SOFTWARE TOOL ARTICLE

DataViz: visualization of high-dimensional data in virtual reality [version 1; referees: awaiting peer review]

Eric Feng [1], Xijin Ge[2]

[1]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, California, 94720, USA
[2]Department of Mathematics and Statistics, South Dakota State University, Brookings, SD, 57007, USA

v1 **First published:** 23 Oct 2018, **7**:1687 (doi: 10.12688/f1000research.16453.1)
Latest published: 23 Oct 2018, **7**:1687 (doi: 10.12688/f1000research.16453.1)

Abstract

Virtual reality (VR) simulations promote interactivity and immersion, and provide an opportunity that may help researchers gain insights from complex datasets. To explore the utility and potential of VR in graphically rendering large datasets, we have developed an application for immersive, 3-dimensional (3D) scatter plots. Developed using the Unity development environment, DataViz enables the visualization of high-dimensional data with the HTC Vive, a relatively inexpensive and modern virtual reality headset available to the general public. DataViz has the following features: (1) principal component analysis (PCA) of the dataset; (2) graphical rendering of said dataset's 3D projection onto its first three principal components; and (3) intuitive controls and instructions for using the application. As a use case, we applied DataViz to visualize a single-cell RNA-Seq dataset. DataViz can help gain insights from complex datasets by enabling interaction with high-dimensional data.

Keywords

Virtual Reality, Principal Component Analysis, Visualization, High-dimensional, Unity

Open Peer Review

Referee Status: *AWAITING PEER REVIEW*

Discuss this article

Comments (0)

Corresponding author: Xijin Ge (Xijin.Ge@sdstate.edu)

Author roles: Feng E: Software, Visualization, Writing – Original Draft Preparation; **Ge X**: Conceptualization, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This material is based upon work supported by the National Science Foundation/EPSCoR Grant Number IIA – 1355423 and by the State of South Dakota. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the National Science Foundation.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Feng E and Ge X. **DataViz: visualization of high-dimensional data in virtual reality [version 1; referees: awaiting peer review]** *F1000Research* 2018, **7**:1687 (doi: 10.12688/f1000research.16453.1)

First published: 23 Oct 2018, **7**:1687 (doi: 10.12688/f1000research.16453.1)

Introduction

Historically, we have heavily relied on 2-dimensional (2D) graphical displays to communicate large amounts of data. These graphs have also been useful in finding patterns within datasets and building intuition for more accurate and meaningful analysis. However, for large and complex datasets containing numerous dimensions, traditional 2D charts and graphs are inadequate in demonstrating the multi-faceted nature of relevant information.

The 3-dimensional (3D) visualization of datasets are valuable because they offer a starting solution to the problem above; the addition of another dimension allows for more information to be presented and thus decreases the potential for misinterpretation while concurrently increasing the possibility of pattern-matching and building intuition.

This paper researches the potential of using virtual reality (VR) as a platform to graphically render datasets in 3D by creating a visualization application. VR is already being used in a variety of fields including flight simulations[1], mental health therapy[2], and even visualizations of molecules and their interactions[3]. In the specific field of data visualization, several applications exist, including a surround-screen, projection-based visualizer named CAVE[4], one developed using OpenGL that visualizes economic data[5], and iViz[6], an efficient and intuitive visualizer using VR that is also the most similar to the application developed in this research. DataViz attempts to make further progress by creating a modern, intuitive, and readily available application.

We continue to explore the potential of VR in the graphical rendering of large datasets; to do so, we have developed a Unity3D VR application for HTC Vive (HTC, New Taipei City, Taiwan) that runs principal component analysis (PCA) on datasets before graphing the subsequent projection into three dimensions. The software was designed to run efficiently with an intuitive interface.

Methods

Implementation

In the design of this application, special consideration was given to the following elements: the method of data analysis, the format of the input data, the limitations in computing power of the selected platform, and the mitigation of motion sickness.

Data analysis

The primary method of data analysis is PCA. The rationale behind this decision is that because humans live in three dimensions, the most intuitive manner of visualization is one that plots in that space. In this sense, PCA is excellent at taking large dimensional data and reducing them to plottable 3D coordinates, making the resulting graph more intuitive, and helping users discover patterns and develop scientific intuition.

Input data

DataViz only accepts data in the table format (CSV or TXT). Occasionally, the user would want to analyze the transpose of the provided data. Although the transpose of a table could easily be found using specialized functions in Numpy or R, we decided to add the transpose functionality into the application.

In addition to transposition, DataViz also allows the user to omit specific columns from the file. This may be due to a variety of reasons including an unwanted dimension of data or column names. This functionality allows researchers to analyze only the columns they are interested in.

The user may also have a column that labels the points. Users can designate a specific column that differentiates the data with various tags, and these groups will show up in a graph legend during runtime.

Limitations in computing power

The engine used in developing this application is Unity®. Unity is one of the most popular platforms for VR development but is not specifically designed for statistical analysis. Therefore, PCA on large datasets may result in slow run times, especially when there is a lack of an appropriate graphics card or other computational power involved. To overcome this limitation, the application can also accept coordinate data derived from PCA or other dimensionality reduction methods such as t-SNE[7]. In this manner, users can circumvent the slower computations associated with Unity.

VR considerations

When implementing the VR aspect of the application, we concentrated on two main considerations: immersion and motion sickness. For the former, the primary goal was to allow the user to focus on the graphical rendering of his/her data without being bothered by the complicated details on how to use the tool. In pursuit of this, we designed an intuitive interface and series of menus, with clear instructions on the associated GitHub page in 'Software Availability'.

Another concern when designing for VR was motion sickness. Motion sickness is a consequence of conflicting input between visual and inner ear senses and is a major problem in current VR simulations[8]. It has been found that motion sickness is a consequence of the action of motion and not displacement itself, and as a result, we designed our movement to be in short bursts of teleportation.

The application is built using the Unity® engine with scripting done in C#. The PCA and transpose implementation is from the Accord.Net 3.8 framework (http://accord-framework.net). The mouse embryonic development data used in the case study is from Ref 9.

Operation

DataViz was designed to be an intuitive application for graphically rendering large datasets. Upon opening the software, a user should follow the onscreen prompts and fill out the appropriate parameters to input their dataset as well as use the extra functionalities described above. DataViz automatically runs PCA on the input dataset according to user configurations. If needed, more detailed instructions can be found on the associated GitHub page.

VR is a resource intensive activity. The following are guidelines for ensuring the quality and performance of DataViz.

System Requirements (https://www.vive.com/us/ready/):

- Processor: Intel i5-4590 / AMD FX 8350 equivalent or greater
- Graphics card: NVIDIA GeForce GTX 1060 or AMD Radeon RX 480, equivalent or better
- Memory: 4 GB RAM or more
- Video output: HDMI 1.4 or DisplayPort 1.2 or newer
- USB: 1x USB 2.0 or newer
- Operating system: Windows 7 SP1 or newer

Use case

Mouse embryo development

The primary goal in the development of this application was to determine the viability of using VR to graphical render and analyze complex data sets. After development, we tested the DataViz by analyzing a high-dimensional dataset regarding mouse embryonic development[9]. Using single-cell RNA sequencing (scRNA-Seq), Deng *et al.* generated hundreds of expression profiles of individual embryonic cells from zygote blastocyst stages.

By graphically rendering the 3D PCA projection of the data and subsequent analysis, we were able to verify an expected trend of embryo development; initial cell division (zygote stage to 16-cell stage) results in large-scale physical changes inside the embryo. This is in contrast to later cell division where the various stages of embryo development are more similar to one another. We can also see the developmental trajectory in the transcriptomic landscape (Figure 1).

This method of analysis has some limitations, the foremost being an inability to account for all the data present. While reducing high-dimensional data to three dimensions simplifies the resulting plot and may help formulate testable hypotheses through further research or build intuition and comprehension regarding the data provided, it is inevitable that we lose some of the variance present in higher dimensions. In this test case, Table 1 reveals the proportion of the data retained per principal component. One way of overcoming this would be to use non-linear dimensionality reduction methods like multidimensional scaling (MDS) or t-SNE.

Despite the shortcomings involved in the provided analysis and plotting approach, DataViz is still useful for categorizing the data into disjoint groups.

Discussion

Two of the primary motivations for using VR to visualize data were the introduction of a third dimension as well as increased interactivity with data. As shown by the Use Case, although the current functionality is limited to PCA, the application is useful in demonstrating the potential that VR has to offer in the analysis and communication of large, complex datasets.

Table 1. The application is unable to account for the full variance in the data. For example, in the test case of mouse embryo development, the resulting three-dimensional graph could only reveal 33% of the original dataset.

Variable	PC 1	PC 2	PC 3
Proportion of variance	0.210	0.088	0.034
Cumulative proportion	0.210	0.300	0.332

PC, principal component.

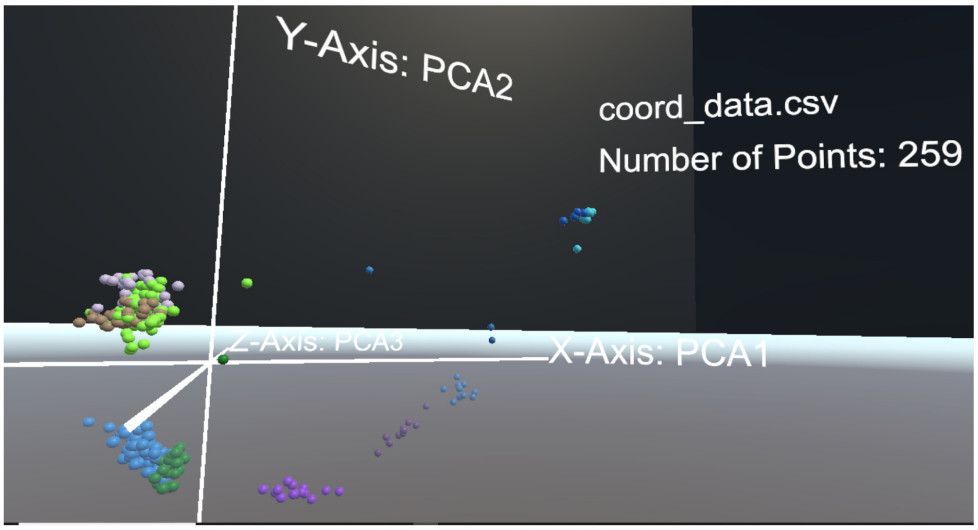

Figure 1. The application when plotting provided mouse embryonic development coordinate data. The graph displays the similarities among the blastocyst stages in comparison to changes in earlier stages of development. We can identify categories and general trends of the data using this method.

To understand this potential further, future research should focus on human trials in determining the statistical difference between the traditional 3D plot on a computer screen and a VR simulation regarding data comprehension and analysis. Additionally, in order to account for more variance in the original dataset, future research should consider other dimensionality reduction methods.

Conclusion

We have developed an application for visualizing high-dimensional data in VR. It reduces high-dimensional data using PCA before generating an immersive 3D scatter plot. It also contains a variety of functionalities including the ability to transpose the given input and to accept raw coordinate data. A major limitation of DataViz is its inability to account for the full variance in the dataset. Also, the amount of benefit that visualization receives from being in VR as opposed to on a 2D monitor is unknown.

Data availability

The data of mouse embryo development can be found in Deng *et al*., 2014[9]

Software availability

Source code and additional instructions available at: https://github.com/thunder2011/DataViz

Archived source code at time of publication: https://doi.org/10.5281/zenodo.1455787[10].

Grant information

This material is based upon work supported by the National Science Foundation/EPSCoR Grant Number IIA – 1355423 and by the State of South Dakota. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the National Science Foundation.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Chittaro L, Buttussi F: **Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety.** *IEEE Trans Vis Comput Graph.* 2015; **21**(4): 529–38.
PubMed Abstract | Publisher Full Text

2. Powers MB, Rothbaum BO: **Recent advances in virtual reality therapy for anxiety and related disorders: Introduction to the special issue.** *J Anxiety Disord.* 2018; pii: S0887-6185(18)30342-6.
PubMed Abstract | Publisher Full Text

3. O'Connor M, Deeks HM, Dawn E, *et al.*: **Sampling molecular conformations and dynamics in a multiuser virtual reality framework.** *Sci Adv.* 2018; **4**(6): eaat2731.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Cruz-Neira C, Sandin D, DeFanti T: **Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE.** Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '93, Anaheim, CA USA. 1993; 135–142.
Publisher Full Text

5. Sullivan P: **Graph-Based Data Visualization In Virtual Reality: A Comparison Of User Experiences.** California Polytechnic State University. 2016.
Publisher Full Text

6. Donalek C, Djorgovski SG, Cioc A, *et al.*: **Immersive and Collaborative Data Visualization Using Virtual Reality Platforms.** *2014 Ieee International Conference on Big Data (Big Data).* 2014; 609–614.
Publisher Full Text

7. van der Maaten L: **Accelerating t-SNE using Tree-Based Algorithms.** *Journal of Machine Learning Research.* 2014; **15**: 3221–3245.
Reference Source

8. Becker J, Ngo T: **Mitigating Visually-Induced Motion Sickness in Virtual Reality.** Stanford University. 2016.
Reference Source

9. Deng Q, Ramsköld D, Reinius B, *et al.*: **Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells.** *Science.* 2014; **343**(6167): 193–6.
PubMed Abstract | Publisher Full Text

10. Feng E: **thunder2011/DataViz: Centralize Package and Executible onto Github (Version v1.0.2).** *Zenodo.* 2018.
http://dx.doi.org/10.5281/zenodo.1455787