the computer.

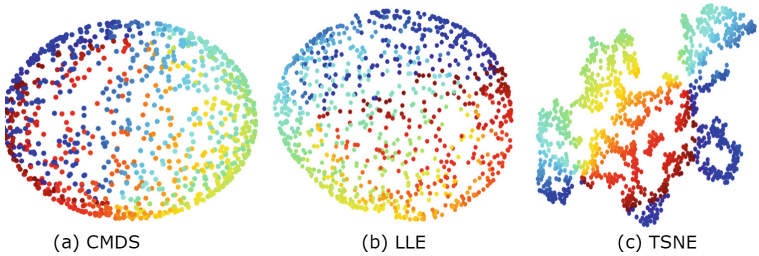

Fig. 4. The effects of dimensionality reduction methods considered on the 3-D sphere. The results are embedded data represented in a bi-dimensional space.

The interface was developed in Processing in virtue of the ease of represent information in a visual way, the interface shows all the content in relation of pixels, of this way all the data points must be modify and change only to positive points. In the Fig. 5 shows the final interactive interface with RGB model.

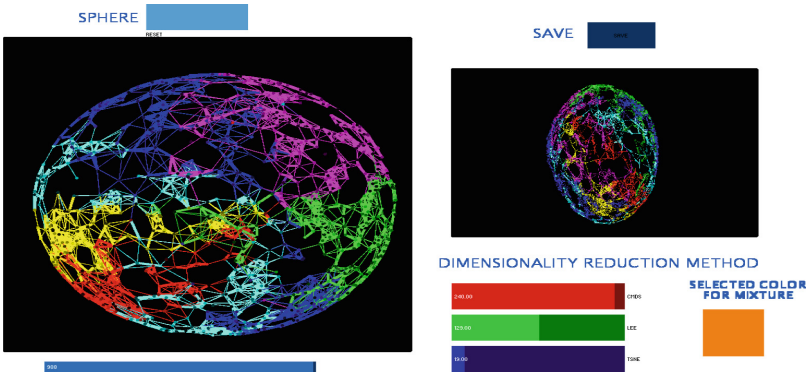

Fig. 5. Finally interface developed in Processing with interactive RGB model. Sample video https://sites.google.com/site/intelligentsystemsrg/home/gallery/

Figure 6 shows the result with the interaction between the user and the interface in three important aspects: RGB mixture color, the 2-D visualization and the mixture performance. As seen, $R_{NX}(K)$ measure allows for assessing both the different mixtures and the methods independently Since the area under its curve represents a representation quality measure of the low-dimensional space, is in turn a visual and intuitive indicator that helps the user to find the best either a single DR method or the proper mixture [9].

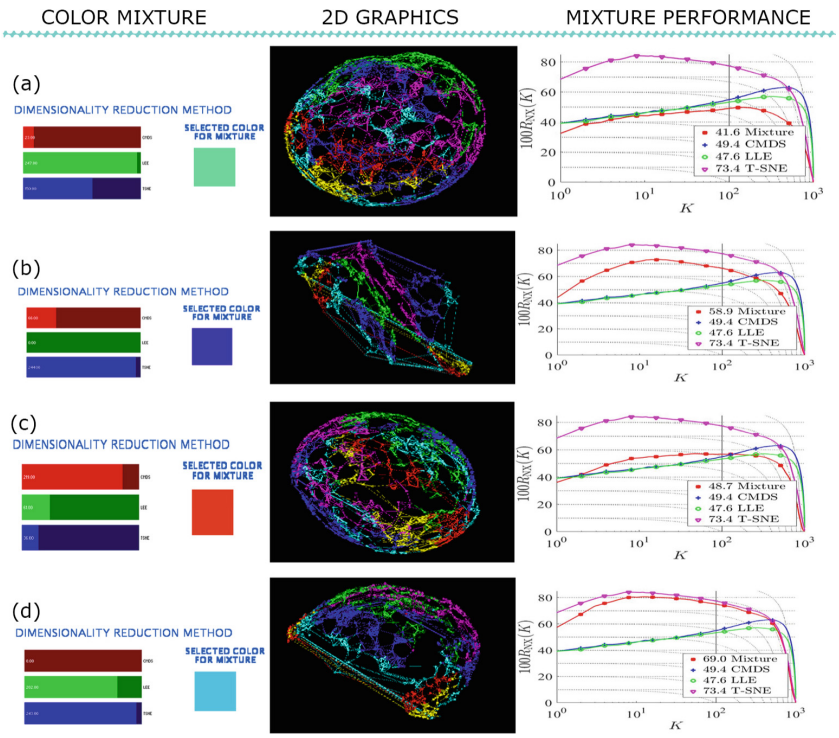

Fig. 6. The picture show four results of the interaction between interface and the user and how they find their mixture, in some cases the mixture had been with good performance and others do not.

As well, the interface incorporates a slider bar to dynamically draw the edges between nodes. This is useful for visual analysis given that it allows to relate the structure of high-dimensional data (original data) within the visualization of the low-dimensional representation space, the thickness line amounts to relation between the points in 2-D and 3-D dimension. Therefore, is easy to see by the user the DR mixture quality, as the picture shows in Fig. 7.

This follows from the interaction between the user and the interface, which shows greater preferences in blends of blue and green in men, while in women it changes its selection to yellow and pink colors. This indicates that the most

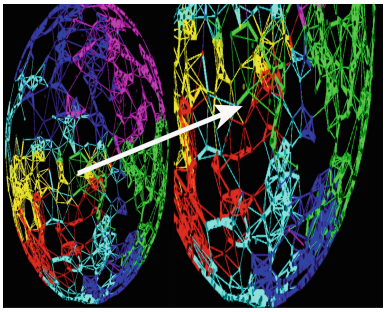

Fig. 7. The figure indicates the affinity of the points by making the relation with the thickness of the lines that join to each other. (Color figure online)

widely used method is CMDS and T-SNE, respectively. In addition, the affinity bar allows the verification of the result of the mixture, giving the opportunity to change the result by observing the distance of the points increases when plotting them with The RGB mix.

6 Conclusions and Future Work

This paper presents an improved visualization method, which is based on the mixture of dimensionality reduction methods by following a color-human-perception criterion and enables users to have mental structures on the performance of the obtained results by visualizing a similarity measure calculated at the high dimension data. Particularly, the mixture is performed as a weighted sum whose weights are defined as the average of the tonality of the primary light colors of RGB.

As a future work, other dimensionality reduction methods are to be integrated into the interface and improve intuitive way of generate mixture DR methods. The interface needs more mathematical developments regarding the way to perform the mixture of DR methods.

Aknowledgments. The authors would like to thank the project "Desarrollo de una metodología de visualización interactiva y eficaz de información en Big Data" supported by VIPRI from Universidad de Nariño - Colombia, as well as Universidad Técnica del Norte - Ecuador.

References

1. Ward, M.O., Grinstein, G., Keim, D.: Interactive Data Visualization: Foundations, Techniques, and Applications. CRC Press, Boca Raton (2010)
2. Salazar-Castro, J., Rosas-Narváez, Y., Pantoja, A., Alvarado-Pérez, J.C., Peluffo-Ordóñez, D.H.: Interactive interface for efficient data visualization via a geometric approach. In: 2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA), pp. 1–6. IEEE (2015)

3. Peña-Unigarro, D.F., Salazar-Castro, J.A., Peluffo-Ordóñez, D.H., Rosero-Montalvo, P.D., Oña-Rocha, O.R., Isaza, A.A., Alvarado-Pérez, J.C., Theron, R.: Interactive visualization methodology of high-dimensional data with a color-based model for dimensionality reduction. In: 2016 XXI Symposium on Signal Processing, Images and Artificial Vision (STSIVA), pp. 1–7, August 2016
4. Alvarado-Pérez, J.C., Peluffo-Ordóñez, D.H., Therón, R.: Visualización y métodos kernel: integrando inteligencia natural y artificial (2016)
5. Dai, W., Hu, P.: Research on personalized behaviors recommendation system based on cloud computing. Indones. J. Electr. Eng. Comput. Sci. **12**(2), 1480–1486 (2013)
6. Dastan, M.: The role of visual perception in data visualization. J. Vis. Lang. Comput. **13**(6), 601–622 (2002)
7. Peluffo-Ordóñez, D.H., Alvarado-Pérez, J.C., Lee, J.A., Verleysen, M., et al.: Geometrical homotopy for data visualization. In: European Symposium on Artificial Neural Networks (ESANN 2015). Computational Intelligence and Machine Learning. (2015)
8. Díaz, I., Cuadrado, A.A., Pérez, D., García, F.J., Verleysen, M.: Interactive dimensionality reduction for visual analytics. In: Proceedings of the 22th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014), pp. 183–188. Citeseer (2014)
9. Rosero-Montalvo, P., Diaz, P., Salazar-Castro, J.A., Peña-Unigarro, D.F., Anaya-Isaza, A.J., Alvarado-Pérez, J.C., Therón, R., Peluffo-Ordóñez, D.H.: Interactive data visualization using dimensionality reduction and similarity-based representations. In: Beltrán-Castañón, C., Nyström, I., Famili, F. (eds.) CIARP 2016. LNCS, vol. 10125, pp. 334–342. Springer, Cham (2017). doi:10.1007/978-3-319-52277-7_41
10. Borg, I., Groenen, P.J.: Modern Multidimensional Scaling: Theory and Applications. Springer Science & Business Media, New York (2005)
11. Peluffo-Ordóñez, D.H., Lee, J.A., Verleysen, M.: Short review of dimensionality reduction methods based on stochastic neighbour embedding. In: Villmann, T., Schleif, F.-M., Kaden, M., Lange, M. (eds.) Advances in Self-Organizing Maps and Learning Vector Quantization. AISC, vol. 295, pp. 65–74. Springer, Cham (2014). doi:10.1007/978-3-319-07695-9_6
12. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
13. Park, Y., Cafarella, M., Mozafari, B.: Visualization-aware sampling for very large databases. In: 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pp. 755–766, May 2016
14. Emberson, L.L., Amso, D.: Learning to sample: eye tracking and fMRI indices of changes in object perception. J. Cogn. Neurosci. **24**(10), 2030–2042 (2012)
15. Bertini, E., Lalanne, D.: Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In: Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration, pp. 12–20. ACM (2009)
16. Peluffo-Ordóñez, D.H., Lee, J.A., Verleysen, M.: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 171–177. IEEE (2014)
17. Levkowitz, H.: Color Theory and Modeling for Computer Graphics, Visualization, and Multimedia Applications. Springer, New York (1997)
18. Dix, A.: Human-Computer Interaction. Springer, New York (2009)
19. Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing **112**, 92–108 (2013)