Data Visualization & Virtual Reality using Big Data

R. Gnana Bharathy
Lecturer Senior Grade
Department of Computer Engineering
Ayya Nadar Janaki Ammal Polytechnic College, India

Abstract— This paper provides a multi-disciplinary overview of the research issues and achievements in the field of Big Data and its visualization techniques and tools. The main aim is to summarize challenges in visualization methods for existing Big Data, as well as to offer novel solutions for issues related to the current state of Big Data Visualization. This paper provides a classification of existing data types, analytical methods, visualization techniques and tools, with a particular emphasis placed on surveying the evolution of visualization methodology over the past years. Based on the results, we reveal disadvantages of existing visualization methods. Despite the technological development of the modern world, human involvement (interaction), judgment and logical thinking are necessary while working with Big Data. Therefore, the role of human perceptional limitations involving large amounts of information is evaluated. Based on the results, a non-traditional approach is proposed: we discuss how the capabilities of Augmented Reality and Virtual Reality could be applied to the field of Big Data Visualization. We discuss the promising utility of Mixed Reality technology integration with applications in Big Data Visualization. Placing the most essential data in the central area of the human visual field in Mixed Reality would allow one to obtain the presented information in a short period of time without significant data losses due to human perceptual issues. Furthermore, we discuss the impacts of new technologies, such as Virtual Reality displays and Augmented Reality helmets on the Big Data visualization as well as to the classification of the main challenges of integrating the technology.

Key words: Big Data, Visualization, Virtual Reality, Augmented Reality, Mixed Reality, Human Interaction, Evolution

I. INTRODUCTION

The whole history of humanity is an enormous accumulation of data. Information has been stored for thousands of years. Data has become an integral part of history, politics, science, economics and business structures, and now even social lives. This trend is clearly visible in social networks such as Facebook, Twitter and Instragram where users produce an enormous stream of different types of information daily (music, pictures, text, etc.) [1]. Now, government, scientific and technical laboratory data as well as space research information are available not only for review, but also for public use. For instance, there is the 1000 Genomes Project [2, 3], which provide 260 terabytes of human genome data. More than 20 terabytes of data are publicly available at Internet Archive [4, 5], ClueWeb09 [6], among others.

This paper covers various issues and topics, but there are three main directions of this survey:

- Human cognitive limitations in terms of Big Data Visualization.
- Applying Augmented and Virtual reality opportunities towards Big Data Visualization.
- Challenges and benefits of the proposed visualization approach.

The rest of paper is organized as follows: The first section provides a definition of Big Data and looks at currently used methods for Big Data processing and their specifications. Also it indicates the main challenges and issues in Big Data analysis. Next, in the section Visualization methods, the historical background of this field is given, modern visualization techniques for massive amounts of information are presented and the evolution of visualization methods is discussed. Further in the last section, Integration with Augmented and Virtual Reality, the history of AR and VR is detailed with respect to its influence on Big Data. These developmental processes are supported by the proposed oncoming Big Data visualization extension for VR and AR, which can solve actual perception and cognition challenges.

II. BIG DATA: AN OVERVIEW

Today large data sources are ubiquitous throughout the world. Data used for processing may be obtained from measuring devices, radio frequency identifiers, social network message flows, meteorological data, remote sensing, location data streams of mobile subscribers and devices, and audio and video recordings. So, as Big Data is more and more used all over the world, a new and important research field is being established.

The mass distribution of the technology and innovative models that utilize these different kinds of devices and services appeared to be a starting point for the penetration of Big Data in almost all areas of human activity, including the commercial sector and public administration [27].

A. Big Data Processing Methods

Currently, there exist many different techniques for data analysis [36], mainly based on tools used in statistics and computer science. The most advanced techniques to analyze large amounts of data include: artificial neural networks [37–39]; models based on the principle of the organization and functioning of biological neural networks [40, 41]; methods of predictive analysis [42]; statistics [43, 44]; Natural Language Processing [45]; etc. Big Data processing methods embrace different disciplines including applied mathematics, statistics, computer science and economics. Those are the basis for data analysis techniques such as Data Mining [39, 46–49], Neural Networks [41, 50–52], Machine Learning [53–55], Signal Processing [56–58] and Visualization Methods [59–61]. Most of these methods are interconnected and used simultaneously during data processing, which increases system utilization tremendously (see Fig. 1). We would like to familiarize reader with the primary methods and techniques in Big Data processing. As this topic is not a focus of the paper, this list is not exhaustive. Nevertheless, the main interconnections between these methods are shown and application examples are given. Optimization methods are mathematical tools for efficient data analysis. Optimization includes numerical analysis focused on problem solving in various Big Data challenges: volume, velocity, variety and veracity [62] that will be discussed in more detail later.

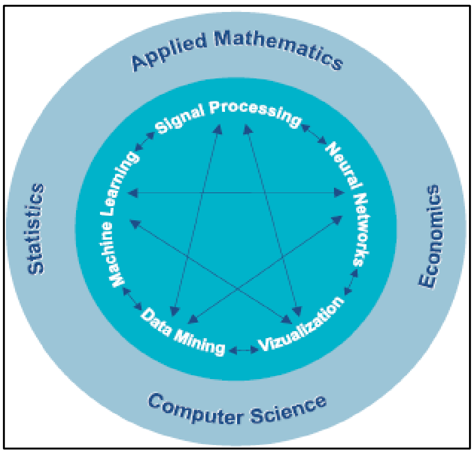

Fig. 1: Big Data Processing Method Interconnection Applied Mathematics Statics Economics And Computer Science Are Foundation Of The Big Data Processing Method. Meanwhile, Data Mining, Signal Processing, Neural Networks, Visualization & Machine Leaning Are Strongly Connected To Each Other

Data mining includes cluster analysis, classification, and regression and association rule learning techniques. This method is aimed at identifying and extracting beneficial information from extensive data or datasets. Cluster analysis [75, 76] is based on principles of similarities to classify objects. This technique belongs to unsupervised learning [77, 78] where training data [79] is used. Classification [80] is a set of techniques which are aimed at recognizing categories with new data points. In contrast to cluster analysis, a classification technique uses training data sets to discover predictive relationships. Regression [81] is a set of statistical techniques that are aimed at determining changes between dependent and independent variables. This technique is mostly used for prediction or forecasting. Association rule learning [82, 83] is set of techniques designed to detect valuable relationships or association rules among variables in databases.

B. *Big Data Challenges*

Big Data has some inherent challenges and problems that can be primarily divided into three groups according to Akerkar et al. [36]: (1) data, (2) processing and (3) management challenges (see Fig. 2). While dealing with large amounts of information we face

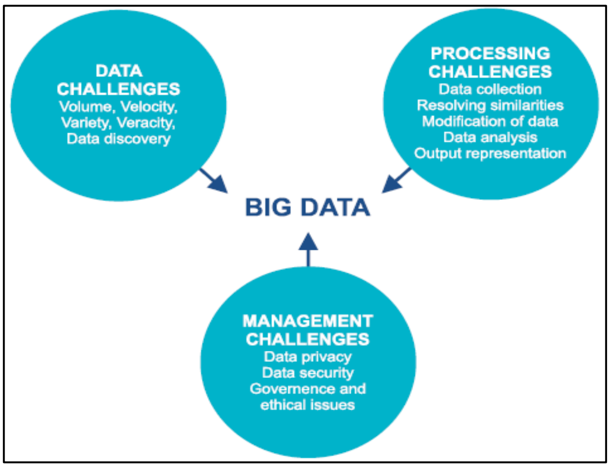

Fig. 2: Big Data Challenges the Picture Illustrates Three Main Categories of Big Data Challenges that are associated with data its Management and Processing Issue

Big Data As those Big Data characteristics are well examined in scientific literature [99– 101] we will only discuss them briefly. Volume refers to the large amount of data, especially, machine-generated. This characteristic defines a size of the data set that makes its storage and analysis problematic utilizing conventional database technology. Variety is related to different types and forms of data sources: structured (e.g. financial data) and unstructured (social media conversations, photos, videos, voice recordings and others). Multiplicity of the various data results in the issue of its handling. Velocity refers to the speed of new data generation and distribution. This characteristic requires the implementation of real-time processing for the streaming data analysis (e.g. on social media, different types of transactions or trading systems, etc.). Veracity refers to the complexity of data which may lead to a lack of quality and accuracy. This characteristic reveals several challenges: uncertainty, imprecision, missing values, misstatement and data availability.

III. VISUALIZATION METHODS

Historically, the primary areas of visualization were Science Visualization and Information Visualization. However, during recent decades, the field of Visual Analytics was actively developing.

As a separate discipline, visualization emerged in 1980 [104] as a reaction to the increasing amount of data generated by computer calculations. It was named Science Visualization [105–107], as it displays data from scientific experiments related to physical processes. This is primarily a realistic three-dimensional visualization, which has been used in architecture, medicine, biology, meteorology, etc. This visualization is also known as Spatial Data visualization, which focuses on the visualization of volumes and surfaces. It utilizes graphics to assist people in comprehending and interpreting data. As it helps to form mental models of the data, for humans it is easier to reveal specific features and patterns of the obtained information. Visual Analytics [112–114] combines visualization and data analysis. It has absorbed features of Information Visualization as well as Science Visualization. The main difference from other fields is the development and provision of visualization technologies and tools.

Visualization methods have evolved much over the last decades (see Fig. 3), the only limit for novel techniques being human imagination. To anticipate the next steps of data visualization development, it is necessary to take into account the successes of the past. It is considered that quantitative data visualization appeared in the field of statistics and analytics quite recently. However, the main precursors were cartography and statistical graphics, created before the 19th century for the expansion of statistical thinking, business planning and other purposes [115]. The evolution in the knowledge of visualization techniques resulted in mathematical and statistical advances as well as in drawing and reproducing images. By the 16th century, tools for accurate observation and measurement were developed. Precisely, in those days the first steps were done in the development of data visualization.

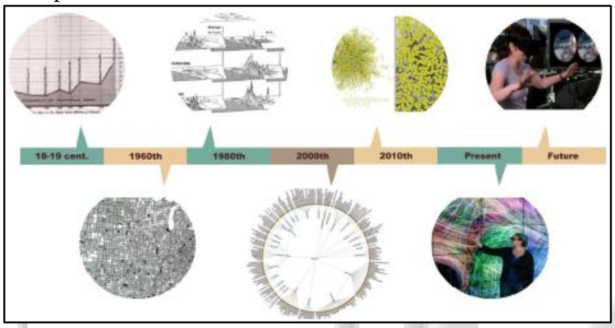

Fig. 3: the Evolution of Visualization Methodology. Development of Visualization Method Originates from 18th Century and it is Rapidly Improving Today due to Technical Sophistication

Nowadays, there are a fairly large number of data visualization tools offering different possibilities. These tools can be classified based on three factors: by the data type, by visualization technique type and by the interoperability. The first refers to the different types of data to be visualized [124]:

– Univar ate data one dimensional arrays, time series, etc.
– Two-dimensional data Point two-dimensional graphs, geographical coordinates, etc.
– Multidimensional data financial indicators, results of experiments, etc.
– Texts and hypertexts Newspaper articles, web documents, etc.
– Hierarchical and links the structure subordination in the organization, e-mails, documents and hyperlinks, etc.
– Algorithms and programs Information flows, debug operations, etc.

The second factor is based on visualization techniques and samples to represent different types of data. Visualization techniques can be both elementary (line graphs, charts, bar charts) and complex (based on the mathematical apparatus). Furthermore, visualization can be performed as a combination of various methods. However, visualized representation of data is abstract and extremely limited by one's perception capabilities and requests (see Fig. 4).

Types of visualization techniques are listed below:

1) 2D/3D standard figure [125]. May be implemented as bars, line graphs, various charts, etc. (see Fig. 5). The main drawback of this type is the complexity of the acceptable visualization for complicated data structures;
2) Geometric transformations [126]. This technique represents information as scatter diagram (see Fig. 6). This type is geared towards a multi-dimensional data set's transformation in order to display it in Cartesian and non-Cartesian geometric spaces. This class includes methods of mathematical statistics;

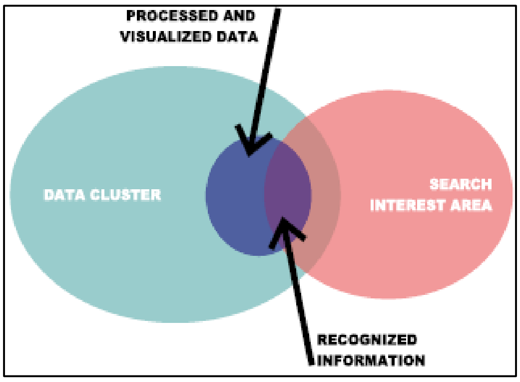

Fig. 4: Human Perception Capability issue. Human perception capabilities are not sufficient to embrace large amount of data

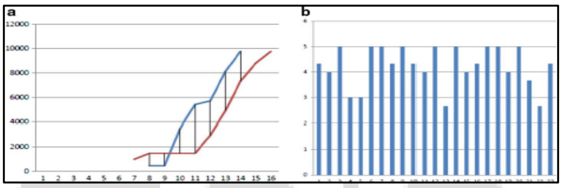

Fig. 5: An Example Pf the 2D/3D Standard Figures Visualization A) The Simple Line Graph & 2) Example of a Bar Chart

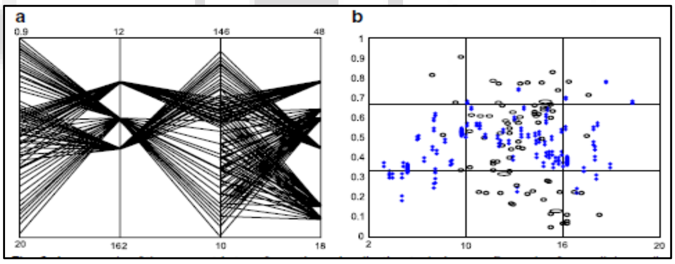

Fig. 6: An Example of the Geometric Transformation Visualization Techniques A) Example of Parallel Coordinates & B) The Scatter Plot

3) Display icons [127]. Basically, this type displays the values of elements of multidimensional data in properties of images (see Fig. 7). Such images may include human faces, arrows, stars, etc. Images can be grouped together for holistic analysis. The result of the visualization is a texture pattern, which varies according to the specific characteristics of the data;
4) Methods focused on the pixels [128]. The main idea is to display the values in each dimension into the colored pixel and to merge some of them according to specific measurements (see Fig. 8). Since one pixel is used to display a single value, therefore visualization of large amounts of data can be reachable with this methodology;
5) Hierarchical images [129]. Tree maps and overlay measurements (see Fig. 9). These type methods are used with the hierarchical structured data.

IV. INTEGRATION WITH AUGMENTED & VIRTUAL REALITY

It is well known that the vision perception capabilities of the human brain are limited [152]. Furthermore, handling a visualization process on currently used screens requires high costs in both time and health. This leads to the need of its proper usage in the case of image interpretation. Nevertheless, the market is in the process of being flooded with countless numbers of wearable devices [153, 154] as well as various display devices [155, 156].

The term Augmented Reality was invented by Tom Caudell and David Mizel in 1992 and meant to describe data produced by a computer that is superimposed to the real world [157]. Nevertheless, Ivan Sutherland created the first AR/VR system already in 1968. He developed the optical see-through head-mounted display that can reveal simple three-dimensional models in real time [158]. This invention was a predecessor to the modern VR displays and AR helmets [159] that seem to be an established research and industrial area for the coming decade [160]. Applications for use have already been found in military [161], education [162], healthcare [163], industry [164] and gaming fields [165]. At the moment, the Oculus Rift [166] helmet gives many opportunities for AR practice. Concretely, it will make it possible to embed virtual content into the physical world. William Steptoe has already done research in this field. The use of it in the visualization area might solve many issues from narrow visual angle, navigation, scaling, etc. For example, offering a way to have a complete 360-degrees view with a helmet can solve an angle problem. On the other hand, a solution can be obtained with help of specific widescreen rooms, which by definition involves enormous budgets. Focusing on the combination of dynamic projection and interactive filtering visualization methods, AR devices in combination with motion recognition tools might solve a significant scaling problem especially for multidimensional representations that comes to this area from the field of Architecture. Speaking more precisely, designers (specialized in 3D-visualization) work with flat projections in order to produce a visual model [167]. However, the only option to present a final image is in moving around it and thus navigation inside the model seems to be another influential issue [168].

V. CONCLUSION

In practice, there are a lot of challenges for Big Data processing and analysis. As all the data is currently visualized by computers, it leads to difficulties in the extraction of data, followed by its perception and cognition. Those tasks are time-consuming and do not always provide correct or acceptable results. In this paper we have obtained relevant Big Data Visualization methods classification and have suggested the modern tendency towards visualization-based tools for business support and other significant fields. Past and current states of data visualization were described and supported by analysis of advantages and disadvantages. The approach of utilizing VR, AR and MR for Big Data Visualization is presented and the advantages, disadvantages and possible optimization strategies of those are discussed. For visualization problems discussed in this work, it is critical to understand the issues related to human perception and limited cognition. Only after that, the field of design can provide more efficient and useful ways to utilize Big Data. It can be concluded that data visualization methodology may be improved by considering fundamental cognitive psychological principles and by implementing most natural interaction with visualized virtual objects. Moreover, extending it with functions to exclude blind spots and decreased vision sectors would highly improve recognition time for people with such a disease. Furthermore, a step towards wireless solutions would extend device battery life in addition to computation and quality improvements.

REFERENCES

[1] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big Data: the next frontier for innovation, competition, and productivity. June Progress Report. McKinsey Global Institute; 2011.
[2] 1000 Genomes: a Deep Catalog of Human Genetic Variation. 2015. http://www.1000genomes.org/.
[3] Via M, Gignoux C, Burchard EG. The 1000 Genomes Project: new opportunities for research and social challenges. Genome Med 2010;2(3) Internet Archive: Internet Archive Wayback Machine. 2015.http://archive.org/web/web.php.
[4] Nielsen J. Comparing content in web archives: differences between the Danish archive Netarkivet and Internet
[5] Archive. In: Two-day Conference at Aarhus University, Denmark. 2015
[6] The Lemur Project: The ClueWeb09 Dataset. 2015. http://lemurproject.org/clueweb09.php/.
[7] Russom P. Managing Big Data. TDWI Best Practices Report, TDWI Research; 2013.
[8] Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView IDC Anal Future. 2012;2007:1–16.
[9] Beyer MA, Laney D. The importance of “Big Data”: a definition. Stamford: Gartner; 2012.
[10] Demchenko Y, Ngo C, Membrey P. Architecture framework and components for the big data ecosystem. J Syst Netw Eng 2013;1–31 [SNE technical report SNE-UVA-2013-02].
[11] Turner V, Reinsel D, Gantz JF, Minton S. The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. IDC Analyze the Future 2014.
[12] Husain SS, Kalinin A, Truong A, Dinov ID. SOCR data dashboard: an integrated Big Data archive mashing medicare, labor, census and econometric information. J Big Data. 2015;2(1):1–18.
[13] Keahey TA. Using visualization to understand Big Data. IBM Business Analytics Advanced Visualisation; 2013.
[14] Microsoft Corporation: Power BI—Microsoft. 2015. https://powerbi.microsoft.com/.
[15] Amazon.com, Inc. Amazone Web Services. 2015. https://aws.amazon.com/.
[16] Google, Inc. Google Cloud Platform. 2015. https://cloud.google.com/.
[17] Socrata. Data to the People. 2015. http://www.socrata.com.

[18] D3.js: D3 Data-Draven Documents. 2015.http://d3js.org.
[19] The Cytoscape Consortium: Network Data Integration, Analysis, and Visualization in Box. 2015. http://www. cytoscape.org.
[20] Tableau—Business Intelligence and Analytics. http://tableau.com/
[21] Kandel S, Paepcke A., Hellerstein J, Heer J. Wrangler: interactive visual specification of data transformation scripts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM; 2011. pp 3363–72.
[22] Schaefer D, Chandramouly A, Carmack B, Kesavamurthy K. Delivering Self-Service BI, Data Visualization, and Big Data Analytics. Intel IT: Business Intelligence; 2013.
[23] Choy J, Chawla V, Whitman L. Data Visualization Techniques: From Basics to Big Data with SAS Visual Analytics. SAS: White Paper; 2013.
[24] Ganore P. Need to know what Big Data is? ESDS—Enabling Futurability. 2012.
[25] Agrawal D, Das S, El Abbadi A. Big Data and cloud computing: current state and future opportunities. In: Proceedings of the 14th International Conference on Extending Database Technology, ACM; 2011. pp 530–3 (2011).
[26] Kaur M. Challanges and issues during visualization of Big Data. Int J Technol Res Eng. 2013;1:174–6.
[27] Childs H, Geveci B, Schroeder W, Meredith J, Moreland K, Sewell C, Kuhlen T, Bethel EW. Research challenges for visualization software. Computer. 2013;46:34–42.

