Multi-scale Quality Assessment in MRI Research Workflows: Challenges and Opportunities

Steven R. Gomez*

Department of Computer Science
Brown University

Abstract

Verifying the quality of brain imaging is often the first step in research workflows that use this data. Despite the importance of pruning out bad data early in these workflows, little work has been done to characterize standards and methods for controlling quality in brain research workfiows. In this paper, we present an analysis of quality assessment in brain research at two scales – the "image level" and the "domain-semantics level" – based on video-recorded interviews conducted with four domain scientists.

Challenges and opportunities for improvement to these workflows are identified by the analysis. One technique identified – using crowd-sourcing to evaluate some image quality aspects – was evaluated using workers on Amazon's Mechanical Turk and shown to be a reliable resource for some quality control tasks. Finally, we characterize the long-term implications of a multi-scale quality-assessment framework for brain imaging workflows.

1 Introduction

An imaging *artifact* is a feature appearing in brain imaging data that does not actually exist in the brain. Artifacts are dangerous in clinical settings, where they may obscure lesions or other structural or functional properties of the brain tissue. In neuroscience research pipelines, imaging is often used to produce other visual representations derived from this data, such as plots of microstructure properties or tensor fields fitted from diffusion-weighted images (DWIs). Here, artifacts may lead researchers to false insights or conceal them altogether.

The goal of this work is to give researchers a small taxonomy for subtasks or scales in quality assessment (QA) – the *numeric scoring* of quality of an image set with respect to artifacts. Through interviews, we found brain researchers that do QA with informal protocols. The proposed taxonomy enables these scientists to better communicate the quality control process used in generating specific insights, and as we show, helps identify potential improvements to the process.

In this paper, we describe interviews with four brain researchers who use magnetic resonance imaging (MRI) and characterize their activities with respect to QA. From those interviews, we identify two scales of assessment and hypothesize about the human resources needed to execute each. The contributions of this work are two-fold:

1. An analysis of two scales of quality assessment – the "image level" (IL) and the "domain-semantics level" (DSL) – as identified through interviews with four brain researchers. Quality checks at the IL target image properties, like noise or contrast, that can be judged simply based on the presence of visual characteristics without knowledge of the underlying data. The DSL requires knowledge about how the imaging signal should looks at specific anatomical areas. Presumably, while IL QA can be learned by viewing example images, DSL requires an analyst to have more expertise in the area.
2. An evaluation of crowdsourcing for QA that shows workers on Mechanical Turk perform some image-level assessments (e.g., contrast and signal/noise scoring) comparably to trained research assistants.

2 Related Work

While the causes of imaging artifacts in MRI and diffusion weighted imaging (DWI) have been have been studied, little work has been done to understand: 1) how these artifacts are identified and evaluated in real research workflows, and 2) the human factors relevant in QA tasks in these workflows.

2.1 Medical Image Perception

Related research in medical image perception has been driven by challenges in clinical settings. Studies have shown that radiologists fail to detect and recognize lesions at non-trivial rates [3, 7, 8]. In fact, Krupinski *et al.* [6] report that radiologists have an estimated 20-30% diagnostic false negative rate and 2-15% false positive rate. Our analysis focuses on artifact perception, which is less well-studied than lesion perception and detection. Furthermore, we are studying quality perception for imaging data that is fed into later computational processes – where artifacts may be disguised or harder to identify if not caught earlier – to produce more complex visualizations, such as 3D models of brain regions-of-interest (ROIs) or tractography.

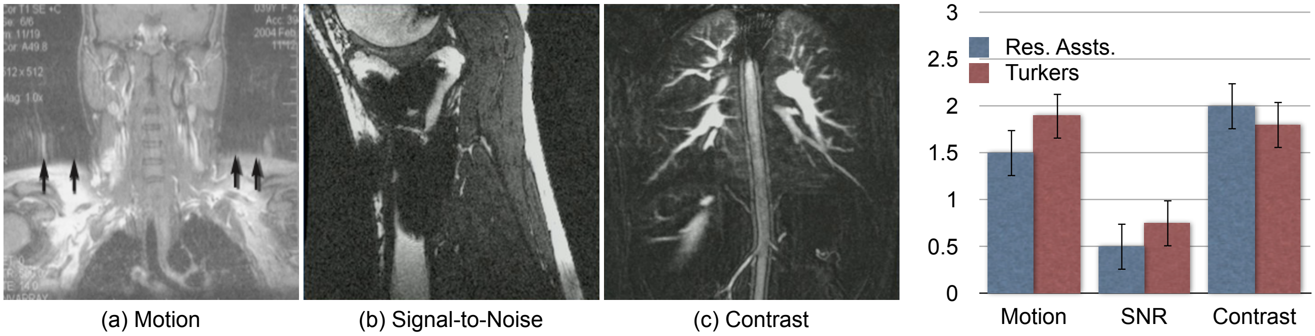
The imaging artifacts explored in the current work are caused by patient factors (e.g., movement during the scan) and by signal collection and processing technologies. The artifacts most dangerous are those that mimic possible pathology or conditions of the brain [4]; in these cases, it may be very difficult or impossible to weed out these scans without additional corroborating evidence. Even features that are unambiguously artifacts, like 'ringing' effects from the subject moving during the scan, may be difficult to identify due to the cognitive and perceptual challenges involved in evaluating them. At the same time, using computer vision to find these artifacts is not a solved problem [2], and in practice, as we found, brain researchers often analyze the imaging data themselves early in their workflows.

2.2 Workflows for Imaging Research

Related work has also examined science workflows that incorporate visual analysis processes. The "data foraging" and analysis steps that led to scientific insight have been characterized by Springmeyer [10], Pirolli [9], and others. At a finer scale, Amar *et al.*[1] gives a taxonomy of analytic "building blocks". We extend this by making data quality assessment an explicit step in the analysis process that is separate from interpreting the meaning of the data.

Furthermore, there has been little work in characterizing the expertise needed to complete certain analytical steps. This is critical as we consider how to leverage new kinds of resources and

*e-mail: steveg@cs.brown.edu



(a) Motion (b) Signal-to-Noise (c) Contrast

Figure 1: [PLACEHOLDER DATA] (a–c) show some *image-level* artifacts that cause characteristics persistent through the image sets. These were evaluated both by research assistants trained to assess quality and by Turkers who were given good and bad example images. Comparative scoring results are shown (d) along with bars indicating the standard error for each task.

infrastructure, like crowd-sourcing, to improve science workflows. Heer *et al.* showed that online workers ("Turkers") on Mechanical Turk can quantitatively evaluate chart information as well as conventional in-person study participants [5]. We extend this by showing that Turkers also score MRIs for image quality comparably to in-person research assistants.

3 Methods

3.1 Interviews with Domain Scientists

R1 is a tenure-track professor of psychiatry and human behavior; R2 and R3 are both doctoral students in computer science studying computational models of brain structure properties; R4 is a professor of psychiatry. Video-recorded interviews lasted 45–60 minutes, and participants were asked to explain how imaging relates to his/her research workflow; how and when quality assessment happens in that workflow; and what challenges in quality control s/he experiences. Because the backgrounds of each participants were different, the interviews were necessarily free-form and contained an opportunity for the participant to demonstrate parts of his/her workflow to the camera.

In the interviews, we found that QA is done at two levels.

3.1.1 Image Level Assessment

- R1 informally uses a 4-point scoring system to judge contrast, noise, and motion artifacts. This is an initial step in the research pipeline that is used to discard MRIs with poor quality. His group views T1-weighted MRIs as 1-2 minute videos of slices. He does not yet do this for DWIs but would if he had the "RA power" to do so.

- R2 and R3 use simulated brain data primary, and do not do IL assessment.

3.1.2 Domain-Semantics Level Assessment

- R1's pipeline involves computing the tensor field of water diffusion in the brain's white matter. After this, he looks at specific areas in the brain with known diffusion direction, like the corpus callosum, and checks the direction. Doing this kind of check requires domain knowledge of the image set.

3.2 Crowdsourcing MRI Quality Assessment

Twenty (20) Turkers were recruited to assess T1-weighted MRI scans in each of 6 test conditions (3 artifact types crossed with 2 quality extremes, "good" and "bad"). Four (4) research assistants who had evaluated similar artifacts previously also each assessed all 6 conditions.

Score for each of 3 categories based on the rubric in R1's lab: motion artifacts, contrast, and noise (SNR).

0	very bad
1	somewhat bad
2	somewhat good
3	very good

Each individual task (a "HIT") consisted of reviewing videos showing the MRI slices from top to bottom, axially. Four example videos are shown; one is perfect (score of 3 in all categories); one has bad (score of 0) motion artifacts; one has bad contrast; and one has bad SNR. Finally, one *unknown* scan is shown, and the subject is asked to score each category.

4 Results

The comparative performance of Turkers and research assistants is shown in Fig. 1. There was no significant difference in scoring between these populations, for all HITs. Among both groups, the SNR HIT had greater variance than the other tasks.

5 Conclusion

In this paper, we analyze multiple scales of QA that happen in real brain-imaging research workflows, as identified through interviews with domain scientists. Implications of this kind of analysis for researchers includes better communication of insight provenance, as well as better matching of human resources, including crowd-sourced labor, to QA tasks that varied skill-sets. We evaluated this latter point and found crowdsourced workers on Amazon Mechanical Turk were able to score image-based QA tasks nearly as consistently (within $x\%$ error, on average) as trained research assistants.

References

[1] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, 2005.

[2] A. W. Anderson and J. C. Gore. Analysis and correction of motion artifacts in diffusion weighted imaging. *Magnetic Resonance in Medicine*, 32(3):379–387, September 1994.

[3] A. E. Burgess. Visual perception studies and observer models in medical images. *Seminars in Nuclear Medicine*, 41(6):419–436, 2011.

[4] L. J. Erasmus, D. Hurter, M. Naude, H. G. Kritzinger, and S. Acho. A short overview of mri artefacts. *SA Journal of Radiology*, pages 13–17, August 2004.

[5] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2010.

[6] E. A. Krupinski. The importance of perception research in medical imaging. *Radiation Med.*, 18(6):329–334, 2000.

[7] H. L. Kundel. History of research in medical image perception. *J Am Coll Radiol.*, 3(6):402–408, 2006.

[8] D. J. Manning, A. Gale, and E. A. Krupinski. Perception research in medical imaging. *British Journal of Radiology*, 78:683–685, 2005.

[9] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive analysis. In *International Conference on Intelligence Analysis*, 2005.

[10] R. R. Springmeyer, M. M. Blattner, and N. L. Max. A characterization of the scientific data analysis process. In *IEEE Visualization*, 1992.