Assessing the intrinsic precision of 3D/3D rigid image registration results for patient setup in the absence of a ground truth

Jian Wu^{a)} and Martin J. Murphy^{b)}

Department of Radiation Oncology, Virginia Commonwealth University, Richmond, Virginia 23298

(Received 10 December 2009; revised 10 March 2010; accepted for publication 5 April 2010; published 11 May 2010)

Purpose: To assess the precision and robustness of patient setup corrections computed from 3D/3D rigid registration methods using image intensity, when no ground truth validation is possible.

Methods: Fifteen pairs of male pelvic CTs were rigidly registered using four different in-house registration methods. Registration results were compared for different resolutions and image content by varying the image down-sampling ratio and by thresholding out soft tissue to isolate bony landmarks. Intrinsic registration precision was investigated by comparing the different methods and by reversing the source and the target roles of the two images being registered.

Results: The translational reversibility errors for successful registrations ranged from 0.0 to 1.69 mm. Rotations were less than 1°. Mutual information failed in most registrations that used only bony landmarks. The magnitude of the reversibility error was strongly correlated with the success/failure of each algorithm to find the global minimum.

Conclusions: Rigid image registrations have an intrinsic uncertainty and robustness that depends on the imaging modality, the registration algorithm, the image resolution, and the image content. In the absence of an absolute ground truth, the variation in the shifts calculated by several different methods provides a useful estimate of that uncertainty. The difference observed by reversing the source and target images can be used as an indication of robust convergence. © 2010 American Association of Physicists in Medicine. [DOI: [10.1118/1.3414041](https://doi.org/10.1118/1.3414041)]

Key words: image registration, image-guided treatment setup, setup uncertainty, registration error, optimization algorithm

I. INTRODUCTION

Historically, patient setup for external-beam radiotherapy has used 2D projection images to measure the translational shifts needed for optimal beam alignment. The shifts can be determined either via 2D/2D registration to simulation images or via 2D/3D registration to the planning CT. Owing to the poor visibility of soft-tissue features in 2D projections, alignment is usually established via bony landmarks or implanted fiducials. The introduction of CT systems into the treatment room¹⁻⁴ allows patient setup via 3D/3D image registration, using bony landmarks, fiducials, or soft tissue. The registration process can be based on the image intensities or extracted features. In all cases, though, there is some uncertainty in the estimated rigid setup correction.

In analyzing the uncertainties in rigid registration results, we will distinguish between intrinsic and extrinsic sources of error. What we define as intrinsic error is associated with the registration process itself, which is an approximate calculation that is usually done via an iterative search to maximize an image similarity measure. In the real world, the registered source and target images will never be a perfect match and thus, different registration methods will converge to different approximate solutions when applied to the same data. Intrinsic errors in 3D/3D CT registration due to respiration blurring and motion artifacts have been investigated by Woodford *et al.*⁵ Nonrigid (i.e., shape) differences in the source and target image features will also introduce intrinsic error.

The presence of rotations that are unaccounted for in the registration can introduce an intrinsic error. Extrinsic errors are associated with the imaging scenario itself, e.g., if the imaging modality or the field of view is changed, then one can expect variations in the resulting setup correction. We are specifically interested in the intrinsic uncertainties in intensity-based 3D/3D rigid registration using only anatomical information (i.e., without fiducials).

There have been many prior studies of rigid registration accuracy but generally, they have either made relative comparisons of specific algorithm components⁶⁻⁸ or they have had some absolute reference available to compare to the registration result. One generally defines a ground truth reference as *a priori* knowledge of the actual setup correction while a gold standard reference is the setup correction calculated by the best available method. A ground truth reference can be established with a phantom, while a gold standard reference can be established from fiducial landmarks in the patient.⁹ Unfortunately, when dealing with anatomy-driven registration in actual clinical usage, it is generally not possible to establish either a ground truth or gold standard to estimate registration errors. In particular, it is often not possible to tell with certainty whether the registration has converged to the neighborhood of the global optimum. Undetected convergence failures can lead to setup errors that greatly exceed the expected uncertainty margin. The purpose of this paper is to examine two ways to infer intrinsic registration errors and convergence failures in the absence of a gold stan-

dard or ground truth for the correct registration.

One way to expose the intrinsic errors associated with the registration process is to register pairs of images using several different algorithms and observe the variation in the results. Another test (which is not practical for most 2D/3D registration techniques) is to make two registrations for each image pair while switching the source and the target images. Unless the registration process is specifically constrained to be inversely consistent,¹⁰ the two trials will follow different paths that do not converge to exactly the same result. These two tests can reveal something about the variability and robustness of the registration result in the absence of ground truth.

We have been careful to distinguish registration precision from accuracy. Precision is the variation in different registration results around the similarity maximum; accuracy is the proximity of the similarity maximum to the true rigid transform. This is a particularly vexing distinction in actual clinical cases because the anatomy is often not rigid (not even the bony landmarks if the images include articulated elements). Consequently, there might not even be a true rigid transform, in which case the only meaningful error estimate is the precision. Precision in clinical cases can be estimated by analyzing the behavior of the similarity surface near its optimum to estimate variability, or by making multiple image-guided setup measurements. Unfortunately, registration accuracy in clinical cases that do not have an independent ground truth is essentially indeterminate.

At present, there are a large number of similarity measures and optimization procedures that can potentially be used in an image registration algorithm.¹¹ Past studies have compared as many as nine similarity measures,⁷ while iterative searches can be made via simulated annealing, the downhill simplex method, gradient driven methods such as steepest descent, Levenberg–Marquardt, conjugate gradient, etc. Furthermore, the algorithms themselves can be started either from zero or from randomized initial guesses for the rigid transformation. It would have been impractical (and unnecessary for our purposes) to test all possible combinations of these algorithm components and initialization conditions. Consequently, we used four representative in-house 3D/3D algorithm configurations from among all of these possibilities and investigated the dependence of the intrinsic registration precision and robustness on the various components. We compared the optimization algorithms while controlling for the similarity measure and compared the similarity measures while controlling for the optimization.

To isolate the intrinsic registration error, we observed the variability in setup shifts determined by the four different registration methods when operating on exactly the same image sets. In our tests we used three CT images for each of five prostate patients. These patients did not have implanted fiducials. The four algorithms were combinations of three different intensity-based similarity measures (two versions of mutual information and the mean-squared intensity difference) and two different iterative search algorithms. For each registration method, we reversed the source and target images to observe the variation in the rigid transformation

FIG. 1. The region of interest for the 3D/3D registration of pelvic images.

when all other facets of the registration process were the same. We looked for a relationship between the magnitude of the reversibility error and the failure of the registration to approach a global minimum. For all of the registrations, we used both (1) the entire imaged region of interest (ROI) and (2) only bony landmarks. Because many clinical image registration tools for patient setup only allow for translational offsets, we computed each registration first for only translational shifts and then again for both translations and rotations. These tests exposed the variations in calculated shifts due to algorithm differences when the image modality and the anatomical ROI were the same. For the registrations that isolated bony landmarks, it could be further assumed that the registered anatomy was perfectly rigid.

II. METHODS AND MATERIALS

The study used three fan-beam CT (FBCT) data sets for each of the five prostate cancer patients, which allowed for three pairwise rigid registrations. The CT data were 512×512 voxels in the slice plane and ranged from 80 to 100 slices for each data set. The voxels were $0.88 \times 0.88 \times 3.0$ mm³. Each registration was based on a user-defined ROI for the target image (i.e., the fixed image). The whole image volume was used for the source image (i.e., the floating image). The target ROIs contained the prostate, bladder, and rectum, as well as bony landmarks, while excluding couch top and other irrelevant features. A typical ROI contained about 40 slices, each of approximately 150×200 pixels. Figure 1 illustrates the ROI in a representative slice. One set of registrations allowed only (x, y, z) translations; the second set of registrations included three

rotational degrees of freedom as well. The rotations were defined in the space-fixed convention, with yaw measured around the patient's anterior/posterior axis, pitch measured around the left/right axis, and roll around the inferior/superior axis. We investigated the precision and robustness of the registrations while varying (1) the similarity measure, (2) the optimization algorithm, (3) the image resolution, and (4) the registered image features.

Three similarity measures, the mean-squared difference (MSD), the normalized mutual information (NMI), and the Mattes mutual information (MMI), were used. The MSD is a widely used similarity measure for registering images acquired with the same modality. The MSD between image A and B is given by

$$\text{MSD}(A, B) = \frac{1}{N} \sum_{i=1}^N (A_i - B_i)^2, \quad (1)$$

where A_i and B_i are the two corresponding voxels in image A and B , respectively. N is the number of voxels in the ROI.

Mutual information was introduced by Shannon¹² in his information theory and is defined as

$$I(A, B) = H(A) + H(B) - H(A, B), \quad (2)$$

where $H(A)$ and $H(B)$ denote the marginal entropies of image A and B , respectively, and $H(A, B)$ is the joint entropy of the two images.

The NMI calculates the amount of information that one image contains about the other, which has been widely used in image registration.¹³ We used the NMI definition introduced by Studholme *et al.*¹⁴

$$\tilde{I}(A, B) = \frac{H(A) + H(B)}{H(A, B)}. \quad (3)$$

The calculation of these entropies requires the estimation of the marginal probability density functions (PDFs) of image A and B and the joint PDF between these two images. In our study, the PDFs were computed from the intensity histograms and the joint histogram, which counted all the voxels within the ROI.

In Mattes' implementation¹⁵ of mutual information (i.e., MMI), kernel density estimation is used to calculate the entropy PDFs. The method uses only a fraction of the voxels in the ROI, which greatly speeds up the calculation. We used 50 000 samples in each ROI, regardless of image resolution and content, and a zero-order B-spline and a cubic B-spline as the smoothing kernels for the target and the source images, respectively.

We tested two gradient-based optimization algorithms: Steepest descent (SD) and regular step gradient descent (RSGD) as implemented by the open source software package—INSIGHT TOOLKIT (www.itk.org). Both algorithms adjust transformation parameters in the direction of the similarity gradient with respect to the degrees of freedom. In RSGD, a bipartition scheme is used to compute the step size. The steepest descent algorithm had no selectable optimization parameters. The registration parameters for the other methods included the number of bins for mutual information

calculation, the interpolation method, and optimization termination conditions. For those algorithms, we used typical parameter settings that have been reported in literature. We did not attempt to improve on those settings.

Prior to registering pairs of patient images, we validated the MSD+SD algorithm by artificially shifting a test image chosen from the data set by known translations and rotations. Under these ideal conditions, the algorithm determined the artificial shifts with essentially perfect accuracy (<0.1 mm, $<0.1^\circ$ error), which demonstrated that it was performing correctly. (When we later compared the registration results of the other three algorithms to MSD+SD, the generally close agreement of the results confirmed that they were also performing correctly.)

We then registered the image pairs using both the whole image content in the ROIs and using only bony landmarks, which were selected via a CT number threshold of 1100. The intensity threshold was applied to the target image but not the source image to avoid influencing the registration by introducing artificial features (i.e., edges) via the threshold.

To study the effect of image resolution, the original CT images were down-sampled in the slice plane by the ratio of 1:2 and 1:4, which resulted in the image size of 256×256 and 128×128 , respectively. No image reduction in the axial direction was performed.

For each algorithm, we made a forward and a backward registration on each image pair by reversing the images used as the source and the target. For registrations using the SD algorithm, one forward and one backward registration were performed for each image pair, starting from zero initial offset. For registrations using the RSGD algorithm, ten repetitions were carried out for each forward and backward registration using randomly selected initial offsets (within 5 mm and 5°) in order to evaluate the method's sensitivity to initial conditions.

III. RESULTS

We first investigated the magnitude of the rotations present in the image pairs. (Their presence can influence the intrinsic accuracy of registrations that ignore them.) The forward and reverse registrations for each image pair provided two different estimates of rotation around each axis (differing by a sign). In the absence of any intrinsic registration error they should agree; any disagreement between the two results gives an indication of the uncertainty in the estimate. We used the MSD+SD registrations as the benchmark for estimating rotations. The mean absolute values and variances of the rotations calculated from each pair of forward and backward registrations were $0.90^\circ \pm 0.65^\circ$, $0.40^\circ \pm 0.30^\circ$, and $0.55^\circ \pm 0.55^\circ$ in yaw, pitch, and roll, respectively. The distributions of rotations were skewed such that the median rotations were 0.60° , 0.30° , and 0.30° , respectively. The average ratio of the forward/backward angle difference to the sum for each rotational axis (which gives some measure of the approximate measurement uncertainty as a fraction of the apparent rotation) was 0.36, 0.75, and 0.84 for yaw, pitch, and roll, respectively. The standard deviation of the forward

TABLE I. Mean reversibility errors (mm) (involving translations only). *Abbreviations:* MSD=mean-squared difference; NMI=normalized mutual information; MMI=Mattes mutual information; SD=steepest descent; RSGD=regular step gradient descent.

Similarity measure + optimization method	Image content	Slice resolution		
		512 × 512	256 × 256	128 × 128
MSD+SD	Entire image	0.49 (0.06–1.69)	0.56 (0.07–1.78)	0.78 (0.07–1.53)
	Bony landmark	0.18 (0.02–0.51)	0.30 (0.04–0.71)	0.62 (0.05–1.42)
MSD+RSGD	Entire image	0.40 (0.04–1.37)	0.48 (0.05–1.34)	0.94 (0.07–4.06)
	Bony landmark	0.23 (0.02–1.63)	0.65 (0.02–2.70)	0.45 (0.06–2.74)
NMI+RSGD	Entire image	0.16 (0.01–0.56)	0.11 (0.02–0.30)	0.14 (0.02–0.53)
	Bony landmark	Failed > 50%	Failed > 50%	Failed > 50%
MMI+RSGD	Entire image	0.14 (0.02–0.41)	0.13 (0.03–0.25)	0.15 (0.01–0.62)
	Bony landmark	Failed > 50%	Failed > 50%	Failed > 50%

and the reverse values for the rotations around each mean rotation averaged out to be 0.41° , 0.17° , and 0.26° in yaw, pitch, and roll. These results indicate that the majority of the pitch and roll rotations were approximately consistent with zero (to within measurement uncertainty), while 90% of the yaw rotations were less than 1.3° . The estimated rotations had a minimal impact on the translational setup shifts, half of which changed by less than 0.25 mm and 90% of which changed by less than 0.5 mm in (x, y, z) when rotations were included among the degrees of freedom.

The next consideration was the intrinsic convergence precision of each registration method. The similarity surfaces are not perfectly smooth and monotonic; hence, registrations are unlikely to arrive at exactly the same spot after following different routes. We relate the intrinsic precision to the observed proximity of registration results after following different paths to convergence. If the registrations are not explicitly constrained to be inversely consistent, then forward and reverse registrations will follow different routes and arrive at different rigid transforms. This provides one way to examine the intrinsic precision. We quantified this by the reversibility error, which was defined to be half the Euclidean distance between the forward and the reverse-shifted position of an anatomical point 3.5 cm from the center of rotation. (This showed the effect of rotational reversibility error for six DOF registrations.) Table I (translation) and Table II (translation and rotation) summarize the results for the different registra-

tion configurations. (Only those registrations that converged were included when calculating the errors.) The dependence of the reversibility error on the location of the selected anatomical point has been tested by setting the computation point at 1.75, 3.5, and 7.0 cm from the center of rotation. Because of the small rotation angle differences between the forward and the inverse registrations, the rotation-induced reversibility errors, which range from 0.01 to 0.03 mm, are negligibly small.

Rigid registration robustness was taken to be the likelihood of convergence of the similarity measure to the approximate global optimum without being directed interactively. Our optimization methods minimized the MSD and the negated NMI and MMI. First we looked at what could be learned from the reversibility error. Figure 2 shows the distribution of reversibility errors using the MSD measure for both the SD and the RSGD methods, also including rotations. There is a clear separation between cases with error < 2 mm and cases with error > 2 mm. We then looked for a relationship between the reversibility error and the value of the final similarity measure as a signal of convergence. Specifically, we looked for instances where the forward and the reverse registrations converged to significantly different final similarity measure values, which we took as an indication that at least one of the registrations did not converge to the true minimum. Figure 3(a) shows the correlation between the reversibility error and the difference of the forward (ε_1) and

TABLE II. Mean reversibility errors (mm) (involving translations and rotations). *Abbreviations:* See Table I.

Similarity measure + optimization method	Image content	Slice resolution		
		512 × 512	256 × 256	128 × 128
MSD+SD	Entire image	0.47 (0.11–1.58)	0.40 (0.06–1.38)	0.58 (0.29–1.27)
	Bony landmark	0.20 (0.03–0.53)	0.22 (0.09–0.34)	0.48 (0.09–1.64)
MSD+RSGD	Entire image	0.43 (0.03–1.34)	0.46 (0.03–1.30)	0.68 (0.10–2.54)
	Bony landmark	0.11 (0.04–0.34)	0.13 (0.06–0.30)	0.16 (0.06–0.35)
NMI+RSGD	Entire image	0.17 (0.02–0.42)	0.14 (0.02–0.38)	0.14 (0.03–0.28)
	Bony landmark	Failed > 50%	Failed > 50%	Failed > 50%
MMI+RSGD	Entire image	0.11 (0.03–0.28)	0.11 (0.02–0.37)	0.11 (0.01–0.43)
	Bony landmark	Failed > 50%	Failed > 50%	Failed > 50%

FIG. 2. The frequency of reversibility errors for the MSD similarity measure. Notice the separate grouping of errors < 2 mm and errors > 2 mm.

reverse (ε_2) similarity measures [specifically, $(\varepsilon_1 - \varepsilon_2)(\varepsilon_1 + \varepsilon_2)$] at convergence for the MSD+SD algorithm using both translations and rotations. The reversibility errors < 2 mm are associated with consistent forward/reverse similarity measures, while larger reversibility errors are associated with convergence failures. Then we looked for significance in the depth of the similarity minima for both registrations. Figure 3(b) shows the correlation of the reversibility error and the average of the forward and the reverse similarity measures. The outliers are the same as in Fig. 3(a). This shows that a small reversibility error does not, by itself, indicate a strong convergence in both directions; it could also correspond to either a shallow global minimum or a case where both registrations got trapped in the same local minimum.

For the mutual information algorithms, the reversibility errors for successful registrations were always less than 1 mm. Table III (translation) and Table IV (translation and rotation) summarize the algorithm configurations that were tested and the percentage success rates for each algorithm.

The MSD+SD method was equally robust for both types of image features and benefited from image down-sampling. The MSD+RSGD method was more robust when registering bony landmarks, regardless of image resolution. Mutual information using only bony landmarks had a poor success rate.

Our final consideration was the variation in the translational shifts obtained from the different algorithms. At full resolution, we calculated for each image pair the mean shift in (x, y, z) from the four methods MSD+SD, MSD+RSGD, NMI+RSGD, and MMI+RSGD, for full image content. For bony features, where NMI and MMI failed, we calculated the mean from the two MSD algorithms. From this we calculated the standard deviations of the translations around the

FIG. 3. (a) The MSD+SD reversibility error versus the difference in the forward and the reverse similarity measures at convergence. (b) The MSD+SD reversibility error versus the average of the forward and the reverse similarity measures at convergence.

mean for all successful registrations. For the entire image content, the standard deviation among the four different methods was 0.8 mm. When only bony landmarks were used, the standard deviation for MSD+SD and MSD+RSGD (NMI and MMI failed) was 0.3 mm. These results are similar to the MSD reversibility errors.

Figures 4 and 5 show examples of the similarity surfaces for (negated) MSD, NMI, and MMI as functions of x and y translations. Both the capture range of the registration and the uncertainty in the final translational shifts increase with the width of the similarity peak.

IV. DISCUSSION

Mutual information algorithms failed more often than those for MSD using bony anatomy. The failures were typically associated with translations > 5 mm. The failure rate increased when rotations were included as degrees of freedom. This underscores the weakness of mutual information

TABLE III. Percentage success rate of registrations (out of 30 registration pairs, involving translations only). Abbreviations: See Table I.

Similarity measure+optimization method	Image content	Slice resolution		
		512 × 512	256 × 256	128 × 128
MSD+SD	Entire image	83.3	96.0	100.0
	Bony landmark	83.3	100.0	100.0
MSD+RSGD	Entire image	90.0	96.3	99.7
	Bony landmark	99.7	100.0	100.0
NMI+RSGD	Entire image	85.7	99.0	99.3
	Bony landmark	47.3	46.3	39.7
MMI+RSGD	Entire image	92.7	98.3	99.3
	Bony landmark	48.0	53.3	57.3

as a similarity measure for limited target image content. Mutual information has also been found to be a suboptimal measure for 2D/3D rigid registration for a small ROI.^{6,16}

Down-sampling the CT slices by a factor of 2 reduced the number of convergence failures while having a negligible effect on the reversibility error. This is not a new finding but supports the use of a multiresolution registration strategy to avoid trapping in local minima. Down-sampled images can be used to initiate the registration and then, once the global minimum has been approached, the original high-resolution images can be restored to obtain optimal precision. Unsurprisingly, registering high contrast bony landmarks provided the most consistent results among the methods we tested.

The rotations for our test cases were close to zero. Consequently, the shifts calculated with and without rotations were very similar. Tables I–IV show that excluding rotations did not have much effect on the reversibility errors or the trends in robustness for the different algorithms (although with rotations included the mutual information registrations failed more often). Therefore, we conclude that the reversibility errors and interalgorithmic variations that we observed in Tables I and III do not include any inaccuracy due to the omission of rotations. Intuitively, though, the symmetry of the forward-backward registration test suggests that the reversibility error might not be sensitive to registration imperfections due to the omission of rotations. Therefore, it may not be a good indicator of errors introduced by unknown

rotations. Because many clinical image registration tools that are used for patient setup do not calculate rotations, it would be useful to find a signal for instances when the omitted rotations are having a significant impact on the setup uncertainty.

Our registration results are consistent with the shape of the similarity surface near the global optimum. Figures 4 and 5 show examples of the similarity surfaces for negated MSD, NMI, and MMI as a function of translational shifts. The uncertainty in finding the minimum increases with the width and flatness of the peak. Thus, one would expect a very small reversibility error for the mutual information results, and a larger reversibility error for the MSD results. This is what we observed. The shape of the similarity surface also determines the capture range, i.e., the maximum distance between source and target landmarks that can be successfully registered. A sharp, highly localized peak usually has a smaller capture range than a broader, more diffuse similarity peak. When mutual information is used for bony landmark registration, Fig. 5 shows that the similarity peak is very narrow and sharply confined within ± 5 mm, leading to the large number of registration failures that we observed. In contrast, the MSD surface has a smooth gradient extending beyond ± 15 mm, resulting in robust (albeit less precise) convergence.

A short capture range is not a fatal flaw for fully automatic registration. If one can start the mutual information

TABLE IV. Percentage success rate of registrations (out of 30 registration pairs, involving translations and rotations). Abbreviations: See Table I.

Similarity measure+optimization method	Image content	Slice resolution		
		512 × 512	256 × 256	128 × 128
MSD+SD	Entire image	83.3	96.0	100.0
	Bony landmark	83.3	100.0	100.0
MSD+RSGD	Entire image	94.0	96.3	97.7
	Bony landmark	91.7	85.3	89.7
NMI+RSGD	Entire image	98.0	96.7	94.0
	Bony landmark	0.0	15.3	10.0
MMI+RSGD	Entire image	92.0	93.7	97.0
	Bony landmark	0.7	4.0	2.0

FIG. 4. Similarity measures as functions of shifts in the left/right and anterior/posterior axes (unit: mm) when no intensity threshold was applied. 3D surface plots are shown in the left column and contour plots the right column. The similarity measures are (a) negated MSD, (b) NMI, and (c) MMI.

(MI) registration close to the actual shift, then it will most likely converge. This can be done by registering in two stages, using a MSD algorithm for the first stage of registration to get a good initial estimate, then repeating with a MI algorithm to get the benefits of both long capture range and high precision.

The test images were all FBCT studies. In many clinical applications, the setup images are cone-beam CTs (CBCTs). In an earlier study,¹⁷ we tested the sensitivity of 3D/3D deformable registration to the presence of cupping and high levels of noise in one of the two CTs, such as one might encounter when registering FBCT to CBCT. We found no significant effect. Rigid registrations should be even less sensitive to image quality differences.

In a clinical application, it is important to know whether a particular registration has converged to its global minimum. However, a global minimum can be comparatively shallow, which means the depth of the minimum relative to the start-

FIG. 5. Similarity measures as functions of shifts in the left/right and anterior/posterior axes (unit: mm) when bony landmark threshold was applied. 3D surface plots are shown in the left column and contour plots the right column. The similarity measures are (a) negated MSD, (b) NMI, and (c) MMI.

ing value is not a reliable indicator by itself. Registering the images with two or more algorithms provides one way to test for convergence failures, but this is generally not possible with commercially available patient setup tools. On the other hand, it is practical (and easy) to apply the reversibility test using any available registration software. If the forward and the reverse minima are distinctly different, this is a good indication of a failure. Applying this test, we found that the reversibility error can also provide an effective means to detect failed registrations. When we registered pelvic fan-beam CT images using the MSD similarity metric, the reversibility errors were tightly clustered below 2 mm [Fig. 3(a)], with the exception of conspicuous and well-separated outliers that were consistently associated with convergence failures. For mutual information, the registration failures had reversibility errors > 1 mm. However, a small reversibility error, even when combined with similar forward/reverse minima [Figs. 3(a) and 3(b)] does not necessarily indicate successful

convergence—It could be that the two registrations landed in the same local minimum. We conclude that a large reversibility error is sufficient evidence of a convergence failure and that a small reversibility error is necessary but not sufficient evidence for global convergence. We have found that combining the depth of the minimum with the reversibility error provides further confidence that a global minimum has been reached, but is still not conclusive.

V. CONCLUSIONS

We have observed that the intrinsic variability (precision) of four representative rigid registration algorithms is on the order of 1 mm or less when applied to pelvic CTs. There is an inverse relationship between capture range and precision that affects an algorithm's robustness in converging to a solution. We found that running a 3D/3D image registration in forward and reverse provides a useful indication of whether or not the registration has arrived at the best solution.

ACKNOWLEDGMENT

This work was supported in part by NIH Grant No. P01CA116602.

^{a)}Author to whom correspondence should be addressed. Electronic addresses: jwu@mcvh-vcu.edu and eewujian@hotmail.com

^{b)}Electronic mail: mmurphy@mcvh-vcu.edu

¹J. R. Wong, C. W. Cheng, L. Grimm, and M. Uematsu, "Clinical implementation of the world's first Primatom, a combination of CT scanner and linear accelerator, for precise tumor targeting and treatment," *Phys. Medica* **17**, 271–276 (2001).

²L. E. Court, L. Dong, N. Taylor, M. Ballo, K. Kitamura, A. K. Lee, J. O'Daniel, R. A. White, R. Cheung, and D. Kuban, "Evaluation of a contour-alignment technique for CT-guided prostate radiotherapy: An intra- and interobserver study," *Int. J. Radiat. Oncol., Biol., Phys.* **59**, 412–418 (2004).

³D. A. Jaffray and J. H. Siewerdsen, "Cone-beam computed tomography with a flat-panel imager: Initial performance characterization," *Med. Phys.* **27**, 1311–1323 (2000).

⁴K. J. Ruchala, G. H. Olivera, E. A. Schloesser, and T. R. Mackie, "Mega-

voltage CT on a tomotherapy system," *Phys. Med. Biol.* **44**, 2597–2621 (1999).

⁵C. Woodford, S. Yartsev, and J. Van Dyk, "Image registration of a moving target phantom with helical tomotherapy: Effect of the CT acquisition technique and action level proposal," *Phys. Med. Biol.* **53**, 5093–5106 (2008).

⁶G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. Hill, and D. J. Hawkes, "A comparison of similarity measures for use in 2-D-3-D medical image registration," *IEEE Trans. Med. Imaging* **17**, 586–595 (1998).

⁷D. Skerl, D. Tomazevic, B. Likar, and F. Pernus, "Evaluation of similarity measures for reconstruction-based registration in image-guided radiotherapy and surgery," *Int. J. Radiat. Oncol., Biol., Phys.* **65**, 943–953 (2006).

⁸D. Škerl, B. Likar, J. M. Fitzpatrick, and F. Pernus, "Comparative evaluation of similarity measures for the rigid registration of multi-modal head images," *Phys. Med. Biol.* **52**, 5587–5601 (2007).

⁹D. B. Russakoff, T. Rohlfing, A. Ho, D. H. Kim, R. Shahidi, J. R. Adler, Jr., and C. R. Maurer, Jr., "Evaluation of intensity-based 2D-3D spine image registration using clinical gold-standard data," in *Biomedical Image Registration, Second International Workshop, WBIR 2003*, 2003 [Lect. Notes Comput. Sci. **2717**, 151–160 (2003)].

¹⁰G. E. Christensen and H. J. Johnson, "Consistent image registration," *IEEE Trans. Med. Imaging* **20**, 568–582 (2001).

¹¹J. B. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image Anal.* **2**, 1–36 (1998).

¹²C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.* **27**, 379–423 (1948).

¹³J. P. Pluim, J. B. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imaging* **22**, 986–1004 (2003).

¹⁴C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recogn.* **32**, 71–86 (1999).

¹⁵D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewell, and W. Eubank, "Nonrigid multimodality image registration," in *Medical Imaging 2001: Image Processing*, Vol. 4322, edited by M. Sonka and K. M. Hanson (SPIE, San Diego, 2001), pp. 1609–1620.

¹⁶J. Wu, M. Kim, J. Peters, H. Chung, and S. S. Samant, "Evaluation of similarity measures for use in the intensity-based rigid 2D-3D registration for patient positioning in radiotherapy," *Med. Phys.* **36**, 5391–5403 (2009).

¹⁷M. J. Murphy, Z. Wei, M. Fatyga, J. Williamson, M. Anscher, T. Wallace, and E. Weiss, "How does CT image noise affect 3D deformable image registration for image-guided radiotherapy planning?," *Med. Phys.* **35**, 1145–1153 (2008).