Available online at www.sciencedirect.com

ScienceDirect

Medical Image Analysis 12 (2008) 42–54

MEDICAL IMAGE ANALYSIS

www.elsevier.com/locate/media

A protocol for evaluation of similarity measures for non-rigid registration

Darko Škerl, Boštjan Likar, Franjo Pernuš *

University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia

Received 8 November 2006; received in revised form 11 May 2007; accepted 6 June 2007
Available online 22 June 2007

Abstract

In this paper, we present a protocol for the evaluation of similarity measures for non-rigid registration. The evaluation is based on five intuitive properties that characterize the behavior of a similarity measure, i.e. the accuracy, capture range, distinctiveness of the optimum, number of local minima, and risk of non-convergence. These five properties are estimated locally from similarity measure values that correspond to a range of systematic local free-form deformations, obtained by displacing control points in random directions from the gold standard position. Global similarity measure properties are obtained by combining the local properties over image regions or over the entire image. The feasibility of the proposed evaluation protocol is demonstrated for three similarity measures: mutual information, normalized mutual information and correlation ratio. The evaluation is carried out on a number of MR and CT images: a pair of simulated MR T1 and MR T2 images of the head, three pairs of real MR T1 and T2 images of the head, six pairs of real MR T1 and CT images of the head, and pairs of MR and CT images of three vertebrae. The protocol may help researchers to select the most appropriate similarity measure for a non-rigid registration task.

Keywords: Non-rigid registration; Evaluation; Similarity measures; Validation

1. Introduction

Medical image registration is important for integration of complementary information on patient anatomy and physiology, for longitudinal studies in which the state of an organ or a tissue is followed in time so as to monitor progression of disease or efficacy of treatment, for comparison of images of individuals to a digital atlas with known statistical properties, for image-guided therapy, and in many other applications. According to the nature of geometrical transformation, registration techniques can be divided into rigid and non-rigid ones. A rigid transformation is composed of rotations and translations, while a non-rigid transformation can be modeled by spline warps, truncated basis function expansions, Navier–Lamé equations, or by a viscous fluid model (Bookstein, 1989; Christensen et al., 1996; Unser, 1999).

Validation and comparison of medical image registration algorithms are two very important issues. Prerequisites for validation is standardization of validation methodology, which should include design of validation data sets, definition of corresponding "gold standard" or "ground truth" and its accuracy, validation protocol, and validation metrics. Two important registration validation projects have been conducted in the past, the "Retrospective Image Registration and Evaluation Project" (West et al., 1997) for evaluating multimodal rigid registration accuracy and the "Retrospective Evaluation of Inter-subject Brain Registration" (Hellier et al., 2003) for evaluating non-rigid registrations. The third project, "Non-rigid Image Registration Evaluation Project" is under way (Christensen et al., 2006). Evaluating the performance of non-rigid registration methods is a much more difficult task than the validation of rigid registrations. A few isolated markers or landmarks,

* Corresponding author. Tel.: +386 1 4768 248; fax: +386 1 4768 279.
E-mail address: franjo.pernus@fe.uni-lj.si (F. Pernuš).

1361-8415/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2007.06.001

that are sufficient for generating "ground truth" rigid transformations, are not sufficient to constitute the "ground truth" that would completely define the correspondence between all image points. Therefore, there is rarely if ever a "ground truth" correspondence map that would enable judging the performance of a non-rigid registration algorithm. As an alternative, a "pseudo gold standard" transformation can be simulated by a transformation that is modeled differently from the one that will be used in validation. The most common approach to non-rigid deformation simulation is to displace a set of points and interpolate the displacement map using splines (Rohr et al., 2003; Unser, 1999). The displaced landmarks can be anatomical, geometrical or simply corresponding to intersections of a regular/irregular grid superimposed on the image. The problem with "pseudo gold standards" is that simulated deformation fields are often not realistic enough. In case that a "gold standard" does not exist and deformation fields cannot be simulated well enough, validations must rely on real data and some metrics derived from the data. The evaluation metrics that have been proposed in the past include the intensity variance, obtained by registration of a population of images with a target image and averaging the intensities of the registered images, the inverse consistency which evaluates the registration performance based on desired transformation properties and transitivity which tells how the pair-wise registrations of the image population satisfy the transitivity property (Christensen et al., 2006; Christensen and Johnson, 2003). Another approach is to use richly annotated real images and the relative overlap of the segmentations alone (D'Agostino et al., 2006) or in conjunction with intensity variance, inverse consistence and transitivity as the metrics (Christensen et al., 2006). The problems with validations that are based on segmented images are that images have to be segmented, which is not a trivial task, that the metrics do not give results in millimeters, and that registration estimations are obtained for the segmented regions as a whole and not for unsegmented structures within these regions.

In intensity-based registrations, the accuracy and robustness of a rigid or non-rigid registration method is directly related to the behavior of the similarity measure as it measures the quality of agreement between the registered images. The behavior of the similarity measure depends on a number of factors, like the imaging modality, image content, image quality, spatial transformation and numerous implementation details. The complex interdependence of these factors makes the assessment of the influence of a particular factor on the similarity measure, and through it on the outcome of the registration, difficult. For rigid registrations, similarity measures may be evaluated by drawing plots or traces, showing their behavior when one image is systematically translated from and/or rotated around the "gold standard" registration position (Jenkinson and Smith, 2001; Maes et al., 2003). However, such an evaluation gives limited information on the behavior of the similarity measure because it is evaluated only at a very small fraction of the parameter space, which in case of non-rigid registration is large. Besides, the information obtained in this way is only qualitative.

Recently, we have presented a protocol for a more thorough optimization-independent evaluation of similarity measures for rigid registration (http://lit.fe.uni-lj.si/Evaluation) (Škerl et al., 2006). In this paper, we propose an extension of the protocol to evaluate similarity measures for non-rigid registrations as well. We use B-splines to model the geometrical deformations and numerous images of different modalities to demonstrate that the protocol is feasible for the evaluation of similarity measures for non-rigid registrations that are based on a set of regularly or irregularly distributed corresponding point pairs, for example, when spatial deformations are modeled by spline warps or by truncated basis functions.

2. Evaluation protocol

The proposed protocol for evaluation of local properties of a similarity measure is based on the assumption that local deformations can be simulated in one of the two registered images by systematically displacing a set of control points, for example, by using B-splines. Another reasonable assumption is that correspondences between sets of control points are provided by "gold standard" registrations. The idea behind the proposed evaluation protocol is to estimate some important properties of similarity measures locally by systematically displacing individual control points from the "gold standard" registration. In this way, local deformations are systematically simulated for an arbitrary number of random displacements in one of the two registered images, while the other image stays fixed. For each displacement of a point that generates a local deformation, a similarity measure value between the deformed and fixed image is calculated and these values are then used for local estimation of similarity measure properties. Details on the evaluation protocol are given in the following subsections.

2.1. Local deformation scheme

To evaluate the behavior of a similarity measure several features that characterize the behavior of a similarity measure as a function of the local image deformation have to be derived. The behavior of a similarity measure is a function of all the parameters of the spatial deformation model. In 3D, each control point can be displaced in x, y and z directions so that the number of parameters of the deformation model is three times the number of control points. As a result, the behavior of a similarity measure cannot be evaluated by an exhaustive search over all possible displacements in such a large parameter space. However, to derive the properties of a similarity measure for a single control point, all the other control points can be fixed, so that the similarity measure can be looked upon as a function of local displacements or deformations of a single anatomical region. In this way, local deformations can be

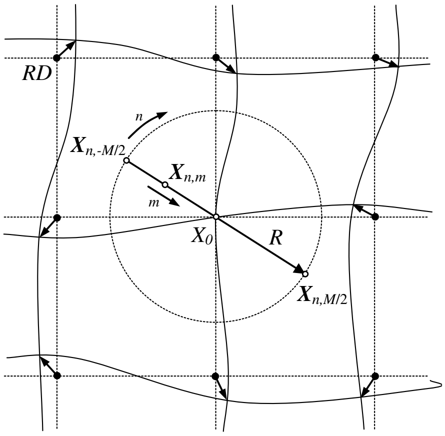

Fig. 1. Systematic displacements of the central control point from $\mathbf{X}_{n,-M/2}$ to $\mathbf{X}_{n,M/2}$ and displacements of the neighboring control points by RD in random directions.

simulated separately for each of the control points in the 3-dimensional parameter space defined by x, y and z directions. Thereby, the evaluation of similarity measure for non-rigid registration becomes more tractable but still practically infeasible by an exhaustive search over all possible local displacements, especially if perturbations of the neighboring control points is, and often should be, considered to simulate more realistic local deformations. To solve this demanding evaluation problem in a practically feasible manner, we propose a statistical evaluation protocol. The protocol is based on systematic simulation of random local deformations of one of the registered images, computing the corresponding similarity measure values between the deformed and fixed image, and statistical estimation of local similarity measure properties.

Let the "gold standard" position of each control point define the origin $\mathbf{X}_0$ of the 3-dimensional parameter space (x, y and z) of displacements related to that point and let SM($\mathbf{X}$) be the value of a similarity measure corresponding to image deformation generated by displacing a control point from the origin to some point $\mathbf{X}$; $\mathbf{X} = [x, y, z]$ within the 3D parameter space. Control points are systematically displaced to M points evenly spaced along N probing lines. Each probing line is defined by a randomly selected starting position $\mathbf{X}_{n,-M/2}$ at a distance R from the origin $\mathbf{X}_0$ and its mirror point $\mathbf{X}_{n,M/2}$. Fig. 1 illustrates the displacement of a control point in two dimensions. For each displacement, a similarity measure value SM($\mathbf{X}_{n,m}$) is computed. Similarity measure values at M points along a probing line define one similarity measure profile. To simulate a more realistic local deformation, all the other control points are displaced for a predefined distance RD in independent random directions. These are repeatedly randomized before computing each of the N similarity measure profiles.

The procedure for systematical displacement of each of the control points and computation of corresponding similarity measure profiles SM($\mathbf{X}_{n,m}$) is described by the following pseudo code:

```
For each control point {
   For each probing line ($n = 1$ to $N$) {
      – Randomly select the direction of the probing line
        ($\mathbf{X}_{n,-M/2} - \mathbf{X}_{n,M/2}$)
      – Randomly select displacement directions for all
        other control points
      – Displace all other control points for distance $RD$
        in the selected directions
      For each point on the probing line ($m = -M/2$ to
      $M/2$) {
         – Displace the control point to the position $\mathbf{X}_{n,m}$
         – Deform the floating image according to the
           positions of all control points
         – Compute the similarity measure value
           SM($\mathbf{X}_{n,m}$) between the floating and the target image
      }
   }
}
```

In this way, the problem of assessing the behavior of a similarity measure within a large parameter space is addressed by sampling similarity measure profiles along an arbitrary number of randomly selected lines. Each profile contains some information on the local behavior of a similarity measure. The behavior of a similarity measure can be estimated already from a few profiles and then improved by sampling an arbitrary number of additional profiles.

2.2. Estimation of similarity measure properties

Similarity measure properties are estimated from all N similarity measure profiles SM($\mathbf{X}_{n,m}$). All similarity measure values are normalized to the interval $[0, 1]$:

$$\mathrm{SM}(\mathbf{X}_{n,m}) \leftarrow \frac{\mathrm{SM}(\mathbf{X}_{n,m}) - \mathrm{SM}_{\min}}{\mathrm{SM}_{\max} - \mathrm{SM}_{\min}} \tag{1}$$

where $\mathrm{SM}_{\min}$ and $\mathrm{SM}_{\max}$ are the minimal and maximal values of $NM + 1$ similarity measure values before normalization, respectively. Let $X_{n,\max}$; $\max \in \{-M/2, \ldots, 0, \ldots, M/2\}$, be the position and SM($\mathbf{X}_{n,\max}$) the value of the global maximum of the similarity measure along line n, and let $\mathbf{X}_{n,\mathrm{loc}}$ be the position of the minimum closest to $\mathbf{X}_{n,\max}$. Looking from the global maximum $\mathbf{X}_{n,\max}$ outwards, let $d_{n,m}$ be the positive similarity measure gradient:

$$d_{n,m} = \begin{cases} \mathrm{SM}(\mathbf{X}_{n,m-1}) - \mathrm{SM}(\mathbf{X}_{n,m}) & \text{if } \mathrm{SM}(\mathbf{X}_{n,m-1}) > \mathrm{SM}(\mathbf{X}_{n,m}) \quad \text{and for all } m \text{ before global maximum} \\ \mathrm{SM}(\mathbf{X}_{n,m+1}) - \mathrm{SM}(\mathbf{X}_{n,m}) & \text{if } \mathrm{SM}(\mathbf{X}_{n,m+1}) > \mathrm{SM}(\mathbf{X}_{n,m}) \quad \text{and for all } m \text{ after global maximum} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Several local properties of a similarity measure can be computed for each control point from the obtained similarity measure profiles $\mathrm{SM}(\mathbf{X}_{n,m})$. In this paper, we will compute five local properties as proposed for similarity measure evaluation for rigid registration in Škerl et al. (2006) and as implemented in the Internet application: http://lit.fe.uni-lj.si/Evaluation. Briefly, the five properties are:

1. Accuracy ACC of a similarity measure is defined as the root mean square of distances between the origin $\mathbf{X}_0$ and the positions $\mathbf{X}_{n,\max}$, $n = 1, 2, \ldots, N$ along the N lines, where the $\mathrm{SM}(\mathbf{X})$ reaches a global maximum.
$$\mathrm{ACC} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\|\mathbf{X}_{n,\max} - \mathbf{X}_0\|^2}\ [\mathrm{mm}] \tag{3}$$
The accuracy ACC actually estimates the mean error between the true gold standard position and the positions of the global maxima exhibited by the similarity measure along the N random probing lines.
2. Distinctiveness of optimum $\mathrm{DO}(r)$ is the estimation of the uncertainty of the location of the global maximum, which is expressed by the average change of the similarity measure value at a distance r from the global maximum:
$$\mathrm{DO}(r) = \frac{1}{2rN}\sum_{n=1}^{N}[2\mathrm{SM}(\mathbf{X}_{n,\max}) - \mathrm{SM}(\mathbf{X}_{n,\max - r/\sigma}) - \mathrm{SM}(\mathbf{X}_{n,\max + r/\sigma})]\ [\mathrm{mm}^{-1}] \tag{4}$$
where $\sigma = 2R/M$, represents the distance between two consecutive points along a probing line. The property DO thus estimates the behavior of the similarity measure in the neighborhood of the global maximum. The similarity measure may rise steeply when approaching the global maximum and then fall quickly after reaching the maximum, causing a sharp (distinctive) peak at the global maximum. Alternatively, a similarity measure can have a flatter maximum, which is reflected in a lower DO value.
3. Capture range CR is defined as the smallest of the N distances between the global maxima and their respective closest minima encountered in the N random probing lines:
$$\mathrm{CR} = \min_n \|\mathbf{X}_{n,\max} - \mathbf{X}_{n,\mathrm{loc}}\|\ [\mathrm{mm}] \tag{5}$$
As such, CR represents the distance at which a local optimization method may converge to a local, instead to the global maximum. This property thus estimates the range of displacements from the true maximum from which the registration process is expected to converge, regardless of the optimization method utilized.
4. Number of minima $\mathrm{NOM}(r)$ is the number of minima of the similarity measure along the N random probing lines and within a certain distance r from the global maximum. $\mathrm{NOM}(r)$ is the cumulative number of minima as a function of distance r. In contrast to CR, which estimates the worst-case scenario by the closest minimum, $\mathrm{NOM}(r)$ is a more integral property, estimating the distribution of similarity measure minima around the global maximum.
5. Risk of non-convergence $\mathrm{RON}(r)$ is another more integral similarity measure property that, besides the number of minima given by the property $\mathrm{NOM}(r)$, estimates the extent or the magnitude of minima within distance r from the global maximum. $\mathrm{RON}(r)$ is defined as the average of all positive gradients $d_{n,m}$ within distance r from the global maximum:
$$\mathrm{RON}(r) = \frac{1}{2rN}\sum_{n=1}^{N}\sum_{m=\max - r/\sigma}^{\max + r/\sigma} d_{n,m}\ [\mathrm{mm}^{-1}] \tag{6}$$
A positive gradient $d_{n,m}$ indicates the increase of similarity measure value, looking from the global maximum outwards, and estimates the risk of convergence to a local maximum. A large value of $\mathrm{RON}(r)$ indicates that a similarity measure has distinctive and/or broader local maxima, which represent a risk for non-convergance to the true global maximum.

The first two properties describe the behavior of a similarity measure close to the "gold standard", while the last three properties estimate the robustness or convergence properties of a similarity measure. The better is a similarity measure, the smaller are the values of the accuracy, number of minima, and risk of non-convergence, and the larger the capture range and distinctiveness of optimum.

All similarity measure properties are statistical estimations, derived from the "gold standard" position $\mathbf{X}_0$, similarity measure values $\mathrm{SM}(\mathbf{X}_{n,m})$ and corresponding positive gradients $d_{n,m}$. As such, the values of similarity measure properties can be considered as random variables with certain variabilities that depend on the number N of similarity measure profiles $\mathrm{SM}(\mathbf{X}_{n,m})$. Since the probability distributions of the similarity measure properties are unknown and depend on specific similarity measures and analyzed images, the exact dependencies of similarity measure properties on the size of N is difficult to asses theoretically. In practice, however, the higher the N, the better the estimation of similarity measure properties. Besides, when comparing different similarity measures or the effects of different implementation parameters, the same randomization scheme can and should be used in all the experiments. In this way, the variability of statistical estimation and its dependence on the number of profiles N can be further reduced.

3. Experiments and results

3.1. Experimental datasets

Four sets of images (Fig. 2) were used in the experiments:

Set 1 One pair of MR T1 and T2 images of the head from McGill BrainWeb simulator (Collins et al., 1998).

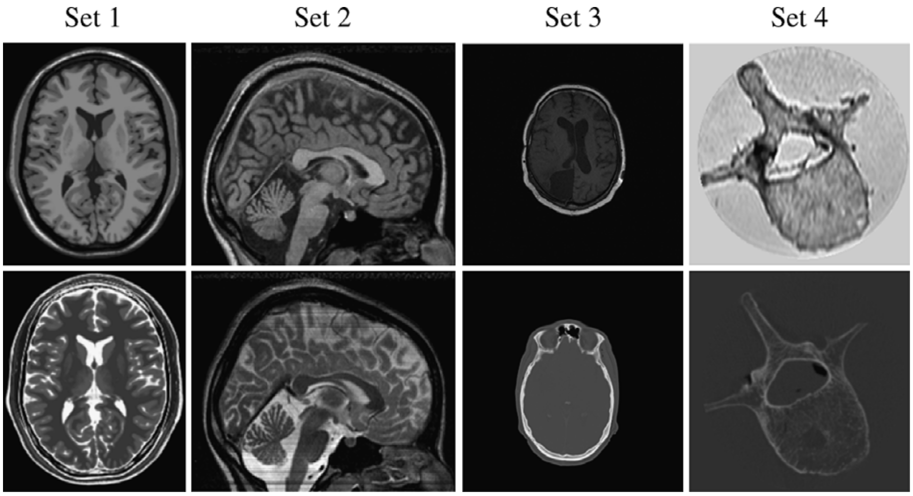

Fig. 2. Slices of the images from Sets 1 to 4.

The images were 217 × 181 × 181 voxels in dimension. Voxel sizes were $1 \times 1 \times 1\ \mathrm{mm}^3$. The two images contained 3% noise, no intensity inhomogeneities and no tumors. The "gold standard" registration for this set was inherently involved in the simulated images.

Set 2 Three MR T1 and T2 images of the head from ICBM database (http://www.loni.ucla.edu/ICBM/). The images were 217 × 181 × 181 voxels in dimension, with voxel sizes of $1 \times 1 \times 1\ \mathrm{mm}^3$. The "gold standard" registrations for this set of real images were obtained by stereotactic frames.

Set 3 Three head image sets of three patients (001, 002 and 003) from the Vanderbilt University RIRE project (http://www.vuse.vanderbilt.edu/~image/registration/), each set comprising MR T1, MR T1 rectified and CT images. Dimensions of the MR images were around 256 × 256 × 23 voxels while dimensions of the CT images were around 512 × 512 × 30 voxels. The MR T1 rectified images were original MR T1 images corrected for spatial distortions (Chang and Fitzpatrick, 1992; Maurer et al., 1996). The "gold standard" registrations for this well-established registration database were obtained by fiducial markers.

Set 4 The last set was comprised of three MR images of vertebrae L2, L3 and L4 and a corresponding CT images (Tomažević et al., 2003). Dimensions of the MR volumes of interest were 238 × 238 × 22 voxels and 512 × 512 × 233 voxels of the CT images. The voxel sizes were $0.4 \times 0.4 \times 1.9\ \mathrm{mm}^3$ for the MR images and $0.3 \times 0.3 \times 1\ \mathrm{mm}^3$ for the CT images. MR images were corrected for intensity non-uniformities (Likar et al., 2001). Rigid "gold standard" registrations for this set of real images were obtained by fiducial marker registration (Tomažević et al., 2004).

3.2. Similarity measures

In this study, we have experimented with the following three similarity measures: the mutual information (MI) (Maes et al., 1997; Wells et al., 1996), normalized mutual information (NMI) (Studholme et al., 1999), and correlation ratio(Roche et al., 1998):

1. Mutual information (MI):
$$\mathrm{MI}(a,b) = H(a) + H(b) - H(a,b) \tag{7}$$
2. Normalized mutual information (NMI):
$$\mathrm{NMI}(a,b) = \frac{H(a) + H(b)}{H(a,b)} \tag{8}$$
3. Correlation ratio (COR):
$$\mathrm{COR}(a|b) = \frac{Var[E(a|b)]}{Var(a)} \tag{9}$$

All similarity measures were computed on 2D joint intensity histograms and corresponding joint probability distributions $p(\cdot)$, obtained by binning the intensity pairs of all the overlapping voxels from the floating a and reference image b. $H(\cdot)$ denotes the marginal or joint entropy: $H(\cdot) = \sum p(\cdot) logp(\cdot)$, $E(\cdot)$ is the mathematical expectation, and $Var(\cdot)$ the variance of the corresponding variables. Partial volume interpolation (Maes et al., 1997) was used to create the joint intensity histograms.

3.3. Deformation model

Numerous spatial deformation models based on regularly or irregularly distributed sets of corresponding points were proposed in the past. Most often, however, thin-plate splines (Bookstein, 1989) and B-splines (Rueckert et al., 1999; Unser, 1999) were used. We have implemented B-splines to model local deformations on a set of regularly distributed control points, named knots. In all the experiments that follow, 3rd degree B-splines and a 3D grid of 7 × 7 × 7 knots were used. Evaluation of the five properties of the three similarity measures was performed on the inner 5 × 5 × 5 knots (125 knots in total, 375 parameters of the deformation model), while the knots on the edges were fixed.

3.4. Implementation parameters

The parameters of the evaluation protocol R, N and M were set to 20 mm, 50 and 80, respectively, yielding the distance $\sigma = 0.5$ mm between two consecutive points along a line. Before computing the similarity measure profile along each of the N probing lines for each knot, all the other knots were randomly displaced from the "gold standard" position for a distance $RD = 1$ mm by which a more realistic simulation of the registration process was achieved (Fig. 3). In all figures and tables, NOM stands for NOM(R)/N and RON stands for RON(R), while DO and RON are given in 10^{-3}/mm and 10^{-6}/mm, respectively.

3.5. Parameter variation

In this section, the sensitivities of the similarity measure properties to the variations of the parameters of the evaluation protocol are studied. Fig. 4 shows how four properties (ACC, DO, NOM and RON) of the mutual information typically change when the number of random lines N is increased. As expected, the similarity measure properties varied more when the number of random lines N was small. The experiments showed that similarity measure properties could be well estimated even with a relatively small value of $N \approx 30$. The value of $N = 50$, used in this paper, yielded relatively accurate estimations of similarity measure properties. The variabilities were further reduced by applying the same randomization scheme in all comparative evaluations.

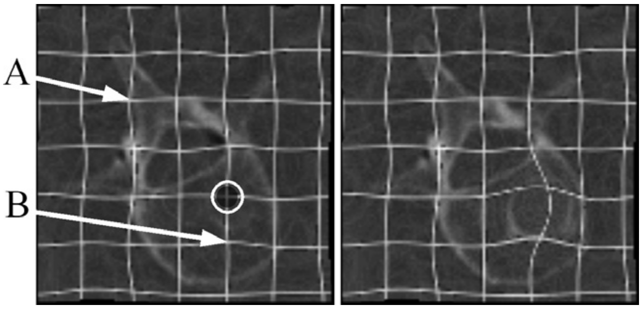

Fig. 3. Left: deformed image for all knots displaced for $RD = 1$ mm, except for the analyzed knot in a circle, which was fixed. Right: the same image with the analyzed knot displaced for 10 mm. The points (A, B) illustrate the knots for which local results are given in Table 4.

Fig. 5 illustrates how the accuracy (ACC) of mutual information (MI) and correlation ration (COR) depends on the number of points M along a probing line. The accuracy of both similarity measures was almost the same when changing M for ±20 around the pre-selected value of 80.

Fig. 6 illustrates how the similarity measure properties ACC and RON depend on the random displacement RD. Large RD yielded lower accuracies. On the other hand, the robustness, estimated by the risk of non-convergence (RON), was much less sensitive to the displacements of the other knots, as long as the displacements RD were within 5 mm. Therefore, small random displacements of other knots had a modest impact on the similarity measure properties, which can be explained by the fact that, for a given deformation model and parameter settings, small displacements of the neighboring points do not significantly locally deform the images around the analyzed knots. This deformation, induced from the neighboring knots, becomes more prominent for $RD > 5$ mm and reflects in a worse behavior of all similarity measures. In comparison to MI and NMI, COR similarity measure is obviously more sensitive to the induced deformation, i.e. to the misregistration of the neighboring points.

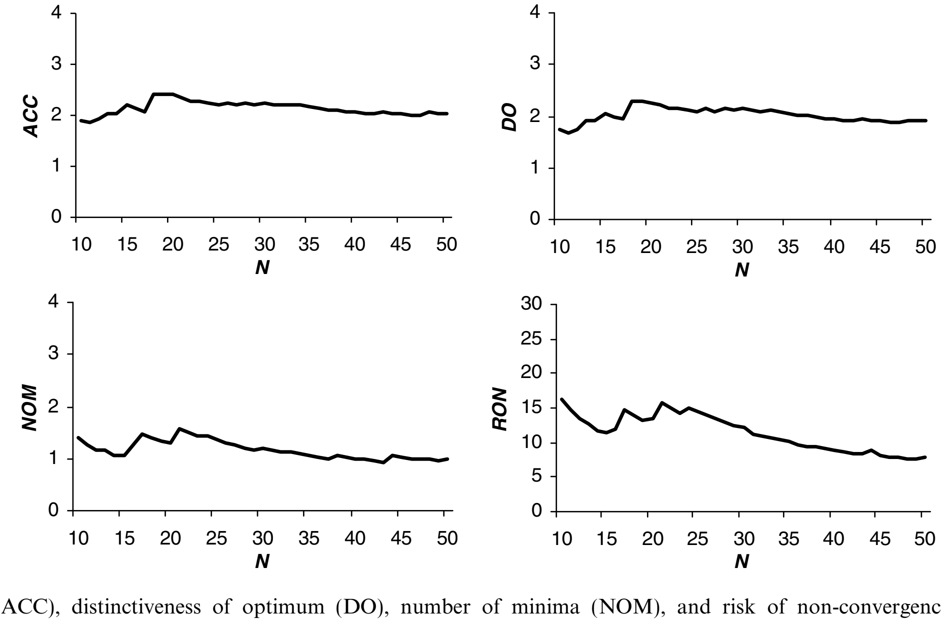

Fig. 4. The accuracy (ACC), distinctiveness of optimum (DO), number of minima (NOM), and risk of non-convergence (RON) of the mutual information, given as a function of the number N of probing lines (for images from Set 3, patient 001).

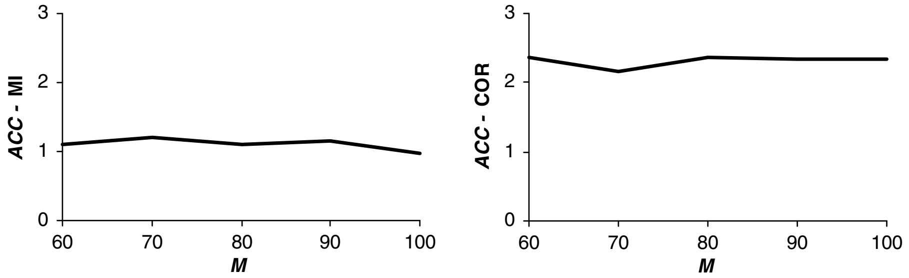

Fig. 5. The accuracy (ACC) of the mutual information (left) and correlation ratio (right), given as a function of the number M of points along a probing line (for images from Set 3, patient 002).

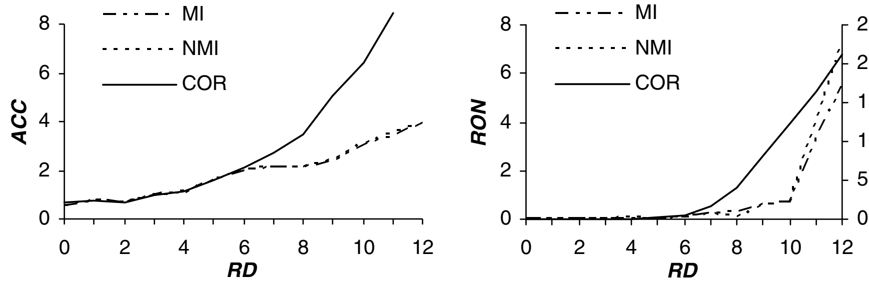

Fig. 6. ACC and RON as a function of displacement RD (for images from Set 3, patient 001). The scale for RON of COR is given on the right side of the right chart.

The last parameter R, representing the maximum local displacement of control points from the "gold standard", has no impact on the first two properties ACC and DO, which describe similarity measure properties close to the "gold standard". The third property CR, estimating the worst-case scenario by the closest minimum, is also independent on parameter R as long as R is larger than the distance to closest local minimum. For the last two similarity measure properties NOM(r) and RON(r), which are actually the cumulative functions of the displacement r from the "gold standard", the parameter R defines the domain of the two property functions NOM(r) and RON(r). As such, these two properties can be assessed at any distance r; $r < R$, independently on the parameter R. On the other hand, given as NOM(R) and RON(R), the absolute values of the two properties directly depend on the parameter R, which is a desired property that enables relative comparisons of different similarity measures within a maximum pre-selected domain R.

Table 1
Means and standard deviations of properties of the three similarity measures for MR T1 and MR T2 images of the head from Set 1

	MR T1–MR T2		
	MI	NMI	COR
ACC	**0.7** (0.2)	**0.7** (0.1)	1.3 (0.4)
DO	**2.8** (0.6)	**2.8** (0.6)	1.3 (0.8)
CR	**15.8** (5.8)	**15.8** (5.8)	10.1 (9.0)
NOM	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
RON	**0.0** (0.1)	**0.0** (0.1)	0.1 (0.1)

The values in bold indicate the best similarity measure with respect to the corresponding property.

Table 2
Means and standard deviations of properties of the three similarity measures for non-rigid registration of three pairs of MR T1 and MR T2 images of the head from Set 2

	MR T1–MR T2		
	MI	NMI	COR
ACC	**2.8** (1.0)	2.9 (1.1)	7.6 (4.1)
DO	1.7 (0.7)	1.7 (0.6)	**2.5** (2.3)
CR	**0.6** (1.7)	0.5 (0.2)	0.5 (0.2)
NOM	2.0 (2.1)	2.0 (2.1)	**1.8** (1.3)
RON	**133** (153)	**133** (149)	1222 (1577)

The values in bold denote the best similarity measure with respect to the corresponding property.

3.6. Results

For each image pair in the four image datasets, values of the five properties of the three similarity measures were determined for each of the 125 control points. Tables 1–4 give the means and standard deviations of the 125 values of the five properties and the three similarity measures. In this way, global estimations of the similarity measure properties, facilitating the comparisons between different similarity measures, were provided. The values of the similarity measure properties at each control point were interpolated over the entire image domains and scaled to image intensities for better visualization of properties (Figs. 7–10). Outlines of some anatomical structures were superimposed

Table 3
Means and standard deviations of the three similarity measures for non-rigid registration of three sets of MR T1, MR T1 rectified and CT images of the head from Set 3

	MR T1–CT					
	MI		NMI		COR	
	Original	Rectified	Original	Rectified	Original	Rectified
ACC	2.0 (0.8)	**1.4** (0.8)	2.2 (0.9)	**1.5** (1.0)	**4.4** (2.1)	4.5 (2.3)
DO	2.2 (0.6)	**2.3** (0.5)	2.2 (0.5)	**2.4** (0.6)	**1.9** (1.1)	1.3 (0.5)
CR	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.2 (0.0)	0.5 (1.4)	**0.6** (1.9)
NOM	1.0 (1.0)	**0.8** (1.0)	0.9 (0.8)	**0.8** (0.9)	0.9 (1.2)	**0.5** (0.4)
RON	12.5 (26.7)	**11.7** (19.9)	11.1 (14.9)	**11.0** (20.0)	144.8 (267.9)	**56.1** (135.7)

The values in bold denote whether a similarity measure performed better on original or on rectified MR T1 images when registered to CT images.

Table 4
Properties of the three similarity measures for three sets of MR and CT images of vertebrae

	MI			NMI			COR		
	L2	L3	L4	L2	L3	L4	L2	L3	L4
Point A (spinous process): MR–CT									
ACC	1.2	**1.4**	0.6	**1.1**	**1.4**	**0.5**	2.5	2.4	2.2
DO	**3.0**	**3.1**	5.2	2.8	**3.1**	**5.4**	2.1	1.4	2.9
CR	**1.0**	0.5	0.4	**1.0**	**0.9**	**7.0**	0.5	0.5	0.4
NOM	**0.1**	0.1	**0.1**	**0.1**	**0.0**	**0.1**	2.1	0.8	2.8
RON	**2.0**	6.1	**6.1**	2.8	**4.3**	7.2	1159.1	294.7	914.7
Point B (vertebral body): MR–CT									
ACC	2.5	6.7	2.2	**2.3**	**6.0**	1.9	3.0	7.0	**1.0**
DO	2.1	**3.4**	3.5	**2.4**	3.0	**3.8**	1.2	1.2	2.2
CR	**0.5**	**0.5**	**0.4**	**0.5**	**0.5**	**0.4**	**0.5**	**0.5**	**0.4**
NOM	**3.4**	3.8	2.7	3.6	4.3	**2.6**	4.6	**2.8**	2.8
RON	1139.2	2376.3	1302.2	1093.5	1901.8	1129.4	**479.0**	**631.3**	**377.2**

The values in bold denote the best similarity measure for each vertebra with respect to the corresponding similarity measure property.

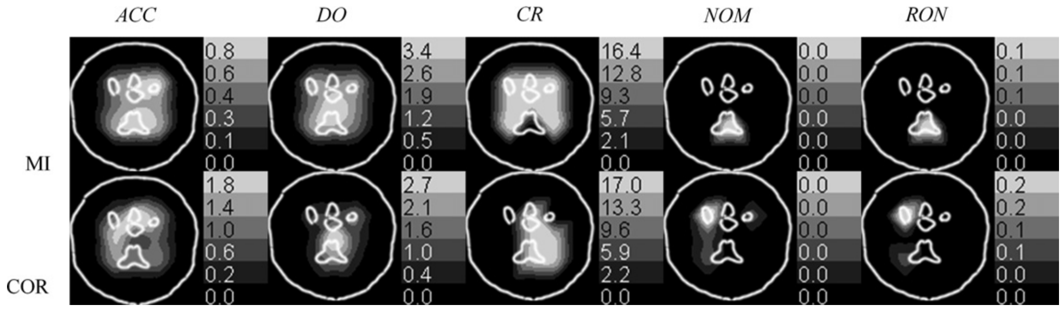

Fig. 7. Values of five properties of MI and COR for MR T1 and T2 images from Set 1.

on the images of the similarity measure properties so that property values could be better related to the underlying anatomy. In Figs. 7–10 the scales of property values are given on the right side of corresponding images.

Fig. 7 shows the five properties of MI and COR for non-rigid registration of artificially generated MR T1 and T2 images of a head from Set 1. The images of property values indicate that MI performed better than COR both in terms of accuracy and robustness. Very similar conclusions can be drawn from Table 1, which gives the means and standard deviations of MI, NMI and COR. All similarity measures were relatively accurate (0.7–1.3 mm), had relatively large capture ranges (10–15 mm) and contained almost no local minima. MI and NMI indicated similar performances, while COR performed less well. However, Fig. 7 shows large spatial variability between local similarity measure properties, indicating that accuracy and convergence of the non-rigid registration are expected to depend heavily on the underlying anatomical structures and on the similarity measures utilized.

Fig. 8 shows the five properties of MI for non-rigid registration of three sets of real MR T1 and MR T2 images from Set 2. The results reveal that mutual information was less accurate and robust in the parietal lobe area. Robustness was also poor in the frontal lobe area. Spatial distributions of the similarity measure properties were

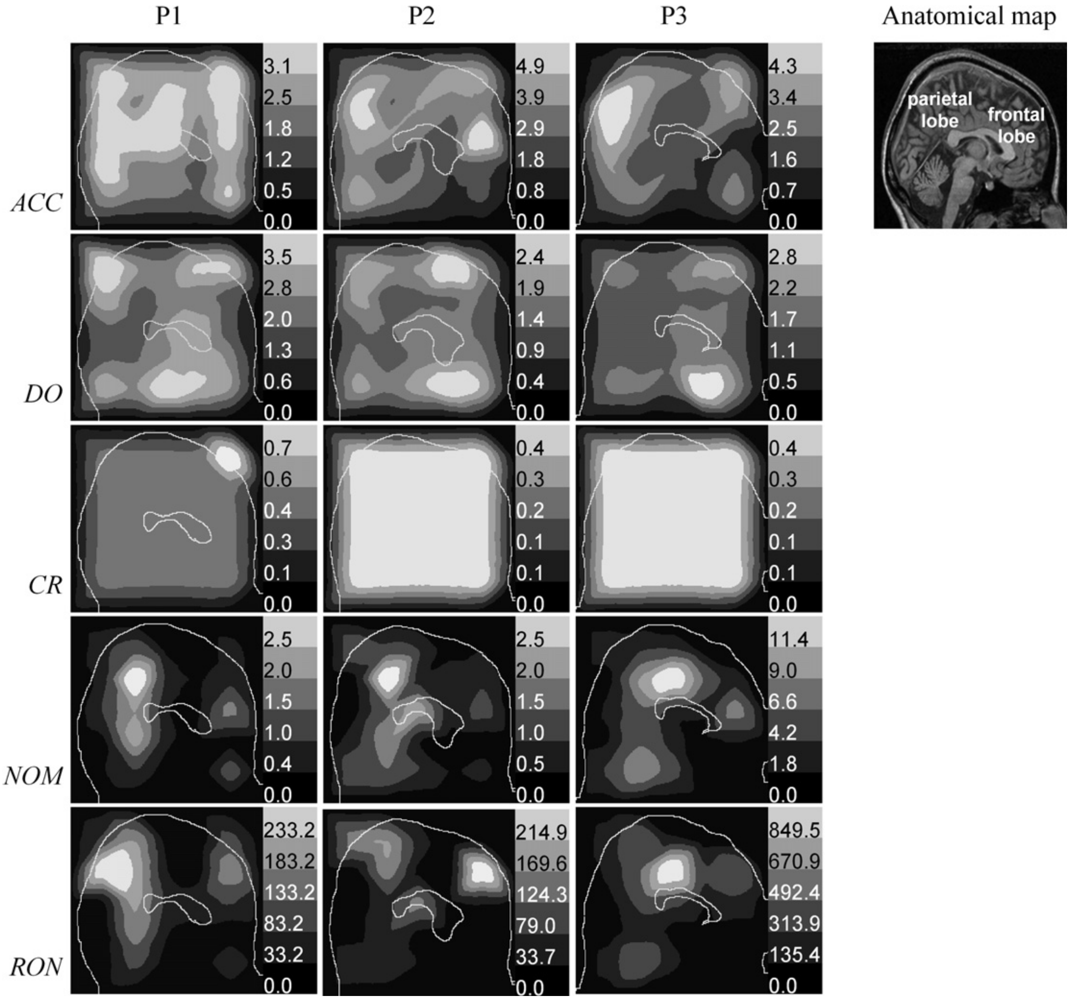

Fig. 8. Values of five properties (from top to bottom) of MI for non-rigid registration of three pairs (P1–P3) of real MR T1 and MR T2 images of the head from Set 2.

relatively consistent among the anatomies of the three patients (P1–P3). Table 2 gives the means and standard deviations of the similarity measure properties for the head images from Set 2. In general, COR yielded smaller accuracy and robustness than MI and NMI, while the differences between MI and NMI were again very small.

Fig. 9 shows how the properties of mutual information (left two columns) and correlation ratio (right two columns) for images from Set 3, patient 002 were spatially distributed. The first and the third columns correspond to the registration of original MR T1 and CT images, while the second and fourth columns correspond to the registration of rectified MR T1 and CT images. It can be seen that the two measures behaved quite differently in different parts of the brain, while the performances on original and rectified images MR images were comparable. The differences between MI and COR similarity measures are especially notable for the ACC, DO and RON properties, which differ both in spatial distribution and in absolute values, while the distributions of CR and NOM properties were similar. Altogether, the MI similarity measure performed better than COR. Both similarity measures yielded better performances in the frontal than in the parietal lobe, which could be attributed to the presence of pathology in this region.

Table 3 shows the global properties of the three similarity measures. The results indicate that MI and NMI were more accurate and had better convergence properties when rectified instead of non-rectified MR images were used as the floating ones. For the COR, better convergence properties were obtained on rectified MR images, while the accuracy was slightly better on the original than on rectified MR images.

Fig. 10 shows the spatial distributions of the five similarity measure properties of MI for vertebrae L2–L4 from Set 4. For all three vertebrae, mutual information was more accurate and robust in the spinous process and in the lamina of the vertebral arch than in the area of the vertebral bodies. In addition, the optimum of MI was more distinct (DO) and the measure was more robust (small NOM and RON) in the spinous process and lamina. Again, the results indicate that the spatial distribution of similarity measure properties were relatively highly consistent among the different anatomies of the three vertebrae (L2–L4).

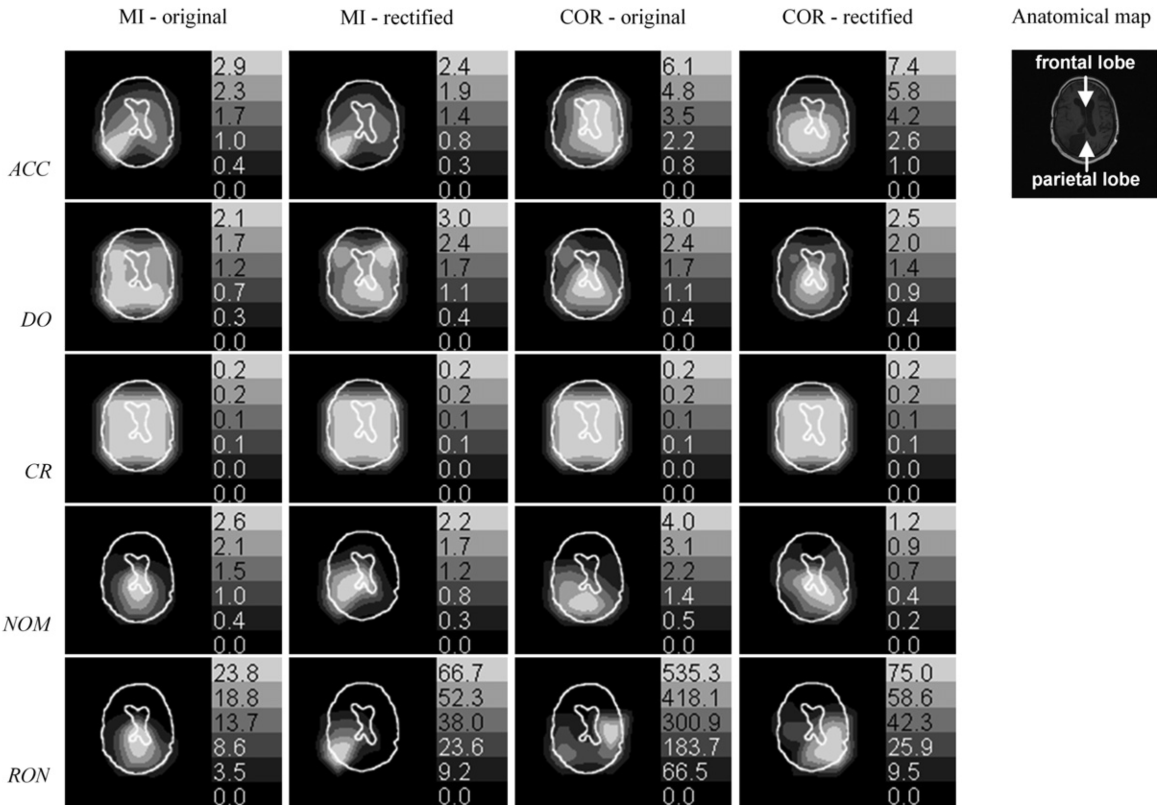

Fig. 9. Values of five properties (from top to bottom) of MI (left two columns) and COR (right two columns) for non-rigid registration of MR T1 and MR T1 rectified image to CT image of patient 002 from Set 3.

Table 4 shows the results of the evaluation of the three similarity measures for non-rigid registration of three sets of MR and CT images of the three vertebrae. Local comparison of the similarity measures is given for points A and B, indicated in Fig. 3. Both MI and NMI were accurate and robust around the spinous process and lamina (Point A). In the body of the vertebra (Point B) MI and NMI were less accurate and less robust, i.e. had larger NOM and RON values, while COR was more robust in the vertebral body area. Global similarity measure properties for the images of the vertebrae from the Set 4 are shown in Table 5. MI and NMI behaved better than COR as all five properties indicated, while the differences between MI and NMI were again very subtle.

4. Discussion

The behavior of a similarity measure is strongly influenced by the imaging modality, image content and differences in image content, geometric transformation, partial image overlap, implementation details and image degradations, such as noise, intensity inhomogeneities and geometrical distortions. Knowing the local and global behavior of similarity measures under different circumstances is important as it might help to select the best similarity measure for a given application and type of images and to predict the outcome of the registration. To obtain valuable information on the behavior of similarity measures for non-rigid registrations we have devised a protocol partially based on the protocol for evaluating similarity measures for rigid registration (Škerl et al., 2006). The protocol for evaluating similarity measures for rigid registrations requires that accurate "gold standard" registrations of image pairs, characteristic for the application, are available. This protocol could not be simply extended to evaluate similarity measures for non-rigid registrations because the parameter space of non-rigid registrations is too large compared to the six-dimensional parameter space of 3D rigid registrations and because "gold standards" for non-rigid registrations are far more difficult to obtain. These two problems were addressed by evaluating similarity measures at control points for which correspondences ("gold standards") could be identified. The points might be anatomical or geometrical landmarks, or simply points regularly or irregularly distributed over the image domain. The high parameter space, characteristic for non-rigid registrations, was avoided by creating local deformations by systematically displacing individual control points in only three dimensions. For each displacement of a point that generated a local B-splines based deformation, a similarity measure value was calculated and these values were used to locally estimate properties that characterize the behavior of a similarity measure at that point. Since local similarity measure properties are estimated for each control point, the properties can be interpolated over the entire image by the deformation model and shown as images of properties. On the

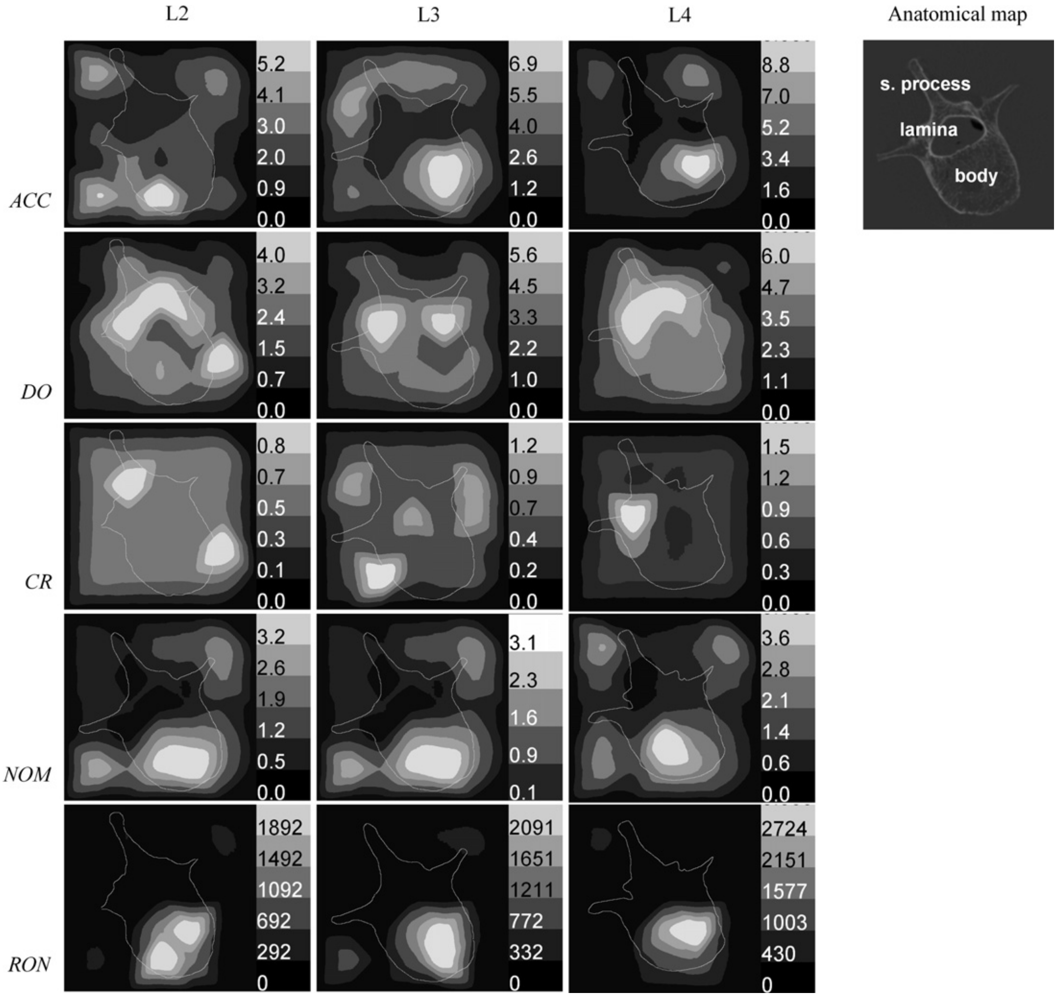

Fig. 10. Values of the five properties (from top to bottom) of MI for non-rigid registration of MR and CT images of vertebrae L2, L3 and L4 from Set 4.

Table 5
Means and standard deviations of the three similarity measures for three sets of MR and CT images of vertebrae from Set 4

	MR–CT		
	MI	NMI	COR
ACC	2.7 (1.7)	**2.6** (1.7)	4.5 (2.7)
DO	**2.8** (1.4)	2.8 (1.3)	1.9 (1.5)
CR	0.5 (0.2)	**0.7** (0.8)	0.5 (0.1)
NOM	**1.4** (1.1)	**1.4** (1.1)	2.9 (1.6)
RON	388 (686)	**326** (554)	641 (795)

The values in bold denote the best similarity measure with respect to the corresponding property.

other hand, more global similarity measure properties can be obtained by combining (e.g. averaging) the local properties over the entire image as in this paper, or over smaller regions that may represent manually of automatically outlined volumes of interest.

The deformation scheme used in this paper was based on popular *B*-splines, frequently utilized in a number of non-rigid registration tasks, being either intra-modality or inter-modality, intra-subject or inter-subject. The popularity of *B*-splines may be attributed to simple and general implementation, allowing free-form deformations of adjustable degree of freedom. The purpose of this paper was not to deal with deformation modeling, which is important and difficult problem per se, but to provide a relatively general tool for the evaluation of the behavior of a similarity measure for non-rigid registration based on control points. Nevertheless, the evaluation protocol or the basic idea of systematic local parameter perturbation and statistical evaluation might potentially be extended to other deformation schemes, for example, to those based on physical modeling. However, it is important to note that the selected deformation scheme should be as realistic as possible, preferably modeling only those deformations that can occur in real images. However, since no such specific models with minimal required degree of freedom exist, some redundant unrealistic deformations that are less likely to occur in real images will always be induced in the registration process during the optimization procedure. The proposed random local perturbation scheme generates a set of possible deformations that might occur during the registration. As such, the protocol enables similarity measure evaluation that is relevant for a given deformation model and that is independent of the optimization procedure.

The evaluation protocol has five parameters that have to be selected for a particular evaluation problem, namely, knot spacing, the maximum local displacement *R*, the distance *RD* of random displacements of the other but analyzed control point, the number *M* of sampling points along a probing line, and the number *N* of probing lines. The reasonable values of these parameters depend on the registration task, image properties, and/or on the temporal and computational limitations for the evaluation. The first parameter, knot spacing or the number of knots, depends on the required degree of freedom for the given registration task and is thus subjected to deformation modeling. The second parameter of the evaluation protocol is the maximum local displacement *R* of control points by which the extent of local deformation is selected. The value of this parameter should be chosen with respect to the expected local deformation that may occur between the images or that may be induced by the optimization procedure. To avoid folding of the image space, the selection of parameter *R* should also reflect the nature of deformation model, the knot spacing, and the distance *RD* of random displacements of other control points. As a rule of thumb, the parameter *RD* should not be larger than 10% of the knot spacing, while the parameter *R* should not be larger than half the knot spacing. Another parameter of the evaluation protocol is the number *M* of sampling points along a probing line, which should yield a distance $\sigma = 2R/M$ between two consecutive points along a line that is smaller than the smallest voxel dimension. A good choice of σ is half of the smallest voxel dimension, although smaller values of σ would yield better estimations of the registration accuracy but at additional computational costs. The last parameter of the evaluation protocol is the number *N* of probing lines. Since the similarity measure properties are statistical estimations the parameter *N* should be as high as possible. Nevertheless, similarity measure evaluation can be carried out already by just a few probing lines and then sequentially improved by arbitrary additional probing lines. This is an important and practical characteristic of the proposed evaluation protocol, enabling constant monitoring of the similarity measure properties during the long-term evaluations and thereby providing a valuable quantitative feedback to the developers of similarity measures, dealing with temporal and computational limitations. Another important issue, associated with the selection of all five parameters of the evaluation protocol, is the question of how parameter selection influences the evaluation results. It was demonstrated, as Figs. 4–6 show, that the similarity measure properties are relatively insensitive to the values of parameters *N*, *M*, and *RD* over a relatively large range, although some variability is always present in such statistical estimations. Nevertheless, the variability of statistical estimation can be further reduced by using the same randomization scheme and implementation parameters in comparative evaluations.

The feasibility of the proposed protocol was demonstrated for three similarity measures on a pair of simulated MR T1 and MR T2 images of the head, three pairs of real MR T1 and T2 images of the head, six pairs of real MR T1 and CT images of the head, and pairs of MR and CT images of three vertebrae. The non-rigidity in head images can reflect the non-rigidity of the soft tissue, e.g. the brain that is often the tissue of interest. Besides, the non-rigidity between MR and CT images of the same, although rigid, objects may emerge also from the imperfections of the imaging modalities, e.g. from spatial distortion of MR images that are typically much higher that those of CT images.

The evaluations of similarity measures on a 3.2 GHz Pentium IV computer required about two months of CPU processing time for image deformation, calculation of similarity measure values, and computing corresponding similarity measure properties. The obtained results indicated that similarity measure properties depend heavily on location, which was to be expected. The obtained spatial distributions of similarity measure properties were highly spatially consistent among different subjects. This confirms the repeatability and feasibility of the proposed evaluation protocol. Further, the obtained results for rectified (distortion corrected) MR images were better than for original MR images, indicating the consistence between the evaluation protocol and the rectification technique. Altogether, the differences between MI and NMI similarity measures were subtle. This was to be expected because the two measures are very similar, although the NMI is supposed to perform better if image overlap is changing heavily, i.e. if the intersection of floating and target image domains depends heavily on the transformation parameters and thereby hampers the estimation of image similarities. However, this adverse phenomenon is more prominent in global rigid registration, while in local non-rigid registration, image overlap is far less dependent on local transformation parameters, especially when edge knots are kept fixed. Nevertheless, MI yielded slightly better results for brain images and NMI for the images of vertebrae. On the other hand, the COR similarity measure was much less accurate and exhibited worse convergence properties in all of the analyzed images. Therefore, for multimodal image registration in which there is mainly statistical relation between the intensities of the images undergoing registration, information theoretic similarity measures are currently the measures of choice.

In conclusion, the proposed protocol may be a valuable tool for: (a) studying the behavior of a similarity measure locally and globally, (b) studying the influence of sampling, interpolation, histogram binning, partial image overlap, and image degradation, such as noise, intensity inhomogeneity, and geometrical distortions, (c) comparing different similarity measures, (d) comparing different implementations of the same measure, (e) evaluating the effects of certain implementation details on the behavior of the similarity measures and (f) comparing the behavior of the similarity measures for different image modalities and image contents, etc. The protocol was tested and proved useful for the evaluation of multimodal intra-subject registration. However, the evaluation of monomodal and, having an appropriate "gold standard", also the evaluation of inter-subject registration is feasible. As such, the protocol may help researchers, confronted with non-rigid registration tasks, to select the most appropriate similarity measures and to design efficient registration schemes.

Acknowledgements

The authors thank Dr. Louis Collins from McGill University for providing the images and gold standard transformations in Set 2. The images and gold standard transformations in Set 3 were provided as part of the "Retrospective Image Registration Evaluation" project, National Institutes of Health, Project Number 8R01EB00 2124-03, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN, USA. The authors also thank Ramya Balachandran and Ning Xu from Vanderbilt University, USA for their help with the gold standard. This work was supported by the Ministry of Higher Education, Science and Technology, Republic of Slovenia under the Grants P2-0232, L2-7381, and L2-9758.

References

Bookstein, F.L., 1989. Thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 567–585.

Chang, H., Fitzpatrick, J.M., 1992. A technique for accurate magnetic resonance imaging in the presence of field inhomogeneities. IEEE Transactions of Image Processing 11 (3), 319–329.

Christensen, G.E., Geng, X., Kuhl, J.G., Bruss, J., Grabowski, T.J., Pirwani, I.A., Vannier, M.W., Allen, J.S., Damasio, H., 2006. Introduction to the Non-rigid Image Registration Evaluation Project (NIREP). Lecture Notes in Computer Science 4057, 128–135.

Christensen, G.E., Johnson, H.J., 2003. Invertibility and transitivity analysis for nonrigid image registration. Journal of Electronic Imaging 12 (1), 106–117.

Christensen, G.E., Rabbitt, R.D., Miller, M.I., 1996. Deformable templates using large deformation kinematics. IEEE Transactions on Image Processing 5 (10), 1147–1435.

Collins, D.L., Zijdenbos, A.P., Kollokian, V., Sled, J.G., Kabani, N.J., Holmes, C.J., Evans, A.C., 1998. Design and construction of a realistic digital brain phantom. IEEE Trans Med Imaging 17 (3), 463–468.

D'Agostino, E., Maes, F., Vandermeulen, D., Suetens, P., 2006. An information theoretic approach for non-rigid image registration using voxel class probabilities. Medical Image Analysis 10 (3), 413–431.

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher, G., Collins, D.L., Evans, A., Malandain, G., Ayache, N., Christensen, G.E., Johnson, H.J., 2003. Retrospective evaluation of intersubject brain registration. IEEE Transactions of Medical Imaging 22 (9), 1120–1130.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Medical Image Analysis 5 (2), 143–156.

Likar, B., Viergever, M.A., Pernus, F., 2001. Retrospective correction of MR intensity inhomogeneity by information minimization. IEEE Transactions of Medical Imaging 20 (12), 1398–1410.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE Transactions of Medical Imaging 16 (2), 187–198.

Maes, F., Vandermeulen, D., Suetens, P., 2003. Medical image registration using mutual information. IEEE Transactions of Medical Imaging 91, 1699–1722.

Maurer Jr., C.R., Aboutanos, G.B., Dawant, B.M., Gadamsetty, S., Margolin, R.A., Maciunas, R.J., Fitzpatrick, J.M., 1996. Effect of geometrical distortion correction in MR on image registration accuracy. Journal of Computed Assisted Tomography 20 (4), 666–679.

Roche, A., Malandain, G., Pennec, X., Ayache, N., 1998. The correlation ratio as a new similarity measure for multimodality image registration. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (Eds.), Medical Image Computing and Computer Assisted Intervention. Springer Verlag, Cambridge, MA, USA, pp. 1115–1124.

Rohr, K., Fornefett, M., Stiehl, H.S., 2003. Spline-based elastic image registration: integration of landmark errors and orientation attributes. Computer Vision And Image Understanding 90 (2), 153–168.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions of Medical Imaging 18 (8), 712–721.

Studholme, C., Hill, D.L., Hawkes, D.J., 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognition 32, 71–86.

Škerl, D., Likar, B., Pernuš, F., 2006. A protocol for evaluation of similarity measures for rigid registration. IEEE Transactions of Medical Imaging 25 (6), 779–791.

Tomaževič, D., Likar, B., Pernuš, F., 2004. "Gold standard" data for evaluation and comparison of 3D/2D registration methods. Computed Aided Surgery 9 (4), 137–144.

Tomaževič, D., Likar, B., Slivnik, T., Pernuš, F., 2003. 3-D/2-D registration of CT and MR to X-ray images. IEEE Transactions of Medical Imaging 22 (11), 1407–1416.

Unser, M., 1999. Splines: a perfect fit for signal and image processing. IEEE Signal Processing Magazine 16 (6), 22–38.

Wells 3rd., W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Medical Image Analysis 1 (1), 35–51.

West, J., Fitzpatrick, J.M., Wang, M.Y., Dawant, B.M., Maurer Jr., C.R., Kessler, R.M., Maciunas, R.J., Barillot, C., Lemoine, D., Collignon, A., Maes, F., Suetens, P., Vandermeulen, D., van den Elsen, P.A., Napel, S., Sumanaweera, T.S., Harkness, B., Hemler, P.F., Hill, D.L., Hawkes, D.J., Studholme, C., Maintz, J.B., Viergever, M.A., Malandain, G., Woods, R.P., et al., 1997. Comparison and evaluation of retrospective intermodality brain image registration techniques. Journal of Computed Assisted Tomography 21 (4), 554–566.