Toward Mixed Method Evaluations of Scientific Visualizations and Design Process as an Evaluation Tool

Bret Jackson[1] Dane Coffey[1] Lauren Thorson[1 3]
David Schroeder[1] Arin Ellingson[2] David Nuckley[2 4] Daniel F. Keefe[1] *

[1]University of Minnesota Department of Computer Science and Engineering
[2]University of Minnesota Department of Biomedical Engineering
[3]University of Minnesota Department of Design, Housing and Apparel
[4]University of Minnesota Program in Physical Therapy

ABSTRACT

In this position paper we discuss successes and limitations of current evaluation strategies for scientific visualizations and argue for embracing a mixed methods strategy of evaluation. The most novel contribution of the approach that we advocate is a new emphasis on employing design processes as practiced in related fields (e.g., graphic design, illustration, architecture) as a formalized mode of evaluation for data visualizations. To motivate this position we describe a series of recent evaluations of scientific visualization interfaces and computer graphics strategies conducted within our research group. Complementing these more traditional evaluations our visualization research group also regularly employs sketching, critique, and other design methods that have been formalized over years of practice in design fields. Our experience has convinced us that these activities are invaluable, often providing much more detailed evaluative feedback about our visualization systems than that obtained via more traditional user studies and the like. We believe that if design-based evaluation methodologies (e.g., ideation, sketching, critique) can be taught and embraced within the visualization community then these may become one of the most effective future strategies for both formative and summative evaluations.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*

Keywords

Visualization, Design, Evaluation

*Emails: {bjackson,coffey,dashroe,keefe}@cs.umn.edu
{lthorson,ellin224,dnuckley}@umn.edu

BELIV 2012 Seattle, Washington USA

1. INTRODUCTION AND THE CASE FOR DESIGN AS AN EVALUATION

Evaluating scientific visualization methods remains one of the most significant and challenging problems in our field. One reason for this is that evaluation is a multifaceted problem, involving understanding both novel technologies and the way that people use these technologies. Within the past decade, several complementary methods of evaluation have been used with some success. For example, studies of human perception have been used to evaluate the success of conveying shape or other data with various 3D computer graphics techniques. Interrante's work, which introduced the use of texture rather than transparency to convey complex 3D shapes [6] was pioneering in this area not only for its contributions to computer graphics methods but also for the rigorous evaluation methodologies that it introduced, which were motivated by and grounded in human perceptual capabilities. Similar to this style of "low-level" evaluation of visuals as motivated by visual perception, the field of scientific visualization has also recently drawn upon methods for low-level evaluation of interaction techniques, as informed by the human-computer interaction (HCI) community [10]. Although it is clear from these examples that quantitative evaluations like these can be successful, we argue that this style of low-level evaluation also has important drawbacks. For example, when advanced computer graphics and interactive techniques are combined together to develop a complete innovative data visualization system, the complexity of the system is often so high and the data that are visualized so complex that it becomes nearly impossible to design a testable task that could be used in the type of low-level human perception or HCI evaluations mentioned above.

This is a major problem for the field of visualization because the advanced interactive visualization systems that we are now creating represent some of the most exciting work done in the field, and these systems are often the ones that have the most potential to positively impact the data analysis processes of our collaborators in science, engineering, and other disciplines. This drives the need for new forms of evaluation that are appropriate for complex, interactive visualization systems.

To this end, Chris North and colleagues have recently introduced an exciting alternative evaluation strategy, insight-based evaluation, which can be viewed as a complement to low-level studies [9]. We are drawn to this methodology because it is so perfectly aligned with the high-level goal of visualization, which is nicely summarized as gaining new insights from data. In the insight-based evaluations that have been conducted to date, full-featured visualization systems are compared to each other using the metric of the number of insights generated on the part of the users. Insights are categorized in terms of their value, for example, a "deep domain insight" is more valuable than a "trivial" insight. Although this type of evaluation has great appeal in the sense that it seems to measure the ultimate effectiveness of visualization systems as directly as possible, it also has shortcomings. Since the methodology is so applied, it is difficult to employ in the early stages of visualization design. A complete system(s) that enables users to work realistically with their data is needed. Unfortunately, once such a complete system exists it typically takes a major effort to redesign the system to make use of any feedback that comes from the evaluation. Thus, this type of evaluation is most useful in a summative rather than a formative role.

The first conclusion we draw from these recent examples in the visualization research community is that it is likely that a mixed-method style of evaluation may be needed. Low-level studies tell us something, and summative evaluations using high-level insight-based methodologies (and/or detailed case studies of visualization applications) tell us additional information. But, is this enough? Can we evaluate visualizations completely through these two methodologies? What about evaluation during the design process rather than simply once a tool is complete?

In our efforts to answer these questions, we have found ourselves returning again and again to the value of the art / design / critique inspired process that we use regularly to develop novel visualizations in our research lab, but which has not yet gained widespread use in the field of scientific visualization. We believe this process may be a key to enabling better evaluations of visualizations if our community can learn to adopt it and to recognize its value as a formal evaluation methodology.

This creative design process involves large amounts of initial ideation through sketching and other forms of prototyping, which is then combined with iterative group critique of the design artifacts. Such an approach has been employed for decades by other disciplines such as architecture and graphic design. More recently, the HCI community has recognized and promoted the value of this process for evaluating user experience [1]. The view that we promote in this paper is that this type of formal design methodology, developed and accepted in a broad range of other visual and creative fields, should be recognized as an important and critical evaluation methodology in our field. We envision a time when published research papers provide detailed evaluation data by describing and picturing the series of iterative ideas that were developed, critiqued, and discarded or accepted throughout the design of the visualization tool. These evaluative data would include specific criticism made by visualization experts, domain scientists, and other stakeholders involved in critique sessions that would span the timeframe for development of the tool. If approached correctly, leveraging the knowledge we have from related disciplines such as graphic design that already make heavy use of this methodology, we believe that this form of evaluation may actually be much more valuable than what we see in current practice in the field. For example, when reading such an evaluation, we would expect to learn not just the percentage of time that a particular technique works or fails but also what other approaches might be considered along with a criticism of these approaches and other analysis.

To better explain the new emphasized role that a design-based evaluation approach can have in our field and how this could complement other evaluation methodologies we continue in the sections below by describing a range of recent evaluative studies conducted in our research lab. We begin with a discussion of the specific roles that traditional quantitative and quantitative evaluation has played in our recent work and highlight some of the limitations that we have encountered. We then present a case study of designing a detailed visualization of the motion of the human spine that illustrates how adopting formalized design processes can be used as a form of evaluation. We conclude with a number of recommendations for making this style of design and evaluation successful.

2. TRADITIONAL EVALUATIONS: SUCCESSES AND SHORTCOMINGS

Traditional evaluation techniques, such as quantitative and qualitative user studies, have been proven to be both successful and useful in the evaluation of visualization techniques. However, they are not without shortcomings, which often seem most apparent in research involving large, complex system design or research that seeks to enable fundamentally new scientific workflows.

A critical challenge that arrises when adopting a traditional quantitative user study methodology to evaluate a large visualization system is selecting an appropriate task that is both testable and also yields some interesting insight into how users work with the system for real data analysis. We faced these challenges in the design and evaluation of Interactive Slice WIM [2], a multi-touch virtual reality system for exploring and interacting with large volumetric datasets. This system combines many novel ideas, each of which could be separately quantitatively tested. For example, it includes multi-touch gestures for interacting with objects floating above the tabletop, a slicing based world-in-miniature metaphor, and a number of interactions for navigating, querying, and selecting data. The evaluation approach that we adopted in the end was to combine a traditional quantitative user study based on a navigation task with high-level qualitative feedback from expert users.

For the quantitative study, we realized we would need to limit the scope of the task in order to make it testable, so we focused on the important visualization sub-task of navigating through complex anatomical environments, such as a 3D reconstruction of the human heart. A clear shortcoming of this approach is that navigation is only one of a number of data analysis tasks supported by the system and required by users to do a full data analysis, so from the beginning it was clear that this quantitative study would not provide a complete evaluation of the system. The specific task required users to navigate through and search inside an isosurface of a human heart extracted from CT imagery. Two search targets were placed in predefined locations inside of the environment. Users were required to locate the two search targets and compare their size to determine which of two targets was larger. The decision of what data to use for this navigation task was also important and challenging.

The interface was designed to overcome the challenges of working in highly complex and organic environments, such as the human body. However, exposing novices to these complex environments introduces confounding factors in the study design, such as familiarity to the human anatomy or even their ability to work in such spatially complex environments. On the other hand, simplifying the environment and data used in the task may not have adequately tested the intended use of the system. Our study design attempted to strike a balance between these issues.

To help to overcome the limited scope of this quantitative study, in our published work, we combined this quantitative evaluation together with a high-level qualitative evaluation, both through an exit questionnaire of the study participants and by our domain science collaborators who helped design and motivate the system. Through the questionnaire we learned that users felt the interface was well suited for the task and that world-in-minaiture metaphors were easily understood. Our collaborators in the field of medical device design provided the most valuable feedback in a number of talk-out-loud sessions during the development of the tool and after its completion, which helped immensely to target to system toward relevant science and engineering problems.

Another recent evaluation that we performed in our research lab took the approach of comparing against an alternative "best of class" technique. In theory, this style of evaluation makes great sense, as our motivation in research is typically to advance beyond the current state-of-the-art techniques; however, identifying the best point of comparison, and especially the task that should be used for that comparison, is often challenging. A recent example comes from our work developing a novel interface for navigating through volume datasets that combines 6-DOF hand-tracking input with multi-touch gestures [3]. The novel aspect of this work is combining the two input modalities (6-DOF tracking and multi-touch). As such, the two "best of class" options we considered for comparison were to compare against the best touch-based visualization system for accomplishing a similar task (we identified FI3D [13] as the state-of-the-art system in this area) or to compare against 6-DOF techniques, such as virtual reality wand-based interfaces [8]. In the end, we reasoned that within the current research environment in which there is great interest in advancing multi-touch technologies that our work might have the most positive impact when viewed as an extension to current multi-touch capabilities; thus, the most appropriate point of comparison was the FI3D tool. With this point of comparison identified, we set about defining an appropriate testable task. In our experience, this is the major challenge with this style of comparative evaluation. The task we selected was a multiple-object docking task, which has the nice property that the accuracy of the docking can be calculated very directly. Unfortunately, it also has a weakness for understanding the use of visualization systems in that it tends to cause the user to focus quite specifically on a solid object in space, whereas our intent for the interface was to enable the user to work more fluidly with volumes of data (e.g., volume renderings) that have many features spread across a volume. To approximate this, we did include multiple objects through the volume that the user was to manipulate, but we noticed that users tended to focus on just one object as they worked with the interface; thus, we believe their mindset was probably more one of manipulating a single object rather than exploring a volume of data. This seemed to cause a number of undesirable artifacts in the results, for example, we believe that the differences we observed in our tests of our interface and FI3D were due mostly to a slight difference between the two techniques: our approach enabled the artists to define an axis of rotation at any 3D depth relative to the touch surface, FI3D uses a heuristic based on the depth of the data volume. The ability to define an arbitrary axis seems to be less desirable in the situation where the user has already clearly identified an object of interest.

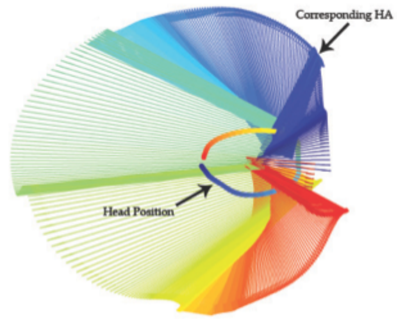

Figure 1: Preliminary top-down visualization of IHA's calculated for a neck circumduction exercise data collected experimentally. Color is mapped to time. The thin colored line labeled "Head Position" shows a trace of the tip of the patient's head over time, and the longer axes radiating outward labeled "Corresponding IHA" show the IHA at each corresponding point in time.

3. DESIGN-BASED EVALUATION

To illustrate how adopting a formalized design process could be used as a form of evaluation, we present a case study describing the development of a visualization showing the detailed motion of the human spine.

3.1 Case Study: Spine Neck Motion

Our collaborators in the field of biomedical engineering are interested in diagnosing and treating neck pain. Chronic neck pain is a frequent symptom of the general population and its causes are currently not well understood. Often, diagnosis relies on planer flexion/extension exercises, moving the head in a nodding motion. Only recently, have researchers started to explore the use of more complex motions such as head circumduction, rolling the head around the neck, as a potentially better way of quantify the kinematics of the complexly coupled vertebrae in the cervical spine.

One of the main challenges to understanding these data is the lack of good ways to visualize complex 3D spatial motions. A mathematical construct that has shown some promise for similar biomechanical data analyses is the helical axis [12], which describes the motion of a rigid body from one pose to another as a combination or rotation about and translation along an unique axis in space. For motion sequences, an instantaneous helical axis (IHA) can be calculated at small time intervals. The set of IHA's that together describe a motion sequence may be a useful construct for

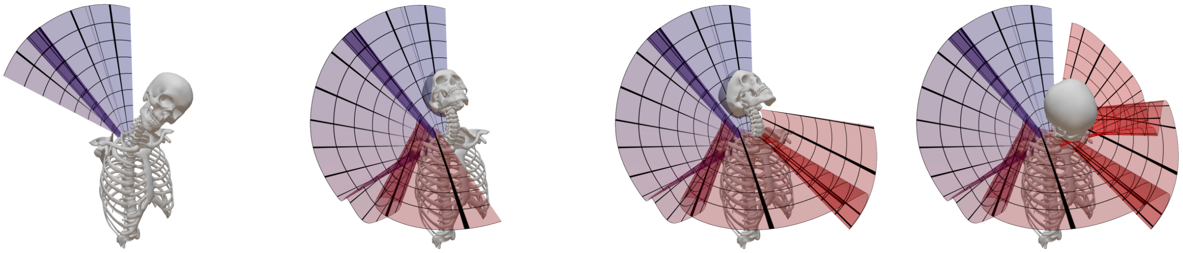

Figure 2: A sequence of 4 shapshots from the interactive animated visualization tool developed. 3D bone geometries are used to provide anatomical context to interpret the data. The axes are re-interpreted as a 3D surface. Effective use of transparency and texture are employed when rendering the surface to aid in perception of it's 3D shape and to emphasize the folds or "kinks" in the surface, which appear to be the most clinically important aspect of the data. Animation is used as shown in these three sequential frames to improve the understanding of the 3D location of the IHA relative to the pose of the bones at each phase of the motion.

evaluating and comparing biomechanics data, but relatively little work has been done previously in visualizing this type of data, so it is unclear how best to present such information to researchers and clinicians.

For neck circumduction exercises, our collaborators found that when the set of IHA's is plotted it forms a pattern that radiates outward from the neck as shown in Figures 1 and 2. One current hypothesis is that discontinuities in this pattern (e.g., kinks) indicate moments where the patient experiences pain and/or locations where disc degeneration is likely. If this proves to be true, then IHA analyses may become a critically important clinical tool. Thus, visualization methods are needed both to help evaluate current hypotheses about the clinical relevance of IHA analyses and to explain to clinicians what these axes mean and how they can be used in terms that they can understand.

We adopted a design-based approach to develop an effective visualization system to accomplish these goals, making heavy use of sketching and critique as means of evaluation throughout the process. Our starting point was the 2D visualization shown in Figure 1, created in matlab to produce a preliminary view of the data. Figure 2 shows a series of snapshots from the interactive, animated visualization tool we developed.

3.2 Ideation and Ideation

Two-dimensional paper sketching has long been a cornerstone of successful design processes. Because sketches are quick to make, disposable, and accessible, they provide an ideal way to generate and communicate many ideas [1].

Sketching as a form of ideation is similar to the methodology of rapid prototyping in software design. Both recognize that programming full implementations is costly, and integrating evaluation and feedback earlier in the process ultimately leads to better solutions.

In our approach, we advocate a sketching methodology for visualization first pioneered by Keefe et al. [5], and further described by Thorson et al. [11]. For instance, during the development of our case study, we started designing the neck motion visualization with the help of a graphic designer working in our lab. Over the course of one month, she created 320 concept sketches and illustrations depicting various ways of showing the motion of the IHA (see several examples in figure 3).

An experienced graphic designer well-versed in illustrative programs (e.g., Adobe Illustrator) is trained to rapidly explore a variety of solutions, which translates into the extremely quick production of vector graphic sketches. The speed and abundance of work is also a result of the graphic designer not constructing data-driven vectors, but illustrations based upon formal concepts that spark discussion of how the data can be applied.

The benefit of creating these varied sketches is based directly upon the standard workflow of a graphic designer; presentation of the visual problem and desired concept to be communicated (e.g., kinks in circumduction motion), ideation specific to this problem, sketching of all possible visual solutions-formally ranging from the traditional to the experimental, followed by a critique of these visual ideas. This process redefines suitable solutions to the problem by coupling more than one sketch, or traditional and abstract formal elements. This can only be achieved by sketching a wide range of solutions to be critiqued as a whole.

3.3 Critique

Complementing sketching and ideation, critiques are an integral part of using the design process for evaluation in design. In its most helpful form, a critique is a meticulous group discussion centered on how well particular aspects or details of a visualization support the intended goal [5]. Critique has been reported to have been employed successfully for visualization design in a few other cases. For example, Jackson et al. [4] used critique to evaluate 2D vector visualizations. Kosara et al. [7] also discuss the role that critique plays in developing visualization, highlighting the requirements of neutral voice, basis in fact, no self-promotion, and clear goals when critiquing one's own work. Although there is some evidence that the value of critique is starting to be recognized within the field of visualization, we believe this role should be further emphasized and, in particular, that robust critique should be acknowledged in our community as a formal mode of evaluation, with results published alongside more traditional quantitative and qualitative evaluations.

Critiques of spine motion visualization ideas were conducted weekly as we developed the visualization system. Early in the process one of the key visual decisions we explored was whether the axes should be depicted as discrete axes in space or abstracted a bit so as to highlight the 3D surface that they sweep out over time. Sketches used in these critiques can be seen to the left and middle of Figure 3. As

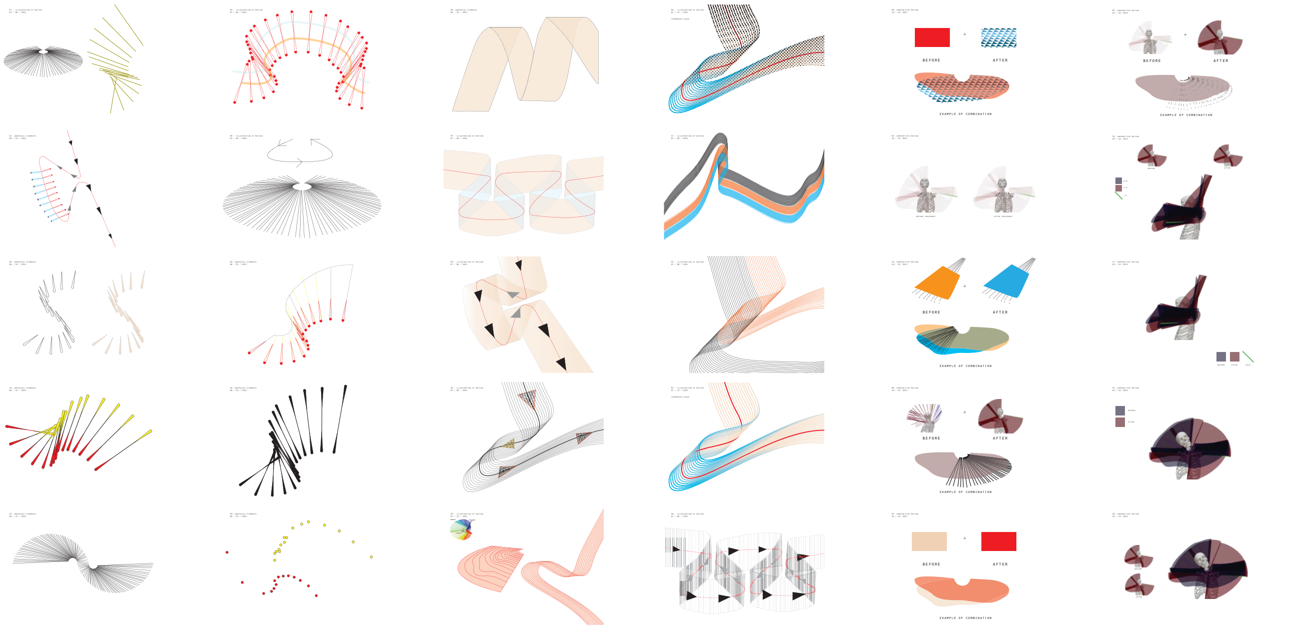

Figure 3: A selection from the set of 320 pages of 'sketches' created by our graphic designer team member. Each sketch, typically created using Adobe Illustrator, depicts a unique visual idea for how IHA data might be mapped to visual form. The sketches explore use of color, texture, transparency, animation, interactive control, narrative, metaphor, and more.

hinted by the selection of sketches shown here, many ideas for depicting the axes as a surface were developed, including strategies for mapping additional data variables onto this surface. We adopted the most successful of these ideas with respect to highlighting overlaps in the surface, as shown in Figure 4, since this was viewed as one of the most important clinical objectives of the visualization.

It is important to note that this process and its results were the product of the insights and extensive experience of the graphic designer and visualization experts involved, including knowledge of prior perceptual studies. Additionally, the domain scientists contributed by making sure that the visualization mapped directly to the clinical objectives. The vast amount of sketches produced in such a short time frame allows the lab to quickly move forward into a more refined visualization, based upon the critique of positive and negative elements seen in the sketches presented.

4. CONCLUSIONS

For a formal visualization design process to be recognized as a form of evaluation within our field, we have several recommendations. First, the design process must include not only an increased role for visual ideation, which might include tools such as sketches and other forms of prototyping, but also critiques where the qualities of the visual ideas are carefully discussed.

Artists, domain scientist, and visualization experts all must be involved at some point within the critique discussions. Each brings something to the table to improve the evaluation process and create better visualizations. For instance, the artists, illustrators, and graphic designers have a well developed sense of aesthetics and ways of creating visual form. Domain scientists provide insights into what aspects of the visualizations effectively show the science they are

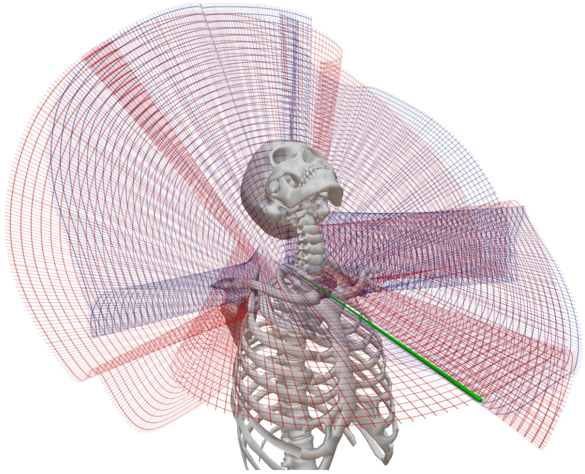

Figure 4: Screenshot from the interactive animated spinal kinematics visualization system, here displaying two sets of IHA data: the blue surface shows data for a healthy motion and the red for unhealthy.

interested in studying. Visualization experts act as facilitators, making sure the designs are able to be implemented programmatically, and bringing their visualization and HCI experience to bear on the evaluation.

Clearly, more traditional quantitative and qualitative evaluation methods still have an important role. Within our own research group, we will certainly continue to develop evaluation methodologies in these areas. However, our recommendation is that the visualization community should recognize and embrace the value of publishing early design ideas and accompanying discussion and criticism gained during critique. This is one of the most valuable forms of evaluation that we could expect in the area of visualization, especially when such evaluations include insights from all stakeholders in the visualization, each of whom is typically highly trained and specialized in his/her own area. Thus, elevating insights of this form to the level of "formal evaluation" within our publications would be beneficial for the new information it would convey, which we believe is often much greater than that which can be obtained through more traditional means of evaluation. In particular, design-based evaluations might help us to overcome a number of the limitations (e.g., defining an appropriate testable task) that are so prevalent when evaluating complex visualization systems.

5. ACKNOWLEDGMENTS

Placeholder

6. REFERENCES

[1] B. Buxton. *Sketching User Experiences: Getting the Design Right and the Right Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007.

[2] D. Coffey, N. Malbraaten, T. Le, I. Borazjani, F. Sotiropoulos, A. Erdman, and D. Keefe. Interactive Slice WIM: Navigating and interrogating volume datasets using a multi-surface, multi-touch VR interface. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1, 2011.

[3] B. Jackson, D. Schroeder, and D. F. Keefe. Nailing down multi-touch: Anchored above the surface interaction for3D modeling and navigation. In *Proceedings of Graphics Interface*, 2012.

[4] C. D. Jackson, D. Acevedo, D. H. Laidlaw, F. Drury, E. Vote, and D. Keefe. Designer-critiqued comparison of 2D vector visualization methods: a pilot study. In *ACM SIGGRAPH 2003 Sketches & Applications*, 2003.

[5] D. F. Keefe, D. Acevedo, J. Miles, F. Drury, S. M. Swartz, and D. H. Laidlaw. Scientific sketching for collaborative VR visualization design. 14(4):835–847, 2008.

[6] S. Kim, H. Hagh-Shenas, and V. Interrante. Conveying three-dimensional shape with texture. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, pages 119–122, 2004.

[7] R. Kosara. Visualization criticism - the missing link between information visualization and art. In *Proceedings of the 11th International Conference Information Visualization*, pages 631–636, 2007.

[8] M. R. Mine, F. P. Brooks, Jr., and C. H. Sequin. Moving objects in space: exploiting proprioception in virtual-environment interaction. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '97, pages 19–26, New York, NY, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[9] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, 2005.

[10] J. Schulze, A. Forsberg, Kleppe, R. Zeleznik, and D. H. Laidlaw. Characterizing the effect of level of immersion on a 3D marking task. In *Proceedings of HCI International*, Las Vegas, NE, July 2005.

[11] L. Thorson, H. Sohn, J. Downing, A. Ellingson, D. Nuckley, and D. F. Keefe. A designerâĂŹs approach to scientific visualization: Visual strategies for illustrating motion datasets. In *Poster Proceedings of IEEE VisWeek*, 2011.

[12] H. Woltring, R. Huiskes, A. de Lange, and F. Veldpaus. Finite centroid and helical axis estimation from noisy landmark measurements in the study of human joint kinematics. *Journal of Biomechanics*, 18(5):379 – 389, 1985.

[13] L. Yu, P. Svetachov, P. Isenberg, M. H. Everts, and T. Isenberg. FI3D: Direct-Touch interaction for the exploration of 3D scientific visualization spaces. *IEEE Transactions on Visualization and Computer Graphics*, 16(6)(November/December 2010), 2010.