Toward a Common Validation Methodology for Segmentation and Registration Algorithms

Terry S. Yoo [1], Michael J. Ackerman [1], Michael Vannier [2]

[1] National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA
{yoo, ackerman}@nlm.nih.gov
[2] Department of Radiology, College of Medicine, University of Iowa, Iowa City, IA, USA
michael.vannier@uiowa.edu

Abstract. The National Library of Medicine and its partners are sponsoring Insight, a public software toolkit for segmentation and registration of high dimensional medical data. An essential element of this initiative is the development of a validation methodology, a common means of comparing the precision, accuracy, and efficiency of segmentation and registration methods. The goal is to make accessible the data, protocol standards, and support software necessary for a common platform for the whole medical image processing community. This paper outlines the issues and design principles for the test and training data and the supporting software that comprise the proposed Insight Validation Suite. We present the methods for establishing the functional design requirements. We also present a framework for the validation of segmentation and registration software and make some suggestions for validation trials. We conclude with some specific recommendations to improve the infrastructure for validating medical image processing research.

Introduction

The National Library of Medicine (NLM), with its partner institutes: the National Institute of Dental and Craniofacial Research (NIDCR), the National Eye Institute (NEI), the National Institute of Mental Health (NIMH), the National Science Foundation (NSF), the National Institute of Deafness and Other Communication Disorders (NIDCD), and the National Cancer Institute (NCI), are sponsoring a program to develop an application programmer interface (API) and first implementation of a segmentation and registration toolkit called Insight [6]. The goal of this initiative is to create a self-sustaining code development effort to support image analysis research in segmentation, classification, and deformable registration of medical data. The intent is to amplify the investment being made through the Visible Human Project and future programs for medical image analysis by reducing the reinvention of basic algorithms. We are also hoping to empower young researchers and small research laboratories with the kernel of an image analysis system in the public domain.

The Insight Software Research Consortium is a team from academia and industry. The prime contractors are: General Electric Corporate R&D, Kitware, Inc., MathSoft, Inc., the Univ. of North Carolina at Chapel Hill, the Univ. of Pennsylvania (the VAST Lab and Radiology), and the Univ. of Tennessee. Subcontracts have been extended

Published in: S. Delp, A. DiGioia, B. Jaramaz, eds., *Medical Image Computing and Computer-Assisted Intervention* (Proceedings of MICCAI 2000, the 3rd International Conference), Lecture Notes in Computer Science, Volume 1935, (Springer-Verlag: Berlin) 2000, pp. 422-431.

to: Harvard Brigham and Women's Hospital, U. Penn's GRASP Lab, the Univ. of Pittsburgh, and Columbia and the Univ. of Utah.

As part of the software initiative, the Insight Team is creating a software toolkit, and an algorithm validation methodology. These developments have begun to raise and consolidate difficult research questions. The medical imaging community as a whole is beginning to recognize the difficulties in obtaining definitive ground truth with regard to questions of segmentation and registration. How do we validate segmentation and registration methods in the absence of firm control data? How do we generate a model for ground truth? In the absence of ground truth, software researchers often compare the output of one algorithm with that of a known algorithm in the literature. What metrics do we choose as comparison points between output images? What relevance do these metrics have on clinical outcomes?

Background

The Visible Human Project was initially formed to collect data from human subjects to serve as a guidebook and baseline dataset in modern anatomy research and education. Data from two subjects, one male and one female, were collected through a variety of methods including the standard radiological techniques of X-ray CT studies, magnetic resonance imaging, and plain film radiographs. In addition to these conventional clinical studies, the subjects were frozen and sectioned at 1 mm (male subject) and 1/3 mm (female subject) intervals. The exposed surfaces were photographed with 35 and 70 mm film and digitized with an electronic camera. The resulting data has entered into broad use in education and in medical research [5].

In February 1998, a workshop sponsored jointly by NLM and NIDCD explored the growing needs of the research and education community for more powerful digital tools and higher resolution models of human anatomy. Among their many recommendations, the workshop participants recommended the pursuit of advanced image analysis software tools to accommodate future higher resolution data [4]. The demand for more powerful segmentation and registration tools was confirmed through panel discussions and other exchanges during the Second Visible Human Conference, held at NIH in Bethesda, MD in October 1998 [1].

The Insight Team convened its first organizational meeting in November of 1999. At that time, the software consortium opened the discussions on the difficult issues of validating segmentation and registration algorithms. A meeting of the Insight Subcommittee on Validation was scheduled for March 2000 to explore the breadth of these topics and draft an approach for solving them.

Design Principles

The first task before the Subcommittee on Validation was to establish some basic design requirements for the Insight Validation Suite. Since the consortium and its related professionals represent a diverse set of users of medical image data, the task of achieving agreement on the basic issues was not simple. The committee used Quality function deployment (QFD), a management tool to capture the needs and priorities of the Insight user. The QFD process, originally applied in industrial circles, helps

streamline production and decrease time-to-market by targeting specific customer desires [2].

Method: QFD Analysis

The QFD process is comprised of several steps, each driving towards the goal of having a list of specific, prioritized features for a segmentation and registration validation suite. In the first step, Insight developers acting as primary users, answered a series of questions. For example, workshop participants answered questions such as, "What do you like about validation software as it exists today?" and "If cost were not an issue, what capability would you ask for and why would you want it?" The questions are designed to provoke brain-storming and to encourage descriptive, free-form responses. Multiple responses to each question are allowed and, in fact, encouraged. Each separate idea is recorded anonymously on a separate index card.

The next step establishes categories into which the users' needs can be classified. All index cards from the previous step are placed face-up on a table, and the participants are asked to help sort the cards into categories with common attributes. For example, cards stating "no technical support necessary" and "complicated to use" could both be grouped together because they both address the ease-of-use of a system. Any participant can group, ungroup, or regroup cards as he or she sees fit, and all grouping is done with minimal discussion. Grouping continues until a small, pre-determined number of categories (or card groupings) is established.

After the groupings are established, they must be named. In this step, it is important to let the users define the name for each category. This allows the users' linguistics to enter the design process from the very beginning. For each stack of index cards grouped together in the previous step, all cards in that stack are read aloud. The participants discuss the group's common characteristics and come to a consensus on what the category name should be. During this process, groupings may be combined or divided as necessary to capture all main ideas presented in the index cards. The category name of each grouping is recorded. In the committee meeting, a reduced set of thirteen groupings defined by the sorting process resulted in eighteen unique categories after the naming process was completed.

Finally, a voting process is used to prioritize the categories. Each participant is given a certain number of ratings to distribute among the categories. The ratings vary in value from nine (the best rating) down to one, and each rating can be given to a fixed number of categories. The votes are tallied, and the categories are ranked from highest to lowest priority according to points received.

Results: Affinity Diagram

The results of the QFD Analysis from the meeting of the Insight Subcommittee on Validation can be seen in Figure 1. The category names were determined by workshop participants and are listed on the left. The right-most column reflects the total score as determined by the voting process.

The principle user of the Insight Toolkit and its validation suite is expected to be the software tool designer. Insight is an application programmer's interface (API), and it will support access and hooks for user interface programming; however, it will

	Category	Score
1 .	Software Issues	144
2 .	Consensus Acceptibility	123
3 .	Statistical Foundation	120
4 .	Ground Truth	110
5 .	Quantitative evaluation	107
6 .	Robustness	101
7 .	Extensible Databases / data quality	68
8 .	Registration	65
9 .	Automation	61
10 .	Efficiency	57
11 .	Application	43
12 .	Multimodality	36
13 .	Resolution	26

Figure 1. Results from QFD Voting process.

not contain tools for user interfaces nor visualization and graphics software. Insight is expected to provide mathematical and statistical rigor to advance medical applications through a compact, portable toolkit. In this context, issues of portability, software extensions, robustness, and other Software Issues are expected to top the list of issues. As a public resource Consensus Acceptability is also expected as a top priority. However, our committee does not explicitly have a professional statistician, yet questions of Statistical Foundations, Ground Truth, and Quantitative evaluation are among the five most important issues reported through the QFD analysis.

While not overly surprising, these results have indicated possible gaps in the expertise of the Insight Team. Moreover, the current initiative does not support a strong statistical component. Recommendations for bridging these gaps are made in a later section.

Experimental Design of Validation Trials

The design of blind evaluation for computer algorithm validation is an area in need of more study. The committee considered the work of Yu as one possible model [7].

Figure 2. shows the basic structure of a blind trial designed by Yu and his team. Separate evaluators were selected, and no evaluator knew whether the output decisions were made by a medical student, by any of eight physicians (including the patient's actual diagnosing physician), or a computer system. Figure 3 shows a modified view of a proposed automated structure for validating segmentation and registration algorithms. The costs of truth generation of the blind evaluation are often prohibitive to the average algorithm designer. Insight is attempting to provide these elements as part of its work.

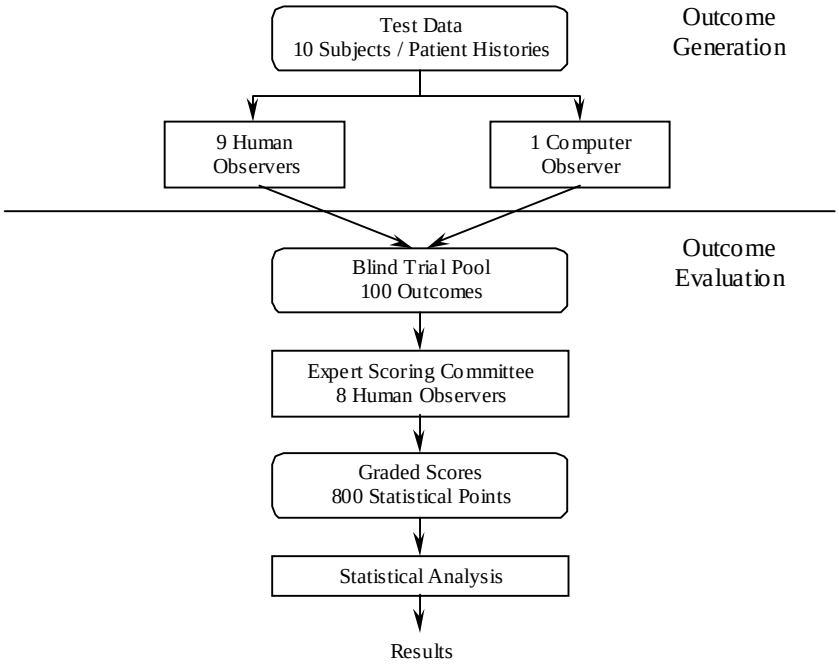

Figure 2. Simplified clinical trial model for evaluating computer performance.

Design elements

Some essential design elements that are common to both structures include *truth generation*, the concept of *blind evaluation or scoring*, the use of *multiple observers* to capture the variation in the human decision process, and the analysis of the output through statistical analysis.

Some of the differences include the absence of multiple input datasets in the Insight model arising from a lack of sufficient data. This poverty of input information may introduce significant bias into the test structure.

The output of the outcome generation is handled differently. In the human trials case, a blind evaluation committee is used. In the Insight model, the blind committee is replaced with an automated scoring system. The input is also modified slightly. While truth is generated from collected human data in both cases, the Insight model attempts to pre-collect truth data for re-use in multiple trials. This must be done carefully, separating test data from training data to keep from affecting the experiment. The output of the automated scoring system will be differences in "figures of merit," a term coined by Udupa. These elements suggest that the automated scoring system itself must be validated and shown to be free of experimental bias. The figures of merit should also be demonstrated to be clinically useful indicators of outcome.

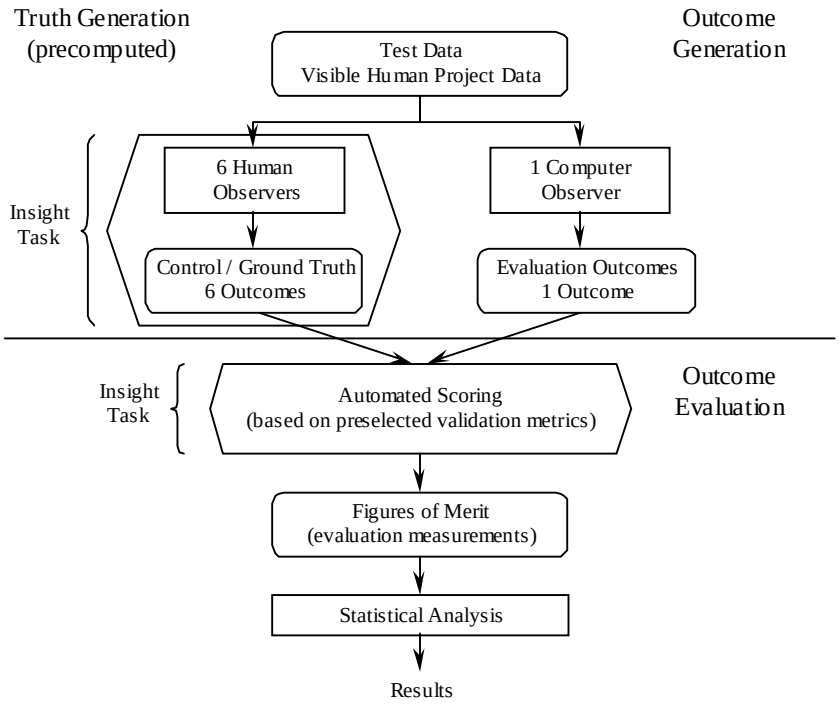

Figure 3. Modified Insight clinical trial model for evaluating segmentation and registration algorithms.

Validation Requirements

The committee iterated over several issues that are considered requirements for a validation suite. They include the removal of experimental bias (including human bias and software training bias). The blind evaluation of the outcomes is also considered essential. The separation of training data from testing data (including training truth and testing truth) is required. Many methods require a large body of training data to begin to make associations, and these data must not be shared with the test data. In addition, one model for software evolution is the tweaking of software parameters. To keep from customizing a method to a particular dataset or group of datasets, a large body of data should be available both for training and testing. Also, the periodic renewal of the testing suite to keep from incremental bias encroaching from repeated testing.

The statistical analysis of segmentation algorithms should study the precision of the method (how well it corresponds with ground truth), the accuracy of the method (how consistent it is), and the efficiency of the method (including the amount of expert user interaction it requires). The committee listed the difference in human vs. computer tasks as an essential element of study. Any validation study should recognize computers are much better at quantitative measurements of images than humans. These effects are of great interest in any validation study undertaken.

Researchers working in the field of segmentation and registration should consider these issues carefully when performing experiments.

Committee Policy Recommendations

The emerging results of this study are beginning to provide a principled infrastructure upon which we can build a validation suite for the evaluation of segmentation and registration software tools. The goal is to archive and make accessible the data, standard protocols, and support software necessary to serve as a foundation for a common platform for the whole medical image processing community.

However, all of the building blocks for such a foundation do not presently exist. There are still significant scientific concerns that must be addressed. To contend with these issues, we are making the following recommendations to the scientists of the medical image processing research community and the programs that support them.

Recommendation: A broad mission for data collection

Common data, accessible across the community, fosters collaborative research and comparative analysis across research groups. The Visible Human Project data is a clear example of how data can create a rendezvous for scientific research. However, the current Visible Human Project data itself is not sufficient to support a common validation effort in segmentation and registration research. The Insight Subcommittee on Validation recommends that the National Institutes of Health and other related agencies embark on a mission to collect comparable data from a wide cross-section of the population as an ongoing effort to support medical imaging into the next century.

Multiple datasets: As a single data point, the Visible Human Project data cannot begin to address the issues of human variation. More importantly, there are dangers in using a limited collection of data to serve for both the training as well as the testing of algorithms and software systems. Significant biases will be introduced into the research and development process; new work will grow to be based on flawed science. The existing data can be used in a limited fashion. For instance, segmentation software can be trained on various portions of the anatomy and tested on others. However, many applications are anatomy specific, and much of the anatomy of humans has only one or two examples per subject.

Multiple modalities: Any data collection initiative should emphasize multiple datasets and multiple modalities. The growth of imaging modalities and the growth of computing in medicine has created an increasing need for the fusion of multiple sets of volume data. The intent is to bring as much information together as possible through advanced computing to create greater insight and better illuminate the human condition.

Clinical Data: The subcommittee recommends that any data collection initiative consider clinical data sources as primary means of obtaining medical information. The existing Visible Human Project data is limited to post mortem images. These data do not accurately reflect the types of images that pervade the field of medical imaging. Moreover, the Visible Human Project as a study in macro anatomy of embalmed cadavers is in some risk of distancing itself from mainstream medicine. New imaging techniques in functional imaging are providing powerful vehicles for studying human physiology at a cellular level.

Registration datasets: Data supporting the validation of registration algorithms should be collected; it may differ from the type of data needed for segmentation experiments. Fiducial information and landmarks play an important role in registration techniques and should be included in any data collection effort.

Continuing effort: The subcommittee recommends that the mission in medical image data collection be framed as a continuing effort. Imaging technology evolves constantly, improving its ability to resolve fine structures as well as visualize previously inaccessible information. Magnetic resonance functional imaging and ultrasound doppler imaging are two examples of new types of information rather than just improvements in existing modalities. Legacy images stored in databases will lose their relevance rapidly due to the pace of change in imaging systems.

The issues of Consensus Acceptability and Extensible Databases were prominent among the elements of the QFD affinity diagram. All data archived in the recommended effort should be publicly available, indexed via a database system, capable of supporting the anticipated growth suggested by the need for a continuing effort in data collection.

Recommendation: A program in statistics

The Insight Subcommittee on Validation recommends that the National Institutes of Health and related funding agencies undertake an initiative in the support of statistical research in validation science. Several issues surrounding experimental design, the control of human and software development bias, and the quantification of ground truth repeatedly appeared in the planning process. The need for a strong statistical foundation was one of the highest priorities discovered in the QFD analysis.

Questionable truth: A comprehensive and consistent treatment of comparing multiple random variables in the absence of imperfect truth would be an important research resource for the imaging community. The intractable issues of Ground Truth were a high priority in the committee's QFD analysis. Unlike many other scientific research areas, most questions of segmentation and registration can only be compared with the performance of skilled human observers. Ground truth assembled from the responses of multiple people, will be subject to the random variations of human decisions. Published methods for receiver operating characteristic (ROC) analysis for segmentation and registration and other common statistical means of managing such data should be sought.

Limiting or understanding the free imaging variables: The committee recommends the study of imaging protocols to quantify and normalize the variation among imaging systems. No two scans taken of the same patient correspond absolutely. Variation among manufacturers, among technique used at different institutions, and even arising from the relative age of different medical scanners yield differences that are difficult to characterize. Without better understanding of these differences, comparing two output segmentations may be subject to independent variables that are not relevant to the clinical questions being asked. The committee suggests further study in how to normalize the differences among scanners, placing some emphasis on common radiology and pathology technique for data acquisition.

Understanding and reducing experimental bias: The committee recommends an exploration of the sources of bias in validation studies. Much of the significance of a scientific study can be lost if overwhelmed by flawed experimental design. In particular, many segmentation techniques such as Bayesian analysis, genetic programming, and neural network processing require that representative training data be provided to create the statistical framework necessary for their computations. It is essential to separate the test data from the training data, otherwise the tests will be biased and the technique being studied will falsely report superior performance. Additionally, the test data itself must periodically be renewed, otherwise the effects bias collected over time through repeated testing with the same data will indirectly adversely influence the development of segmentation and registration methods. Human bias must also be considered since the human visual system will be used to generate much of the ground truth used in performance experiments. There is much uncharted scientific study in how to create good experiments in imaging software validation.

Figures of merit – evaluation and grading vs. clinical relevance: The committee has also recommended that a study of validation metrics be undertaken with regard to their clinical relevance. Specifically, we recognize that different metrics (e.g., minimum cross sectional area, total surface area, or total volume to name a few) may have different levels of importance depending on the presented pathology, the procedure under consideration, the age of the subject, or any of a number of variables. Characterization of the validation metrics in how well the reflect clinical conditions is necessary to enable useful predictions and directions for future research. What is desired is a measurement that is easy to quantify in a digital imaging context that reflects clinical outcome. Such metrics can then become valid measurements of performance for our segmentation and registration systems.

Conclusions and Future Work

The Insight Software Consortium Subcommittee on Validation is assembling the foundations of a public resource for validating algorithms for segmentation and validation of medical image data. A requirements analysis process was used to generate some common ground among the variety of interests represented among the consortium of anatomists, computer scientists, surgeons, radiologists, psychologists and other related professionals. A basic framework has been proposed as a model for discussing and refining the design of validation experiments and the infrastructure necessary to support them. Based on its findings, the committee is recommending to the Visible Human Project software sponsors and other NIH programs the creation of a long term data collection initiative. In addition, the committee recommends a related program to establish a strong statistical foundation for clinical trials and validation studies in segmentation and registration research.

The QFD Voting process forced the participants to work together to define and name categories of importance from a medical point-of-view, ensuring that the focus of the analysis was correctly targeted to the end-users. Data for a separate analysis based on the Kano method [3] was collected during the same workshop; that data is under study and the results will be available at MICCAI 2000.

Acknowledgements

This work would also not be possible without the support and dedication of the sponsoring agencies and the program officers whose vision launched and continue to sustain this initiative. This distinguished group includes Dr.'s Lindeberg,. Slavkin,. Oberdorfer, Jacquet, Griffin, Clarke, Croft, Gulya, and Huerta. We would like to thank the Insight Team for their continuing participation in this development effort. Many of the ideas published here reflect the collective thinking of the group rather than the particular expertise of the authors. In addition, we would especially like to thank Dr.'s Molholt and Imielinska and Columbia University for organizing and hosting the Insight Software Consortium Subcommittee Meeting on Validation. We'd also like to thank Hillary Schmidt for both her participation and her particular contribution to the validation model presented in this paper.

References

1. R. A. Banvard and P. Cerveri, eds., 1998, Proceedings of the Second Visible Human Project Conference. October 1-2, 1998, Bethesda, MD: US Dept. of Health and Human Services, Public Health Service, NIH.
2. J. Hauser and D. Clausing, 1988, The House of Quality, Harvard Business Review, May-June 1988, pp. 63-73.
3. N. Kano, 1993, A Perspective on Quality Activities in American Firms, California Management Review, 35/3 (Spring 1993): 12-31.
4. NIDCR, 1998, Summary Report: Virtual Head and Neck Anatomy Workshop. February 25, 1998, Bethesda, MD: US Dept. of Health and Human Services, Public Health Service, NIH. (http://www.nidcr.nih.gov/news/strat%2Dplan/headneck/contents.htm).
5. V. Spitzer, M. J. Ackerman, A. L. Scherzinger, and D. Whitlock, 1996, The Visible Human Male: A Technical Report, J. of the Am. Medical Informatics Assoc., 3(2) 118-130.
6. T. S. Yoo and M. J. Ackerman, 2000, A New Program in Medical Image Data Processing, *Medicine Meets Virtual Reality 2000* (Proceedings of the 8th Annual Medicine Meets Virtual Reality Conference), J. Westwood, *et al*. eds., IOS Press, Amsterdam: pp. 385-391.
7. V. L. Yu, *et al*.. 1979, Antimicrobial Selection by a Computer, J. of the Am. Medical Assoc., 242(12) 1279-1282..