A Framework for Comparison and Evaluation of Nonlinear Intra-Subject Image Registration Algorithms

Release 0.2

Martin Urschler[1], Stefan Kluckner[1] and Horst Bischof[1]

October 24, 2007

[1]Institute for Computer Graphics and Vision, Graz University of Technology, Austria
Contact: urschler@icg.tu-graz.ac.at

Abstract

Performance validation of nonlinear registration algorithms is a difficult problem due to the lack of a suitable ground truth in most applications. However, the ill-posed nature of the nonlinear registration problem and the large space of possible solutions makes the quantitative evaluation of algorithms extremely important. We argue that finding a standardized way of performing evaluation and comparing existing and new algorithms currently is more important than inventing novel methods. While there are already existing evaluation frameworks for nonlinear inter-subject brain registration applications [8, 2], there is still a lack of protocols for intra-subject studies or soft tissue organs. In this work we present such a framework which is designed in an ”open-source” and ”open-data” manner around the Insight Segmentation & Registration Toolkit to aid in intra-subject, intra-modality registration evaluation. The goal of our work is to provide the research community with the basis framework that should be extended by interested people in a community effort to gain importance for evaluation studies. We demonstrate our proposed framework on a sample evaluation and release its implementation and associated tools to the public domain.

Contents

1 Introduction

Nonlinear (non-rigid) image registration is an important field for medical computer vision applications [4]. Since it resembles an ill-posed problem, it is also a very complex topic from the theoretical point of view. Unfortunately, validation of nonlinear image registration algorithms tends to be quite difficult as well, so we are facing a major challenge if quantitative statements about performance and comparison of algorithms are required.

On the one hand, the problem of validating nonlinear registration stems from the lack of ground-truth data which in general can not be derived from realistic, clinical data sets under investigation. On the other hand, the definition of suitable quantitative measures to compare different algorithms or an algorithm result to a synthetic ground truth also is a challenge. Many quantitative similarity metrics (which are frequently used in publications to compare registration results), like e.g. the sum-of-squared differences, are only judging outcomes of the registration process and they are often biased to algorithms which use the same or a related similarity measure in the cost function of their underlying optimization scheme.

Due to the problems outlined above and the inherently available influence of noise, partial volume effect, limited numerical precision, interpolation schemes, etc. on quantitative measurements we refer to algorithm evaluation (in accordance with [8]), i.e. measuring ”relative” algorithm performance, in contrast to validation or the measurement of ”absolute” performance.

We argue that the current state of the art in nonlinear registration algorithm evaluation is inadequate. For registration algorithms it is only possible to gain practical, clinical acceptance if the research community builds a standardized protocol for evaluation. Therefore, in this contribution we propose a modular evaluation framework for objectively comparing algorithms which is designed with the spirit of open-source, open-access, open-data and open protocols in mind, according to the basic ideas and principles from the Insight Segmentation & Registration Toolkit (ITK [9]) and the Insight Software Consortium.

The building blocks of our framework are:

- Freely available data sets, e.g. from the NLM data collection project [12], the MIDAS project from the Insight Software Consortium [3] or the NCIA project [11].
- A number of freely available algorithms for nonlinear registration, like e.g. the ITK is able to provide and continuously gathers.
- A set of quantitative measures on which the research community has agreed upon, with public domain implementations residing in a central repository.
- A pool of synthetic deformation models, again with the agreement of the research community, either defined analytically or derived by using the existing pool of registration algorithms.
- An evaluation framework in an easily customizable scripting language that forms the glue for the rest of the components.

In the paper we focus on the description of the generic framework and we show an example implementation that focuses on intra-subject, intra-modality nonlinear CT image registration focusing on differences in breathing states. However, our framework can (and should!) easily be extended to more fields of application, be it inter-modal, inter-subject or any kind of inter-modal registration different from CT, by implementing appropriate components in addition to the existing ones.

2 Related Work

Given the current state of the art in nonlinear intra-subject, intra-modality registration we consider it more important to perform research on the evaluation of algorithms than inventing a novel one. In literature one finds many publications on algorithms which do not focus on a thorough evaluation and there are only few publications which are entirely dedicated to evaluation.

A common approach to perform evaluation is to identify landmark or region correspondences independently of the registration and to measure how well they are aligned after registration. While this is a sufficient method for rigid registration [26] and also for some nonlinear registration applications [8] it obviously has drawbacks in the latter case since the error in regions far from the landmarks in general can not be measured.

In literature we can find several similar attempts to create evaluation projects. Examples are the retrospective rigid registration evaluation project from West et al. [25], the VALMET project for evaluation of segmentation results [6] or more recently the NIREP project for evaluation of nonlinear inter-subject registration algorithms [2]. Additionally the success of community-based evaluation efforts are also obvious in related computer vision fields like stereo correspondence algorithms [14] and multi-view 3D reconstruction [16]. However, currently there are no projects that focus on a comparison of the large number of available algorithms for intra-subject nonlinear registration.

Next we give some examples for works that deal with nonlinear registration evaluation. In [8] an evaluation study on brain data sets is presented where six inter-subject registration methods are investigated and compared using a number of local and global measures. The main focus lies on the correct alignment of those brain landmarks which are assumed to be present in all individuals. In [15] a finite element method is used to create a physically plausible synthetic deformation model that might be used for comparison with registration results. Their focus lies on the evaluation of B-Spline grid based methods [13] applied to mammography MR input data. The work of [24] presents an evaluation of the Demons [19] algorithm applied to prostate CT images. They use synthetic experiments to argue on the suitability of Demons for this specific application. The focus of [17] is on the comparison of different similarity measures in the context of nonlinear registration. Their protocol includes a number of optimization-independent, statistical measures describing capture range, number, location and extent of local optima as well as the accuracy and distinctiveness of optima of a cost function derived from a similarity measure. In [2] a framework for the evaluation and comparison of inter-subject brain registration algorithms is presented. Their project NIREP (Nonlinear Image Registration Evaluation Project) is most related to our work, however, its dedication to inter-subject registration and its incompatibility with the idea of ”open-source” and ”open-data” are the major restrictions.

3 A Nonlinear Intra-Subject Registration Evaluation Framework

The main problem in nonlinear image registration evaluation is the lack of ground truth information from clinical data sets. Researchers therefore have to restrict themselves to define practically relevant synthetic experiments. There are several groups of synthetic experiments possible, i.e. (1) synthetically deformed synthetic data, (2) synthetically deformed phantom data and (3) synthetically deformed real data. In this paper we concentrate on the third kind of experiments, however, the framework allows the usage of the first two input types as well. The choice of synthetic deformations is crucial in an evaluation procedure. If the chosen transformation is too simple or restrictive, then the derived quantitative measures are restricted to a very coarse approximation of the originally investigated problem. This fact especially complicates the task of nonlinear registration.

Our approach is to use a combination of simple synthetic transformation methods to derive quantitative measures on the performance of several nonlinear registration algorithms. We assume that the calculation of many different synthetic deformations, with each of them testing different behavior, along with a large number of different evaluation measures leads to a method for thorough evaluation and comparison. We further hypothesize in this paper that one can make use of a ”pool” of different, sufficiently dissimilar methods to create a number of synthetic deformations. If this ”pool” of methods is large enough one can assume that the evaluation converges to a ”bronze standard” evaluation, a term which was borrowed from a publication of Glatard et al. [7] who describe a way of establishing a ”ground truth” in rigid registration evaluation by combining a large number of data sets with a number of algorithms and analyzing all possible combinations of registration results. One of our most important assumptions is, that a community effort is needed to increase the number of algorithms, measures and synthetic transformations in order to arrive at a practically relevant evaluation protocol.

We have built an evaluation framework to assist in the quantitative comparison of different algorithms over a variety of synthetic transformations. This framework uses Python as high-level scripting language to call the C++ methods that perform the synthetic transformation calculation, call the registration algorithms and the computation of the quantitative evaluation measures. Input images are synthetically transformed with the help of a configuration file. A list of nonlinear registration algorithms along with its parameterizations may be specified to perform the computations. Finally the synthetic ground-truth deformation fields and the original moving images are compared with the outcomes of the registration algorithms and log files with the quantitative evaluation results are written. Note, that all of these processes are performed in an easily customizable fully automatic fashion. Another important aspect this evaluation framework allows, is to automatically test algorithms with different parameterizations to determine optimal parameters with respect to the evaluation measures.

Our basic strategy for the evaluation is in all cases identical. We take an original image, apply a synthetic transformation and store the synthetically warped image and the resulting displacement field. The displacement field will be our ground truth to compare to. Now each of the investigated nonlinear registration algorithms gets the synthetically warped image as fixed input and the original image as moving input, i.e. we try to find a displacement field that warps the original image to the synthetically transformed. In this way we can finally compare the warped image with the synthetically warped image and the ground truth displacement field with the calculated displacement field. Note that only this way it can be guaranteed that the displacement fields represent transformations in the same directions (from fixed to moving image). Compare Figure 1 for an overview of this strategy. This figure also clearly shows that the nonlinear registration algorithms have no insights on the working of the synthetic transformations, making their behavior completely independent from these transformations. A possible bias that one method might have concerning a certain synthetic transformation may be removed by increasing the number of investigated synthetic deformations.

3.1 Compared Algorithms

In this section we list the nonlinear registration algorithms from the ITK that are currently used in the framework. We shortly describe each method and the parameters that have to be chosen for the evaluation

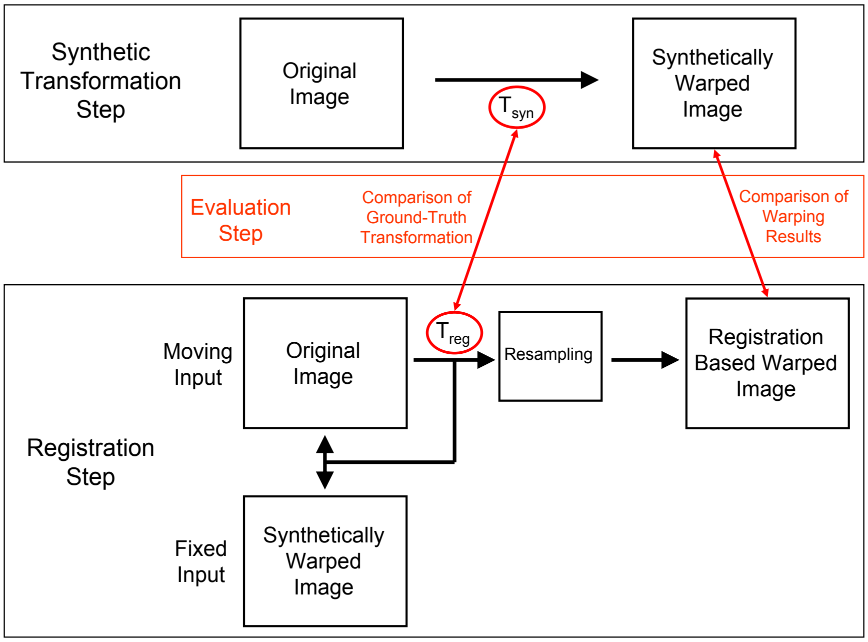
Synthetic Transformation Step
Original Image
T_{syn}
Synthetically Warped Image
Evaluation Step
Comparison of Ground-Truth Transformation
Comparison of Warping Results
Moving Input
Original Image
T_{reg}
Resampling
Registration Based Warped Image
Registration Step
Fixed Input
Synthetically Warped Image

Figure 1: The basic setup for the synthetic transformation evaluation experiments. The synthetic transformations are applied to the original image, the nonlinear registration of original and synthetically transformed image leads to a transformation and a warped image which can be compared.

procedure. There are two parameters that are required in all of the algorithms, the number of levels in the Gaussian pyramid approach and the number of iterations per pyramid level.

Demons Registration - ”demons” is an approximation to the elastic or fluid deformation model, where the computation of the similarity metric and the regularization of the displacement field are decoupled and the latter part is approximated by a Gaussian smoothing [19]. The only additionally required parameter is the standard deviation sigma of the displacement field smoothing step.

Symmetric Demons Registration - ”symdemons” builds upon the same principles as Demons, however, the gradient term in the similarity metric computation uses information from both images to be registered, thereby increasing robustness. The only required parameter is the standard deviation sigma of the displacement field smoothing step.

Level Set Motion Registration - ”levelsetmotion” is a very efficient registration algorithm that interprets the image intensity as isocontours of a level set and tries to match those iso-contours [21]. It gains its efficiency from the fact that no regularization is used.

Curvature Registration - ”curvature” uses a regularization term that penalizes the norms of the second derivatives of the displacement field components to perform registration [5]. In this implementation the variational formulation for combining similarity and regularization term is used [10] and numerically solved using a fast fourier transform. Additional parameters are the time-step τ of the semi-implicit scheme and the regularization weight α.

We propose to add two more algorithms to our framework, which currently are not included in the official ITK release, but were submitted to the 2007 MICCAI Open Science Workshop. Here obviously the open manner of this workshop shows its clear advantage compared to traditional workshops.

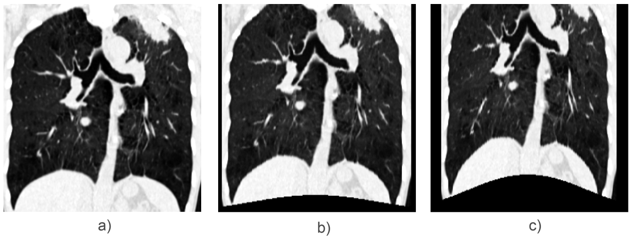

Figure 2: Synthetic transformation - Simulated breathing (*simbr*). a) original data set, b) effect of $t_{vertical} = 25mm$ and $t_{inward} = 10mm$ and c) effect of $t_{vertical} = 55mm$ and $t_{inward} = 25mm$.

Fast Block Matching Registration - ”fastblock” is a 2007 MICCAI Open Science workshop contribution by Suarez-Santana et al [18]. It is a simple algorithm relying on a pyramidal block-matching scheme and a local entropy-based similarity measure. There are two parameters required in this algorithm which define regularization and similarity.

Diffeomorphic Demons Registration - ”diffdemons” is another 2007 MICCAI Open Science workshop contribution by Vercauteren et al [23, 22]. It describes a diffeomorphic extension of the Demons algorithm in order to increase its suitability for inter-subject registration. Although inter-subject registration is not in the focus of this work, we chose to include this paper in order to enlargen the pool of available algorithms. There are two additionally required parameters, the standard deviation sigma of the displacement field smoothing and the maximum step length of displacement field updates.

The reader may note that we decided not to include the B-Spline based algorithm using a MeanSquares metric or the Finite Element based elastic registration. The B-Spline method in our experience requires too much computation time, we have not yet managed to bring down the computational effort near the times of the other ITK algorithms. The decision to ignore the Finite Element based registration was made due to the very complex way this algorithm has to be set up. We invite the community to add these implementations to the framework.

3.2 Synthetic Deformations

We are using several different kinds of synthetic transformations. Note that in this paper the choice of synthetic deformations is guided by the intended application of evaluating thoracic soft-tissue data sets. However, the framework certainly allows to add and remove any kind of synthetic transformation to tailor the evaluation to the intended application area.

Our first synthetic model is a very simple simulated breathing transformation. It depends on two parameters $t_{vertical}$ and t_{inward} which provide a means for a simple approximation of diaphragm and rib-cage movement. Further, a slight intensity variation is applied to the interior of the lung, to simulate the partial volume effect. A more detailed description of this transformation can be found in [20]. We will denote it *simbr*. Examples for this transformation are given in Figure 2, for the experiments we use two instances of this transformation (*simbr-25-10*, *simbr-55-25*).

The second synthetic transformation consists of a regular grid with random deformations, interpolated by a thin plate spline. There are two parameters that may be varied, the grid size of the points to be placed and the possible extent of the random deformation that is applied per grid-point. It will be denoted *grid* from now on. The parameters of this transformation are *gridSize* and *maxDeviation*. Figure 3 gives some examples for this transformation, later on we will be using two instances of this transformation (*grid-32-4*, *grid-32-8*).

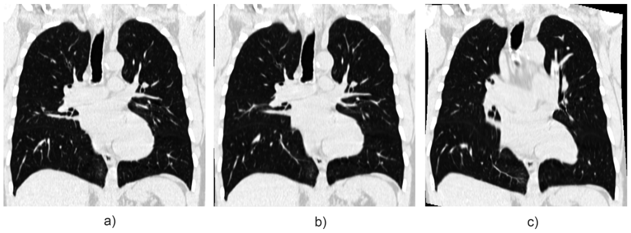

Figure 3: Synthetic transformation - grid based (*grid*). a) original data set, b) effect of $gridSize = 32$ and $maxDeviation = 2$ and c) effect of $gridSize = 32$ and $maxDeviation = 4$.

The third synthetic transformation is a uniform periodic one which we took from [27]. Here the displacement field follows a cosinus distribution with a given amplitude and a given phase. We use a single instance of this transformation with amplitude 5 and phase 32 (*uniform-5-32*).

Our final form of synthetic transformations makes use of the pool of nonlinear registration algorithms described above. Here, given two input images that differ in a nonlinear fashion, we calculate the nonlinear registration of these two data sets with each of the algorithms. For n algorithms this gives us n displacement fields. Now it is possible to use each of these n displacement fields as synthetic deformations to synthetically warp one of the input images. Thus we get a pair of images to be registered by all of the available algorithms.

Note that we currently do not simulate noise in these synthetic deformation models, however, this could easily be incorporated by modifying the programs for calculating the synthetic deformations.

3.3 Quantitative Evaluation Measures

Two kinds of measures are used in our evaluation framework. The first one is used to compare displacement field ground truth data with displacement field results calculated during the registration. The second kind of measures compares two sets of images, before and after registration. This comparison step is performed on fixed and moving input image and on fixed and warped moving image. After successful registration, warping the moving image should result in an increase of similarity with the fixed image.

We always compute the evaluation measures only on the overlapping regions of the input image data sets and the deformation fields, i.e. background regions in either of the data set are excluded from the computation. This is especially important in the case of synthetic transformations, where certain areas might vanish due to a shrinking like behavior, we always mark these regions with special values during synthetic transformation and we omit those marked regions from the evaluation process.

We have to remark here that all of these measures possess one inherent shortcoming. Data sets with a size of $256 \times 256 \times 256$ consist of a total around 16 million voxels and each measure of interest will try to reduce this huge amount of information to one single number. Although some of our measures are inspired by robust statistics, their information value has to be at least discussed and also questioned. One should always keep this in mind when looking at quantitative evaluation results.

Displacement Field Measures

To assess the similarity of a synthetic ground-truth deformation field φ_{syn} and a deformation field computed by a nonlinear registration algorithm φ_{reg}, we use:

Root-Mean-Square of Displacement Field RMS_{disp} The Root-Mean-Square of the displacement field interprets the two displacement fields of interest as feature vectors of dimension $3 \times N_1 \times N_2 \times N_3$, where N_i is the number of voxels on the input image grid in the according dimension. The measure reads

$$RMS_{disp} := \sqrt{\frac{1}{3N_1N_2N_3} \sum_{\mathbf{x}\in\Omega} \sum_{i=1}^{3} \left(\varphi_{syn}(\mathbf{x}_i) - \varphi_{reg}(\mathbf{x}_i)\right)^2}$$

Median Absolute Deviation of Displacement Field MAD_{disp} This measure is similar to the Root-Mean-Square, however it is a more robust variant in the presence of outliers. The median absolute deviation is defined as

$$MAD_{disp} := \text{Median}\left(|m_i - \text{Median}\left(m_i\right)|\right)$$

with $m_i = |\varphi_{syn}(\mathbf{x}_i) - \varphi_{reg}(\mathbf{x}_i)|$ and $i = 1 \cdots 3$.

Maximum Deviation of Displacement Field MAX_{disp} The maximum deviation of the displacement field states the maximal difference of all components of the two displacement fields $\varphi_{syn}$and φ_{reg}. We use a more robust maximum, i.e. our maximum difference is defined as the difference that is larger than 95% of all other values.

Jacobian of Displacement Field JAC_{disp} The Jacobian of the displacement field gives information about the transformations consistency. Especially it identifies possible singularities of the field. The Jacobian is defined as the determinant of the first partial derivatives of the transformation (compare the definition in [8]), negative values of the determinant resemble singularities, i.e. foldings. The measure JAC_{disp}scans the deformation field for negative values and reports its number as a percentage of the total number of voxels, ideally it would be 0.

Image Similarity Measures

To assess the similarity of images (fixed image $I_F(\mathbf{x})$, moving image $I_M(\mathbf{x})$) before and after registration we use:

Root-Mean-Square of Intensity Differences RMS_{int} The Root-Mean-Square of the intensity differences needs as its input the pixel-wise intensity differences over the overlapping region of the image domain. It is defined as

$$RMS_{int} := \sqrt{\frac{1}{N_1N_2N_3} \sum_{\mathbf{x}\in\Omega} d(\mathbf{x})^2}$$

with

$$d(\mathbf{x})^2 := \begin{cases} 100 & \text{if}\, (I_F(\mathbf{x}) - I_M(\mathbf{x}))^2 > 100 \\ (I_F(\mathbf{x}) - I_M(\mathbf{x}))^2 & \text{otherwise} \end{cases}$$

We decided to clamp intensity differences that are larger than 100 Hounsfield Units (a number that was chosen empirically) and assign the absolute difference of 100 Hounsfield Units to all of these differences. This makes the measure more sensitive to lower intensity differences, the unmodified Root-Mean-Square measure would weight outliers too strong. Note that we perform this clamping consistently for comparing all image pairs, either original fixed and original moving or original fixed and warped moving.

Median Absolute Deviation of Intensity Differences MAD_{int} This measure is similar to the Root-Mean-Square, however it is a more robust variant in the presence of outliers. Therefore no clamping is necessary within this measure. The median absolute deviation is defined as

$$MAD_{int} := \text{Median}\left(|d(\mathbf{x}) - \text{Median}\left(d(\mathbf{x})\right)|\right)$$

with $d(\mathbf{x}) = |I_F(\mathbf{x}) - I_M(\mathbf{x})|$.

Maximum Intensity Difference MAX_{int} The maximal intensity differences on the overlap region of the image grid. We use a robust maximum, i.e. our maximum intensity difference is defined as the intensity difference that is larger than 95% of all other values. Here of course no clamping (like for the Root-Mean-Square measure) of the intensity differences is performed.

Normalized Mutual Information NMI_{int} The normalized mutual information is a measure from information theory, which relates the information content of two images by probability distributions of the gray values. We use the formulation from the ITK normalized mutual information histogram metric that divides twice the mutual information by the sum of the individual entropies resulting in a measure between 0 and 1.

Edge Overlap $EDGE_{int}$ We are using a simple scheme to extract strong gradients from both images, based on the Canny edge detector [1]. The Canny edge detector is used with a σ of 3 times the voxel spacing, the result is a binary image with edges marked with a 1. The sum of the absolute differences divided by the number of voxels in the overlap region is used to compare the binary images. This function lies between 0 and 1 with 0 denoting optimal overlap.

4 A Sample Evaluation

To show the applicability of our evaluation framework we now present an evaluation example. The intention is to test different algorithms for the purpose of registering intra-modality thoracic CT data sets under the influence of breathing differences. Therefore, we are using two different publicly available data sets. First, we took the data set ”NormalChestCTNoContrast” from the NLM (National Library of Medicine) Data Collection Project [12]. This data set was used for the anlytically defined synthetic deformation experiments (see Figure 4a) and is abbreviated *nlm*. Second, we downloaded a pair of data sets from the NCIA (National Cancer Imaging Archives) repository [11] that differ by some breathing motion (see Figure 4b,c). We used it for the creation of synthetic deformation fields from the different registration algorithms. We refer to this data as *ncia*, it can be found in the LIDC section of the archives with the identifier ”30047”.

All of the data sets were resampled to a resolution of 256 × 256 × 256 voxels to limit the run-time of some of the algorithms to a reasonable time-span. We performed our experiments on 64 bit AMD Opteron processors with 2.4GHz and 8GB of RAM running Linux. Since algorithms tend to need a large amount of system memory the choice to use a 64 bit system was necessary.

The necessary parameters for the different algorithms were chosen as follows. For the two algorithms that were taken from the MICCAI Open Science workshop contributions we directly took the proposed default arguments without tuning. The other algorithms were set up with 4 shrink levels in the Gaussian pyramid, and a number of iterations that was calculated from two parameters $iter1 = 35$ and $iter2 = 25$ as $iter = iter1 * shrinkLevel + iter2$ in order to perform many more iterations on higher levels in the pyramid where calculation is cheap. The *sigma* parameter for demons and symmetric demons was chosen to be 1 voxel. τ and α in the curvatrue based algorithm were both set to 1.0. After this choice all of the parameters remained fixed during the evaluations.

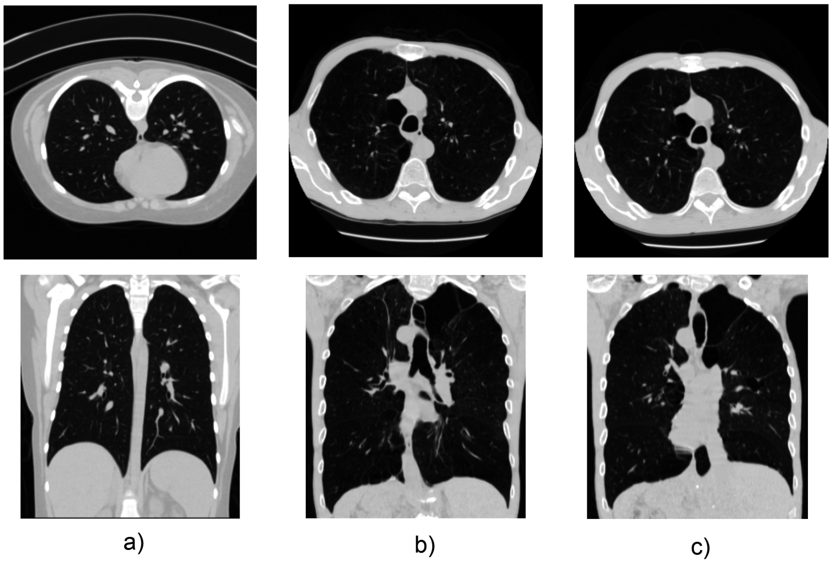

Figure 4: Our input data. a) Data set ”nlm”. b), c) Data set ”ncia” at two different breathing states.

4.1 Discussion of the Evaluation Results

The results of our evaluation can be found in Tables 2, 3, 4, 5. Due to the large number of quantitative measures resulting from the framework we decided to discuss it only in an overall manner.

What one can immediately see from these results, is that there are two algorithms which are not performing well on the synthetic data. These are the level set motion registration and the fast block matching registration. We assume that the level set motion method fails due to its lack of regularization and therefore do not suggest that it should be used. Especially the large percentage of negative values of the jacobian determinant of the displacement field are a problem. The problems are also visible on result images like in Figure 5c. The problems of the fast block matching algorithm are in the border region (see Figure 5a). We are not yet sure if this is an implementation problem or a problem of the method. This algorithm should be adapted and evaluated anew.

The overall best results on the evaluation experiments is given by the diffeomorphic Demons algorithm. Also the classical Demons algorithm performs very well, however on the important displacement field comparisons diffeomorphic Demons behaves better. The percentage of negative jacobian determinants is zero for diffeomorphic Demons as could be expected. The curvature algorithm wins some of the performance measures as well, however there is no clear benefit compared to diffeomorphic Demons and the run-times are approximately equal. As for the symmetric Demons implementation, it performs quite well, however, there are some situations where the quantitative measures grow very high. This corresponds to misregistered data sets (see Figure 5b), we suppose that there might be an implementation problem in the itkFastSymmetricDemonsFunction.

We omit presenting output images of well registered data sets, since the difference to the original image is seldom visible with the eye. However, we invite the reader to perform the same experiments using our implementation of the framework.

The run-times of the different algorithms on the $256 \times 256 \times 256$ can be found in Table 1. Note that these are only approximate values. We can see that the practically quite well performing Demons algorithm also

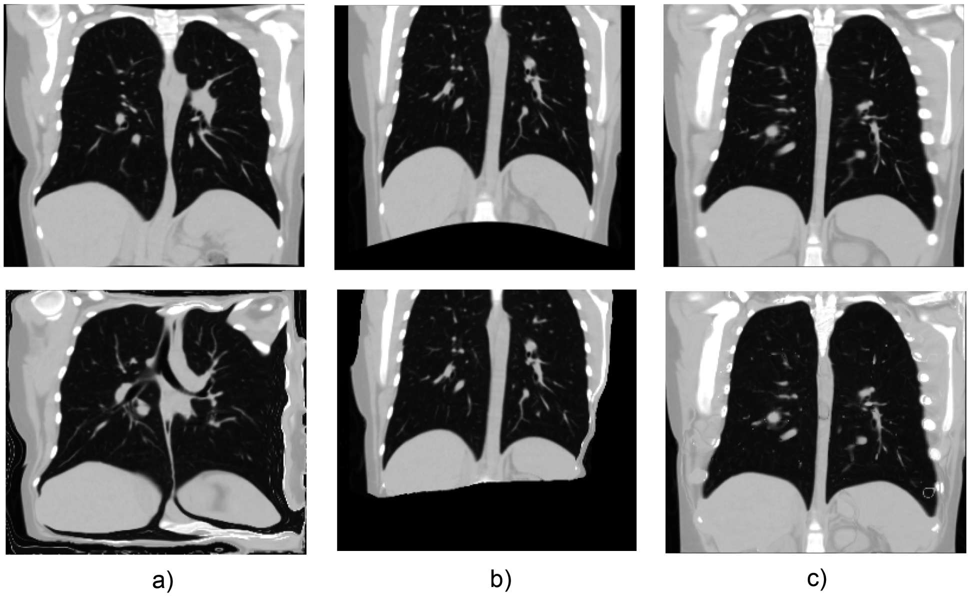

Figure 5: Problems of some registration algorithms. Top row is the target image and bottom row the registration result of a) fast block matching, b) symmetric Demons and c) level set motion algorithm.

Algorithm	demons	symdemons	levelset	curvature	fastblock	diffdemons
Runtime [s]	800	2800	850	3500	1700	3100

Table 1: Computational effort of the various investigated algorithms.

offers fast computation times.

To conclude we currently would suggest using Demons for time-critical registration of thoracic CT data and diffeomorphic Demons for the most accurate registration.

5 Conclusion & Future Work

In this paper we have presented an evaluation framework for evaluating and comparing nonlinear registration algorithms with the special intent on intra-subject applications. We have successfully shown how it is possible to automatically compare open-source algorithms on public domain data using our tools which we provide in conjunction with this submission. Our main goal is to create a standardized way of evaluating algorithms maintained at a centralized and open repository, e.g. the Insight Consortium server structure. In order to gain importance we invite the research community to contribute and/or discuss algorithms, quantitative measures and synthetic transformations. We know that our sample evaluation is not yet representative, but we hope to converge towards some kind of ”bronze standard” when collecting more and more modules.

Further work beside the implementation of the framework on a central repository should definitely include to upgrade the Python based framework to automatically distribute the computations among different computers in a cluster. The inclusion of methodologies to provide a framework for real data experiments or different setups (inter-modal, inter-subject) should be a straight-forward task by implementing additional registration methods, synthetic deformations and accompanying similarity measures. Finally in accordance to a fruitful reviewer comment, different noise models should be incorporated into the synthetic deformations.

Measure			simbr-25-10	simbr-55-25	grid-32-4	grid-32-8	uniform-5-32
RMS_{disp}	*initial*	[mm]	7.937	18.40	1.930	3.782	3.848
	demons	[mm]	1.642	7.880	0.915	1.748	2.811
	symdemons	[mm]	7.470	21.52	1.686	3.968	3.153
	levelset	[mm]	18.52	25.30	11.86	12.61	10.63
	curvature	[mm]	5.731	10.29	3.450	4.052	2.666
	fastblock	[mm]	65.89	70.67	56.42	nan	75.26
	diffdemons	[mm]	**1.017**	**7.757**	**0.222**	**0.638**	**1.689**
MAD_{disp}	*initial*	[mm]	2.479	5.945	0.806	1.524	2.541
	demons	[mm]	0.055	0.500	0.043	0.076	0.187
	symdemons	[mm]	0.081	0.967	0.054	0.114	0.165
	levelset	[mm]	4.670	7.764	2.486	2.983	3.115
	curvature	[mm]	0.048	**0.079**	**0.020**	**0.027**	**0.037**
	fastblock	[mm]	11.26	19.61	2.706	5.929	16.33
	diffdemons	[mm]	**0.024**	0.310	0.054	0.116	0.365
MAX_{disp}	*initial*	[mm]	18.34	41.39	3.628	7.386	6.608
	demons	[mm]	3.596	22.16	0.956	3.525	6.112
	symdemons	[mm]	7.004	47.51	2.517	8.836	7.205
	levelset	[mm]	41.27	55.28	26.08	27.44	24.03
	curvature	[mm]	7.281	**16.95**	4.113	5.728	6.168
	fastblock	[mm]	166.6	168.7	153.3	152.2	178.8
	diffdemons	[mm]	**0.658**	21.97	**0.452**	**1.267**	**4.200**
JAC_{disp}	*demons*	[%]	**0**	**0**	**0**	**0**	**4.58956e-06**
	symdemons	[%]	0.0282316	0.123886	0.00223255	0.00272346	0.011709
	levelset	[%]	0.271759	0.365262	0.195936	0.20556	0.191423
	curvature	[%]	0.0569056	0.224993	0.00568753	0.0122414	0.0112995
	fastblock	[%]	0.445634	0.529313	0.391153	0.377419	0.475234
	diffdemons	[%]	**0**	**0**	**0**	**0**	**0**

Table 2: Registration results for the analytically defined synthetic experiments. Displacement field measures.

Measure			demons	symdemons	levelset	curvature	fastblock	diffdemons
RMS_{disp}	*initial*	[mm]	10.68	11.36	20.00	11.76	64.96	10.23
	demons	[mm]	2.115	**3.166**	22.04	6.182	61.74	1.424
	symdemons	[mm]	6.152	6.589	22.30	8.190	61.81	5.259
	levelset	[mm]	17.89	17.91	**15.40**	17.93	66.34	17.39
	curvature	[mm]	7.160	5.966	21.47	6.804	60.82	6.298
	fastblock	[mm]	71.44	73.48	77.31	74.63	**36.82**	71.15
	diffdemons	[mm]	**1.563**	3.355	21.53	**5.154**	59.61	**0.716**
MAD_{disp}	*initial*	[mm]	3.810	4.216	6.486	4.434	14.19	3.799
	demons	[mm]	**0.222**	0.657	7.192	1.026	9.844	**0.077**
	symdemons	[mm]	0.318	**0.554**	7.916	0.743	7.469	0.106
	levelset	[mm]	5.598	5.680	**4.111**	5.900	16.14	5.456
	curvature	[mm]	0.325	0.588	7.290	**0.270**	6.391	0.073
	fastblock	[mm]	13.58	14.64	24.10	15.45	**2.977**	13.78
	diffdemons	[mm]	0.306	0.732	6.622	1.022	6.853	0.125
MAX_{disp}	*initial*	[mm]	22.80	23.87	43.97	24.94	152.2	21.99
	demons	[mm]	4.020	7.388	48.74	14.29	147.7	2.883
	symdemons	[mm]	13.04	14.89	47.91	20.07	149.1	11.77
	levelset	[mm]	37.87	37.86	**34.95**	38.24	154.1	37.05
	curvature	[mm]	14.83	13.75	47.07	16.80	146.4	13.02
	fastblock	[mm]	174.3	178.9	180.6	181.2	**87.78**	173.3
	diffdemons	[mm]	**2.802**	**6.873**	48.26	**11.23**	144.0	**1.359**
JAC_{disp}	*demons*	[%]	0.00162363	0.0216962	0.00681263	0.0185669	0.0226597	8.34465e-06
	symdemons	[%]	0.0601146	0.127387	0.117009	0.129581	0.078536	0.0621149
	levelset	[%]	0.240135	0.237093	0.209323	0.246824	0.348112	0.234164
	curvature	[%]	0.06316	0.120942	0.123614	0.131058	0.142232	0.0565321
	fastblock	[%]	0.437119	0.432593	0.432847	0.436029	0.284698	0.440297
	diffdemons	[%]	**3.54052e-05**	**2.98023e-05**	**4.79817e-05**	**1.20997e-05**	**1.29938e-05**	**0**

Table 3: Registration results for the synthetic experiments using a pool of algorithms. Displacement field measures.

Measure			simbr-25-10	simbr-55-25	grid-32-4	grid-32-8	uniform-5-32
RMS_{int}	*affine*	[HU]	60.77	73.82	43.97	50.87	52.32
	demons	[HU]	10.57	22.05	**7.496**	**9.133**	16.18
	symdemons	[HU]	26.29	43.07	13.94	17.23	13.47
	levelset	[HU]	34.63	36.47	32.97	33.44	31.52
	curvature	[HU]	21.22	28.70	13.47	18.30	**13.13**
	fastblock	[HU]	60.16	67.61	51.62	54.28	65.38
	diffdemons	[HU]	**8.022**	**24.80**	8.817	15.00	23.60
MAD_{int}	*affine*	[HU]	22	76	7	10	10
	demons	[HU]	1	1	1	1	1
	symdemons	[HU]	1	1	1	1	1
	levelset	[HU]	7	8	6	7	6
	curvature	[HU]	1	1	1	1	1
	fastblock	[HU]	17	43	8	12	31
	diffdemons	[HU]	1	2	1	1	3
NMI_{int}	*affine*		0.212	0.131	0.365	0.293	0.301
	demons		**0.861**	**0.781**	**0.873**	**0.842**	0.753
	symdemons		0.761	0.622	0.835	0.797	**0.772**
	levelset		0.418	0.393	0.446	0.438	0.454
	curvature		0.765	0.684	0.801	0.753	0.761
	fastblock		0.231	0.212	0.318	0.284	0.195
	diffdemons		0.860	0.676	0.796	0.721	0.599
$EDGE_{int}$	*affine*		0.199	0.192	0.173	0.184	0.180
	demons		0.012	0.038	**0.012**	**0.016**	0.037
	symdemons		0.016	0.035	0.018	0.022	0.033
	levelset		0.125	0.123	0.122	0.122	0.116
	curvature		0.014	**0.019**	0.019	0.021	**0.019**
	fastblock		0.132	0.150	0.120	0.130	0.150
	diffdemons		**0.010**	0.052	0.019	0.030	0.068
MAX_{int}	*affine*	[HU]	835	923	361	571	661
	demons	[HU]	**5**	**11**	**5**	**7**	**13**
	symdemons	[HU]	8	21	7	10	14
	levelset	[HU]	98	105	92	94	81
	curvature	[HU]	15	22	16	17	20
	fastblock	[HU]	994	998	923	928	1020
	diffdemons	[HU]	8	34	15	29	61

Table 4: Registration results for the analytically defined synthetic experiments. Intensity difference measures.

Measure			demons	symdemons	levelset	curvature	fastblock	diffdemons
RMS_{int}	*affine*	[HU]	62.16	61.04	61.20	61.09	72.07	62.15
	demons	[HU]	**11.06**	23.92	40.99	21.78	51.01	**6.429**
	symdemons	[HU]	16.34	**15.99**	30.40	**13.47**	41.33	13.35
	levelset	[HU]	32.77	32.28	**29.20**	32.27	**37.83**	33.41
	curvature	[HU]	25.31	25.49	38.13	16.73	40.68	17.69
	fastblock	[HU]	59.17	59.38	62.71	59.37	47.75	58.24
	diffdemons	[HU]	28.81	36.06	47.16	33.59	51.80	19.23
MAD_{int}	*affine*	[HU]	27	24	26	24	65	27
	demons	[HU]	**1**	2	7	**1**	5	**1**
	symdemons	[HU]	**1**	**1**	**4**	**1**	**2**	**1**
	levelset	[HU]	9	8	7	8	13	9
	curvature	[HU]	2	4	7	2	4	**1**
	fastblock	[HU]	23	23	30	22	12	21
	diffdemons	[HU]	4	6	12	5	11	**1**
NMI_{int}	*affine*		0.204	0.215	0.208	0.212	0.126	0.201
	demons		**0.789**	0.667	0.439	0.700	0.382	**0.841**
	symdemons		0.749	**0.730**	**0.522**	**0.753**	**0.494**	0.800
	levelset		0.471	0.479	0.519	0.462	0.407	0.458
	curvature		0.638	0.606	0.463	0.714	0.454	0.806
	fastblock		0.201	0.188	0.160	0.189	0.318	0.211
	diffdemons		0.609	0.519	0.389	0.537	0.350	0.740
$EDGE_{int}$	*affine*		0.181	0.185	0.187	0.184	0.190	0.185
	demons		**0.045**	0.088	0.138	0.083	0.102	**0.041**
	symdemons		0.055	**0.073**	**0.124**	**0.074**	**0.084**	0.049
	levelset		0.133	0.143	0.135	0.139	0.123	0.136
	curvature		0.068	0.094	0.137	0.079	0.094	0.052
	fastblock		0.164	0.172	0.182	0.167	0.136	0.161
	diffdemons		0.086	0.116	0.150	0.110	0.126	0.056
MAX_{int}	*affine*	[HU]	1004	1007	995	1007	1060	999
	demons	[HU]	**13**	30	128	22	843	**4**
	symdemons	[HU]	16	**21**	88	**16**	204	6
	levelset	[HU]	87	85	**73**	85	**106**	91
	curvature	[HU]	49	60	137	27	245	13
	fastblock	[HU]	1057	1056	1055	1056	817	1051
	diffdemons	[HU]	86	104	202	89	661	38

Table 5: Registration results for the synthetic experiments using a pool of algorithms. Intensity difference measures.

References

[1] John Canny. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986.

[2] G. E. Christensen, X. Geng, J. G. Kuhl, J. Bruss, T. J. Grabowski, I. A. Pirwani, M. W. Vannier, J. S. Allen, and H. Damasio. Introduction to the Non-rigid Image Registration Evaluation Project (NIREP). In *Proc Workshop on Biomedical Image Registration*, volume LNCS 4057, pages 128–135. Springer Verlag, 2006.

[3] Insight Software Consortium. MIDAS - Multi Format Image and Data Assimilation System. http://www.insight-journal.org/dspace/community-list/, 2007.

[4] W. R. Crum, L. D. Griffin, D. L. G. Hill, and D. J. Hawkes. Zen and the Art of Medical Image Registration: Correspondence, Homology, and Quality. *NeuroImage*, 20(3):1425–1437, 2003.

[5] B. Fischer and J. Modersitzki. Curvature based image registration. *Journal of Mathematical Imaging and Vision*, 18(1):81–85, 2003.

[6] G. Gerig, M. Jomier, and M. Chakos. Valmet: A new validation tool for assessing and improving 3D object segmentation. In *Proc Intern Conf on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume LNCS 2208, pages 516–528, 2001.

[7] T. Glatard, X. Pennec, and J. Montagnat. Performance Evaluation of Grid-Enabled Registration Algorithms Using Bronze-Standards. In R. Larsen, M. Nielsen, and J. Sporring, editors, *Proc Intern Conf on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 4191 of *LNCS*, Copenhagen, Denmark, 2006. Springer.

[8] P. Hellier, C. Barillot, I. 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 Intersubject Brain Registration. *IEEE Transactions on Medical Imaging*, 22(9):1120–1130, 2003.

[9] ITK. Insight Software Consortium Segmentation and Registration Toolkit. http://www.itk.org, 2006.

[10] J. Modersitzki. *Numerical Methods for Image Registration*. Oxford University Press, 2004.

[11] National Cancer Institute (NCI). National Cancer Imaging Archives. https://imaging.nci.nih.gov/ncia/faces/baseDef.tiles, 2007.

[12] NLM. National Library of Medicine Image Data Collection Project. http://nova.nlm.nih.gov/Mayo/, 2006.

[13] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, August 1999.

[14] D. Scharstein and R. Szeliski. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, 47(1/2/3):7–42, 2002.

[15] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith, A. Degenhard, M. O. Leach, D. R. Hose, D. L. G. Hill, and D. J. Hawkes. Validation of Nonrigid Image Registration Using Finite-Element Methods: Application to Breast MR Images. *IEEE Transactions on Medical Imaging*, 22(2):238–247, 2003.

[16] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *Proc Conf on Computer Vision and Pattern Recognition (CVPR)*, pages 519–526, 2006.

[17] D. Skerl, B. Likar, and F. Pernus. Evaluation of Similarity Measures for Non-Rigid Registration. In J. P. W. Pluim, B. Likar, and F. A. Gerritsen, editors, *Workshop on Biomedical Image Registration*, volume LNCS 4057, pages 160–168. Springer Verlag, 2006.

[18] E. Suarez-Santana, R. Nebot, C.-F. Westin, and J. Ruiz-Alzola. Fast BlockMatching Registration with Entropy-based Similarity. *Insight Journal – ISC/NA-MIC Workshop on Open Science at MICCAI 2007*, 2007. available online @ http://hdl.handle.net/1926/549.

[19] J.-P. Thirion. Image matching as a diffusion process: An analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[20] M. Urschler, C. Zach, H. Ditt, and H. Bischof. Automatic Point Landmark Matching for Regularizing Nonlinear Intensity Registration: Application to Thoracic CT Images. In R. Larsen, M. Nielsen, and J. Sporring, editors, *Proc Intern Conf on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 4191 of *LNCS*, pages 710–717, Copenhagen, Denmark, 2006. Springer.

[21] B. C. Vemuri, J. Ye, Y. Chen, and C. M. Leonard. Image registration via level-set motion: Applications to atlas-based segmentation. *Medical Image Analysis*, 7(1):1–20, 2003.

[22] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Diffeomorphic Demons Using ITK's Finite Difference Solver Hierarchy. *Insight Journal – ISC/NA-MIC Workshop on Open Science at MICCAI 2007*, 2007. available online @ http://hdl.handle.net/1926/510.

[23] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Non-parametric Diffeomorphic Image Registration with the Demons Algorithm. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI'07)*, Brisbane, Australia, Oct 2007.

[24] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung. Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy. *Phys. Med. Biol.*, 50(12):2887–2905, 2005.

[25] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer Jr., R. M. Kessler, and R. J. Maciunas. Retrospective Intermodality Registration Techniques for Images of the Head: Surface-Based Versus Volume-Based. *IEEE Transactions on Medical Imaging*, 18(2):144–150, February 1999.

[26] R. P. Woods, S. T. Grafton, J. D. G. Watson, N. L. Sicotte, and J. C. Mazziotta. Automated Image Registration: II. Intersubject Validation of Linear and Nonlinear Models. *Journal of Computer Assisted Tomography*, 22(1):153–165, 1998.

[27] Z. Zhang, Y. Jiang, and H. Tsui. Consistent multi-modal non-rigid registration based on a variational approach. *Pattern Recognition Letters*, 27:715–725, 2006.