A Methodology for Supporting Collaborative Exploratory Analysis of Massive Data Sets in Tele-Immersive Environments

Jason Leigh (spiff@evl.uic.edu),
Andrew E. Johnson, Thomas A. DeFanti
Electronic Visualization Laboratory,
University of Illinois at Chicago

Stuart Bailey (sbailey@eecs.uic.edu),
Robert Grossman
National Center for Data Mining

Keywords: tele-immersion, collaborative virtual reality, data mining

Abstract

This paper proposes a methodology for employing collaborative, immersive virtual environments as a high-end visualization interface for massive data-sets. The methodology employs feature detection, partitioning, summarization and decimation to significantly cull massive data-sets. These reduced data-sets are then distributed to the remote CAVEs, ImmersaDesks and desktop workstations for viewing. The paper also discusses novel techniques for collaborative visualization and meta-data creation.

1 Introduction

In 1997 a series of National Science Foundation (NSF) and Department of Energy (DOE) sponsored workshops brought together computer scientists specializing in high-performance computing and scientific visualization, and domain scientists in physics, chemistry, materials science, and engineering. Their goal was to assess the needs of the scientific and engineering community; to identify current and projected computational capabilities; and to outline a federal research and development agenda in scientific visualization, human interface, and the manipulation of massive scientific data-sets[1].

The findings of the workshops indicated a clear trend- that the amount of data collected and generated through scientific simulations was growing dramatically and that currently available technologies for interpreting this data are becoming increasingly inadequate. It was estimated that in 1999 a typical data query will access between 3-30 tera-bytes of data. This is expected to increase to 30-300 tera-bytes in 2001, and 1 peta-byte in 2004. Progress in data mining can help significantly in finding the gems of information buried in the data, however scientists are currently still unable to articulate sufficiently smart algorithms that can reliably find relevant features or draw correct and relevant conclusions on their own. Visualization on the other-hand transforms data into graphical representations that exploit the high-bandwidth channel of the human visual system, leveraging the brain’s remarkable ability to detect patterns and draw inferences. Hence human expertise is central to any process that requires understanding. Unfortunately the visualization algorithms and the high-performance display hardware and software on which they depend, have not kept pace with the sheer amount of data that needs to be visualized. Today's most advanced graphics engines are able to render 3 million shaded, stereoscopic triangles / second at a resolution of 1920x1024. But today's scientific applications need to be able to render 160 million triangles / second. In 2004 these graphics systems will be able to render 15 million triangles/second at a resolution of 4000x3000, but by then scientific applications will require the ability to render 19.2 giga triangles / second at resolutions of 8000x8000[1].

The recommendation made to NSF and DOE was that a new generation of data-access, data mining, visualization, and networking tools need to be developed to match the growing requirements of scientific inquiry. However it was emphasized that these tools could no longer work in isolation- they must be interoperable with each other to allow seamless manipulation and visualization of the data; and they must support multi-user access to encourage regular and long-term collaboration between scientists.

The work-in-progress described in this paper represents a collaboration between experts in advanced collaborative visualization at the Electronic Visualization Laboratory and experts in data mining at the National Center for Data mining. The goal is to develop the Tele-Immersive Data Exploration environment (TIDE)- a collaborative virtual environment for the exploration of massive data-sets.

Tele-Immersion (TI) is defined as the integration of audio and video conferencing, via image-based modeling, with collaborative virtual reality (CVR) in the context of data-mining and significant computation. The ultimate goal of TI is not merely to reproduce a real face-to-face meeting in every detail, but to provide the “next generation” interface for collaborators, world-wide, to work together in a virtual environment that is seamlessly enhanced by computation and large databases.

When participants are tele-immersed, they are able to see and interact with each other and objects in a shared virtual environment. Their presence will be depicted by life-like representations of themselves (avatars) that are generated by real-time, image capture, and modeling techniques. The environment will persist even when all the participants have left it. The environment may autonomously control supercomputing computations, query databases and gather the results for visualization when the participants return. Participants may even leave messages for their colleagues who can then replay them as a full audio, video and gestural stream.

Tele-Immersion has entered the Next Generation Internet (NGI) (www.ngi.gov) and Internet2 (www.Internet2.edu) vocabulary. In the applications section of the Computing Research Association's "Research Challenges for the Next Generation Internet," five key enabling technologies were identified as common to the future use of the NGI [2]: Database Access, Audio and Video, Real-Time and Delayed Collaboration, Distributed Computing, and Tele-Immersion.

The goal of TIDE is to employ Tele-Immersion techniques to create a persistent environment in which collaborators around the world can engage in long-term exploration and analysis of massive scientific data-sets. TIDE's research foci seeks to develop: new human-factors techniques and technologies for multi-dimensional visualization; new technologies for sustaining long-term collaborative data exploration; and new algorithms for partitioning, summarization and decimation of massive data-sets.
TIDE will engage users in CAVEs, ImmersaDesks and desktop workstations around the world connected by the Science and Technology Transit Access Point (STARTAP) - a system of high speed national (vBNS, MREN, ESNet) and international (SingAREN, CANARIE, SurfNET) networks[3]. This gathering of networks connects national sites such as the National Center for Supercomputing Applications and Argonne National Laboratory with international sites such as the Cooperative Research Center for Advanced Computational Systems in Australia (ACSys), the Institute of High Performance Computing in Singapore (IHPC), and Stichting Academisch Rekencentrum Amsterdam in the Netherlands (SARA).

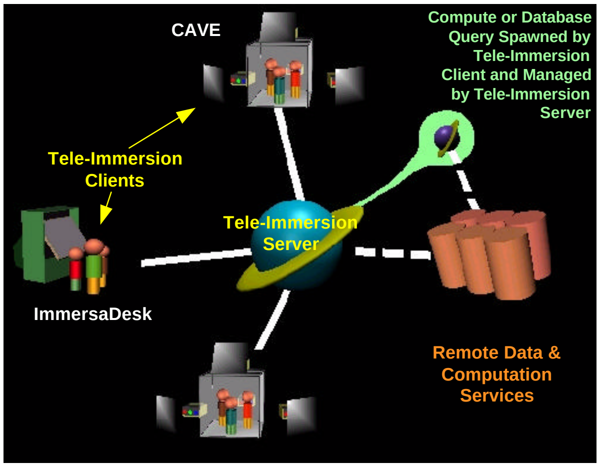
CAVE
Compute or Database Query Spawned by Tele-Immersion Client and Managed by Tele-Immersion Server
Tele-Immersion Clients
Tele-Immersion Server
ImmersaDesk
Remote Data & Computation Services

Figure 1. The Tele-Immersion Data Exploration Environment

2 TIDE : a Tele-Immersive Data Exploration environment

Envision a scenario in which three users- one in a CAVE, another on an ImmersaDesk and yet another on a desktop workstation are all engaged in a routine data exploration enterprise within a virtual laboratory. The CAVE virtual reality system is a 10 foot-cubed room that is projected with stereoscopic images creating the illusion that objects appear to co-exist with the user in the room. The ImmersaDesk is a smaller, drafting-table-like system also capable of projecting stereoscopic images. All users are separated by hundreds of miles but appear co-located-able to see each other as either a video image or as a simplified virtual representation (commonly known as an avatar). Each avatar has arms and hands so that they may convey natural gesture such as pointing at areas of interest in the visualization. Digital audio is streamed between the sites to allow them to speak to each other. The desktop workstation displays a data-flow model that can be used to construct the visualization that is shared between all three display devices. The participants in the VR displays can use three-dimensional tools to directly manipulate the visualization- for example in the CAVE a user is changing the isosurface value in the data-set. These changes are automatically propagated to all the other visualization displays. In the meantime the ImmersaDesk user, noticing an anomaly in the data-set, inserts an annotation in the data-set as a reminder to return to more closely examine the region. Closer examination of the region is achieved by instructing a remote rendering server consisting of multiple giga-bytes of memory and tera-bytes of disk space, to render the images in full detail as a stereoscopic anima-

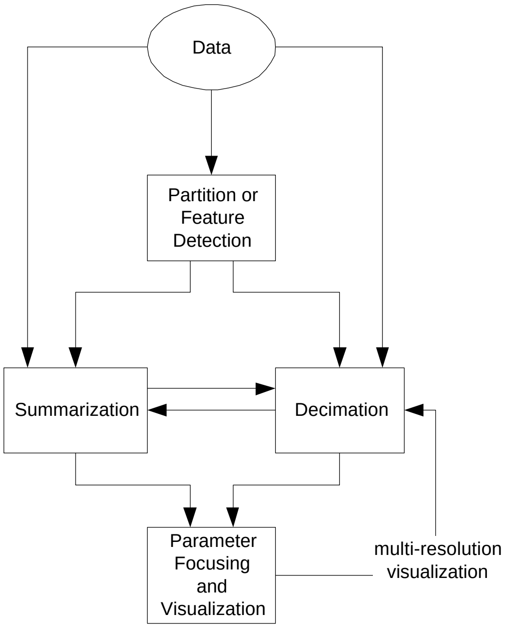
Data
Partition or Feature Detection
Summarization
Decimation
Parameter Focusing and Visualization
multi-resolution visualization

Figure 2. Methodology for processing massive data-sets for visualization.

tion sequence. These animations will take some time to generate and so the users continue to examine other aspects of the data-set. Eventually the rendering is complete and the remote server streams the animation to each of the visualization clients for viewing.

The conceptual organization of TIDE is diagrammed in Figure 1. All the Tele-Immersion clients (TICs) are synchronized by the Tele-Immersion Server (TIS). The TIS is connected in turn to the various remote data and computation services to mediate interaction between these services and the TIC's.

2.1 Remote Data & Computation Services

Remote Data and Computation Services refer to external databases and/or simulations/compute-intensive tasks running on supercomputers or compute clusters that may be called upon to participate in a TIDE work session. The databases may house raw data, or data generated as a result of computations- either as data sets representing each time step of a simulation or as checkpoint files containing intermittent snapshots of the simulations. In most cases the data-sets contain too many dimensions and are too large to visualize within core memory or within the real-time constraints required by Tele-Immersion. Figure 2 outlines the steps by which these large data-sets may be significantly reduced for tele-immersive visualization. In the description of each step we will use the following motivating example taken from atmospheric research data: consider global data on a uniform latitude and longitude based grid. Each latitude and longitude point on the grid maps to a feature vector with the following 12 attributes:

GND temperature, LA temperature, HA temperature, GND pressure, LA pressure, HA pressure, GND wind velocity, LA wind velocity, HA wind velocity, GND humidity, LA humidity, HA humidity;

where GND = ground level, LA = low atmosphere, and HA = high atmosphere.

2.1.1 Partition / Feature Detection

Initially a feature detection step or a partitioning step may be applied to the massive data-set. Feature detection may be used to search for e.g. storm formations. Feature detection is useful when there is apriori knowledge of the conditions that result in the feature to be detected. Otherwise partitioning may be used if less knowledge is available- as is usually the case in the early steps of data exploration. Partitioning involves dividing a large data set into smaller non-isotropic/non-uniform sub-regions or grids which contain a similar degree of complexity. For example in searching for temperature/humidity patterns in potential tornado hot spots one may partition the atmospheric data-set on HA temperature and HA wind velocity where variation in HA temperature is less than 0.5 degrees and variation in HA wind velocity is less than 5 knots. Hence the result is a "map" which is divided into regions which may have significantly differing temperatures and wind velocities between regions but have small variations within each region.

2.1.2 Summarization

The summarize function provides the option of processing the data before it is visualized. At one extreme, the data can be passed through without any processing; at the other extreme, the data can be smoothed or filtered in various ways. In addition a derived value may be computed from other derived or non-derived attributes. This in essence, is a way to reduce the dimensionality of the data-set. For example, for each partition, summarize GND wind velocity and LA wind velocity by averaging their values.

2.1.3 Decimation

This decimate function reduces the size of a large set by discarding data. More generally, one can sample data from a partition with a non-uniform distribution, giving more weight to some data points. In some cases, it is important to discard some data points and to over sample others. For example, when trying to understand the nature of rare events in large data sets, it may be useful to over sample the rare events and down sample the others. In our atmospheric data we might decimate by choosing 5% of all partitions whose GND wind velocity is greater than 60 knots and LA wind velocity is less than GND wind velocity.

The summarization and decimation steps may be applied in conjunction with eym 0 T¯Tc 0•5 TŽCae t

Figure 3. Tele-Immersed Collaborators engaged in a Collaborative Work Environment built with CAVERNsoft. On each end are the avatars of participants manipulating the data in the environment.

overall efficiency of the data analysis process. One

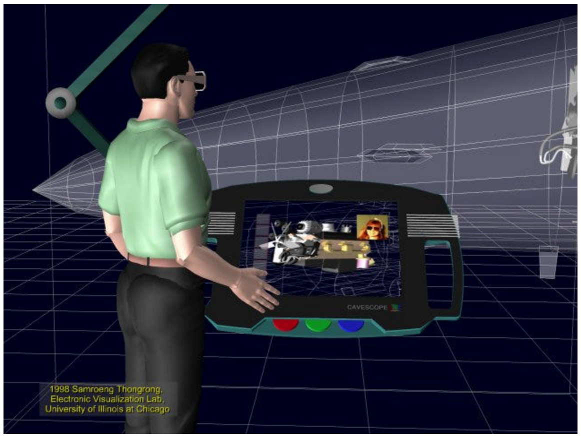
1998 Samroeng Thongrong,
Electronic Visualization Lab,
University of Illinois at Chicago
CAVESCOPE

Figure 4. The CAVEScope

b) Strategy (a) only works if the size of the polygon list is within the memory capacity and the real-time rendering capacity of the client's graphics engine. Also there are rendering algorithms (such as raycasting and raytracing) which produce dramatically better visual quality than those generated by real-time graphics algorithms that cannot be rendered in real-time. For these the user can use a highly decimated model as the placeholder from which the desired viewpoints are selected. The remote rendering server can then be commanded to render these viewpoints as a sequence of stereoscopic images or animations that can then compressed and streamed back to the visualization clients. As the image generation process is unlikely to occur in real-time the compressed images may have to be gathered at the local disk of the rendering server or the Tele-Immersion server and streamed to the client at a later time. As the Tele-Immersion server is persistent, collaborators can enter the environment at any time to review the results of the rendering process.

Finally since it will take considerable time for visualization tools in VR to match the flexibility and depth of tools that are provided by existing Problem Solving Environments (PSEs) such as AVS, IRIS Explorer, and SCIrun, modules for these popular systems need to be built to allow them to seamlessly deliver visualizations to the Tele-Immersion Server[12]. The ultimate goal is to provide a visual programming interface within the VR environment that will allow collaborators to build complex visualizations with the same or greater ease than those provided by existing desktop PSEs. This is a non-trivial problem that is unlikely to be solved in the near future.

2.3.3 Meta Data Creation

Most computational scientists agree that a crucial part of the exploratory data analysis process includes the creation of snapshots and annotations to track the progress of the exploration and to record discoveries that are made[13]. On desktop PSE's these annotations (meta-data: data about the data) are typically entered in text windows. This common mode of data-entry however is problematic as well as limiting in VR. All existing VR displays lack the resolution to display text clearly in a virtual window- for example a 1x1 foot region of a 10x10 foot CAVE screen has a resolution of only 128x102 pixels. VR systems such as Head-Mounted Displays essentially blind the user to the outside world making it difficult to operate a keyboard. In the CAVE or ImmersaDesk a keyboard can be placed nearby however both these systems still suffer from the low display resolution problem. There are three ways to address this:

a) Increase the resolution of the VR displays. Work in this area is actively being conducted at Argonne National Laboratory as well as Princeton University to tile a large wall with high resolution projectors. These tiled displays often have a resolution of 6400x3072 and are driven either by multi-pipe Onyx2s or clusters of Linux or NT PCs. Although this work is presented here in the context of its usefulness for meta-data creation, the added resolution is in general invaluable for scientists to be able to view their data in greater detail.

b) Provide a high resolution flat-panel touch-sensitive display that can be used as an ancillary input and display device. Such a device would be well suited to allow the user to operate the Problem Solving Environments described above. Furthermore a tracking system can be mounted on the device to allow it to serve as a window that can display localized or filtered information as the display is moved through the virtual environment. Such a device (called the CAVEscope- shown in (Figure 4) was conceived by Tom DeFanti at EVL in 1996. We are aware that a vendor by the name of Virtual Research Systems, Inc. now manufactures a similar device.

c) Provide annotation tools that take advantage of VR. For CAVERNsoft we have built a plug-in module that will allow a participant to record audio and gesture as an annotation that can then be attached as a virtual post-it to objects or states of the environment. When an annotation is re-played an avatar materializes to re-enact the recorded message. This is particularly effective in VR because it allows the user to point and gesture at the area of interest in the environment while the annotation is being recorded. This is similar to recording the real world with a video camera- the difference however is that VR recordings have the added benefit that the playback can re-create all the attributes of the virtual world and situate you in the world so that you can view the playback from any vantage point. Furthermore since the state of the world and the avatar are all captured as discrete data rather than individual images, the annotations can be queried. This concept is a generalization of the Virtual Mail (Vmail) system [14], a tool for supporting asynchronous communication with users that are many timezones away (for example in collaborations between the U.S. and Japan.) In Vmail it was observed that users viewing the messages would tend to react to the avatars as if the original participant was actually in the environment with them- forgetting that they were actually viewing a recording that was made hours ago.

3 Closing Remarks

This paper has outlined a methodology for using tele-immersive systems as a high-end visualization interface for exploring massive data-sets. TIDE's research foci seeks to develop: new human-factors techniques and technologies for multi-dimensional visualization; new technologies for sustaining long-term collaborative data exploration; and new algorithms for partitioning, summarization and decimation of massive data-sets.

We are only in the architectural design phases of TIDE. However as TIDE's underlying architecture will be based on CAVERNsoft much of the architecture for supporting Tele-Immersion is already in place. We anticipate that a proof-of-concept will be built by the end of 1999 in which it will already be useful for visualizing a variety of data-sets.

We have just completed a pilot study to understand how multiple collaborative representations may be employed by participants in a scientific visualization exercise using CAVE6D- a collaborative tool for viewing multi-dimensional atmospheric and oceanographic data[11].

In the future, the progress of the TIDE project can be tracked from the CAVERN web site at www.evl.uic.edu/cavern.

4 Acknowledgements

Major funding is provided by the National Science Foundation (CDA-9303433.) The virtual reality research, collaborations, and outreach programs at EVL are made possible through major funding from the National Science Foundation, the Defense Advanced Research Projects Agency, and the US Department of Energy; specifically NSF awards CDA-9303433, CDA-9512272, NCR-9712283, CDA-9720351, and the NSF ASC Partnerships for Advanced Computational Infrastructure program. The CAVE and ImmersaDesk are trademarks of the Board of Trustees of the University of Illinois.

5 References

[1] Smith, P. and Van Rosendale, J. editors, Data and Visualization Corridors: Report on the 1998 DVC Workshop Series, DOE and NSF Sponsored, 1998.

[2] Smith, J. and Weingarten, F. (eds.), Research Challenges for the Next Generation Internet, Computing Research Association, 1997, p. 20.

[3] DeFanti, T. A. and S. Goldstein. The STAR TAP web site, http://www.startap.net, 1998.

[5] Leigh, J. Johnson, A. DeFanti, T., Brown, M., et al. A Review of Tele-Immersive Applications in the CAVE Research Network, Proc. IEEE Virtual Reality 1999, pp 180-187, Houston, Texas, Mar 14 - Mar 17, 1999.

[4] Leigh, J., Johnson, A., DeFanti, T., CAVERN: A Distributed Architecture for Supporting Scalable Persistence and Interoperability in Collaborative Virtual Environments. In Virtual Reality: Research, Development and Applications, Vol 2.2, December 1997 (1996), Pp 217-237.

[6] Leigh, J., Johnson, A., Vasilakis, C., DeFanti, T., Multi-perspective Collaborative Design in Persistent Networked Virtual Environments. Proc. IEEE Virtual Reality Annual International Symposium '96. Santa Clara, California, Mar. 20 - Apr. 3, 1996, Pp 253-260, 271-272.

[7] Larkin, J.H. & Simon, H. A. Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99. 1987.

[8] Bibby, P. A., & Payne, S. J., Internalization and the use specificity of device knowledge. Human-Computer Interaction, 8(1), 25-56, 1993.

[9] Salzman, M. Dede, C. Loftin, B., & Ash, K. VR's Frames of Reference: A visualization technique for

mastering abstract information spaces. In Proceedings of the Third International Conference on Learning Sciences, pp. 249-255. Charlottesville, VA: Association for the Advancement of Computers in Education., 1998.

[10] Ainsworth, S.E., Wood, D. J., & Bibby, P.A. Evaluating principles for multi-representational learning environments. 7th EARLI conference, Athens, 1997.

[11] Lascara, C., Wheless, G., Cox, D., Patterson, R., Levy, S., Johnson, A., Leigh J., TeleImmersive Virtual Environments for Collaborative Knowledge Discovery to appear in the proceedings of the Advanced Simulation Technologies Conference '99 San Diego CA, April 11-15, 1999.

[12] Problem Solving Environments- Projects, Products, Applications and Tools: www.cs.purdue.edu/research/cse/pses/research.html.

[13] Springmeyer, R., Werner N., Long, J. Mining Scientific Data Archives through Metadata Generation First IEEE Metadata Conference, April 16-18, 1996, NOAA Auditorium, Silver Spring, Maryland.

[14] Imai T., The Virtual Mail System, Proc. IEEE Virtual Reality 1999, Houston, Texas, Mar 14 - Mar 17, 1999.

[15] Nakayama, K. and Silverman, G. H. Serial and Parallel Processing of Visual Feature Conjunctions. *Nature 320*, pp. 264-265, 1986.

[16] Roy, T.et al., Cosmic Worm in the CAVE: Steering a High-Performance Computing Application from a Virtual Environment, PRESENCE, Vol. 4, Issue 2, 1995, pp.103-109. 1995.