LA-UR-98-3100

Approved for public release;
distribution is unlimited.

Title: Multi-Source Data Analysis Challenges

Author(s): James P. Ahrens 113788 CCS-1
Sam Uselton, MRJ Technology Solutions at NASA Ames
Wes Bethel, E. O. Lawrence Berkeley National Laboratory
Lloyd Treinish, IBM T. J. Watson Research Center
Andrei State, University of North Carolina at Chapel Hill

Intended for: IEEE VISUALIZATION '98 ; 199810 ; RESEARCH TRAINING

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (7/06)

Multi-Source Data Analysis Challenges

Organizer
Sam Uselton, MRJ Technology Solutions at NASA Ames

Panelists
Jim Ahrens, Los Alamos National Laboratory
Wes Bethel, E. O. Lawrence Berkeley National Laboratory
Lloyd Treinish, IBM T. J. Watson Research Center
Andrei State, University of North Carolina at Chapel Hill

INTRODUCTION

At least three factors are converging to make multi-source data analysis pervasive in the near future. Digital data acquisition is becoming easier and cheaper. Computational simulations are gaining fidelity and detail while becoming more practical to compute. And everything is becoming networked so data from many sources can be reached by a single user or application. From cross validation of computational and experimental models to steering computational simulations with real world observations, bringing data from multiple sources together is much more powerful than using each source separately. And computer systems can provide support for users in situations where they would be overwhelmed by volume or complexity without the support. But multi-source data analysis is harder than single source data analysis, and designing, building and deploying tools for others to use for it is very hard.

The varieties of data available and the needs of those analyzing it vary greatly between application areas, and so do the challenges of analyzing this multi-source data. The panelists have been selected to represent a variety of areas in which multi-source data is common. Each one has accepted the mission of convincing the audience that his application has the greatest challenges to be overcome. This structure is designed to expose a large number and variety of research problems and motivate work on these problems.

Jim Ahrens	The Accelerated Strategic Computing Initiative
Wes Bethel	Geosciences
Lloyd Treinish	Atmospheric Sciences
Andrei State	Medicine
Sam Uselton	Aerospace Engineering Design

Table 1: Application areas for panelists.

POSITION STATEMENTS

James P. Ahrens

The Accelerated Strategic Computing Initiative (ASCI) will provide simulation capabilities needed to predict the performance, safety, reliability and manufacturability of nuclear weapons systems. The requirements in simulation and modeling are driven by two fundamental changes in the nuclear weapons landscape: (1) The Comprehensive Test Ban Treaty and (2) The Stockpile Life Extension Program, which extends weapon lifetimes well beyond their anticipated field lifetimes.

The move from confidence based on nuclear testing to confidence based on computer-based comparison and analysis of multi-source simulation and experimental data is both a profound change in the assessment process and a significant challenge. Data sources include the results of computational simulations, data from new non-nuclear experiments and from historic nuclear tests. The data sources can be of differing spatial and temporal resolutions, dimensions, and grid types introducing registration and granularity problems. ASCI simulations create massive data source sizes further complicating the visualization and analysis process. Additional complexity is introduced because some data, such as data from historical nuclear tests, cannot be re-gathered. Visualization tools play a key role in the comparison process both qualitatively, to assess the agreement of global data features, as well as quantitatively, to compare the agreement of specific data points. The ASCI multi-source visualization and analysis problem is the hardest problem on the panel because we compelled to "get it right" since it is at the heart of the current nuclear weapons safety and reliability assessment process.

Wes Bethel

In the practice of visualization and analysis in the geosciences, we are confronted with the most treacherous of conditions. Like many other disciplines, our coordinate systems span the range from angstroms to light years. Also like other disciplines, we are faced with the challenges posed by integrating observed and simulated data.

The "easy" applications are those which have the benefit of "spatial registration," or "one true coordinate system." Atmospheric and oceanographic sciences have the benefit of a reference grid (the planet). Aerospace has the benefit of a reference grid: a body being modeled. Computer assisted medical procedures require careful registration of instruments, patient and other data, but again, the reference grid is well defined, often found within restraining devices on the table. We all know that space-to-space mapping is straightforward, although not always easy. In fact, the DOE-ASCI program has solved this problem by adopting a "Vector Bundles" data model for their computational program. The DOE-ASCI multi-source data problems are already solved by the nature of the data model. Related analysis issues are similarly straightforward: fluid flow models can be mapped from one space to another so long as there is conservation of mass, velocity, or other appropriate physical property. Mapping from space-to-space or mass-to-mass is a well-defined and solved problem.

Site modeling and characterization, by definition, requires integration and analysis of grid-ful and grid-less data. For a wide class of site modeling problems, the overall geologic characteristics of the site must be established prior to any analysis. Unfortunately, geologic classification and characterization is not an exact science due to subsurface heterogeneity and the discontinuous nature of geology. Given that an entire site cannot be "dug up" and analyzed and tested cubic foot by cubic foot, estimates and "best guesses" are made based upon field observations, experience with the site and "gut feel.." From these estimates, quantitative and qualitative models are derived, and used to predict subsurface groundwater movement or the existence of a body of ore.

Proceedings of the IEEE Visualization 98 (VIZ'98)
0-8186-9176-x/98 $10.00 © 1998 IEEE

IEEE COMPUTER SOCIETY

Given that space-to-space and mass-to-mass transformations are well defined (and solved), what remains to be done? Since the goal of visualization is insight, and assuming that the path from confusion to insight may not always be linear, we would like to provide our users with tools which allow them to freely navigate, explore and query. Given the amount of published material on the subject, navigation is largely a solved problem.

Exploration poses interesting challenges. One of the most difficult tasks I've ever had was assigned to me in high school biology class. In this assignment, I had to inventory and classify *everything* within a one-square-meter plot of ground next to the Platt River in Central Colorado. At first, it seemed that I would be done in ten minutes. But then, the more closely I looked, the more "stuff" there was to see, and before I knew it, several hours had passed and I was no closer to inventorying "everything" than I was when I started. Human nature is to stand back and "get the big picture," then to move in close and scrutinize. Multi-resolution modeling and presentation techniques help, but may not include "grid-less data" which may be very important to a researcher.

The geosciences present the most difficult challenges: integration of grid-ful and grid-less data, analysis and visualization of sparsely and irregularly sampled data, and meso- to micro-scale coordinate systems.

Lloyd A. Treinish

Visualization and analysis in the atmospheric and space sciences is a challenging task. The myriad of available data from observation (in situ and remotely sensed) and computation (simulations and empirical models) are complex and large in volume. These data are multidimensional (spatial, spectral), dynamic and consist of many physical variables. There is enormous variation in the instrumentation used to observe and the computations employed to model the Earth that have consequences in the data topologies (e.g., coordinate systems, meshes), sampling (spatially, temporally and spectrally) and error characteristics. With the former, the variation is often compounded by inconsistencies in the data gathering process, especially for long-term monitoring. These measurements each relate some aspect of the physical phenomena under observation. Typically, they must be combined in order to glean some knowledge of the data. Furthermore, they are often used in conjunction with simulations to verify theory or as initial or boundary conditions for empirical models.

Thus, one must consider a notion of data "fusion", by which these disparate sources can be utilized simultaneously. For example, NASA's Earth Observing System (EOS), whose initial deployment will be within a year, will have to receive, process and store about one terabyte (TB) of data per day, for over a decade from a number of instruments. These data will be compared and used with many models. One long-term goal of such efforts is to view the Earth as an integrated system — to merge and define the interactions between the near-space environment, the atmosphere, the oceans, the land (both surface and subsurface), etc. An additional aspect of such work is the evaluation of the environmental effects of anthropomorphic activities. These data require care in their presentation so that artifacts due to the visualization process are not introduced and erroneously interpreted as features in the data. For example, the provided form of these data may be ill-suited for the study of phenomena that occur continuously over a nominally spherical surface (i.e., it tears the data). In addition, they may not be uniformly available for the entire Earth or at least spatial domains being examined. Each of the data sets to be "fused" are generally not geographically co-registered and are defined on differing geometric structures. Further, the coordinate system for visualization and interaction may need to differ from those native to the data sets of interest.

Therefore, one must approach visualization from the perspective of data management by introducing an uniform data model that is matched to the structure of the data as well how such data are used. The implementation of such a model effectively decouples the management of and access to the data from the actual application and is as important a component of a visualization system, for example, as underlying graphics and imaging technology.

One consequence of such a data (model)-centric approach is that the same operation(s) can be applied to data sets that need to be visually fused or correlated (i.e., displayed and interacted together) without introducing interpolation or resampling to a common mesh. The latter process implies a modification to the data, whose impact could be hidden in subsequent visualization. Further, if a specific visualization task requires a cartographic projection, then these data sets can be independently warped by the prerequisite transformation. Any geometric distortion that is introduced is due only to the actual projection since the data and topology remain invariant through such a transformation. It is also independent of the choice of realization or rendering technique or cartographic projection, and hence, provides a framework for experimenting with different visualization strategies. This enables correlative visualization for visual fusion from two perspectives. First is the capability to look at multiple sets of data in exactly the same fashion (i.e., visual comparison within a common framework). Second is the capability to utilize a variety of visualization strategies within the chosen coordinate system, for examining a single set of parameters from one source or many parameters from multiple sources. The specific choices can be dictated by the goal of the visualization task(s) as defined by the scientist studying the data.

Andrei State

The current research in augmented reality at UNC focuses on medical applications. We are developing real-time augmented reality (AR) systems to assist physicians in performing certain kinds of surgical interventions. Currently we are only targeting ultrasound-guided and laparoscopic procedures, but we believe that AR has significant potential in many areas of interventional medicine.

Our modest task at hand is to display live ultrasound data and laparoscopic range data in real time and properly registered to the patient. We use video-see-through head-mounted displays and high-performance graphics computers to provide the user with a dynamic stereoscopic view of the scanning data inside the patient. The visualized data must appear to stay in place within the patient as the user moves around the patient for visual inspection and/or surgical intervention.

To achieve this, our system must acquire data from multiple sources: stereo video streams from the head-mounted stereo cameras, a video stream from the ultrasound scanner, range data (video and depth) from the 3D laparoscopic scanner, as well as position and orientation data from the trackers attached to the scanner and to the user's head. Eventually we also plan to use pre-acquired imaging data such as CT (computer tomography) or MRI (magnetic resonance imaging). The pre-op data would have to be adapted to the current shape of the patient's body, since human bodies are deformable. Deformation or even drastic discontinuous changes also occur during and as a result of intervention. Such changes in body or organ topology would have to be tracked and used to warp the pre-acquired data sets "in place." Otherwise these will not be useful in guiding interventions or may even prove harmful by "mis-guiding" the intervention.

0-8186-9176-x/98 $10.00 © 1998 IEEE

Integrating the heterogeneous sample streams acquired from all these data sources into coherent, understandable, medically useful, synthetically enhanced imagery to be presented to the surgeon at real-time frame rates is a daunting task. Each of the above mentioned data sources has its own coordinate system, sampling rate, grid type, resolution, as well as characteristic latency when delivering streaming data. The data sources will have to be calibrated to each other geometrically and temporally. So far we have developed simple methods to statically pre-calibrate some of the geometric and temporal relationships between a few of the sources only. However, while the tracking devices and algorithms register stationary synthetic and real imagery well, registration is degraded considerably in dynamic environments, even though we use a variety of techniques such as prediction, interpolation of past readings, and reordering of computation, in order to reduce apparent latency and thus dynamic registration errors.

One of the problems is that static pre-calibration implies static geometry and latency relationships. We are of course aware that the temporal relationships mentioned above (and if we are unlucky, the spatial ones as well) are in fact dynamically changing—along with pretty much everything else in our envisioned systems. This means we will have to develop methods to dynamically measure (without introducing additional lag) these relationships in order to continuously compensate for temporal drift.

Even if spatio-temporal integration could be achieved, we still don't know how to visually present the merged information to the surgeon except in the simple scenarios we have addressed so far. The visual presentation will have to be done such that the surgeon continuously receives visual information relevant to the current status of the patient and the intervention that is being performed. The AR display must never burden or encumber the surgeon. For example, what is the correct way of mixing the surgeon's view of live tissue in the operative field with imaging data, such that relevant information in the imaging data—better avoid that vital artery that's just below this tissue—does not obstruct the surgeon's view of the real live tissue being operated on?

Finally, all this is supposed to happen live, in real-time, and has limited rehearsal potential (every intervention is different). To top it all, human life may directly or indirectly depend on our system! I find it hard to imagine a more difficult multi-source visualization task than having to address all of the above issues with an integrated hardware/software system which must perform reliably under such critical conditions and subject to such stringent demands. We are still many years away from a system that could routinely be used by a surgeon on human patients.

Sam Uselton

Aerospace design and engineering contains the biggest challenges in multi-source data analysis, and the biggest payoffs for success. Design of modern aircraft requires the coordinated efforts of large groups of people from a diverse range of specialties. Design choices of one group define constraints to other groups. The traditional method of assuring that all the constraints are met involved linearizing the design process, so that a design moved in sequence from one group to another. This process is slow and design "sweet spots" may be missed due to early decisions limiting later possibilities. Discovering an insurmountable obstacle late in the process means starting again from an earlier stage, which is too expensive. So designs tend to be conservative, again missing opportunities for improvements.

Concurrent engineering offers the possibility of shorter time to completion while improving the end product. Automated tools are necessary to enable concurrent engineering. They are widely used in manufacturing and in the later design steps preparing for production. Automated support for design processes is harder, especially in a concurrent setting. Much of the relevant data is complex and fits databases poorly. The data is as diverse and heterogeneous as the disciplines creating it: computational and experimental aerodynamics, structural mechanics and dynamics, propulsion, control systems, manufacturability and maintainability, and even market assessment. Some of the individual data sets are extremely large, and the number of data items relevant to a single project is astounding. Data is continuously created and modified during design activities, and several incomplete and incompatible designs may be in simultaneous investigation for the same project, adding considerably to the problems. And the team working on such a project is likely to be distributed all over the US and might even include overseas groups and a variety of corporate entities.

The data analysis challenges are many and diverse. Since the users and their needs vary, the same data needs to be presented in different ways, so a wide variety of analysis tools are appropriate. The first step in data fusion is registration and normalization of the data from multiple sources. Independent groups have conflicting defaults for everything from data formats to coordinate systems to units of measurement. Data is taken on regular grids, structured grids, unstructured grids, hybrid grids, multi-block grid systems, grids moving relative to each other, grids which evolve over (simulated) time, and with no grids. Sample spacing varies widely between data sets and within the same dataset. The same simulation may be run at different resolutions or using different computational models. Different wind tunnel models of different scales may be used in the same project.

Models (that is, descriptions of the design) developed for different purposes have differing details: a structural model finds the overlap of metal panels and the spacing of rivets in a wing important, while fluid dynamicists are concerned only with a smooth approximation to the aerodynamic shape. Wind tunnel models (and flight test models) deform under their loads, in ways that may or may not match computational models of the deformation. Yet inspecting the pressures from a CFD simulation in the context of the stress analysis from an FEA analysis is clearly desirable to specialists from both areas. And these difficulties just touch the surface of the collection of problems that are opportunities for research.

Aerospace design applications are the best candidates for developing solutions to these difficulties not only for the myriad complexities of the problems but also because it has a large, economically significant base of support. And the applications to space transportation make it the most exciting application too.

BIOGRAPHIES

James P. Ahrens is technical staff member at Los Alamos National Laboratory. He works in the Advanced Computing Laboratory as part of the Visualization Group. In addition to the Accelerated Strategic Computing Initiative he also works with scientists from Energy Research projects to visualize their simulation results including the global climate, accelerator physics and wildfire modeling projects. He received a B.S. degree in Computer Science from the University of Massachusetts at Amherst in 1989. He attended the University of Washington in Seattle, receiving his M.S. in 1992 and Ph.D. in 1996 in Computer Science. His research interests include parallel and distributed visualization algorithms and frameworks, visualization systems and multi-source visualization techniques. He is a member of the IEEE Computer Society and co-chair of the Parallel Visualization and Graphics Symposium (the successor to the Parallel Rendering Symposium) to be held in 1999 in conjunction with IEEE

0-8186-9176-x/98 $10.00 © 1998 IEEE

Visualization '99.

Wes Bethel is part of the Visualization Group at E. O. Lawrence Berkeley National Laboratory. Their bifurcated mission is to bring the best of visualization technology to ongoing discipline-specific scientific research programs within the Energy Research community, as well as research and development of visualization technology. Bethel has received numerous awards for service to the scientific visualization community, the most meaningful of which was "AVS Hero," bestowed by a shall-remain-nameless luminary in the visualization community. Bethel earned an MS in Computer Science at the University of Tulsa in 1986. Current research interests include visualization of large data in distributed memory environments, techniques for remote visualization, and scientific computing.

Lloyd A. Treinish is a research staff member in the Visual Analysis group at the IBM T. J. Watson Research Center in Yorktown Heights, NY. He works on techniques, architectures and applications of data visualization and methods of data management, which includes a focus on earth, space and environmental sciences. His research interests range from visualization systems, data models, and perceptual rule-based tools to study of atmospheric and space physics phenomena, and cartography. Earlier he did related work for over a decade at NASA/Goddard Space Flight Center in Greenbelt, MD. A 1978 graduate of MIT with an S.M. and an S.B. in physics, and an S.B. in earth and planetary sciences, he has been at IBM since April 1990.

Andrei State is a researcher in the Department of Computer Science at the University of North Carolina (UNC) at Chapel Hill. He has held this position since 1991. From 1991 to 1993 he was a graphics system designer for the VISTAnet Gigabit Network project. Prior to that he worked as a software engineer at Thomson Digital Image and on CATIA at Dassault Systemes, both in Paris, France. He received his Dipl.-Ing in Aerospace Engineering in 1988 from the University of Stuttgart and his MS in Computer Science in 1991 from UNC. Andrei is the principal designer of the UNC augmented reality visualization systems. His current research interests include AR technology and 3D morphing techniques.

Sam Uselton is a researcher in visualization and computer graphics. He has been a contractor with the NAS Systems Division at NASA Ames Research Center since 1989. He received his BA from the University of Texas (Austin) and his MS and PhD from the University of Texas at Dallas. Before moving to NASA, he taught computer science full time and consulted in industry part time for ten years. Sam has worked with many scientists from industry and academia to visualize data from a variety of fields, with particular emphases in medicine, oil exploration and production, and aeronautics. His current research interests include visualization of very large, multi-source data sets, visualization quality, parallel algorithms for visualization and graphics, direct volume rendering, and realistic image synthesis.

0-8186-9176-x/98 $10.00 © 1998 **IEEE**