

# Introduction to the Non-Rigid Image Registration Evaluation Project (NIREP)

Gary E. Christensen<sup>1</sup>, Xiujuan Geng<sup>1</sup>, Jon G. Kuhl<sup>1</sup>, Joel Bruss<sup>2</sup>,  
Thomas J. Grabowski<sup>2,3</sup>, Imran A. Pirwani<sup>4</sup>, Michael W. Vannier<sup>5</sup>,  
John S. Allen<sup>6,7</sup>, and Hanna Damasio<sup>6</sup>

<sup>1</sup>Department of Electrical and Computer Engineering, The University of Iowa

<sup>2</sup>Department of Neurology, The University of Iowa

<sup>3</sup>Department of Radiology, The University of Iowa

<sup>4</sup>Department of Computer Science, The University of Iowa

<sup>5</sup>Department of Radiology, University of Chicago

<sup>6</sup>The Dornsife Cognitive Neuroscience Imaging Center, Univ. of Southern California

<sup>7</sup>Department of Anthropology, Univ. of Southern California

**Abstract.** Non-rigid image registration (NIR) is an essential tool for morphologic comparisons in the presence of intra- and inter-individual anatomic variations. Many NIR methods have been developed, but are especially difficult to evaluate since point-wise inter-image correspondence is usually unknown, i.e., there is no “Gold Standard” to evaluate performance. The Non-rigid Image Registration Evaluation Project (NIREP) has been started to develop, establish, maintain, and endorse a standardized set of relevant benchmarks and metrics for performance evaluation of nonrigid image registration algorithms. This paper describes the basic framework of the project.

## 1 Introduction

Image registration is important for many applications, including longitudinal evaluations in individuals, comparison between individuals, creation of population atlases, use of atlas-linked information in individual cases, delivery of precision therapies, and many others. Non-rigid image registration is a more general approach than the widely used affine and rigid methods, but requires more complex methodology and computational effort to implement. Evaluating the performance of non-rigid image registration algorithms is a difficult task since point-wise correspondence from one image to another is not unique. That is there is rarely if ever a ground truth correspondence map to judge the performance of a registration algorithm.

We have started the Non-rigid Image Registration Evaluation Project (NIREP) to develop software tools and provide shared image validation databases for rigorous testing of non-rigid image registration algorithms. NIREP will extend the scope of prior validation projects by developing evaluation criteria and metrics using large image populations, using richly annotated image databases, using computer simulated data, and increasing the number and types of evaluation criteria.

The goal of this project is to establish, maintain, and endorse a standardized set of relevant benchmarks and metrics for performance evaluation of nonrigid image registration algorithms. Furthermore, these standards will be incorporated into an exportable computer program to automatically evaluate the registration accuracy of nonrigid image registration algorithms.

## 2 Prior Studies

To date, few attempts have been made to objectively evaluate and compare the performance of image registration algorithms using standard evaluation criteria. Two projects that stand out in this regard are the “Retrospective Image Registration and Evaluation Project” [1] led by J. Michael Fitzpatrick of Vanderbilt University for evaluating multimodality rigid registration accuracy and the non-rigid registration evaluation project entitled “Retrospective Evaluation of Inter-subject Brain Registration” [2] led by Christian Barillot of IRISA/INRIA-CNRS Rennes, France. In both of these projects, a common set of images was used to evaluate the performance of registration algorithms. Developers from around the world participated in these projects by registering the images with their own registration algorithms and sending the resulting transformations back to the home site for analysis. The benefits of involving external participants include eliminating implementation biases, distributing the processing load, and providing an incentive to produce good results.

Another important validation/evaluation project is the VALMET software tool for assessing and improving 2D and 3D object segmentation developed by Guido Gerig et al. [3] ([www.ia.unc.edu/public/valmet/](http://www.ia.unc.edu/public/valmet/)). The VALMET software was the first publicly available software tool for measuring and visualizing the differences between multiple corresponding medical image segmentations. It includes four algorithms for comparing segmentations: overlap ratio, Hausdorff distance, surface distance, and probabilistic overlap. The NIREP software evaluates image registration algorithm performance similar to the way the VALMET software evaluates image segmentation performance. In the future, all the VALMET evaluation metrics will be incorporated in to the NIREP software since automatic image segmentation produced from image registration is often used to evaluate registration performance.

## 3 NIREP Framework

The Nonrigid Image Registration Evaluation Project (NIREP) will build upon these earlier projects and provide a rich set of metrics for evaluating registration performance. We are building an exportable software tool that has nonrigid image registration evaluation metrics built into it. This is in contrast to previous projects that collected registration transformations from external participants and performed the analysis internally.

Under our model, NIREP software users will process their own data and evaluate the performance of different nonrigid registration algorithms on this

data. The tests will be fixed in the program such that the users will not be able to affect the results. This model has the advantage of standardizing the evaluation metrics and distributing the processing load. It will provide 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. The full disclosure of all validation tests performed as a result of NIREP project will be collected and analyzed. The results and analysis will be disseminated through publications and a central web site ([www.nirep.org](http://www.nirep.org)).

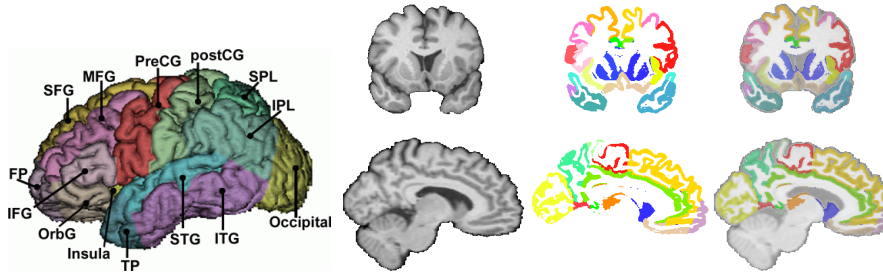
Evaluating the performance of nonrigid image registration algorithms is a difficult task since point-wise correspondence from one image to another is not known. In the absence of the truth, a diverse set of evaluation metrics will be used to evaluate registration performance on multiple well documented image databases. These tests evaluate the performance of image registration algorithms with respect to their transformation properties, agreement with human experts, agreement with computer simulated data, and other indirect performance tests. Complete and accurate reporting of validation tests will be published so that others will be able to detect potential bias (internal validity) and assess the generalizability and applicability of the results (external validity).

The NIREP software evaluates image registration algorithms in a similar fashion to the model used to evaluate the performance of computer hardware and systems. Computer systems are evaluated and compared using not one but many diverse criteria. The diversity of the tests serves both to evaluate system performance and to demonstrate the trade-offs between various systems. Popular examples of groups that evaluate computer hardware performance include the Standard Performance Evaluation Corporation ([www.specbench.org](http://www.specbench.org)) and Tom's Hardware ([www.tomshardware.com](http://www.tomshardware.com)). The diversity of the image registration metrics developed in NIREP evaluate registration algorithm performance and demonstrate trade-offs between different algorithms.

## 4 Evaluation Database

Our initial evaluation database consists of a population of 16 richly annotated 3D MR image volumes corresponding to 8 normal adult males and 8 females. These data sets were selected from a database of healthy right-handed individuals acquired in the Human Neuroanatomy and Neuroimaging Laboratory, The University of Iowa, and now part of the Laboratory of Computational Neuroimaging, The University of Iowa. The males have a mean age of 32.5 years, standard deviation of 8.4 years and range in age from 25 to 48. The females have a mean age of 29.8 years, standard deviation of 5.8 and range in age from 24 to 41. The complete population will be used to evaluate the non-rigid image registration performance for complexly shaped neuroanatomical structures.

The 16 MR data sets have been segmented into 32 gray matter regions of interest (ROIs). Fig. 1 shows an example of the segmentations associated with the MR data sets. Some of these data sets and their segmentations appear in a



**Fig. 1.** Shown are the segmentations available in the neuroanatomy 1 evaluation database. The segmentations include gray matter regions in the frontal, parietal, temporal, and occipital lobes; cingulate gyrus; and insula. The cerebellum, hypothalamus, and brain stem are currently not segmented currently.

new atlas by Hanna Damasio that illustrates the normal neuroanatomy of the human brain, and have been used in several publications about the morphometric analysis of the normal human brain [4, 5, 6].

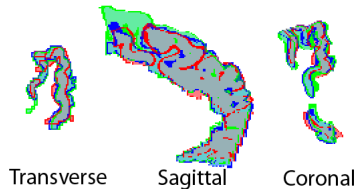
The brains were initially segmented with Brainvox [7] using the criteria described in papers by John Allen et al. [4, 5, 6]. The resulting segmentations partitioned the brain in to regions that contained both gray and white matter. Although the image volumes were carefully segmented, the segmentation process was done in 2D. As a result, the segmentations were smooth in the plane of segmentation but had rough edges when viewed in oblique slices. In addition, many of the segmentations had to have arbitrary boundaries within the white matter since region boundaries are well defined at sulci level but have to rely on “connecting lines” between the depth of the sulci within the white matter. These initial segmentations were then restricted to the gray matter to fix the boundary problems in the white matter. Gray matter segmentations were generated using the approach described in Grabowski et al. [8]. The gray matter segmentations were applied to the regional segmentations to remove the white matter from the segmentations. This produced gray matter ROIs with smooth boundaries at the outer surface of the cortex and at the gray/white interface. The segmentations were then hand edited using the Analyze<sup>TM</sup> software (Mayo Clinic, Rochester Minnesota) to produce the final gray matter segmentations.

## 5 Evaluation Criteria

Since there is rarely if ever a “Gold Standard” to evaluate image registration results, no metric alone is sufficient to evaluate the performance of a nonrigid registration algorithm. However, using information from many different diverse metrics will provide a good indication of the non-rigid image registration performance. In this initial phase of the project, we have included the following four metrics: (1) relative overlap, (2) intensity variance, (3) inverse consistency and (4) transitivity.

**Relative Overlap Metric:** The alignment of the regions of interest (ROIs) is a good indicator of how well two images are registered. The relative overlap of the segmentations is a measure of how well two corresponding segmented regions agree with each other. The Relative Overlap metric is given by  $RO(P, S) = \frac{volume(P \cap S)}{volume(P \cup S)}$  where  $P$  and  $S$  are two corresponding segmentations. In the context of image registration,  $P$  corresponds to a segmentation transformed from image  $i$  to  $j$  compared to the corresponding segmentation  $S$  defined in image  $j$ .

The NIREP software generates a variety of images for evaluation of local spatial registration performance and summarizes relevant performance statistics for each region of interest with tables and plots. Figure 2 shows an example of relative overlap performance for the superior frontal gyrus (SFG) and Table 1 shows statistics for the relative overlap before and after registration for the left and right SFG (LSFG, RSFG), the left and right temporal pole (LTP, RTP), and the left and right orbito-frontal gyrus (LOFG, ROFG). As Fig. 2 illustrates, images provide the user with visual feedback for evaluating how well the superior frontal gyrus was registered before and after registration. Similarly, numerical data such as shown in Table 1 provides the user with quantitative feedback of algorithmic performance. Note that the 32 gray matter segmentations currently included in the neuroanatomy 1 database have relatively small volumes compared to the brain volume and have large surface areas compared to their volume. Segmentations with small volumes and large surface areas typically have small relative overlaps even for good registrations. These segmentations should provide a challenging goal for registration.



**Fig. 2.** Relative overlap images for the right Superior Frontal Gyrus before and after intersubject registration. Red - target, Green - before registration, Blue - after registration.

**Intensity Variance Metric:** A common method used to measure image registration performance is to register a population of images with a target image and average the intensities of the registered images. The idea is that the better the registration algorithm is, the closer each registered image looks to the target image and the sharper the intensity average image. One way to measure the sharpness of the intensity average image is to compute the variance of the registered intensity images. The voxel-wise intensity variance (IV) of a population of  $M$  images registered to image  $j$  is computed as

$$IV_j(x) = \frac{1}{M-1} \sum_{i=1}^M (T_i(h_{ij}(x)) - ave_j(x))^2 \quad \text{where} \quad ave_j(x) = \frac{1}{M} \sum_{i=1}^M T_i(h_{ij}(x)), \quad (1)$$

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

**Table 1.** Relative overlap (RO) summary table for 132 ( $2 \times (11 + \dots + 1)$ ) transformations registered a population of 12 out of the 16 data sets using small deformation inverse-consistent linear-elastic (SICLE) registration method [9]. The voxel size is  $0.7 \times 0.7 \times 0.7 \text{ mm}^3$ . Due to lack of space, we only report the RO of 6 regions of interest (ROIs) as a representative sample of the RO of all 32 ROIs. The 6 ROIs include: Left/Right Temporal Poles (LTP, RTP), Left/Right Superior Frontal Gyrus (LSFG, RSFG), and Left/Right Orbital Frontal Gyrus (LOFG, ROFG).

	LTP	RTP	LSFG	RSFG	LOFG	ROFG
Ave Vol ( $\text{voxel}^3$ )	27900	30900	79600	76700	45700	47300
Std Dev Vol	4850	4320	9960	12200	3250	5080
Ave Surface Area ( $\text{voxel}^2$ )	11100	12200	36900	37000	21800	22600
Std Dev Surface Area	1700	1360	4160	4700	1700	2610
Ave RO: before registration	0.365	0.237	0.267	0.263	0.338	0.194
Ave RO: after SICLE	0.542	0.334	0.364	0.360	0.469	0.290
RO Std dev: before registration	0.0733	0.0632	0.0399	0.0613	0.0747	0.0505
RO Std dev: after SICLE	0.0503	0.0495	0.0341	0.0438	0.0295	0.0460

**Inverse Consistency Metric:** The inverse consistency metrics evaluates registration performance based on desired transformation properties [9, 10, 11]. The inverse consistency metric measures the inverse consistency error between a forward and reverse transformation between two images. Ideally the forward transformation equals the inverse of the reverse transformation implying a consistent definition of correspondence between two images, i.e., correspondence defined by the forward transformation should be the same as that defined by the reverse transformations. Thus, composing the forward and reverse transformations together produces the identity map when there is no inverse consistency error. The inverse consistency error is defined as the squared difference between the composition of the forward and reverse transformations and the identity mapping.

The voxel-wise cumulative inverse consistency error (CICE) with respect to template image  $j$  is computed as

$$CICE_j(x) = \frac{1}{M} \sum_{i=1}^M \|h_{ji}(h_{ij}(x)) - x\|^2 \quad (2)$$

where  $h_{ij}$  is the transformation from image  $i$  to  $j$ ,  $M$  is the number of images in the evaluation population and  $\|\cdot\|$  is the standard Euclidean norm.

The CICE is an example of a necessary evaluation metric for evaluating registration performance but is not a sufficient evaluation metric to guarantee good correspondence. For example, two identity transformations have zero inverse consistency error but in general provide poor correspondence between two images. However, a pair of transformations that provide good correspondence between images should have zero CICE.

**Transitivity Metric:** The transitivity metric [11] evaluates how well all the pairwise registrations of the image population satisfy the transitivity property.

The transitivity property is important to minimize correspondence errors when two transformations are composed together. Ideally, transformations that define correspondence between three images should project a point from image  $A$  to  $B$  to  $C$  to  $A$  back to the original position. The transitivity error for a set of transformations is defined as the squared error difference between the composition of the transformations between three images and the identity map.

The voxel-wise cumulative transitivity error (CTE) with respect to template image  $j$  is computed as

$$CTE_k(x) = \frac{1}{(M-1)(M-2)} \sum_{\substack{i=1 \\ i \neq k}}^M \sum_{\substack{j=1 \\ j \neq i \\ j \neq k}}^M \|h_{ki}(h_{ij}(h_{jk}(x))) - x\|^2. \quad (3)$$

The CTE is another example of a necessary evaluation metric but is not a sufficient evaluation metric for guaranteeing good correspondence. For example, a set of transformations that have zero transitivity error does not imply good correspondence as can be seen with a set of identity transformations. However, a set of transformations that provide good correspondence between images in a population should have zero transitivity error.

## 6 Future Work

In the future, additional evaluation databases and metrics will be added to NIREP. The NIREP project will serve as a repository for evaluation databases and metrics developed at The University of Iowa and contributed from researchers at other institutions. Extensive documentation will be collected describing the strengths and weaknesses of evaluation databases, metrics, and registration algorithms. The NIREP project, with the help of the medical imaging community, will establish, maintain, and endorse a standardized set of relevant benchmarks and metrics for performance evaluation of nonrigid image registration algorithms.

## Acknowledgments

This work was supported by NIH grants EB004126, CA096679 and HL64368.

## References

1. Jay West, J. Michael Fitzpatrick, et al., "Comparison and evaluation of retrospective intermodality brain image registration techniques", vol. 21, pp. 554–566, 1997.
2. P. Hellier, C. Barillot, L. Corouge, B. Gibaud, G. Le Goualher, D.L. Collins, A. Evans, G. Malandain, N. Ayache, G.E. Christensen, and H.J. Johnson, "Retrospective evaluation of inter-subject brain registration", *IEEE Transactions on Medical Imaging*, vol. 22, pp. 1120–1130, 2003.

3. Guido Gerig, Matthieu Jomier, and Miranda Chakos, "Valmet: A new validation tool for assessing and improving 3d object segmentation", in Wiro J. Niessen and Max A. Viergever, editors, *MICCAI 2001*, vol. LNCS 2208, pp. 516–528. Springer, 2001.
4. Hanna Damasio, *Human brain anatomy in computerized images*, Oxford University Press, New York, 2nd edition, 2005.
5. Allen JS, Damasio H, and Grabowski TJ, "Normal neuroanatomical variation in the human brain: an mri-volumetric study", *Am J Phys Anthropol*, vol. 118, pp. 341–58, 2002.
6. Allen JS, Damasio H, Grabowski TJ, Bruss J, and Zhang W, "Sexual dimorphism and asymmetries in the gray white composition of the human cerebrum", *NeuroImage*, vol. 18, pp. 880–894, 2003.
7. R.J. Frank, H. Damasio, and T.J. Grabowski, "Brainvox: an interactive, multi-model, visualization and analysis system for neuroanatomical imaging", *NeuroImage*, vol. 5, pp. 13–30, 1997.
8. T.J. Grabowski, R.J. Frank, N.R. Szumski C.K. Brown, and H. Damasio, "Validation of partial tissue segmentation of single-channel magnetic resonance images of the brain", *NeuroImage*, vol. 12, pp. 640–656, 2000.
9. G.E. Christensen and H.J. Johnson, "Consistent image registration", vol. 20, pp. 568–582, July 2001.
10. H.J. Johnson and G.E. Christensen, "Consistent landmark and intensity-based image registration", vol. 21, pp. 450–461, 2002.
11. G.E. Christensen and H.J. Johnson, "Invertibility and transitivity analysis for nonrigid image registration", *Journal of Electronic Imaging*, vol. 12, pp. 106–117, Jan. 2003.