Editor:
Theresa-Marie Rhyne

Machine Learning to Boost the Next Generation of Visualization Technology

Kwan-Liu Ma
University of California at Davis

Visualization has become an indispensable tool in many areas of science and engineering. In particular, the advances made in the field of visualization over the past 20 years have turned visualization from a presentation tool to a discovery tool. Besides our improved understanding of perceptual aspects of visualization design, our ability to create and integrate novel interaction metaphors and hardware-accelerated rendering algorithms that enable interactive visualization has been key to this success. If you asked different visualization researchers about their opinions on the next key advance to boost the capability of the next generation of visualization technology, you'd likely get a variety of answers. The responses might include coupling quantitative analysis, integrating visualization into the overall workflow of scientific study, or extensive support for collaborative visualization. I would like to add another: incorporation of machine learning into the visualization process. As the large data problem drives many of the remaining challenges in visualization, we often find ourselves buried in data mining tasks. Machine learning has received great success in both data mining[1] and computer graphics;[2] surprisingly, the study of systematic ways to employ machine learning in making visualization is meager.

Machine learning is a well-established field of study. Like human learning, we can make a computer program learn from previous input data to optimize its performance on processing new data. In the context of visualization, the use of machine learning can potentially free us from manually sifting through all the data. However, if we would treat challenges that we face simply as conventional data mining jobs, then we might as well hand our problems to the data mining research community. Visual-based data mining uses visualization to guide data mining and has demonstrated successes in several application areas. Here, I discuss a new approach to making future visualization systems that go beyond visual-based data mining.

In addition to the large data challenges, sophisticated visualization tasks and algorithms require a mastery of the algorithm's details, the data's properties, and the hardware's capabilities. These hurdles often discourage those most knowledgeable of the underlying problem from driving the visual exploration process. As a result, the visualization's potential is limited, possibly reducing the extent of scientific discovery. If we can abstract away sophisticated algorithms and hardware behind a simple and intuitive user interface, then the user is left with only high-level decision making for guiding the data visualization and discovery process. It's viable to integrate intelligence (the ability to learn) into visualization systems to achieve this goal—that is, to eventually remove the need of a human user to handle tedious or repetitive tasks by learning from previous sessions and input data. The problem would then become designing an appropriate user interface for each visualization task. Since visualization is an effective means for both inputting domain knowledge and interpreting complex information, a promising design would realize intelligence augmentation in a way that visualization becomes the user interface of the visualization tool.[3] All tasks are then performed by operating directly on visualizations. To explain what I mean, I will describe intelligent visualization designs for three different applications.

Volume classification and visualization

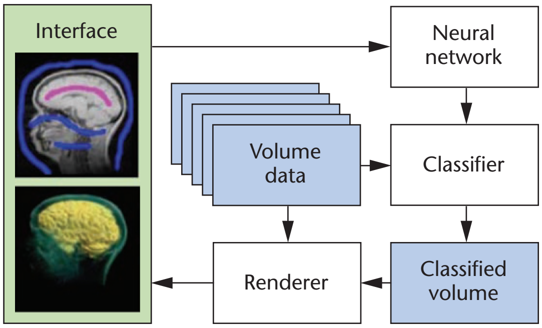
In volume rendering, a critical task is the classification of different materials in a volume according to the visualization's purpose. A transfer function maps voxel values to color and opacity, and the user usually defines it through an interactive graphical editor. Guidelines exist for making color transfer functions. Designing opacity transfer functions for volume rendering is less intuitive for the users since they must work in some derived transfer function space. Furthermore, the conventional 1D transfer function is of limited effectiveness in performing the actual classification. For example, in an MRI head data set, specifying a 1D transfer function that can correctly differentiate the brain and the region near the skull might be difficult since the intensity values of the two regions are so similar. As such, the resulting visualization would show both materials together; in this case, the outer layer might obscure the brain material of interest. The user could reduce the outer layer's opacity to make the brain more visible, but the brain would simultaneously become more transparent and difficult to see. Trying to find a transfer function that is a compromise between the two is a nontrivial task without the appropriate visual interface support. To obtain better classification results, we need higher dimensional transfer functions that take into account more data properties—such as gradient, neighboring texture, and

0272-1716/07/$25.00 © 2007 IEEE

position. A great deal of research exists devoted to the generation of transfer functions for volume visualization.[4] Two-dimensional transfer functions have proved effective, but the complexity of the conventional user interface rises with the dimensionality of transfer functions. Higher dimensional transfer functions are too confusing for the user to edit; when a transfer function has five, ten, or even more dimensions, it's nearly impossible for the user to directly define the function.

One intelligent interface that proved effective for specifying high-dimensional classification functions is a painting interface.[5] Users interactively paint directly on the volume-rendered images or selected cross sections of the volume. The users are given full control of what materials to classify by applying one paint color to parts of the volume representing materials of interest, and another paint color to regions they don't desire. Abstracted from the user is the generation of a high-dimensional classification function using a supervised machine learning technique, such as artificial neural network or support vector machines. Figure 1 shows such an intelligent visualization system. The system uses the painted regions (a small subset of the volume) as training data to learn how to classify the whole volume. The classification step maps each voxel into a value indicating the likelihood that the voxel is part of the material of interest. The system can then map this uncertainty to opacity for rendering.

In practice, users might want to classify more than a single object in a volume. For example, they might want to show more than one material at a time or a certain organ with a high opacity value and other regions with low opacity to provide context. Figure 2 shows the results of the classification of multiple materials. These images would be difficult, if not impossible, to generate with a 1D or 2D transfer function. For example, the cerebrum and the cerebellum of the brain have similar density values. Users can separate these two regions using texture and position information. To classify more than one material at a time requires multiple networks; however, only one neural network is trained at a time. Rendering occurs using multiple passes, one for each of the material classes. A good strategy would be to employ the more expensive hardware neural network renderer only when displaying the most recent material class, with the previous materials rendered from precomputed classified volumes. The ability to learn is powerful since the system can apply what it has learned to performing similar tasks, possibly in a fully automated fashion. An example of the strategy is to reuse a network trained to classify the brain from a scan of a patient for new scans of the same patient.[5]

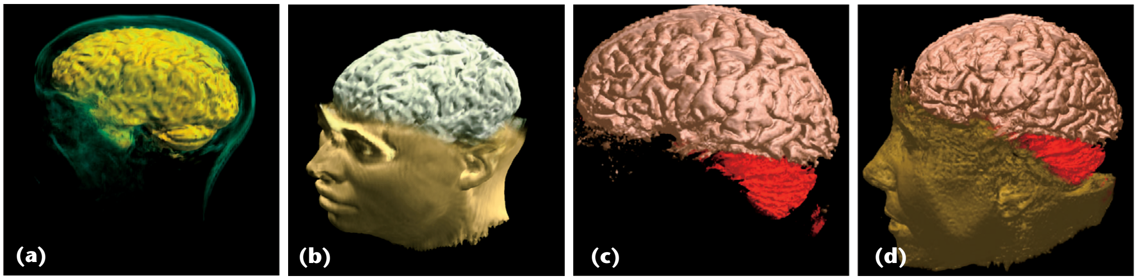


1 Intelligent, high-dimensional volume classification.[5] A painting-based interface lets the user specify regions of interest by brushing. The user trains the neural network to classify wanted and unwanted materials using the voxels (marked by the red and blue strokes).

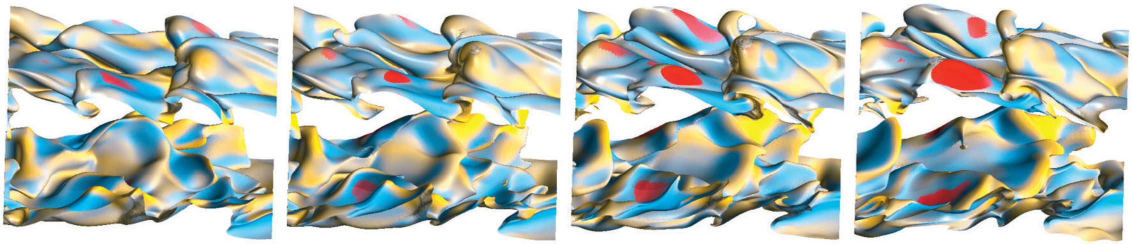
4D flow feature extraction and tracking

In the study of time-varying flow data, feature extraction and tracking is mostly done by explicitly separating the feature of interest from the raw data and creating (and following) either a volumetric or geometric representation of the extracted feature. A data set typically consists of hundreds to thousands of time steps. A single transfer function could not accommodate the generally varying dynamic range of data values over time. Furthermore, some of the features can only be defined by the relationship between multiple variables, or by its size, shape, location, and neighborhood, suggesting that the feature extraction must occur in a high-dimensional space, similar to the aforementioned volume classification problem.

Conventional feature extraction methods are mainly based on an analytical description of the feature of interest. In the case that the properties cannot be easily defined and are sometimes unknown, feature extraction and tracking become a manual-driven and trial-and-error process. The same intelligent system introduced for classifying materials can be applied to 4D feature extraction and tracking in the context of flow visualization. In other words, it's possible for a visualization system to learn to extract and track features in



2 These images demonstrate classification of multiple materials: (a) the brain rendered opaque and the skull and skin semitransparent. (b) This image uncovers the brain while keeping the bottom half of the head. (c-d) These images highlight the cerebrum and cerebellum.



3 Learning to extract and track time-varying flow features.[6] In modeling turbulent mixing on reacting layers, scientists want to verify that for high mixing rates the flame can become locally extinguished, as highlighted in red. Using the intelligent interface, the scientist extracted the desired feature for a few selected time steps and then applied the learned network to the hundreds of other time steps.

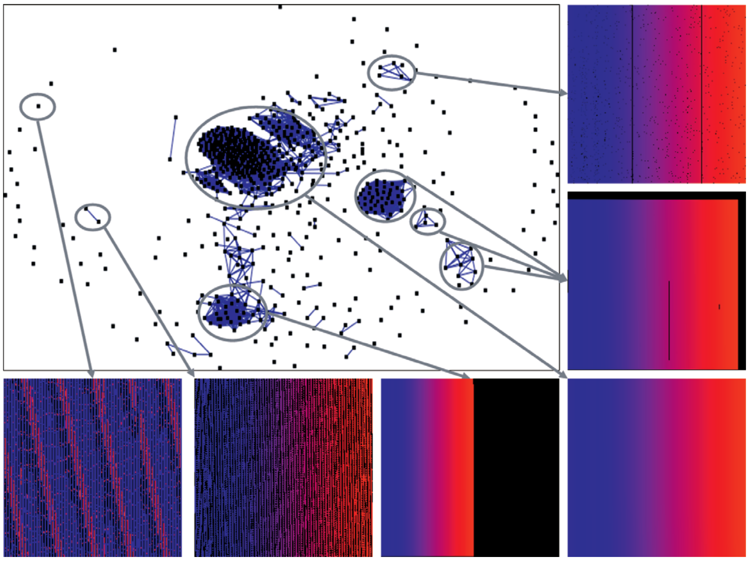
complex 4D flow fields according to their visual properties, location, shape, and size.[6] Again, such an intelligent system approach is powerful because it lets us extract and track a feature of interest in a high-dimensional space without explicitly specifying the relations between those dimensions, resulting in a greatly simplified and intuitive visualization interface.

According to the properties of the feature of interest, the extraction and tracking can occur in either the transfer function space or the data space. In the transfer function space, the user provides good transfer functions for a small number of selected time steps, and the system generates transfer functions for other time steps, which can be in the number of hundreds or even thousands. In the data space, the user specifies values of interest for selected variables, so the system can construct a feature vector for training. Figure 3 shows an example of data-space extraction and tracking using four variables.

Network scan characterization

Another example occurs from network security. Scanning a network is the first step in a network attack attempt. To attempt to make a scan anonymous, an attacker can use a variety of techniques, such as coming from different source addresses or scanning in a random order. Certain variations in arrival time of the scanning connections can help identify such an attacker. However, because these variations are rather chaotic, statistical methods alone are not enough. Analysts need enhanced capabilities for understanding subtle timing characteristics in high-volume Internet activities—for example, hostile probes—as these activities exhibit a fascinating degree of structural detail. They need to differentiate productive activities such as Web crawlers from malicious activities such as worms.

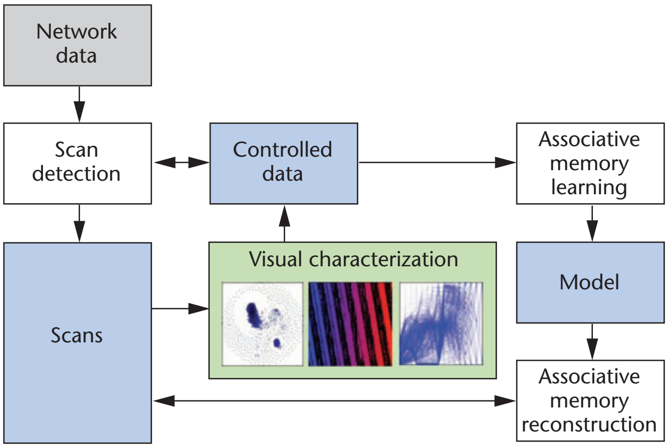
One way to characterize network scans is to analyze destination IP addresses and packet arrival timing information retrieved from the network scans. A visual representation of different metrics of the arrival time at each destination IP address can show structures that help analysts characterize large numbers of network scans. We created a visualization system to quickly and readily classify and compare a large number of scans.[7] Figure 4 shows one view of the system that uses clustering. A problem with the raw network scans is the prevalence of noisy and distorted data, which often makes direct comparison of the data patterns difficult and inaccurate. Thus, we need a way to remove the noise and restore the original scan activity pattern. Machine learning is excellent for such a task. One method of choice is the associative-memory neural network, which mimics how human memory works and is particularly effective at pattern recognition tasks even when the patterns are distorted or incomplete. Figure 5 shows a proposed network scan analysis system that uses both associative memory learning and interactive visualization to characterize scans. A set of controlled data (that is, known scan patterns) helps train the system. The trained system can then correctly fix and classify most of the incomplete or distorted scans in the newly collected data. The system provides the analysts with both overviews (for example, a graph as scan clusters) and detailed views of the scans to verify and refine the classification. The classification not only allows the analysts to examine scans as groups rather than a large amount of individual scans, but also single out scans that require attention.



4 Clustering of scans and some of the representative scan patterns.[7] Color represents various metrics based on arrival time at each destination IP address. In this picture, blue and red indicate arriving early and late, respectively, in the scan.

Conclusion

I anticipate seeing a growing interest in the development of intelligent interfaces for data visualization. Such systems will replace the current clutter of hardware- and algorithm-specific controls with a simple and intuitive interface supported by an invisible layer of complex intelligent algorithms. The user need only make high-level, goal-oriented decisions, making cutting-edge data visualization technology directly accessible to a wide range of application scientists. This exciting direction will draw new attention to several other research areas. One subject of study will be the visualization of machine learning operations.[8] When an artificial-intelligence system can learn a visualization task, presenting the user with the reasoning behind the learning might help with the user interface's transparency. An interface that provides this learning process could also help further refine and optimize the visualization process. A second subject of research will be the development of models of intelligent interfaces for data visualization. In this article, I've shown several supervised intelligent systems. Unsupervised learning is also promising and can lead to new interface designs.[9] Another subject of relevant research is how to visually assess and communicate the uncertainty of an intelligent system's outputs. No real data is perfect, and the learning-based classification and visualization process can introduce an additional layer of ambiguity.[8,10] Providing feedback about the quality of classification and visualization results will be an important part of designing a reliable, intelligent visualization system. Finally, the effort to couple machine learning with interactive visualization and to evaluate the effectiveness of the resulting interface designs should occur using a variety of applications. These studies will pave the way to the creation of next-generation visualization technology. This type of technology will be built upon further exploitation of human perception to simplify visualization; advanced hardware features to accelerate visualization calculations; and machine learning to reduce the complexity, size, and high-dimensionality of data. ■



5 Intelligent visual characterization of network scans. The intelligent system can not only reconstruct those incomplete or distorted scans but also classify them to facilitate subsequent analysis tasks.

Acknowledgments

The reported intelligent visualization research is supported in part by the National Science Foundation (grant IIS-0552334), and by the Department of Energy's SciDAC program. I am grateful for the reviewers' constructive suggestions, which helped improve the final article.

References

1. P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2005.
2. A. Hertzmann, "Machine Learning for Computer Graphics: A Manifesto and Tutorial," *Proc. Pacific Graphics Conf.*, IEEE CS Press, 2003, pp. 22-36.
3. K.-L. Ma, "Visualizing Visualization: User Interfaces for Managing and Exploring Scientific Visualization Data," *IEEE Computer Graphics and Applications*, vol. 20, no. 5, 2000, pp. 16-19.
4. H. Pfister et al., "The Transfer Function Bake-Off," *IEEE Computer Graphics and Applications*, vol. 21, no. 3, 2001, pp. 16-23.
5. F.-Y. Tzeng, E.B. Lum, and K.-L. Ma, "An Intelligent System Approach to Higher-Dimensional Classification of Volume Data," *IEEE Trans. Visualization and Computer Graphics*, vol. 11, no. 3, 2005, pp. 273-284.
6. F.-Y. Tzeng and K.-L. Ma, "Intelligent Feature Extraction and Tracking for Large-Scale 4D Flow Simulations," *Proc. Int'l Conf. High Performance Computing, Networking, Storage and Analysis*, IEEE CS Press, 2005.
7. C. Muelder, K.-L. Ma, and T. Bartoletti, "A Visualization Methodology for Characterization of Network Scans," *Proc. Workshop Visualization for Computer Security* (VizSEC), IEEE CS Press, 2005, pp. 29-38.
8. F.-Y. Tzeng and K.-L. Ma, "Opening the Black Box—Data Driven Visualization of Neural Networks," *Proc. IEEE Visualization Conf.*, IEEE CS Press, 2005, pp. 383-390.
9. F.-Y. Tzeng and K.-L. Ma, "A Cluster-Space Visual Interface for Arbitrary Dimensional Classification of Volume Data," *Proc. Joint Eurographics, IEEE TCVG Symp. Visualization*, Eurographics Assoc., 2004, pp. 17-24.
10. J. Kniss et al., "Statistically Quantitative Volume Visualization," *Proc. Visualization Conf.*, IEEE CS Press, 2005, pp. 287-294.

Contact author Kwan-Liu Ma at ma@cs.ucdavis.edu.
Contact editor Theresa-Marie Rhyne at tmrhyne@ncsu.edu.