A Tele-Immersive Environment for Collaborative Exploratory Analysis of Massive Data Sets

"The purpose of computing is insight—not numbers."
-- R. W. Hamming

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 is a white paper outlining a methodology for employing collaborative, immersive virtual environments as a high-end visualization interface for massive data-sets.

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 techniques for the interactive exploration 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) 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).

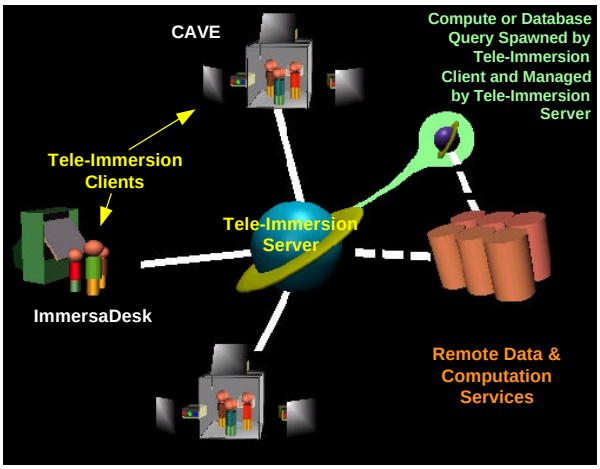

Figure 1. The Tele-Immersion Data Exploration Environment

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 typical data exploration exercise 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 RAM and tera-bytes of disk space, to render the images in full detail

as a stereoscopic animation 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 overall TIDE architecture 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 external data servers and remote rendering servers to mediate interaction between these servers and the TIC's.

2.1 The Tele-Immersion Server

The Tele-Immersion Server's primary responsibility is to create a persistent entry point for the TICs. That is, when a client is connected to the TIS, a user can work synchronously or asynchronously with other users. The environment will persist even when all participants have left it. The server also maintains the consistent state that is shared across all participating TICs. Finally the TIS stores the data subsets that are extracted from the external data sources. The data subsets may consist of raw and derived data sets, three dimensional models or images.

2.2 The Tele-Immersion Client

The Tele-Immersion Client (TIC) consists of the VR display device (either CAVE, ImmersaDesk, etc) and the software tools necessary to allow "human-in-the-loop" computational steering, retrieval, visualization, and annotation of the data. The TIC also provides the basic capabilities for streaming audio and video, and for rendering avatars to allow participants to communicate effectively with one another while they are immersed in the environment. These capabilities come as part of EVL's Tele-Immersion software framework called CAVERNsoft (Figure 2) [4]. CAVERNsoft has already been used in over a dozen collaborative applications ranging from art and design, engineering and science, and education and training; and by clients such as General Motors, Searle/Monsanto, Nippon Telephone and Telegraph, the Center for Coastal and Physical Oceanography, and the Naval Research Lab [5]. We believe it is well suited as the base architecture for the TIC.

2.2.1 Collaborative Interactive Visualization

TIDE's main focus is in supporting collaborative visualization of large data sets. To support

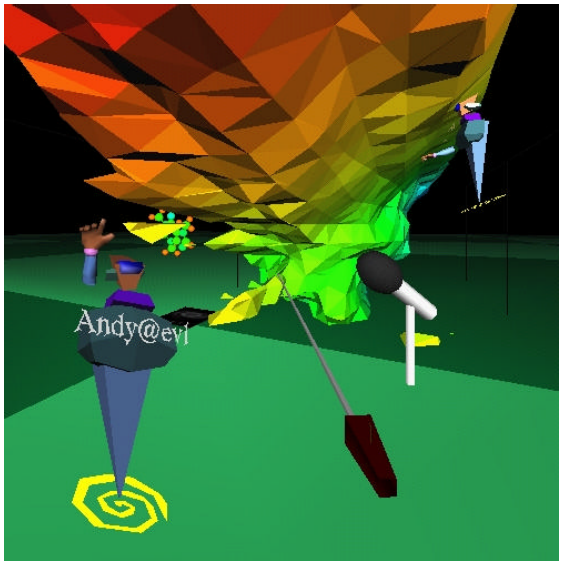

Figure 2. Tele-Immersed Collaborators engaged in a Collaborative Work Environment built with CAVERNsoft. On each end are the avatars of participants manipulating the data in the environment.

collaborative visualization effectively one must consider how adding collaboration can improve the overall efficiency of the data analysis process. One aspect that we are particularly interested in developing in TIDE is the concept of Multiple Collaborative Representations (MCR.) In the real world, individuals who are trying to solve a common problem gather (in workshops, for example) in the hopes that their individual experiences and expertises will contribute new perspectives and solutions to the problem. In many existing collaborative VR applications, participants typically all view and modify the same representation of the data they are viewing. It is our belief that a greater benefit will be derived if the participants are given the power to create and modify their own representations, based on their particular areas of interest and expertise [6].

Recent work in providing multiple representations to enhance learning have implied that this is a non-trivial problem [7, 8]. It has been shown that students perform better in tests when they learn a concept given more than one representation than students given only a single representation [9]. It has also been shown that this is not necessarily the case; instead it was found that multiple representations increased the cognitive load on a learner at the expense of learning [10]. These contradictory findings would suggest that multiple representations help rather than hinder when the benefit of the multiple representations is offset by the increase in the cognitive load incurred in interpreting these representations. This cognitive load may be lessened

if the proper tools are provided to coordinate the correspondences between the representations. This could also be stated as 'If a picture is worth a thousand words, then ten pictures are worth a million.'

We envision a potential application of MCRs in the visualization of multi-dimensional data-sets. Here a large number of dimensions may be partitioned across multiple users to assist in reducing the overall complexity of the content being visualized. The goal of research in MCR is to develop tools to allow participants to coordinate their interpretations of each representation to form a more efficient collective understanding of the data being explored. One example of a tool that has incorporated this technique is CAVE6D [11] developed in collaboration with the Center for Coastal and Physical Oceanography. CAVE6D is a tool for collaboratively viewing multi-dimensional numerical data from atmospheric, oceanographic, and other similar models, including isosurfaces, contour slices, volume visualization, wind/trajectory vectors, and various image projection formats. We are currently in the process of performing user-studies to observe how users take advantage of MCRs, and in what situations MCRs are effective or a hindrance. Furthermore these user studies will allow us to form ideas on the kinds of tools that will be needed to help participants coordinate the MCRs. For example, one extension to traditional visualization tools is to provide the ability to track the dimensions and regions that other participants are simultaneously viewing/brushing. This also introduces a new concept in VR awareness. VR awareness refers to the problem of locating other participants in a large collaborative virtual environment. Whereas all previous work in these awareness tools amount to three-dimensional radars, the kind of awareness tool needed for MCR is a *multi-dimensional* radar.

MCR will be one of the computer-supported-cooperative-work strategies for guiding the design of TIDE's collaborative visualization tools. TIDE will provide a suite of network-aware visualization tools for viewing stereotypical data-sets, such as vector fields and volumes. These tools are intended for visualizing highly decimated data-sets that can be rendered in real-time in the current generation of graphics engines. Note however that graphics rendering power is not necessarily the limitation here. The visualization endpoints may not have the memory and disk capacity to hold the data from which the visualizations are generated even though the eventual visualization is displayable in real-time. To visualize massive data-sets two strategies will be used:

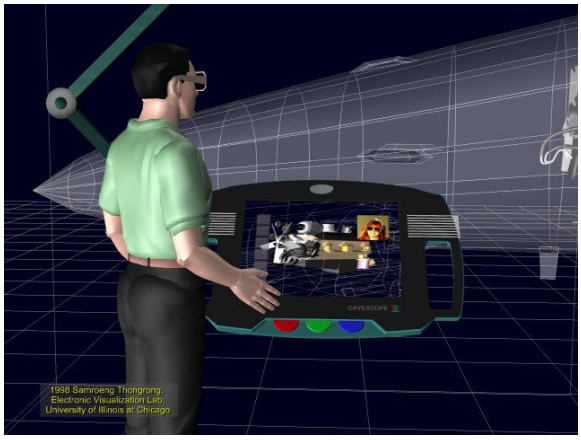

Figure 3. The CAVEScope

a) Generate the geometry for the visualization on remote servers that do have the memory and disk capacity to operate on the data. For example the visualization client would set an isosurface threshold and allow the remote server to perform the marching cubes algorithm to generate the polygon list of the isosurface. This polygon list is then sent back to the client for viewing.
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.

2.2.2 Meta Data Gathering- making annotations and recording discoveries

Springmeyer [13] notes that a crucial part of the exploratory data analysis process involves the creation of snapshots and annotations to track the progress of the exploration and to record discoveries that are made. 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. 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 two ways to address this:

a) 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 3) 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.
b) 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 re-creates all the attributes of the virtual world and situates you in the world to allow you to view the playback from any viewpoint. Furthermore since the state of the world and the avatar are all captured as discrete data rather than individual images, the annotations can easily 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.

2.3 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. In most cases the data-sets contain too many dimensions and are much too large to visualize entirely. However data mining may be employed to clean the data, to detect specific features in the data, or to extract trends from the data. In some cases as the data mining processes may generate models of the data, the models can be used to make predictions on missing data points. Furthermore the models can be used to determine which attributes in a multi-dimensional data-set are the most significant. This is particularly valuable for visualization because the ability to fill missing data points means a more accurate estimate of the missing data can be made than by simple graphical interpolation. In addition by being able to isolate the most significant attributes, a viewer can prioritize the attributes that they assign to visual features (such as hue, intensity, shape etc) in the visualization. For example Nakayama and Silverman [15] have shown that stereoscopic depth is the most powerful, pre-attentively detected visual feature as compared to other features such as intensity and hue (the features most commonly used in scientific visualizations.) This is a particularly

interesting finding for VR because the medium in which VR resides is inherently stereoscopic.

In TIDE the approach taken is to employ data mining algorithms where appropriate as a means to partition space non-isotropically; to exclude attributes with low significance; to "smart" average attribute values to "summarize" a number of attributes into a single attribute (as a means to reduce dimensionality); and to decimate the data based on the limits of the VR visualization system. Initially many of these processes will be controlled on desktop interfaces of PSEs and the resulting decimated data is distributed amongst the collaborators via the Tele-Immersion server. However over time we will gradually allow an increasing number of these functions to be controlled directly from within the Tele-Immersion environment using three-dimensional interfaces.

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 techniques for the interactive exploration 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 functioning proof-of-concept will be built by the end of 1999 in which it will already be useful for visualization a variety of data-sets collaboratively.

We are currently in the midst of performing a pilot study to understand how multiple collaborative representations are employed by participants in a scientific visualization exercise using CAVE6D [11].

In the future, the progress of the TIDE project as well as downloadable versions of the software 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] P. Smith and J. Van Rosendale, editors, "Data and Visualization Corridors: Report on the 1998 DVC Workshop Series, DOE and NSF Sponsored, 1998.

[2] J. Smith and F. Weingarten (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. (1986). Serial and Parallel Processing of Visual Feature Conjunctions. *Nature 320*, 264-265.