Validation of a Non-rigid Registration Algorithm for Multi-modal Data

Peter Rogelj[a], Stanislav Kovačič[a] and James C. Gee[b]

[a]Faculty of Electrical Engineering, University of Ljubljana, Slovenia
[b]Department of Radiology, University of Pennsylvania, Philadelphia, USA

ABSTRACT

We describe the evaluation of a non-rigid image registration method for multi-modal data. The evaluation is made difficult by the absence of gold standard test data, for which the true transformation from one image to another is known. Different approaches have been used to deal with this deficiency, e.g., by using synthetically warped data, by comparison of anatomic regions of interest identified either manually or automatically, and by direct comparison of the registered data. Each of these approaches are limited and in this paper, we illustrate some of the problems that arise based on their application to the evaluation of our multi-modal non-rigid registration method.

Keywords: Non-rigid registration, multi-modal, similarity measure, validation, brain, thoracic

1. INTRODUCTION

The object of medical image registration is to bring anatomically homologous points in different images into correspondence. In this way, the anatomy in one image is geometrically transformed into spatial alignment with that in a second image. One consequence is that the two images are made visually similar. A large number of registration methods exist,[1] and several schemes for their validation have been proposed,[2–4] including a highly successful effort on the evaluation of rigid registration techniques.[5] In contrast, the assessment of non-rigid image registration methods has been problematic.

The difficulty arises because anatomic correspondence can only be established for a very limited set of image pairs. In most situations, the imaged anatomies are different from one another (to a greater or lesser extent) and thus a unique anatomic alignment does not exist. Consequently, most investigators have considered gross parcellations of anatomy that are known to be shared across individuals, but the definition of these regions presents its own difficulties, which we discuss further in the text.

A second line of validation strategies for non-rigid registration methods focus on the evaluation of image correspondence. The implicit assumption is that improved image alignment reflects better anatomic registration, which may not be true. Nevertheless, since a majority of non-rigid registration algorithms only optimize the apparent similarity between the images of interest, direct comparisons of the registered data have been pursued.[6–9] Finally, evaluation of image as well as anatomic correspondence have been used with artificially transformed data, for which the true alignment transformation is available and synthetically generated. In this work, the role and limitations of each of these validation methods are discussed within the context of the evaluation of a non-rigid registration technique for multi-modal data.

Corresponding author:
E-mail: peter.rogelj@fe.uni-lj.si
Address: Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, 1001 Ljubljana, Slovenia

2. A SYSTEM FOR NON-RIGID MULTI-MODAL REGISTRATION

The results shown in this article are obtained by a non-rigid registration approach based on a new class of multi-modal similarity measures.[10] These similarity measures are constructed from information over the images but can be applied pointwise.

The basic scheme of our registration method is shown in Fig. 1. It consists of two stages: estimation of external forces $\boldsymbol{F}$ that drive the registration of image B to A, and application of a spatial deformation model that defines the relationship between these forces and the deformation of B, the latter described by a displacement field $\boldsymbol{U}$. Upon convergence of the method, the deformed version of B (now in register with A) is obtained by applying the displacements $\boldsymbol{U}$, as shown in Fig. 1 (b).
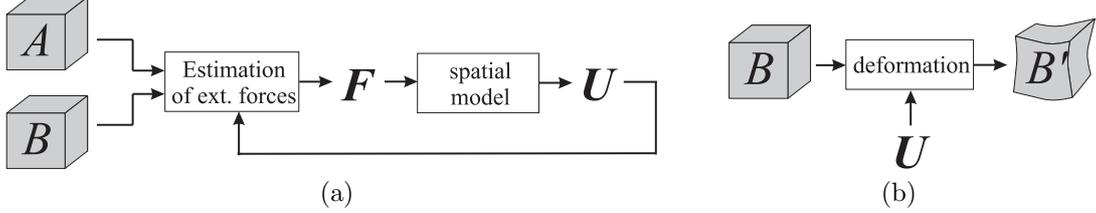
Figure 1. Multi-modal non-rigid registration scheme for matching source image B to target image A (a). Deformation of the source image B (b).

2.1. Estimation of External Forces

External forces $\boldsymbol{F}$ are estimated for each individual voxel of the source image. They consist of two terms: forward force $\boldsymbol{F}_F$ and reverse force $\boldsymbol{F}_R$, such that

$$\boldsymbol{F} = \boldsymbol{F}_F - \boldsymbol{F}_R. \tag{1}$$

Forward forces $\boldsymbol{F}_F$ are forces that drive the registration in a way of moving each point of source image to best match the reference. Reverse forces $\boldsymbol{F}_R$ try to move source image such that certain point of reference image matches the source image best. It is also possible to use either only forward or reverse forces, but using both makes registration more consistent.[8] In our case forces are calculated as follows:

$$\boldsymbol{F}_F(\mathbf{x}) = \left.\frac{\partial}{\partial \boldsymbol{d}}\right|_{d=0} S\left(A\left(\mathbf{x} + \boldsymbol{U}\left(\mathbf{x}\right) + \boldsymbol{d}\right), B\left(\mathbf{x}\right)\right), \tag{2}$$

$$\boldsymbol{F}_R(\mathbf{x}) = \left.\frac{\partial}{\partial \boldsymbol{d}}\right|_{d=0} S\left(A\left(\mathbf{x} + \boldsymbol{U}\left(\mathbf{x}\right)\right), B\left(\mathbf{x} + \boldsymbol{d}\right)\right), \tag{3}$$

where $\mathbf{x} = [x_1, x_2, x_3]^T$ denotes position of undeformed source image, and $S(a, b)$ measures the similarity between point a in the reference and point b in the source image.[10] Generally, this point-based similarity measure requires global knowledge, which is usually derived from global joint intensity distribution. In our experiments three different measures are used. The first one is entropy based:

$$S_H = \log\left(p\left(\mathbf{i}\right)\right), \tag{4}$$

where $\mathbf{i} = [i_A, i_B]^T$ is the intensity pair that consists of point intensities on both images, and $p(\mathbf{i})$ is its joint probability estimated from the whole images. The second similarity measure is based on segmentation[10]:

$$S_S = p(\mathcal{C}_S|\mathbf{i}), \tag{5}$$

where $p(\mathcal{C}_S|\mathbf{i})$ is probability that intensity pair $\mathbf{i}$ belongs to one of the intensity classes $C_i \in \mathcal{C}_S$, which correspond to correct tissue matching. The last measure is derived from previous ones by using each intensity pair as a different intensity class:

$$S_{UH} = \log\left(p\left(i_A|i_B\right) p\left(i_B|i_A\right)\right), \tag{6}$$

where $p(i_A|i_B)$ and $p(i_B|i_A)$ are conditional intensity probabilities.

2.2. Spatial Deformation Model

The design of spatial deformation model follows the idea that linear model can be separated into two parts:

$$\boldsymbol{U}_F = k_E \boldsymbol{F}, \tag{7}$$

$$\boldsymbol{U}(\mathbf{x}) = \int \boldsymbol{U}_F(\mathbf{x} - \mathbf{r})\mathbf{G}(\mathbf{r})d\mathbf{r} = (\boldsymbol{U}_F \otimes \mathbf{G})(\mathbf{x}) = ((k_E \boldsymbol{U}) \otimes \mathbf{G})(\mathbf{x}). \tag{8}$$

The first part (7) is Hooke's law, which says that the point moves proportionally to the applied force $\boldsymbol{F}$ for displacement $\boldsymbol{U}_F$. The second part (8) is a spatial convolution filter and models interdependence of points. The kernel for elastic media is described by Bro-Nielsen,[11] but for simplicity we use Gaussian or exponential filtering. Due to imperfect matches in previous iterations the forces $\boldsymbol{F}$ are applied again in the next iteration and therefore displacements $\boldsymbol{U}_F$ in each iteration are just the improvements of already obtained configuration.

By the principle of linearity final displacements can be calculated by summing up partial displacements,

$$\boldsymbol{U}^{(t)} = \boldsymbol{U}^{(t-1)} + \boldsymbol{U}_F^{(t)} \otimes \mathbf{G}, \tag{9}$$

where t denotes iteration number. This is the concept of incremental deformations[12] used to accommodate large nonlinear deformations, but because external forces $\boldsymbol{F}(x)$ are highly nonlinear function of displacements $\boldsymbol{U}(x)$,[13] the resulting deformation does not necessarily follow the selected (linear) spatial deformation model. In our case this is compensated by introducing point dependence on the accumulated displacements:

$$\boldsymbol{U}^{(t)} = (\boldsymbol{U}^{(t-1)} + \boldsymbol{U}_F^{(t)}) \otimes \mathbf{G} \tag{10}$$

In this way the displacements are filtered throughout the iteration process, such that old forces contribute less than later ones. The undesirable side effect of this filtering is that as the external forces go to zero, the image gradually returns back to its undeformed configuration. Thus, additional external forces are needed to sustain the deformed condition. Thus, we simply combine the two spatial deformation models (Fig. 2):

$$\boldsymbol{U}^{(t)} = (\boldsymbol{U}^{(t-1)} + \boldsymbol{U}_F^{(t)} \otimes \mathbf{G}_1) \otimes \mathbf{G}_2 \tag{11}$$

The first filter ($\mathbf{G}_1$) enables large deformations and precise registration while the second one ($\mathbf{G}_2$) serves to improve the linearity of the results. The coefficient k_e controls registration speed and changes during the registration for each point independently, to allow fast convergence as well as precise, though small registration steps.
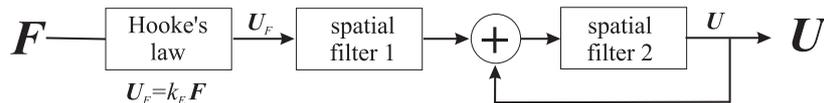
Figure 2: Spatial deformation model.

The registration uses multi-resolution approach, which starts at lowest image resolutions that remove image differences with large spatial extent, and continues with higher resolutions that remove more and more detailed image discrepancies.

3. EVALUATION

Ideally, our evaluation would judge the registration results by measuring the residual difference between corresponding anatomies over the image domain. As mentioned at the outset, such a correspondence does not exist in the majority of practical instances, and instead we focus only on a limited set of anatomical features for which homologies can be established between the image pair (with a certain amount of confidence). For these features, comparisons can be made based on their obtained and "true" mappings or displacements between the images. Alternatively, the spatial mapping can be manufactured, enabling the evaluation to be conducted over all image points in these synthetic data sets.

3.1. Recovering Synthetic Deformations

Figure 3 illustrates a general experimental design for the construction of synthetic test data, for which the correspondence and hence transformation between the images is known. The evaluation procedure makes use of two images, A and B, that are already in register, e.g., acquired at the same time but using different acquisition modalities or protocols.

It is also common to use the same image for A and B, but the results may be biased in favor of the algorithm because of correlated image noise or artifacts. Image B is deformed using some known transformation U_0 to obtain image B', which is then registered to image A. Ideally, the obtained transformation U_R should equal U_0^{-1}. The quality of the registration can thus be measured by the RMS (root-mean-squared) residual difference e_{RMS} between the transformations $U_0^{-1}(\mathbf{x})$ and $U_R(\mathbf{x})$, or, equivalently, their corresponding displacements, over pixels $\mathbf{x}$:

$$e_{RMS} = \sqrt{\frac{1}{N_\Omega} \sum_{\mathbf{x} \in \Omega} \left(U_0^{-1}(\mathbf{x}) - U_R(\mathbf{x})\right)^2} \quad , \quad \mathbf{x} = [x_1, x_2, x_3]^T. \tag{12}$$

Non-object voxels are excluded in the evaluation through the specification of the object domain Ω (in practice, a binary mask defined over the image), where N_Ω denotes the number of voxels in Ω.
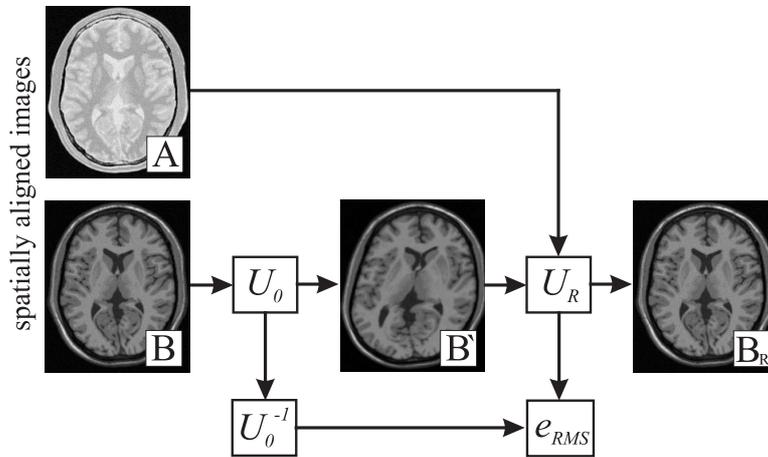
Figure 3: Evaluation scheme for non-rigid registration, based on synthetically deformed images.

As described in Section 2, our non-rigid registration consists of two independent parts, estimation of external forces that are derived from evaluating the relevant similarity measure, and the application of a spatial deformation model.

This modularity makes the synthetic deformation approach suitable for comparing similarity measures. Namely, ranking of comparison results is invariant to actual deformation as well as spatial deformation model. The same ranking is expected for real deformations.

By varying U_0, the use of synthetic test data allows systematic exploration of the method's performance under different kinds and degrees of distortions between images. Nevertheless, the major drawback of synthetic data is their lack of sufficient realism; in certain instances, synthetic data is not even an option because it is impossible to construct a phantom modeling the clinical situation of interest. Moreover, any evaluation based on synthetic data must be careful of the introduction of bias, in which the algorithm may systematically favor the class of simulated deformations chosen for the study. This validation approach can thus provide only limited information on methods for intersubject registration.

The results for the similarity measures S_H, S_S and S_{UH} are shown in Table 1. The results were obtained by cross modality MRI T1-PD registration, using Brainweb simulated brain images[14] and real data. Three

different types of Brainweb images were used: normal images with 9% of noise and images with 40% intensity shading, with $1 \times 1 \times 1$ mm voxel size. For registration of real images we used MRI T1 as a reference with voxel size $0.86 \times 0.86 \times 0.99$ mm, while the source MRI PD image had $0.98 \times 0.98 \times 1.1$ mm voxel size. The synthetic deformation was generated as a sum of Gaussian functions, in our case six functions with standard deviation ranging between 15 and 60 mm were used, which resulted in initial displacement error $e_{rms} = 6.90$ mm for simulated Brainweb images, and $e_{rms} = 14.15$ mm for real images. Results show that registration error e_{RMS}

Table 1: RMS registration error e_{RMS} (mm) for different similarity measures.

similarity measure	Brainweb normal	Brainweb 9% noise	Brainweb 40% shading	real images
S_H	0.64	0.64	3.12	2.37
S_S	0.42	0.34	1.50	1.61
S_{UH}	0.39	0.38	0.89	2.08

is relatively small for all the similarity measures and all the images. This indicates that spatial deformation model used suits to applied deformations, but it does not necessarily mean that registration is appropriate for solving real problems, as synthetic deformation was not designed according to real tissue properties. However, comparison of results shows that measures S_S and S_{UH} are better than S_H measure. They both yield similar results. Measure S_S is supposed to be better for initial steps as it better distinguishes between intensity classes that represent correct and incorrect matching. On the other hand it does not model partial volume voxels, which makes it less appropriate for final registration steps than measure S_{UH}. Results for real images are worse than results of simulated images, because of worse image quality and higher intensity variation of surrounding tissues.

3.2. Segmentation-based Evaluation

The gold standard for evaluating registration is the correspondence of anatomic features between images. For that purpose images must first be segmented to identify the location of corresponding features in the images. Different types of features may be extracted, e.g., point landmarks, segmented structures or regions of interest (ROIs), and classified tissue types.

The drawback associated with the first two kinds of features is that they invariably require human judgement, which will be subjective, difficult to reproduce and possibly erroneous.[2] Nevertheless, expert defined anatomic landmarks or regions of interest remain the gold standard for evaluation of registration accuracy. The correspondence between segmented regions can be measured using the same methods as those used in the evaluation of image segmentation methods.[15]

A commonly applied measure of regional overlap, S_R, is the ratio of the intersection of corresponding regions in both images (R_A and R_B) and their union:

$$S_R = \frac{R_A \cap R_B}{R_A \cup R_B}, \tag{13}$$

where $S_R = 1$ indicates perfect agreement and $S_R = 0$, complete disagreement. However, evaluation based on segmented structures can only evaluate the boundary location of regions and cannot detect incorrect matching within those regions. Furthermore, it is not possible to distinguish between registration errors and true morphological variability.

In its most simple form, the correspondence between anatomies can be evaluated at the level of tissue type of a voxel; that is, for brain images gray matter should be matched to gray matter, white matter to white matter, etc. The associated regions of interest, in contrast to the structures considered above, can be extracted with significantly less or more reproducible human involvement. The drawback here, as above, is that only limited information can be inferred about the accuracy of the anatomic alignment within the regions.

It is evident that the viability of the current validation approach depends critically on the quality of the ground-truth segmentation. Since human intervention is typically necessary to establish ground truth, the resultant uncertainty in the segmentations limits the precision with which the evaluation can be performed.

We next illustrate the segmentation-based approach for evaluation of intersubject registration by our method on MRI T1 brain images. Our dataset contains some two-dimensional segmented regions. We show position of contours of these regions overlaid on registered image. Such visualization can be used for evaluation performed by medical experts (see Fig. 4). The voxel size of images used in this experiment was $0.94 \times 0.94 \times 1$ mm.

(a)   (b)

(c)   (d)
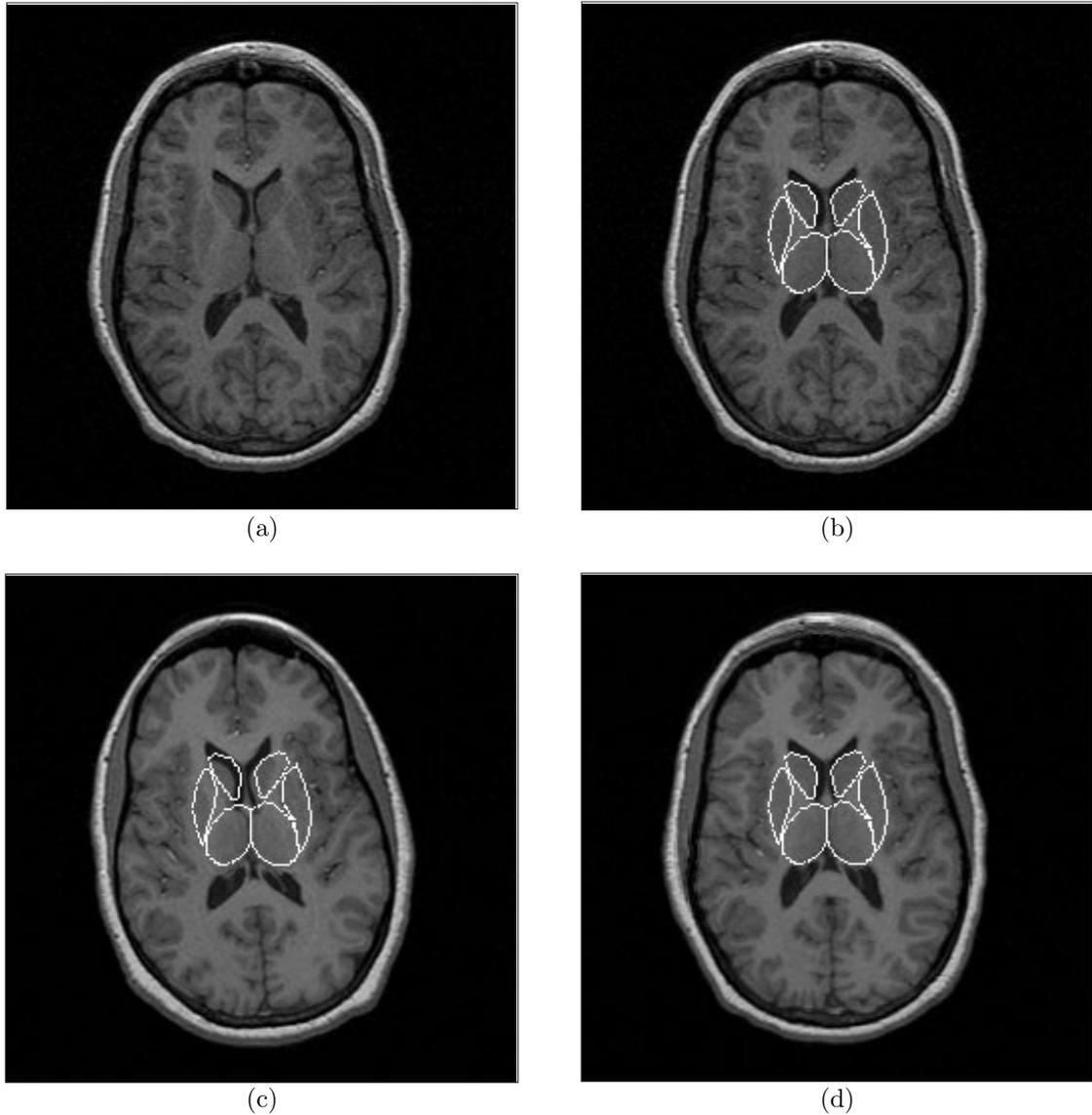
Figure 4. Intersubject registration of MRI images of the brain. (a) Target image, (b) overlaid with ROIs for several subcortical structures. (c) Rigidly registered source image, overlaid with subcortical ROIs from the target image. (d) Non-rigidly registered source image, overlaid with subcortical ROIs from the target image.

We also evaluated the intersubject registration by comparing the overlap of different tissue types. We obtained 60% overlap of white matter and 54% of gray matter. We attribute such low overlap mostly to the

quality of segmentation and also to large anatomical differences. For comparison, the tissue overlap for the best result achieved using synthetic deformations, $e_{RMS} = 0.34$ (see Table 1), was 96% for gray matter and 0.97% for white matter.

3.3. Direct Comparison of Registered Data

Owing to the lack of validation data with which to test a method's accuracy in anatomic registration, an alternative approach is to examine the quality of image alignment obtained with the method. It should be recognized from the outset that such evaluations are clearly far from optimal. They test a different objective—image correspondence—than that—anatomic correspondence—which we are interested in. Consequently, algorithms can yield correspondences that are anatomically incorrect but yet produce registered data that appear highly similar. Nevertheless, when combined even with a very limited evaluation based on anatomic features, the quality of image alignment can provide a useful assessment of the registration performance of an algorithm.

Evaluation of the quality of image alignment is generally conducted visually by experts but in a qualitative fashion. Results can be shown as an overlay of one image on top of the other, either by using different methods for displaying the entire image (e.g., image difference, "chessboard" image), or by only illustrating certain features, such as edges or points.

Figure 5 demonstrates the approach on the application of our method to PET-CT registration of thoracic images. In this case, the acquisitions were made at different phases of the respiratory cycle and, consequently, there are large differences between the imaged anatomic configurations. In addition, the images have relatively low resolution: the voxel size of the PET images as well as the CT image was $4 \times 4 \times 4$ mm.

4. CONCLUSION

In this paper we have illustrated some of the problems pertaining to evaluation, and discuss their origins. Because of unknown transformation that would ideally register real medical images, evaluation must be accomplished using other approaches, based on synthetic deformation, image segmentation or direct image comparison. However, all these approaches can lead to certain evaluation error, as no method can reliably deal with all kinds of misregistration.

Evaluation based on synthetic deformation is less appropriate for overall evaluation of registration, because it favors deformations that are similar to the synthetic ones. It is still appropriate for comparison of methods that are not related to the spatial deformation model as well as for testing overall registration capabilities. Segmentation based evaluation cannot detect misregistration within segmented features and furthermore do not distinguish between registration errors and true variability. However, by using a large number of correctly segmented smaller features, such evaluation converges towards the ideal evaluation. Evaluation based only on the images is the least reliable.

There is also a question of registering morphologically different structures. Should morphological differences remain or not? Most evaluation methods cannot distinguish between registration errors and true variability, so "good" registration results can be obtained only by changing source image morphology. If morphological differences are supposed to be preserved, registration can be correctly evaluated only when it is controlled by medical experts.

The evaluation results shown in this paper must be understood only as a starting point towards extensive evaluation of the method that we plan for the future.

(a)   (b)

(c)   (d)
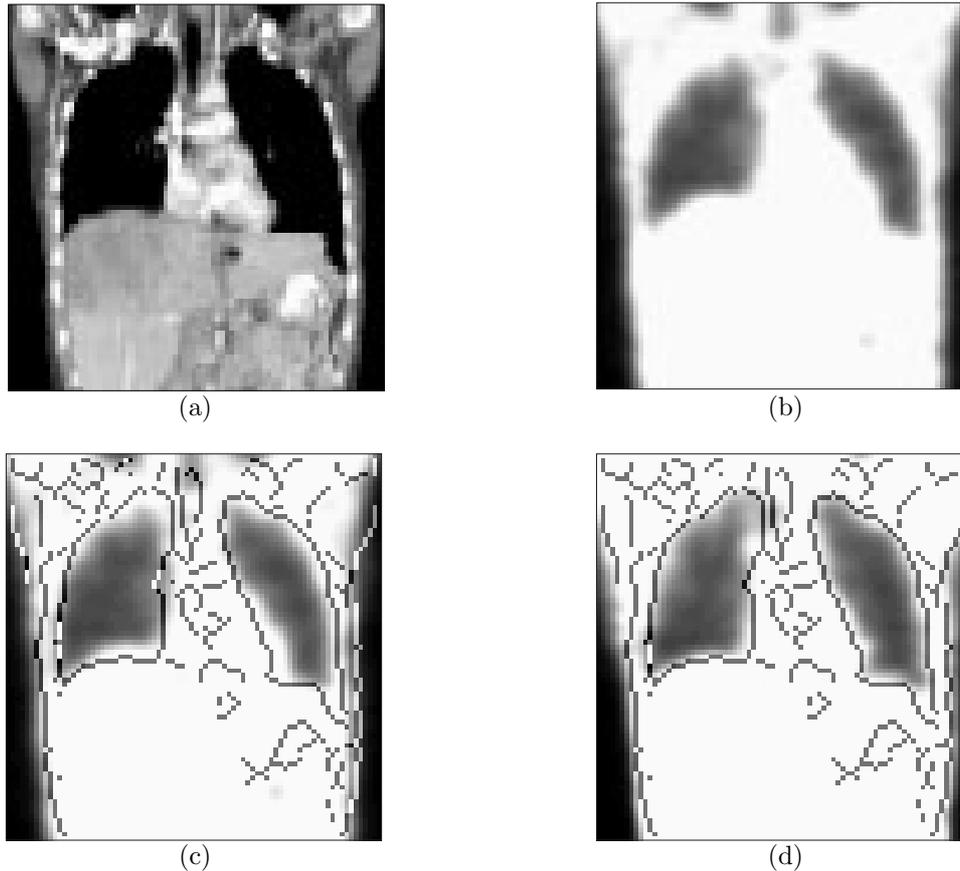
Figure 5. PET-CT registration of thoracic images. (a) Target CT image. (b) Source PET transmission image. (c) Rigidly and (d) non-rigidly registered PET images, overlaid with contours derived from the CT image.

REFERENCES

1. P. Thompson and A. W. Toga, "Warping strategies for intersubject registration," in *Handbook of Medical Image Processing*, I. Bankman, ed., Academic Press, 1999.

2. R. Woods, "Validation of registration accuracy," in *Handbook of Medical Imaging Processing and Analysis*, I. Bankman, ed., pp. 491–497, Academic Press, 1999.

3. P. Hellier, C. Barillot, I. Corouge, B. Giraud, G. Le Goualher, L. Collins, A. Evans, G. Malandain, and N. Ayache, "Retrospective evaluation of inter-subject brain registration," in *4th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention (MICCAI'01)*, W. Niessen and M. Viergever, eds., *LNCS* **2208**, pp. 258–265, (Utrecht, The Netherlands), October 2001.

4. J. Gee, "Performance evaluation of medical image processing algorithms," in *Medical Imaging 2001: Image Processing*, K. Hanson, ed., *Proc. SPIE Medical Imaging 2000* **3979**, pp. 19–27, SPIE Press, Bellingham, WA, 2000.

5. J. West, J. Fitzpatrick, M. Wang, B. Dawant, C. Maurer, R. Kessler, R. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. van den Elsen, S. Napel, T. Sumanaweera, B. Harkness, P. Hemler, D. Hill, D. Hawkes, C. Studholme, J. Maintz, M. Viergever, G. Malandain, X. Pennec, M. Noz, G. Maguire, M. Pollack, C. Pelizzari, R. Robb, and R. Hanson, D. Woods, "Comparison and evaluation of retrospective intermodality image registration techniques," *Journal of Computer Assisted Tomography* **21**(4), pp. 554–566, 1997.

6. A. Guimond, A. Roche, N. Ayache, and J. Meunier, "Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections," *IEEE Transactions on Medical Imaging* **20**, pp. 58–69, January 2001.

7. P. Hellier, C. Barillot, E. Memin, and P. Perez, "Hierarchical estimation of a dense deformation field for 3-d robust registration," *IEEE Transactions on Medical Imaging* **20**, pp. 388–402, May 2001.

8. G. Christensen and H. Johnson, "Consistent image registration," *IEEE Transactions on Medical Imaging* **20**, pp. 568–582, July 2001.

9. P. J. Kostelec, J. B. Weaver, and D. M. Healy Jr., "Multiresolution elastic image registration," *Medical Physics* **25**, pp. 1593–1604, September 1998.

10. P. Rogelj and S. Kovačič, "Similarity measures for non-rigid registration," in *Medical Imaging 2001: Image Processing*, M. Sonka and K. Hanson, eds., *Proc. SPIE* **4322**, pp. 569–578, SPIE Press, Bellingham, WA, 2001.

11. M. Bro-Nielsen, *Medical Image Registration and Surgery Simulation*. PhD thesis, Department of Mathematical Modelling, Technical University of Denmark, 1981.

12. R. Bajcsy and S. Kovačič, "Multiresolution elastic matching," *Computer Vision, Graphics and Image Processing* **46**, pp. 1–21, April 1989.

13. J. C. Gee, "On matching brain volumes," *Pattern Recognition* **32**(1), pp. 99–111, 1999.

14. R.-S. Kwan, A. Evans, and G. B. Pike, *An Extensible MRI Simulator for Post-Processing Evaluation*, vol. 1131 of *Lecture Notes in Computer Science*, pp. 135–140. Springer-Verlag, May 1996.

15. G. Gerig, M. Jomier, and M. Chakos, "Non-rigid multimodal image registration using mutual information," in *Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI'2001*, W. Niessen and M. Viergever, eds., *Lecture Notes in Computer Science* **2208**, pp. 616–523, Springer, October 2001.