Editor: Theresa-Marie Rhyne

User Studies: Why, How, and When?

Robert Kosara
VRVis Research Center, Austria

Christopher G. Healey
North Carolina State University

Victoria Interrante
University of Minnesota

David H. Laidlaw
Brown University

Colin Ware
University of New Hampshire

In crafting today's visualizations, we often design and evaluate methods by presenting results informally to potential users. No matter how efficient a visualization technique may be, or how well motivated from theory, if it doesn't convey information effectively, it's of little use.

Why conduct user studies?

User studies offer a scientifically sound method to measure a visualization's performance. The reasons abound to pursue user studies, particularly when evaluating the strengths and weaknesses of different visualization techniques. For example, in Figure 1 Laidlaw compared six methods for visualizing 2D vector fields.[1] His experiments measured user performance on three flow-related tasks for each of the six methods. He used the results to identify what makes a 2D vector field visualization effective.

Studies can show that a new visualization technique is useful in a practical sense, according to some objective criteria, for a specific task. Even more exciting are studies (like Laidlaw's) that show that a new technique is more effective than an existing technique for an important task. User studies can objectively establish which method is most appropriate for a given situation.

A more fundamental goal of conducting user studies is to seek insight into why a particular technique is effective. This can guide future efforts to improve existing techniques. We want to understand what types of tasks and conditions yield high-quality results for a particular method. This knowledge is critical because different analysis tasks require different visualization techniques.

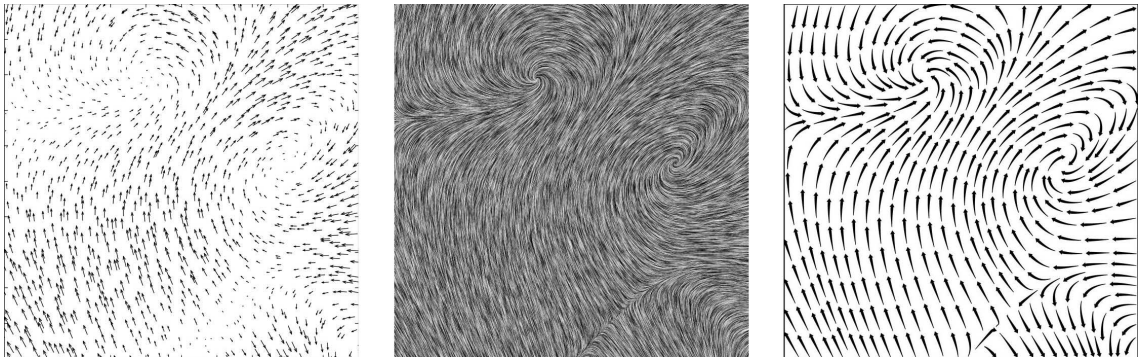
A final use for studies in visualization is to show that an abstract theory applies under certain practical conditions. For example, results from psychophysics or computer vision may not extend to a visualization environment. We can run user studies to test this hypothesis. Results can show when the theories hold and how they need to be modified to function correctly for real-world data and tasks.

A good starting point in any study is the scientific or visual design question to be examined. This drives the process of experimental design. A poorly designed experiment will yield results of only limited value. Although a comprehensive discussion of experimental design is beyond the scope of this article, we offer some suggestions and lessons learned in the "Basics of User Study Design" sidebar. We also describe in the sidebar how we designed experiments to answer important questions from our own research.

Color sequences

One reason for conducting studies is to determine if we can apply theoretical principles derived from other disciplines (such as psychophysics) to visualization design. Researchers have studied human color vision theory for more than a century. Results from this work provide a solid foundation for using color in visualization. However, choosing colors for a particular visualization problem is normally very different from the extremely simple displays used by experimental psychologists. We need to experiment to bridge this gap between theory and practice.

Consider the problem of designing pseudocolor sequences for scientific images. We have a continuous data field over a plane (for example, an energy or density distribution), and we want to use color to illustrate features in the data. Human vision theory dictates that neural signals from the rods and cones in the retina are



1 Three of six visualization methods[1] compared with a user study. Each method shows the same vector field. User performance on different tasks provides quantitative comparisons of the methods.

Published by the IEEE Computer Society

Basics of User Study Design

While a complete tutorial on user studies is beyond the scope of a short article, we hope to share some useful lessons we've learned.

The approach we advocate is a form of applied perception research. Proper use of this technique requires an understanding of how to build experiments that include human participants. It's challenging to design an experiment that will give robust answers to the questions of interest. A typical study might ask, Which prospective method is most promising? Do any of these methods perform better than the best available alternative? Unfortunately, many problems can compromise a study's validity or make it difficult to draw useful insights from the results. Is the task appropriate? Is it possible that participants were using cues other than the ones being examined to perform the task? Is there a control condition to provide a baseline for comparison between different methods? Do all participants have a correct and equivalent understanding of the task? Are all participants sufficiently willing and able to perform the task? Is there a learning effect, wherein the participant performs the task better because he or she has already solved a similar task before?

We can address these problems by

- testing participants for adequate spatial acuity, stereo ability, and absence of color blindness;
- randomizing the presentation order of the trials, by using written instructions;
- letting participants rest during the experiment to avoid becoming fatigued;
- devising robust methods to identify when participants are giving garbage answers; and
- asking participants to successfully complete a training task before proceeding to the recorded trials.

Because of the significant costs associated with running an experiment, it's often valuable to conduct a pilot study with one or two viewers. This allows testing and refining the experimental design before starting a full-fledged study with numerous participants.

A wide range of experimental methods may be appropriate. At one end of the spectrum is the rigorous application of signal detection methods.[1] We can use these to assess the detectability of a target structure from a background of noise. A more common experiment type is the evaluation of a number of different visual features. For example, a study might address the question of how well motion parallax, stereoscopic depth, and surface texture contribute to the perception of surface shape. Such an experiment calls for a factorial design with analysis of variance (ANOVA) to evaluate the results.

Another concern is how many participants to use. The answer depends on what's being studied. For psychophysical experiments that measure low-level visual phenomena, it's acceptable to use only a few participants. This is because there's little variation in viewer's reactions. These experiments contain numerous repeated measures (that is, multiple trials with the same experimental conditions) to ensure a sufficient total number of trials. If cognitive (as opposed to purely perceptual) processes are involved, more participants are normally required. Counterbalancing participants based on characteristics like gender, age, or experience may also be necessary. A detailed description of both participants and methods is an essential component for any publication involving user studies.

Finally, researchers at US universities should be aware that they may be required to obtain prior approval (or exemption) from the Institutional Review Board at their institution before conducting any work involving human subjects. In other countries, similar requirements may apply.

In all cases, consulting with an expert on experiments can be invaluable. This will help with design and in applying appropriate statistical analyses to study the experimental results.

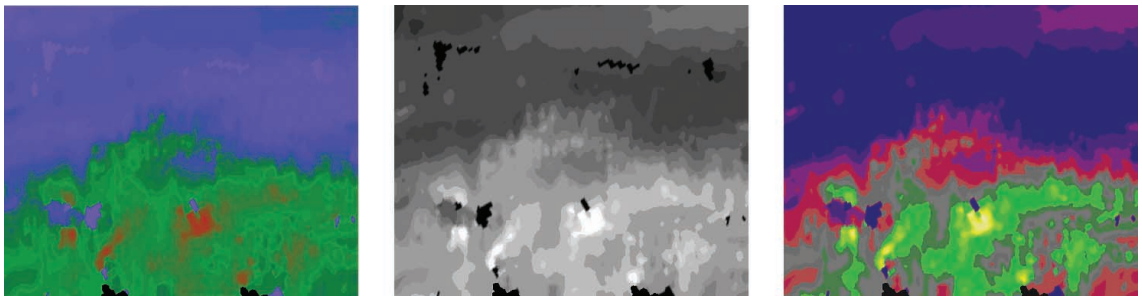
Reference

1. J.A. Swets and R.M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press, 1982.

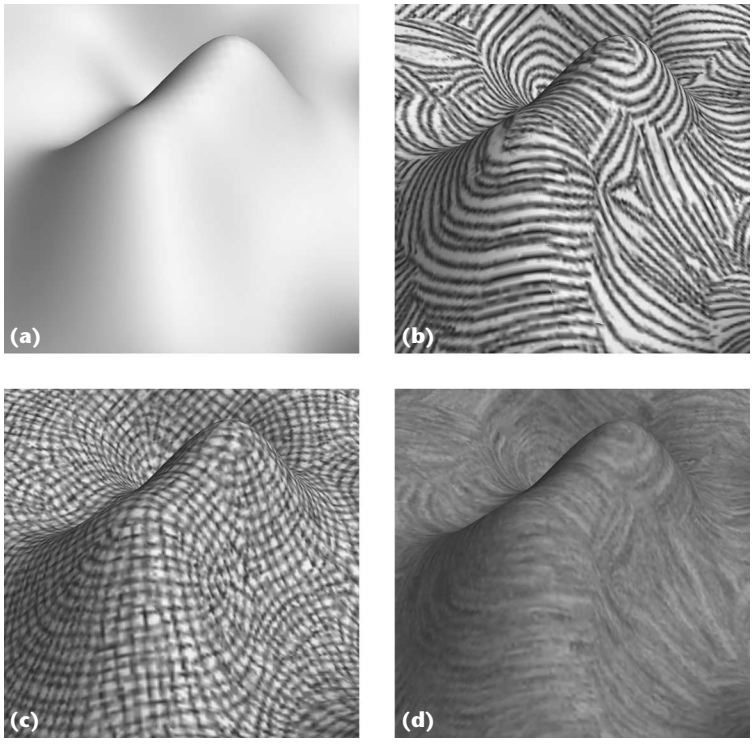
transformed by neural connections in the visual cortex into three opponent color channels: a luminance channel (black–white) and two chromatic channels (red–green and yellow–blue). The luminance channel conveys the most information, letting us see form, shape, and detailed patterns to a much greater extent than the chromatic channels. Perception in the chromatic channels tends to be categorical. That is, we tend to place colors into categories like red, green, yellow, and blue. However, we see hues such as turquoise or lime green more ambiguously. Another relevant theoretical point is that *simultaneous contrast* (the phenomenon by which perceived color is affected by surrounding colors) occurs in all three opponent channels. This can cause large errors when viewers try to read values in the data based on color.

We can use these theories to draw some conclusions regarding the design of color sequences:

- If we want our color sequence to reveal form (such as local maxima, minima, and ridges), or if we need to display detailed patterns, then we should use a sequence with a substantial luminance component.



2 Three color sequences: (a) a chromatic sequence, good for representing categories; (b) a luminance sequence, good for representing form; and (c) a combined chromatic-luminance sequence, good for representing both categories and form.



3 Four examples from a study testing different methods to enhance shape perception: (a) Phong shading, (b) one principal direction, (c) two principal directions, and (d) a line integral convolution.

- If we want to display categories of information—for example, the classification of a terrain into regions of different geological type—then we should use a chromatic sequence.
- If we want to minimize errors from contrast effects, then we should arrange a sequence to cycle through many colors.

We can also construct a general solution that cycles through many colors (to allow categorization) while continuously increasing luminance.

Figure 2 illustrates three different color sequences selected to emphasize a different aspect of the underlying data. Experimental studies have verified that these theoretical predictions apply in the case of color sequences.[2] This demonstrates the use of well-established theories to build design guidelines, together with experiments that validate the guidelines in an applied setting.

Shape from texture

Numerous applications in scientific visualization involve the computation and display of arbitrarily shaped, smoothly curving surfaces. A common case is level surfaces in volume data. By default, the standard practice is to render these surfaces with a smooth, Phong-shaded finish. One important question that arises is, Can we better convey the 3D shape by rendering the surface as if it were made from a subtly textured material, rather than polished plastic? Ample evidence from psychophysics[3] suggests that certain kinds of surface texture can facilitate shape perception (see Figure 3).[4] Unfortunately, the exact mechanisms by which surface texture affects shape perception—and hence the specific characteristics of texture patterns that best show shape—remain unknown. Complicating any naive attempt to use texture to enhance shape appearance is the complementary evidence that under many conditions texture can camouflage surface shape features.[5]

Through carefully designed experiments, it's possible to gain concrete insights into how we might use texture most effectively to support accurate shape perception. More specifically, we can start to answer the question, If we want to design the ideal texture that best conveys the shape of a smoothly curving surface, what should the characteristics be? Visualization researchers' user studies are essential in this endeavor for several reasons.

First, traditional vision researchers are primarily concerned with elucidating the neural processes involved in the perception of shape from texture, and their investigations don't fully encompass the scope of questions that we'd like to ask.

Second, there's a limit to the depth of understanding we can derive purely from introspection and informal empirical comparison. In the absence of a clear task, view-

ers may adopt differing opinions about which textures are most effective. Without concrete experimental evidence, it may be impossible to sort out these differences.

Furthermore, complex problems rarely yield simple answers. If texturing can help, it's unlikely that any method we initially attempt will turn out to be the best in all cases. We expect to discover complicated interactions between surface texture and shading, between texture orientation and surface geometry, and between aesthetics and convention. We may also find numerous task dependencies. This suggests that we'll need to iterate to achieve progressively more effective methods for different purposes. These goals are best achieved through carefully controlled, quantitative user studies that objectively assess the impact of particular texture pattern characteristics on the accuracy of performance on specific tasks.

Perceptual textures

One key issue we must address when we design an experiment is which conditions to study. As the number of conditions (and the interaction between conditions) grows, so does the number of trials needed to test each condition properly. Therefore, we often restrict experiments to the most important conditions.

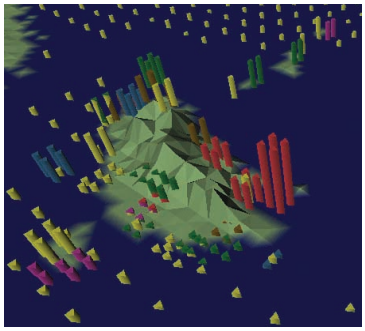
Understanding how we see the basic properties of an image lets us create representations that take advantage of the human visual system. An important discovery in psychophysics from the past 25 years is that human vision doesn't resemble the largely passive process of modern photography. A much better metaphor is a dynamic and ongoing construction project, in which the products are short-lived models of the external world specifically designed for the viewer's current visual tasks. Harnessing human vision for visualization therefore requires that we construct images that draw attention to their important parts.

Previous work in computer vision and psychophysics decomposed texture patterns into a number of basic texture dimensions like size, contrast, regularity, and directionality. Based on this, we wondered whether we could use individual texture dimensions to display multiple attribute values. Controlled experiments offer a way to answer this question.

We showed viewers regularly spaced 20×15 arrays of perceptual texture elements (or *pexels*) that look like upright paper strips. The pexels allow multiple texture dimension variations including height, density, and regularity of placement. Viewers saw the pexel grid for a short duration. We then asked whether a group of pexels with a particular target value was present or absent. Our experiment tested five different conditions selected from models of human vision and from texture segmentation and classification experiments in computer vision.

We varied

- *target type* (target pexels were defined by height, density, or spatial regularity),
- *target–background pairing* (different types of targets—for example, both medium and tall targets),
- *display duration* (the amount of time the viewer saw the pexel array),
- *target patch size* (the number of pexels used as targets), and
- *background texture pattern* (whether nontarget texture properties were held constant or varied randomly).



4 Perceptual texture elements (pexels) used to visualize a typhoon striking the island of Taiwan: pexel height represents wind speed (taller for stronger winds), density represents pressure (denser for lower pressure), and color represents precipitation (blue and green for light rainfall to purple and red for heavy rainfall; yellow indicates an unknown rainfall amount).

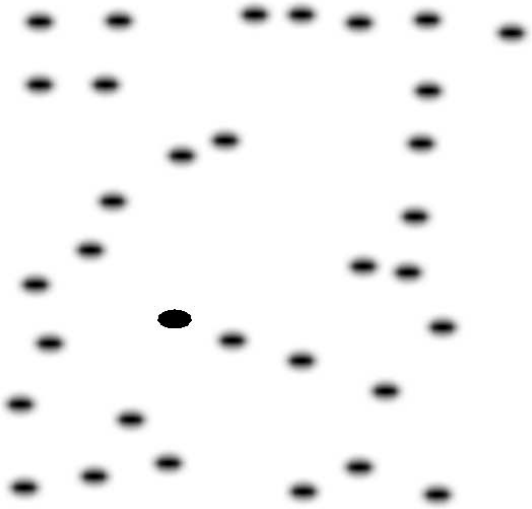
Each condition served a specific function. Target type let us test three different texture dimensions. Target–background pairing searched for differences in performance based on the target dimension's value. Display duration measured the time needed to perform a target detection task. Target patch size asked whether smaller texture patches were harder to identify. Finally, background texture pattern tested for visual interference when secondary texture dimensions varied randomly across the display. Even these basic conditions produced 108 different display types (three target types by two target–background pairings by three display durations by two patch sizes by three background patterns). Each viewer who participated during the experiment observed 576 trials from one target type (16 repetitions of a target's 36 different display types). We randomly selected eight trials (from the 16 repetitions) in each display type to contain a target patch; the remaining eight did not.

Results from the experiment showed a preference for target type (taller targets were easier to identify than shorter, denser, and sparser targets, which were themselves easier to identify than irregular or regular targets). High accuracy was possible for many target types, even for display durations of 150 ms or less. Finally, variations in regularity interfered with the identification of shorter, sparser, and denser targets (but not taller ones). A complete description of the experiment's results is available in Healey and Enns.[6] We applied these results as guidelines for using texture in multidimensional visualizations. Figure 4 shows an example of using pexels to visualize typhoon activity in southeast Asia.

Usability testing

We designed much of the work presented in this article to test basic perceptual features or visualization techniques. We've found, however, that visualization applications have important aspects that we should study within the application's context.

The approach to this type of study is quite different from basic perception experiments. Participants must solve a relatively complex task, where there's a greater freedom of actions and a higher potential for mistakes.



5 An image from the semantic depth of field study. The image was displayed for 200 ms, after which participants were asked to point at the quadrant with the sharp object.

Studying a technique in an application setting (as opposed to an artificially simple environment) is critical because we can't assume that low-level results automatically apply to more complex displays.

Comments from participants are often more important than the other data we collect because they provide valuable hints about what's happening during the experiment. Close observation of the participants can also offer information about experiment details that possibly weren't part of the original hypotheses.

An example of this type of study is the evaluation of semantic depth of field (SDOF),[7] a technique for guiding a viewer to specific information in an image. SDOF is based on the depth-of-field effect from photography, where different parts of a picture are in or out of focus based on their distance from the focal point of the lens. SDOF generalizes this concept. An object's sharpness depends not on its physical position, but on its relevance. Viewers are immediately drawn to the sharp (that is, highly relevant) parts of the image, but they can still choose to look at other, out-of-focus objects (see Figure 5). We designed an experiment that contained both basic perception and application components. The perception studies produced significant results, which were close to what we expected to find. The application findings, however, were much less conclusive.

One application was a map viewer that presented users with a map containing nine layers of information (for example, roads, elevations, and cities). We asked them to position a project (for example, a factory) based on three very important and three somewhat important factors. Users could reorder the layers by selecting which layer was on top. The layers were displayed in three different ways: opaque, semitransparent, and SDOF (the top layer was sharp and underlying layers were increasingly blurred). The hypothesis was that SDOF would make it easier to stack the layers in order of importance, and thus to answer more quickly and correctly.

While some useful results were identified during the application study, we didn't find statistically significant results in either response time or correctness. We concluded there were two problems with the study. First, the maps we used were visually too simple. Second, the number of tasks was too small; more examples per user might lead to significant results. We plan to consider these ideas in future work on SDOF.

When do user studies help?

While user studies are an important tool for visualization design, they aren't the proper choice in every situation. Experiments don't always work as expected and other techniques are available.

Other techniques

It's important to consider other options before designing and running a user study. Studies are time consuming to design, implement, run, and analyze. Typically, we can only use them to answer small questions, and any larger conclusions rely on generalizations that might not be valid. Often, measures that are less precise, quantitative, and objective may provide sufficient insight about a visualization question to let us move forward.

In our investigation of virtual reality tools for archaeological analysis,[8] we labored long and hard to design a good user study to test the system we developed. However, the experimental design eluded us. In the end, we videotaped a pair of archaeologists using the system to evaluate some of their scientific hypotheses. They also generated several new ideas, some of which would have been difficult to generate with other analysis methods. This approach was sufficient to demonstrate the visualization application's use.

In another context, we can also transcend the traditional user study. Artists and designers have been creating visualizations for centuries and have invented effective methods. User studies come from science—in fact, they embody the scientific method of posing hypotheses, taking measurements, analyzing them, and iterating to gain insight. For the scientific study of low-level vision, the methodology works, but as we rise up to the level of a scientific visualization application, it might not be possible to use these techniques to answer important questions.

Can we replace some parts of user testing with expert visual designers? This is a conjecture we can likely test (not surprisingly) with a user study, comparing results of a standard user study with expert visual designer input. Preliminary results suggest that visual designers can replicate some user study results more quickly and with more insight about why differences occur. However, we still have much to learn about the space between perceptual psychology and visual design.

When things go wrong

In some studies, experimental design may lead to results that aren't statistically significant. For example, in a recent study we hypothesized that users would perform differently for a visual search task in virtual reality if the virtual environment were different. In fact, we found that statistically there was no significant difference. Perhaps our conjecture was wrong, but it's also possible that our choice of task or other parts of the experimental design misled us. The virtual environment

may really matter in some cases. We continue to think about how the virtual environment might make a difference, particularly since visual context is important in 2D visual search tasks. Some studies aren't published because of null results, or because the results are inconclusive or uncompelling.

Null results are completely natural because they show that the original hypothesis wasn't supported by the data. This can be because the difference is too small for the amount of data collected, but most often it's because the hypothesized difference is insignificant. This is why the study was done in the first place and it should therefore not be considered a failure. In visualization, we can't publish null results (at least not on their own) easily. Nevertheless, the results can provide insight about which directions of research to pursue and which to abandon.

Inconclusive results are a much more serious problem. They usually mean that there was a design error in the study and that it must be run again. Usually, however, this affects only one part of a study, so the effort is considerably smaller the second time. Also, we can test additional hypotheses emerging from the successful parts of the study.

Uncompelling results can result from choosing the wrong task or measuring the wrong performance quantity. For example, in "The Great Potato Search,"[9] we chose a 3D visual search task. Unfortunately, it was a task that involved looking inward at a relatively small model. We believe a task that involved searching more broadly around the user might have shown important performance differences correlated with changes in the virtual context. While we can (and will) go on to test that new hypothesis, if we had chosen a different task in the first place, we would have been better off. There's always a tension between executing an experiment quickly and spending time on design. Practice can help reduce or alleviate these types of mistakes.

Conclusions

In this article we tried to advance the current state of the art in two ways:

- Promote evaluating visualization methods with user studies. This is being done in certain cases, but it's still far from standard practice in our field.
- Ask where user studies might be useful and where other techniques might be more appropriate (such as ideas from the visual arts).

User studies can improve the quality of our research. Although it's difficult to design a good experiment and the relevant skills require substantial study tempered with experience, a well-conducted study is usually worth the effort. The results can ultimately have a considerable impact and potentially contribute to the discipline's scientific foundations.

Even though we advocate more user studies, we recognize that other methods may be more appropriate in certain situations. Designers should be aware of these methods to select the best tool for the problem at hand. One reason visualization is such a fascinating part of computer science is because so many other fields (such as psychology and the visual arts) overlap with our research. ■

Acknowledgments

We'd like to thank Helwig Hauser, who played a key role in proposing the idea for this article.

We completed this work in part as a component of the basic research on visualization (http://www.VRVis.at/vis/) at the VRVis Research Center in Vienna, which is funded by the Austrian research program Kplus. The US National Science Foundation also supported our work (ACI-0083421, CCR-0086065, CCR-0093238).

References

1. D.H. Laidlaw et al., "Quantitative Comparative Evaluation of 2D Vector Field Visualization Methods," *Proc. IEEE Visualization 2001*, IEEE CS Press, 2001, pp. 143-150.
2. C. Ware, "Color Sequences for Univariate Maps: Theory, Experiments and Principles," *IEEE Computer Graphics and Applications*, vol. 8, no. 5, Sept./Oct. 1998, pp. 41-49.
3. J.T. Todd and F.D. Reichel, "Visual Perception of Smoothly Curved Surfaces from Double-Projected Contour Patterns," *J. Experimental Psychology: Human Perception and Performance*, vol. 16, no. 3, 1990, pp. 665-674.
4. S. Kim, H. Hagh-Shenas, and V. Interrante, "Showing Shape with Texture: Two Directions are Better than One," *Human Vision and Electronic Imaging VIII*, SPIE, no. 5007, July 2003.
5. J.A. Ferwerda et al., "A Model of Visual Masking for Computer Graphics," *Proc. Siggraph*, ACM Press, 1997, pp. 143-152.
6. C.G. Healey and J. T. Enns, "Large Datasets at a Glance: Combining Textures and Colors in Scientific Visualization," *IEEE Trans. Visualization and Computer Graphics*, vol. 5, no. 2, Apr.–June 1999, 145-167.
7. R. Kosara et al., "Useful Properties of Semantic Depth of Field for Better F + C Visualization," *Proc. Joint Eurographics and IEEE Trans. Visualization and Computer Graphics Symp. Visualization* (VisSym 2002), IEEE CS Press, 2002, pp. 205-210.
8. E. Vote et al., "Discovering Petra: Archaeological Analysis in VR," *IEEE Computer Graphics and Applications*, vol. 22, no. 5, Sept./Oct. 2002, pp. 38-50.
9. C.D. Jackson et al., "The Great Potato Search: The Effects of Visual Context on Users' Feature Search and Recognition Abilities in an IVR Scene," *Poster Proc. IEEE Visualization 2002*, http://www.cs.brown.edu/research/vis/lists/pdf/Jackson:2002:TGP.pdf.

Readers may contact Robert Kosara by email at Kosara@VRVis.at.

Contact department editor Theresa-Marie Rhyne by email at tmrhyne@ncsu.edu.