Evaluating Image Registration Using NIREP

Joo Hyun Song, Gary E. Christensen, Jeffrey A. Hawley,
Ying Wei, and Jon G. Kuhl

Electrical and Computer Engineering,
and Iowa Institute for Biomedical Imaging The University of Iowa, Iowa City, IA
52246, USA
{joohyun-song,gary-christensen,jeffrey-hawley,
ying-wei,jon-kuhl}@uiowa.edu
http://www.nirep.org/

Abstract. This paper describes the functionality and use of the Non-rigid Image Registration Evaluation Program (NIREP) that was developed to make qualitative and quantitative performance comparisons between one or more image registration algorithms. Registration performance is evaluated using common evaluation databases. An evaluation database consists of groups of registered medical images (e.g., one or more MRI modalities, CT, etc.) and annotations (e.g., segmentations, landmarks, contours, etc.) identified by their common image coordinate system. Prior to analysis with NIREP, each algorithm is used to generate pair-wise correspondence maps/transformations between image coordinate systems. NIREP has a highly customizable graphical user interface for displaying images, transformations, segmentations, overlays, differences between images, and differences between transformations. Evaluation statistics built into NIREP are used to compute quantitative algorithm performance reports that include region of interest overlap, intensity variance of images mapped to a reference coordinate system, inverse consistency error and transitivity error.

Keywords: NIREP, evaluation, non-rigid image registration, transformation, medical imaging.

1 Introduction

Image registration is important for many medical image applications including longitudinal evaluations within the same individual, comparison across individuals, creation of population atlases, computer aided diagnosis, computer aided treatment, evaluation of outcomes and many others. Unfortunately, evaluating non-rigid image registration algorithm performance is difficult since there is rarely if ever ground truth correspondence to judge the performance.

The Non-rigid Image Registration Evaluation Project (NIREP) was established to develop software tools and provide shared image validation databases for rigorous testing of non-rigid image registration algorithms. This paper reports on progress developing the Non-rigid Image Registration Evaluation Pro-

B. Fischer, B. Dawant, and C. Lorenz (Eds.): WBIR 2010, LNCS 6204, pp. 140–150, 2010.

gram (NIREP) for evaluating registration accuracy of nonrigid image registration algorithms.[1]

Under the NIREP model, users process data on their own and evaluate the performance of different nonrigid registration algorithms using evaluation criteria that are built into NIREP. The data can be the user's own data or data downloaded from the central database repository on the NIREP website. This model has the advantage of standardizing the evaluation metrics and distributing the processing load. It provides researchers with a tool to compare the performance of multiple registration algorithms on their own data so they can make an informed decision regarding the best algorithm for their specific application. It also provides researchers with a tool to validate their research results.

NIREP uses a diverse set of evaluation metrics to evaluate registration performance on well documented evaluation image databases. These tests evaluate the performance of image registration algorithms with respect to their transformation properties, agreement with human experts, and other indirect performance tests.

2 Methods

2.1 Evaluation Database

A critical step in making unbiased comparisons of algorithm performance is to evaluate registration algorithms on the same population of images. NIREP assumes that the registration algorithms to be analyzed have been used to register images contained in a common evaluation database. An evaluation database consists of groups of registered images and annotations identified by their common image coordinate system.

An evaluation database consists of a set of images to be registered and associated data for accessing the registration results. An example of an evaluation database would be a set of N 3D MRI images of the brain and N expertly labeled segmentations–i.e., one segmentation per 3D image volume. The registration algorithms are used to register the MRI data and the segmentations are used assess registration performance. The segmentations could be used to assess performance by examining the overlap of a deformed source segmentation with the segmentation of the target. In this example, each MRI image and its associated segmentation image constitutes one entry in the evaluation database.

In general, evaluation database entries consist of multiple image modalities (e.g., one or more MRI modalities collected in register, CT image, etc.) and many different types of data for assessing registration performance (e.g., expertly labeled segmentations, landmarks, contours, surfaces, etc.). All the data associated with one entry in the evaluation database are indexed by their common coordinate system. Indexing database entries by coordinate systems provides a

[1] In this paper, the acronym NIREP is used to refer to both the evaluation project and to the evaluation software program. It is our hope that it is clear what NIREP means from the context that it is used. Sometimes, both meanings are appropriate.

vocabulary for describing transformations between coordinate systems. For example, we say that the transformation $h_{i,j}$ is used to transform a segmentation from coordinate system i to coordinate system j.

2.2 Evaluation Statistics

Evaluation statistics are criteria that quantify image registration performance based on particular features of the evaluation data. Examples of evaluation statistics include image intensity difference, landmark distance error, and overlap error. NIREP currently implements four evaluation statistics: 1) relative overlap, 2) inverse consistency error, 3) transitivity error, and 4) intensity variance. The transformations required to compute some of these statistics are performed by NIREP, using the interpolation scheme designated by the user.

Relative Overlap. Relative overlap (RO) assesses how well two equally likely segmentations of the same region of interest (ROI) agree or disagree with each other. For an image pair S and T, relative overlap is defined as

$$RO_i(S_i, T_i) = \frac{|S_i \cap T_i|}{|S_i \cup T_i|} \tag{1}$$

where $|S_i \cap T_i|$ is the volume of voxels that intersect between the i^{th} region of interest of images S and T.

Inverse Consistency Error. Inverse consistency error (ICE) evaluates registration performance based on desired transformation properties [1,2,3,4]. It is a common assumption in image registration that the correspondence mapping between two anatomical images is unique–i.e., each point in the source image S is mapped to its corresponding point in the target image T and vice versa. However, in practice, the forward mapping from S to T and the reverse mapping from T to S are not necessarily inverses of each other for most image registration algorithms. This inconsistency reflects an algorithm's inability to uniquely describe the correspondence between two images [5]. Inverse consistency is defined as the mapping of point x in S to a point in T and subsequently being mapped back to the original point. Then inverse consistency error is defined as the distance between the original point x and its transformed point x', which can be expressed in two different ways as

$$ICE1_j(x) = \|h_{ji}(h_{ij}(x)) - x\|^2 \tag{2}$$

or

$$ICE2_j(x) = \left\|h_{ij}(x) - h_{ji}^{-1}(x)\right\|^2 \tag{3}$$

where h_{ij} is the forward transformation from image S to T, h_{ji} the reverse transformation from image T to S, and $\|\cdot\|$ the standard Euclidean norm. Note that the transformations are defined in the Eulerian coordinate system–i.e., defined with respect to the target frame of reference. Equation 3 gives another interpretation of inverse consistency which is computed using the inverse of the reverse transformation. Note that $ICE1_j(x)$ and $ICE2_j(x)$ show the inverse consistency error with respect to the coordinate system of image T. While inverse consistency does not measure the accuracy of the transformation, it measures the consistency of the correspondence defined by forward and reverse transformations between two coordinate systems [1]. It is important to note that zero inverse consistency does not imply accuracy of the mapping. For example, an identity mapping between two images have perfect inverse consistency, but the correspondence is inaccurate for non-identical images.

Transitivity Error. Transitivity error (TE) evaluates how well the registration transformation satisfies the transitivity property [2,6]. Similarly to inverse consistency error, transitivity error is a measure of consistency of the correspondence defined by compositions of transformations. More specifically, transitivity error measures the distance of point x in image A to its mapped point in B, which is then subsequently mapped to image C, and finally back to point x' in image A. Another interpretation of transitivity is the difference of the composition of transformation AB with BC to transformation AC. These two definitions are expressed as follows:

$$TE2_k(x) = \|h_{ki}(h_{ij}(h_{jk}(x))) - x\|^2 \tag{4}$$

and

$$TE2_k(x) = \|(h_{ij}(h_{jk}(x)) - h_{ik}(x)\|^2 \tag{5}$$

where h_{ij} is the transformation from image A to B, h_{jk} the transformation from image B to C, h_{ki} the transformation from image C to A, h_{ik} the transformation from image A to C, and $\|\cdot\|$ the standard Euclidean norm. Similar to that of ICE, $TE1_k(x)$ and $TE2_k(x)$ show the transitivity error with respect to the coordinate system of image C.

Intensity Variance. Intensity variance is a measure of similarity between a population of images based on voxel intensity difference. In image registration applications driven by voxel intensity features, the ideal registration should result in zero voxel intensity difference between the registered images. Intensity variance is a population study based on this characteristic, where the voxel-wise intensity variance (IV) of a population of M images registered to image j is computed as:

$$IV_j(x) = \frac{1}{M-1}\sum_{i=1}^{M}(T_i(h_{ij}(x)) - ave_j(x))^2 \tag{6}$$

where

$$ave_j(x) = \frac{1}{M}\sum_{i=1}^{M} T_i(h_{ij}(x)) \tag{7}$$

and T_i is the i^{th} image of the population, and $h_{ij}(x)$ is the transformation from image i to j in a Eulerian coordinate system.

2.3 Non-rigid Image Registration Evaluation Program (NIREP)

The NIREP image registration evaluation software integrates the evaluation statistics defined previously. The built-in evaluation statistics allow users to evaluate registration performance themselves, without having to make submissions to external evaluators. This allows users to have immediate access to registration evaluation results and to use evaluation feedback to tune their algorithms and improve performance.

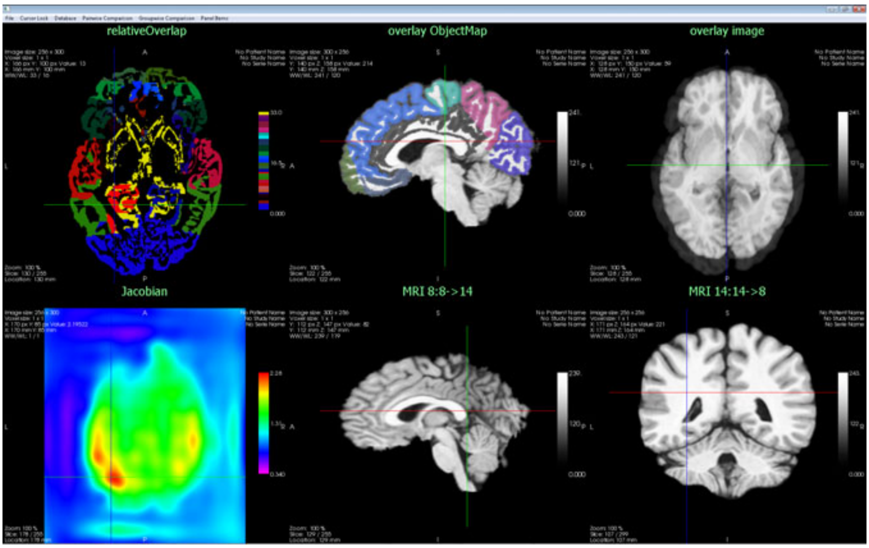

Fig. 1. A typical view of NIREP display showing a 2×3 grid of widget panels. The display widgets support various color schemes and transverse, sagittal and coronal planar views. Cursors of each panel can be locked with cursors of other panels, allowing point-to-point comparisons of data in multiple panels. The top-left panel shows the overlay of object maps from the source and the target coordinate systems. The top-center panel shows the object map of the source overlaid on top of the source image. The top-right panel shows the overlay of the source image on top of the target image. The bottom-left panel shows the Jacobian of the forward transformation. The bottom-center and bottom-right panels show the forward and reverse deformed images, respectively.

The primary user interface of NIREP is a display organized as a rectangular grid of panels, as illustrated in Figure 1. The dimensions (number of rows, columns) of the display are user configurable. Each panel can display different visual or textual information (images, evaluation metrics, etc.) and the contents of each can be controlled independently or locked together. The NIREP can display images, difference images, checkerboard and wipe images, several varieties of overlays, and textual information. Examples of evaluation statistics that provide both visual and quantitative textual analysis include the relative overlap, intensity variance, inverse consistency error, and transitivity error statistics.

Characteristics of multiple panels can be locked together so that they all change together. This, for instance, allows a user to set up an evaluation or algorithm comparison scenario, displaying desired evaluation/comparison statistics in visual or textual form. Once such a scenario has been set up, the user can step through different data sets with the contents of all panels automatically updated to reflect each new data set.

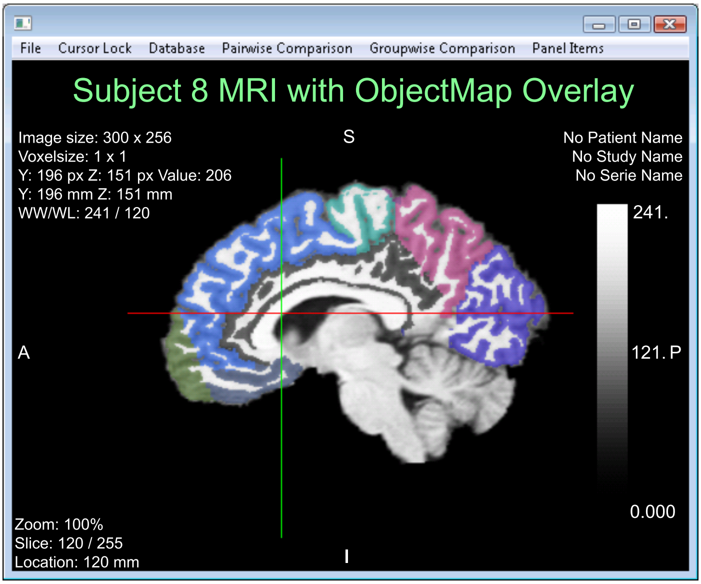

Fig. 2. A NIREP display showing an object map overlay on top of its corresponding MRI data. This display panel widget illustrates the typical information displayed in each panel. The image dimension, voxel size, and voxel intensity at the current cursor location are shown in the top-left corner. The image zoom, slice number, and slice location are shown in the bottom-left corner. The widget popup menu allows users to change widgets, data displayed in the widget, color schemes, the level/window, the zoom factor, and edit titles. The color bar can be positioned anywhere in the panel.

The NIREP software is divided into three main components: the Data Manager, the Evaluator, and the Display Manager. All of these components are managed by their respective configuration files, which will be described below.

Display Manager. The Display Manager is responsible for controlling the content displayed in each panel of the display. Each panel is controlled by a "display widget" which is tailored to the specific type of content to be displayed in that panel. The display panel widgets were adapted from the vtkINRIA3D library developed by INRIA, France (`http://www-sop.inria.fr/asclepios/software/vtkINRIA3D/`). A human readable "Display Description" specifies the layout and content of the display panels. This description specifies the row/column dimensions of the display and contains a "Widget List" that specifies the type of display widget associated with each panel. The Display Description also contains an "Evaluator List" that describes the specific data that needs to be supplied to the widgets. This may include base and precomputed data (e.g., transformations) from the NIREP image database, as well as results computed on-the-fly by NIREP (e.g., Jacobians). The Display Description is parameterized so that users can switch multiple Evaluator operations and display panels with a single variable change. During start-up, the NIREP software reads the initial Display Description from a file. Fig. 1 shows a typical NIREP display with a 2×3 panel configuration.

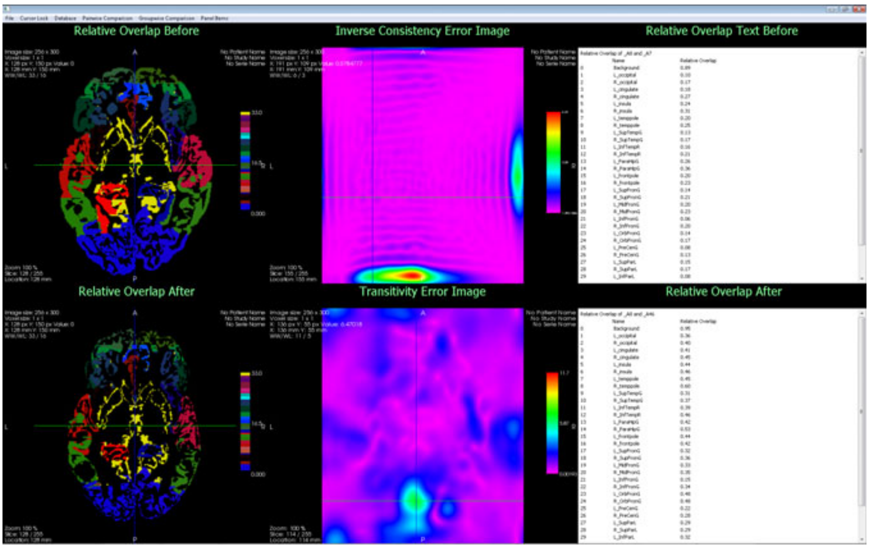

Fig. 3. The display panel can also display text information as shown above. The two right text panels are displaying the relative overlap values for each regions of interest, before and after registration.

Evaluator. The Evaluator is the central processing unit of the NIREP software where all requests and data processing are handled. The display widgets in the Display Manager contact the Evaluator to obtain the necessary data for their respective display functions. The Evaluator is responsible for obtaining data needed by display widgets from the Data Manager and computing new results. The types of data that the Evaluator provides to various display widgets includes images, object maps (annotated segmentation masks), landmarks, contours, surfaces, text tables, and graphs. The Evaluator is designed such that any new data operation or evaluation statistic may be added as a module. The Evaluator uses the Evaluator List, described previously, to determine what data is needed by display widgets. The advantage of a human-readable Evaluator List is that the user can pre-configure desired Evaluator operations by hand before running the NIREP software. This mechanism allows scripting of a large number of operations so that the job can be run in batch mode. An example Display Description is shown in Fig. 4. This description generates the display that visualizes the inverse consistency error for the DEMONS and SICLE registration algorithms.

```
columnSize=2
rowSize=1
Begin WidgetList
  W1,1 = view(iceSicle8-14,ICE 8-14 SICLE)
  W1,2 = view(iceDemons8-14,ICE 8-14 Demons)
End WidgetList
Begin EvaluatorList
  demons8-14    = Transformation(008,014,Demons)
  demons14-8    = Transformation(014,008,Demons)
  iceDemons8-14 = inverseConsistencyErrorImage(demons8-14,
                                               demons14-8,comp)
  sicle8-14     = Transformation(008,014,SICLE_param2)
  sicle14-8     = Transformation(014,008,SICLE_param2)
  iceSicle8-14  = inverseConsistencyErrorImage(sicle8-14,
                                               sicle14-8,comp)
End EvaluatorList
```

Fig. 4. Display Description to generate and visualize the inverse consistency error for the DEMONS and SICLE registration algorithms

Data Manager. The Data Manager manages the loading and storing of data from/to the specified evaluation and algorithm database(s). The Data Manager is responsible for intelligently managing memory by removing data from emery when it is no longer needed by the Evaluator or Display Manager. The Data Manager handles all images supported by ITK (`http://www.itk.org`) and the Analyze 7.5 (Mayo Clinic, Rochester, MN) format. During start-up, the NIREP software reads an evaluation database resource file, algorithm resource file(s), and optionally, persistent data (pre-computed evaluation data saved to disk) resource file, which contain an exhaustive list of data available for evaluation. The Data Manager provides data to the Evaluator and all data generated by the Evaluator is managed by the Data Manager. This memory management schema stores computed results so they do not need to be recomputed if needed in the future.

3 Results

To demonstrate the NIREP software in evaluating image registration performance, an experiment was performed to compare registration results of Thirion

Demons [7], [8] and SICLE [4], [1], [2], [3] algorithms using the NA0 database. The NA0 database consists of a population of 16 annotated 3D MRI volumes corresponding to eight normal adult males and eight females acquired in the Human Neuroanatomy and Neuroimaging (HNN) Laboratory, The University of Iowa, and each data was segmented into 32 gray matter regions of interests.

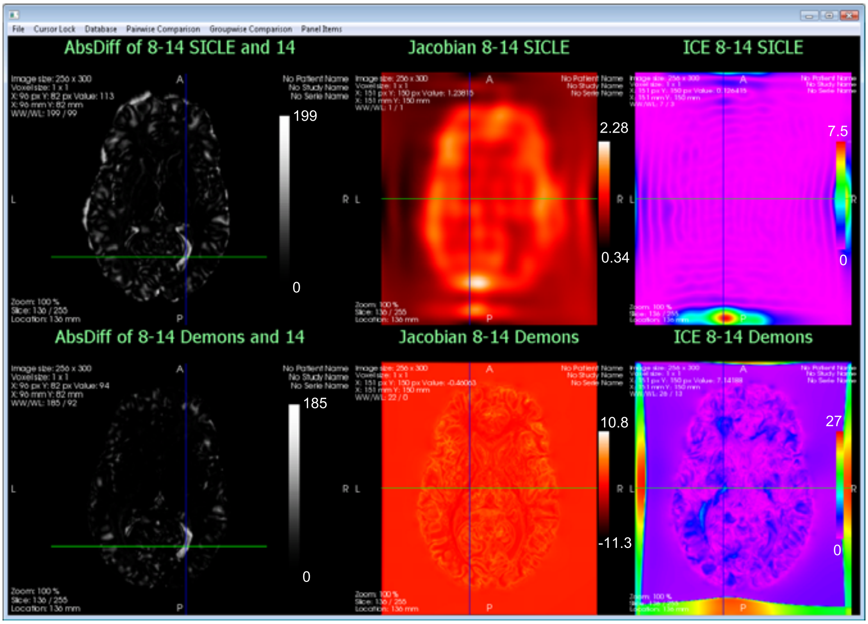

Fig. 5. A display window comparing the registration performances of SICLE and Demons algorithms on subject 8 and 14 of NA0. The left column shows the absolute intensity difference of the deformed source with the target image; the center column shows the Jacobian image of the transformations; and the right column shows the inverse consistency error images of the transformations. Note that the color scales on each of the images are different. The Jacobian for SICLE does not have any zero-crossings (i.e., singularity in transformation), whereas the Jacobian for Demons has many singularities. The cursor location indicates a Jacobian of -0.432701 for Demons and 1.40771 for SICLE. The transformation generated by Demons produced large ICE values, while SICLE produced generally low ICE values inside the brain region. The location pointed by the cursor indicates ICE 8.93467 for Demons, and 0.11022 for SICLE.

Fig. 5 shows a NIREP display window showing the registration results of SICLE and Demons for the registration of subject 8 to subject 14. The left two panels show visually that the absolute intensity difference of the deformed source and the target is smaller for Demons than SICLE. Table 1 shows a portion of the relative overlap values obtained for each region of interest. These results show that the Demons algorithm outperformed the SICLE algorithm based on intensity difference and relative overlap.

Table 1. A subselection of the relative overlap statistics table shown in Fig. 3. The Demons outperformed the SICLE algorithm with respect to relative overlap for this experiment.

Region Name	Relative Overlap (SICLE)	Relative Overlap (Demons)
Background	0.95	0.96
L_occipital	0.36	0.42
R_occipital	0.40	0.48
L_cingulate	0.41	0.48
R_cingulate	0.45	0.53
L_insula	0.44	0.61
R_insula	0.46	0.60
L_temppole	0.45	0.48

However, further evaluation reveals that Demons did not perform as well as SICLE with respect to other evaluation criteria, such as inverse consistency error and distortion measured by the Jacobian of the transformation. Taking advantage of the cursor-lock capability of the NIREP software, it can be seen side-by-side as in the right column of Fig. 5, that the registration made by Demons produced a transformation with high inverse consistency error. In contrast, SICLE, which enforces inverse consistency during registration, had low overall inverse consistency error inside the object boundary as expected.

In addition to inverse consistency error, Demons also performed poorly with the Jacobian of the transformation, with many spots with negative Jacobian values, indicating singularities in the transformation. On the other hand, the SICLE transformation did not contain any singularities.

4 Conclusion

The Non-rigid Image Registration Evaluation Program (NIREP) is a specialized program that makes it easy and intuitive to manipulate data associated with image registration. It differs from other medical image visualization tools in that it can manipulate large amounts of data such as images, deformed images, deformed comparisons specific to image registration. NIREP provides a standard set of evaluation criteria and evaluation databases so that meaningful comparisons between registration algorithms can be made. This standardization ensures that differences in performance are due solely to the algorithms being compared and not other confounding issues. NIREP provides the flexibility for users to create and use their own evaluation databases so that they can investigate how different registration algorithms perform on their own data for their own specific task. Moreover, NIREP provides users the ability make comparisons of their own without being constrained to a predetermined layout. Not only can users compare between any algorithms of choice, the users can make customizations to what kind of comparisons are to be made. The use of a human-readable and user-editable Display Description configuration file for NIREP offers great flexibility to users as to how the display panels are arranged, what types of information are to be displayed, what types of criterion are to be evaluated, what algorithms are to be compared, etc. Another key feature of NIREP is its ability to compute data using transformations of various formats, rather than merely visualizing data. Particularly, NIREP's ability to concatenate multiple transformations and compute inverses of transformations allows users to perform tasks such as computing average shapes based on transformations on the fly.

Acknowledgments

This work is supported in part by NIH grant EB004126.

References

1. Christensen, G.E., Johnson, H.J.: Consistent image registration. IEEE Trans. Med. Imaging 20(7), 568–582 (2001)
2. Christensen, G.E., Johnson, H.J.: Invertibility and transitivity analysis for nonrigid image registration. Journal of Electronic Imaging 12(1), 106–117 (2003)
3. Johnson, H.J., Christensen, G.E.: Consistent landmark and intensity-based image registration. IEEE Trans. Med. Imaging 21(5), 450–461 (2002)
4. Christensen, G.E.: Consistent linear-elastic transformations for image matching. In: Kuba, A., Sámal, M. (eds.) IPMI 1999. LNCS, vol. 1613, pp. 224–237. Springer, Heidelberg (1999)
5. Cao, K.: Local lung tissue expansion analysis based on inverse consistent image registration. Master's thesis, Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, IA 52242 (May 2008)
6. Geng, X., Kumar, D., Christensen, G.E.: Transitive inverse-consistent manifold registration. In: Christensen, G.E., Sonka, M. (eds.) Information Processing in Medical Imaging. LNCS, vol. 3564, pp. 468–479. Springer, Berlin (2005)
7. Thirion, J.P.: Fast non-rigid matching of 3d medical images. Report 2547, Institut National De Recherche en Informatique Et En Automatique (May 1995)
8. Thirion, J.: Image matching as a diffusion process: an analogy with maxwell's demons. Medical Image Analysis 2, 243–260 (1998)