Issues in Multimodality Medical Image Registration

Murray H. Loew

TWO 1-day workshops are being held in 1997 to convene experts able to address several aspects of the issues surrounding the question of multimodality medical image registration. The goals of the first workshop are to produce an agenda that defines the significant problems, and to assign them to participants for a 4-month study. The second workshop is to reconvene the group to report their conclusions and to assemble a final report for widespread distribution.

BACKGROUND

Multimodality medical imaging is finding increasing use in diagnosis, treatment planning, and the conduct of treatment, as physicians combine information about form and function to sharpen their understanding of location and pathology in clinical situations. Registration of images is a prerequisite to the effective use of several sources of image data because anatomic landmarks that are important to physicians may be absent in some of the modalities. Often, the images are of different sizes, resolutions (in space, time, and/or intensity), and orientations. Nonrigid deformations occur frequently, further complicating the problem. In recent years a variety of registration algorithms have been developed and used by researchers for their own purposes; in some cases those algorithms were used by others (perhaps after modification), possibly for different purposes. New applicable techniques are being developed as well in nonmedical areas including defense and nondestructive testing.

Developers and users of registration methods have measured the accuracies and precisions of their methods in many ways. The multiplicity of anatomic sites, imaging parameters, sample sizes, computation strategies, and intended uses of the

registered images have made it nearly impossible to compare the methods quantitatively, or even to characterize them. Thus, it is difficult to provide a principled basis for the selection of a registration method for use in a given application.

Within the past several years, a group of laboratories from several countries participated in an experiment to evaluate the accuracies of a variety of registration methods. Images were distributed and the participants asked to register them and then to return to the organizers a tabulation of the transformations necessary to register the images. The results indicated that some combinations of laboratory and algorithm worked well; others did not. The experiment raised many questions about the kinds of data sets, the evaluation methods, reporting formats, policies regarding errors in reporting (as opposed to computation), and the "boundaries" of algorithms.

Magnetic resonance imaging (MRI) offers an example of that last issue. Because most MRIs suffer from some distortion (due to coil nonlinearities and earth's magnetic field inhomogeneities), a question could be raised as to whether correction of the distortion should be performed at the time of image generation. If not, then should a laboratory's ability to perform such corrections affect the evaluation of its registration algorithm?

Segmentation is an important step in most registration methods, but it does not always identify boundaries (eg, of skin) correctly. Many approaches, therefore, include an interactive (manual) step to allow the user to edit the boundaries or other fiducial loci detected by the algorithm. Is it important to measure the amount of manipulation performed on an image, and to incorporate that into a measure of overall algorithm performance?

Other issues include the following:

Intermodality versus intramodality—are there differences that must be accounted for explicitly?

How do you compare techniques that require fiducials to those that do not?

How do you define "modalities"?—Do we, eg, distinguish between the various MRI pulse sequences?

How do you consider rigid-body as compared to soft-tissue registration?

From the Institute for Medical Imaging and Image Analysis, Department of Electrical Engineering and Computer Science, George Washington University, Washington, DC.

Supported by the Ballistic Missile Defense Organization and the National Technology Transfer Center.

Address reprint requests to Murray H. Loew, Institute for Medical Imaging and Image Analysis, Department of Electrical Engineering and Computer Science, George Washington University, Washington, DC 20052.

*Copyright © 1997 by W.B. Saunders Company
0897-1889/97/1003-1012\$3.00/0*

How do you compare fairly those metrics that use regions of interest with those that use the entire image?

How do you incorporate the effects of various resampling methods for scale correction?

Should there be standard sets of data for testing algorithms? If so, how do you define them?

What are the clinical needs? How can they be included, to help determine, eg, the maximum acceptable error? What should be the role, if any, of observer studies in evaluating registration techniques?

What are the regulatory issues?

What are the effects of, and possible impacts on, the DICOM standard?

Is the registration process automatic or interactive? Is it important to measure the amount and kind of interactivity?

What measure(s) of error are most useful?

How should algorithms be described and/or distributed?

THE WORKSHOPS

The above examples of issues illustrate the need for a structured discussion aimed at generating recommendations for evaluation protocols. It is likely that there will be several, to account for unavoidable variations in implementation.

The first workshop assembled a group of 35

experts in the field, from nine countries. The participants came from the basic, applied, and computational sciences, and included clinicians as well. They identified the central issues, and chose working subgroups to address them. The questions, some of which are listed above, were divided into three areas: clinical uses of registration, metrics for evaluation registration methods, and data sets for evaluation. The goals of the working subgroups are to define the characteristics of, and criteria for evaluating, the metrics and proposed data sets.

The second workshop will receive the recommendations of the subgroups, and issue a set of summary statements indicating the degree of unanimity reached for each recommendation. The subgroups will not produce metrics or data sets themselves; rather, it is hoped that the June consensus will encourage others to develop new, or re-evaluate old metrics and data sets in light of the recommendations.

Detailed presentations of the June results will be submitted to several journals for dissemination to those working in this area. It is hoped that constructive criticism from a large audience will lead to a set of procedures that will result in algorithms, data, and results that can be uniformly used, understood, and compared by all who are interested. There should be substantial benefits both in the laboratory and in the clinic.