EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (2016)
K. Lawonn, M. Hlawitschka, and P. Rosenthal (Editors)

An Introduction to Evaluation in Medical Visualization

Noeska Smit[1] and Kai Lawonn[2]

[1]Delft University of Technology, The Netherlands
[2]University of Koblenz - Landau, Germany

Abstract

Medical visualization papers often deal with data that is interpreted by medical domain experts in a research or clinical context. Since visualizations are by definition designed to be interpreted by a human observer, often an evaluation is performed to confirm the utility of a presented method. The exact type of evaluation required is not always clear, especially to new researchers. With this paper, we hope to clarify the different types of evaluation methods that exist and provide practical guidelines to choose the most suitable evaluation method to increase the value of the work.

Categories and Subject Descriptors (according to ACM CCS): I.3.m [Computer Graphics]: Miscellaneous—

1. Introduction

Medical visualization is by its nature a field that involves domain knowledge from the medical field. Whether the work is technique- or application-oriented, visualizations are developed with medical data or a medical user in mind. In technique papers, the emphasis is often on increasing performance through reducing computation times or memory footprint. Evaluation is therefore done by performance measurements on measurable quantities, such as memory usage and timings of newly developed algorithms in comparison to existing methods. Besides this, a visual comparison can be made, in which results of previous techniques are compared with the current outcome. In a case study, the visualization technique is applied to real world or simulated data and insights gained from the visualizations are described. When presenting techniques, the evaluation methods are often clearly defined depending on the contribution, and domain expert users involvement is not always necessary.

In application-oriented works, various evaluation methods are possible, but often medical domain experts are expected to be involved. While users may vary from researchers to clinicians, the evaluation methods are largely similar. As with technical papers, a case study can be performed. However, since good application papers rely on a thorough requirements elicitation and justification of the design decisions in relation to these requirements, the only way to evaluate the success of the work is to involve domain expert users. When existing visualization methods or applications are currently used, the application can be evaluated using a task comparison, comparing the old method to the new application in order to assess user performance. If such an existing system is not available, a semi-structured interview can elicit responses from the domain experts to see how they value the application in comparison to their current workflow to assess user experience. Forms can be employed to add numerical values to their opinions, often asking experts to state their level of agreement with various positive and negative statements on a Likert scale.

In this paper, we provide an overview of evaluation methods, challenges, and practical tips on performing evaluations to strengthen medical visualization papers. Furthermore, we provide a discussion on the merits of different evaluation types and recommendations for choosing a suitable evaluation in relation to the presented work.

2. Related Work

Previous works have discussed evaluation types in the field of visualization. Munzner proposed a nested model for the visualization design and validation [Mun09]. She discusses which evaluation methods are appropriate depending on the level the contribution is made. Isenberg et al. conducted a systematic review of ten years of evaluations in papers published at IEEE VIS [IIC*13]. They concluded that there was an emphasis on evaluations of algorithmic performance and qualitative result inspections through images. Furthermore, they notice an increasing trend in the evaluation of user experience and user performance.

3. Evaluation methods

Depending on the contributions and paper type, several types of evaluations can be performed.

3.1. Performance evaluation

In case the primary contribution of the work is a novel algorithm, mostly benchmarking is performed. In benchmarking, the performance of the new technique is compared to the performance of previous techniques. The goal is to prove the technique has improved performance over existing methods. Mostly, measurable quantities such as rendering time, processing time, or framerates are compared. In performance evaluations, it is paramount to compare the technique on the same datasets and using the same hardware to prevent unfair comparisons. The evaluation is strengthened by applying the techniques to different types of data to make sure it is not optimized for one dataset and performs worse for others. When a technique improves the visualization in terms of visual quality, a visual comparison can also be made, in which the same scene is rendered side-by-side using different techniques. The work by Ropinksi et al. [RDRS10], for example, features both framerate performance comparisons as well as visual comparisons.

3.2. Case study

Both technique and application papers can benefit from a case study to demonstrate the effectiveness of a visualization on either real-world or simulated data. In a case study, the method is applied to different datasets and the results are described in terms of what is shown or what insights can be gained from the visualization. For instance, a new visualization technique might highlight interesting blood flow patterns that were not visible before and this can be demonstrated in several datasets. Due to the nature of medical data, describing these interesting features of the data might only be possible in collaboration with a domain expert that can describe the exact details. If such experts are not available, these studies should be named usage scenarios rather than case studies [SMM12]. Guidelines for performing a case study have been proposed by Yin in his book on case study research design and methods [Yin94]. An example of a case study for a visualization application that has been performed according to these guidelines can be found in the work by Dzyubachyk et al. [DBB*13]. In their work, they evaluate the user experience of four radiologists using four clinical datasets.

3.3. User study

When the paper is more application oriented, intended end-users of the work need to be involved in the evaluation. In those cases, a user study is performed in which the users are shown or interact with the visualization application directly. In user studies, a distinction can be made between several subcategories of methods that can be employed, such as observation, task comparison, a (semi-)structured interview, or an evaluation form. Often in practice, several of these subcategories are combined to form a thorough evaluation procedure, such as in the work by Lawonn et al. [LLPH15].

Observation: Sometimes it is worthwhile to have the participants interact with the application themselves while observing them in order to not influence their experience with the application. In this case, their interactions may be recorded or they may be encouraged to think out loud and comment on what they are experiencing.

Task comparison: If an application is developed to perform tasks that were previously done using another application, a task comparison can be made between the newly developed and traditional approach. In this task comparison, participants perform tasks in both systems and their correctness, time and confidence in the result can be assessed. It is important to do multiple tasks and switch up the order between participants to prevent bias. It is also not desirable to perform the exact same task on the same dataset twice or more for one participant, since the knowledge gained from the first task may influence the performance on the second.

As an example, Baer et al. [BGCP11] compared different shading techniques for vessels and integrated blood flow. The users were asked to assess the depth of different vessel parts as well as to adjust a normal. Here, they evaluated which method has the best depth and spatial perception according to the user performance.

Interview: When a direct comparison between applications is not possible, often a semi-structured interview is performed to elicit direct feedback on the application from the participants. Several open questions are asked and depending on the responses of the participants, follow-up questions may be posed to get more insight into the opinion of the users.

Evaluation form: Besides an interview, participants may also be asked to respond to an application by filling out a form. A form is a convenient way to get a consistent response from all participants and particularly suitable to have participants assign numerical values to their opinions. For instance, a form could contain 30 positive and negative statements about the application in which the participants can respond on a 5-point Likert scale to indicate their level of agreement. To prevent participant bias, the statement order should be randomized and there should be an equal number of positively- and negatively-phrased statements [SMP03]. The evaluation results can then be presented in a table to provide an overview of all responses and conclusions that can be drawn from these. While this seems like a quantitative method, care has to be taken with drawing statistical conclusions based on the outcomes. Often the number of participants is too limited for a valid statistical analysis.

4. Users

If users need to be involved in the evaluation, questions often arise on how many users are needed and what their background should be. The number of participants required is ideally as high as possible, but in practice depends on the level of expertise. For example, getting five neurosurgeons for a study is an impressive number, while five computer science students is typically not considered to be a sufficient validation. If the numbers are low, care has to be taken with statistical analysis methods, since they often do not hold for the smaller numbers we are typically dealing with.

The type of users is strongly related to the target audience the application is aimed at. For instance, if a paper claims benefits for clinical practice, it is not sufficient to evaluate with medical researchers, but actual clinicians need to be involved. Preferably, the evaluation is performed with experts who are not also coauthors of the paper, to prevent a conflict of interests and bias. If a participant is a coauthor, however, this should be stated in the paper. It might be valuable to ask the users several background questions, in order to afterwards comment on the outcome of the evaluation, for instance, years of experience.

Depending on the level of specialization of the target audience, e.g., medical students versus trained surgical oncologists, the availability of experts may be limited. When it is not possible to get enough domain experts to evaluate the application, often researchers are tempted to involve non-domain-experts, such as computer science (PhD) students, since they are more readily available. It is, however, not trivial to translate medically-oriented tasks to non-medical users. Therefore such an evaluation needs to be used as a last resort, or in addition to a domain expert evaluation to further strengthen a specific component of the proposed method. If this is the only option to get enough evaluation participants, a justification needs to be made in the paper to explain why the evaluation results are valid for medical users as well.

5. Discussion

While algorithmic performance evaluation and qualitative assessment of image results are beneficial in technique papers and can be performed by non-domain experts, application papers need to involve domain experts to validate the utility of the approach. Since there are various types of evaluation possible involving these users, it is not straight-forward to define a good evaluation protocol that works for all applications. Often several types are combined to form an as thorough evaluation as possible, for instance an interactive session in which the participant is observed, followed by a semi-structured interview and finalized with a written response via a form.

A challenge in evaluating medical visualization applications with domain experts is the limited availability of experts. Since their expertise is needed to validate the applications, there are several ways to attempt to improve the situation. If the application can be made available via a web interface, the evaluation can be performed with participants remotely. If the interaction part of the application is not the target of the evaluation, authors can also consider evaluating their application via a video recording. In earlier work, by combining an elaborate video with a Google form, we were able to get eleven medical doctors to participate in our user study [LSPV15]. We would have not been able to get this many participants if we had focused our evaluation on our local institutions.

6. Conclusion

With this work, we provide an overview of evaluation methods for medical visualization applications. We provided an overview of evaluation techniques as well as some practical guidelines on which evaluation method is most suitable for several paper types common in medical visualization. We illustrated every evaluation type with a concrete example, that demonstrate how such an evaluation can be performed in practice.

In the future, we would like to extend this work by providing more examples of best practices for each of the evaluation types. Furthermore, we would like to offer precise guidelines for evaluation type selection and what kind of questions may be suitable in a questionnaire. Similar to how performance evaluations should be applied on the same datasets as well as on the same hardware, we would like to propose a guideline for a standarized questionnaire for use in uestions that should be asked and questions that should be avoided. medical visualization domain expert evaluations. In this way, user studies could become standardized such that results are easier to compare.

References

[BGCP11] Baer A., Gasteiger R., Cunningham D., Preim B.: Perceptual evaluation of ghosted view techniques for the exploration of vascular structures and embedded flow. *Computer Graphics Forum 30*, 3 (2011), 811–820. 2

[DBB*13] Dzyubachyk O., Blaas J., Botha C. P., Staring M., Reijnierse M., Bloem J. L., van der Geest R. J., Lelieveldt B. P.: Comparative exploration of whole-body MR through locally rigid transforms. *International journal of computer assisted radiology and surgery 8*, 4 (2013), 635–647. 2

[IIC*13] Isenberg T., Isenberg P., Chen J., Sedlmair M., Moller T.: A systematic review on the practice of evaluating visualization. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2818–2827. 1

[LLPH15] Lawonn K., Luz M., Preim B., Hansen C.: Illustrative visualization of vascular models for static 2D representations. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer, 2015, pp. 399–406. 2

[LSPV15] Lawonn K., Smit N., Preim B., Vilanova A.: Illustrative multi-volume rendering for PET/CT scans. In *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine* (2015), Eurographics Association, pp. 103–112. 3

[Mun09] Munzner T.: A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE Transactions on 15*, 6 (2009), 921–928. 1

[RDRS10] Ropinski T., Döring C., Rezk-Salama C.: Interactive volumetric lighting simulating scattering and shadowing. In *Visualization Symposium (PacificVis), 2010 IEEE Pacific* (2010), IEEE, pp. 169–176. 2

[SMM12] Sedlmair M., Meyer M., Munzner T.: Design study methodology: Reflections from the trenches and the stacks. *Visualization and Computer Graphics, IEEE Transactions on 18*, 12 (2012), 2431–2440. 2

[SMP03] Sansone C., Morf C. C., Panter A. T.: *The Sage handbook of methods in social psychology*. Sage Publications, 2003. 2

[Yin94] Yin R. K.: *Case Study Research: Design and Methods*. Thousand Oaks, CA: Sage, 1994. 2