A neural network based 3D/3D image registration quality evaluator for the head-and-neck 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 28 April 2010; revised 19 August 2010; accepted for publication 27 September 2010; published 14 October 2010)

Purpose: To develop a neural network based registration quality evaluator (RQE) that can identify unsuccessful 3D/3D image registrations for the head-and-neck patient setup in radiotherapy.

Methods: A two-layer feed-forward neural network was used as a RQE to classify 3D/3D rigid registration solutions as successful or unsuccessful based on the features of the similarity surface near the point-of-solution. The supervised training and test data sets were generated by rigidly registering daily cone-beam CTs to the treatment planning fan-beam CTs of six patients with head-and-neck tumors. Two different similarity metrics (mutual information and mean-squared intensity difference) and two different types of image content (entire image versus bony landmarks) were used. The best solution for each registration pair was selected from 50 optimizing attempts that differed only by the initial transformation parameters. The distance from each individual solution to the best solution in the normalized parametrical space was compared to a user-defined error threshold to determine whether that solution was successful or not. The supervised training was then used to train the RQE. The performance of the RQE was evaluated using the test data set that consisted of registration results that were not used in training.

Results: The RQE constructed using the mutual information had very good performance when tested using the test data sets, yielding the sensitivity, the specificity, the positive predictive value, and the negative predictive value in the ranges of 0.960–1.000, 0.993–1.000, 0.983–1.000, and 0.909–1.000, respectively. Adding a RQE into a conventional 3D/3D image registration system incurs only about 10%–20% increase of the overall processing time.

Conclusions: The authors' patient study has demonstrated very good performance of the proposed RQE when used with the mutual information in identifying unsuccessful 3D/3D registrations for daily patient setup. The classifier had very good generality and required only to be trained once for each implementation. When the RQE is incorporated with an automated 3D/3D image registration system, it can improve the robustness of the system. © 2010 American Association of Physicists in Medicine. [DOI: [10.1118/1.3502756](https://doi.org/10.1118/1.3502756)]

Key words: image registration, neural network, image-guided treatment setup, registration error, registration quality evaluation

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 room^{1–4} 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. Because clinical applications of rigid image registration for patient setup have no ground truth for validation, one requires some other means to identify poor registration results and correspondingly large setup errors. In the present study, we demonstrate the performance of a neural network for automatically classifying registration results.

An intensity-based automated 3D/3D image registration system uses an optimization method to iteratively adjust transformation parameters through maximizing a similarity measure function (or minimizing a cost function). The accuracy of such a system depends on the performance of the optimization method and the accuracy and robustness of the similarity measure function. It also depends on the possible presence of nonrigid elements in the registered images. Although deformable changes in the patient anatomy cannot be corrected by a patient setup system that only adjusts rigid transformation parameters, in this paper, we assume that the deformation is relatively small.

The robustness and accuracy of these automated 3D/3D registration systems have been investigated by studying various similarity measures^{5,6} and optimization methods.⁷ Although in ideal conditions, a well-tuned 3D/3D registration system can achieve near 100% success in reaching the global optimum and subvoxel accuracy,⁸ a real system in a clinic environment still suffers from inaccuracy and can fail because of image content changes, organ deformation, the pres-

ence of foreign objects, and imaging system instability. A major reason for registration failure is trapping in a local minimum during the optimization process, which results in a less optimum solution that usually corresponds to a large registration error. Improper settings of termination conditions or the tradeoff between speed and accuracy for practical reasons may also lead to premature terminations, which correspond to less accurate registration solutions. Clinically, the detection of these inaccuracy and errors is performed by radiation therapists or radiation oncologists by visually comparing overlapped slices of the source and the target images shown on the screen. Although many computer displaying techniques^{9–11} have been developed to facilitate this process, this work remains time-consuming, subjective, and inaccurate.

In this paper, we propose to use an algorithm-based registration quality evaluator (RQE) to automatically detect unsuccessful registrations. This would provide us with an automated and objective way to evaluate registration accuracy and reduce the workload of therapists and physicians. When incorporated into an automated 3D/3D registration method, the RQE can significantly increase the robustness of the overall system with only a small increase in overall computation time. The use of a RQE as a pattern classifier to automatically identify unsuccessful registrations was first proposed by Wu and Samant¹² in their application of an automated 2D/2D image registration system for patient setup. In that study, the RQE yielded good performance based on phantom experiments and preliminary clinical studies. The similarity measure value was used as the only feature for the pattern classifier and an optimum cutoff value was calculated to discriminate the good from the bad results. The method was easy to implement. However, the absolute value of a similarity function can be affected by the imaging system stability, the presence of foreign objects, imaging object deformation, and changes of image contents. Moreover, this method is patient and region-of-interest (ROI) specific. The RQE constructed for a particular patient and particular imaging ROI cannot be used for other patients and ROIs. In this paper, we propose to use features that describe the local geometrical properties of a similarity function. These features are less sensitive to the stability of image systems and the changes of imaging contents. We will also show that the RQE constructed and trained with a subset of patients can be used for the future patients.

II. METHODS AND MATERIALS

II.A. Image registration methods

The study used the image sets of six head-and-neck patients. For each patient, a treatment planning fan-beam CT (FBCT) was acquired in the supine position with 3 mm slice thickness. The matrix size in the slice plane was 512×512 for all patients and the voxel varies from 0.29 to 0.59 mm. For each patient, four daily setup cone-beam CTs (CBCT) were used in this study. The CBCTs were acquired using a Varian On-Board imager installed on the Trilogy linear accelerator (Varian Medical Systems, Palo Alto, CA). The

FIG. 1. The axial and the sagittal views of the FBCT and the CBCT of a head-and-neck patient. The FBCT and the CBCT images are shown in the top and the bottom panels, respectively. The ROI is indicated by the white boxes and the isocenter as the circled cross hairs located in the ROI.

CBCT images were resampled to match the voxel sizes of their corresponding FBCTs before the registrations. For each patient, each daily CBCT (i.e., the source images) was registered to the planning FBCT (i.e., the target image), which allowed for four rigid registrations for each patient and a total of 24 registrations for all six patients.

Each registration was based on a user-defined ROI for the target image (i.e., the FBCT). The whole image volume was used for the source image (i.e., the CBCT). The rectangular cuboid ROIs were defined around the treatment isocenter and included cervical vertebrae and adjacent soft tissues. A typical ROI contained about 20 slices, each with approximately 200×200 pixels. Figure 1 shows the axial and the sagittal views of the images from one of the patients in this study. The FBCT and the CBCT images are shown in the top and the bottom panels, respectively. The ROI is indicated by the white boxes and the isocenter by the circled cross hairs located in the ROI. The rigid registrations computed three translational and three rotational parameters. The rotations were defined in the space-fixed convention, with the yaw measured around the patient's anterior/posterior axis, the pitch measured around the left/right axis, and the roll around the inferior/superior axis.

Two similarity measures, mean-squared difference (MSD) and mutual information (MI), were used as the objective functions for the registration. The MSD is a widely used similarity measure for registering images acquired with the same modality. It calculates the average sum-of-square differences of the gray levels between corresponding voxels. For MI calculation, we used the approach proposed by Mattes *et al.*,¹³ in which 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 and a zero-order B-spline and a cubic B-spline as the smoothing kernels for the target and the source images, respectively. Sixty four histogram bins were used for the MI calculation and the linear interpolation method was used to determine the gray levels of the corresponding voxels in the source image.

A gradient-based optimization algorithm, the regular step gradient descent (RSGD) as implemented by the open source software package INSIGHT TOOLKIT (www.itk.org), was used as our optimizer. The RSGD algorithm advances transformation parameters in the direction of the objective function gradient and a bipartition scheme is used to compute the step size.

We then registered the image pairs using both the whole image content in the ROIs and using only voxels representing bony landmarks, which were selected via a Hounsfield unit threshold of 300. 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. For registrations using MI, no downsampling was performed. But for registration using MSD, a 2:1 in-slice downsampling was carried out. This speeded up the registrations by a factor of 4 and had no significant effect on the registration accuracy.¹⁴

II.B. Training and testing data preparation

The task of identifying unsuccessful registration solutions, which are caused by local minimum trapping or early termination, can be considered as a pattern recognition problem. That is, for a given solution, we classify the registration result as one of the two patterns, a successful registration or an unsuccessful registration, according to a number of features that can be calculated from the given solution. Before this pattern classifier, or the RQE, can be put into use, the RQE must be properly trained using the training data. For our application, each entry (or training vector) of the training data set is associated with a registration solution. It consists of a number of features that are calculated based on that particular solution and its true category—A successful one or an unsuccessful registration.

II.B.1. Feature calculations

The similarity measure value was used as the sole feature in a RQE for a 2D image registration system as reported by Wu and Samant.¹⁵ The performance of the RQE using this method could be affected by the imaging system stability, the presence of foreign objects, imaging object deformation, and the change of imaging contents. In this paper, we proposed to use the distinctiveness of optimum⁶ (DO) and the mirror symmetry¹⁶ (MS) to describe the local geometrical properties of the similarity measure function. The DO is based on the assumption that a global minimum generally has a deeper valley compared to that for a local minimum. The MS is based on the assumption that a global minimum has a relatively symmetrical profile centered at its minimum. The pro-

FIG. 2. Definitions of the distinctiveness of the optimum and the mirror symmetry features.

file of a local optimum, on the other hand, is usually more irregular because of its limited capture range. The MS is also very effective in identifying prematurely terminated solutions, in which the solution point is not located at the center of the symmetric profile. Note that for a pattern classifier to work, it is not required that both assumptions be held. A neural network based classifier can automatically learn the patterns from the training data and use it effectively to classify new data if the patterns remain the same.

The calculations of both features require the similarity function be sampled along the x , y , and z axes for given step sizes. For each axis, the sampled points were centered at the current registration solution and two sampling step sizes were used. The smaller step size (i.e., 1 mm) was used to describe the geometrical properties in a smaller region; and the larger step size (i.e., 5 mm) was used to describe the properties in a larger region. The two features are described as follows.

II.B.1.a. DO. The definition of the DO in this paper was adapted from Skerl's definition. As shown in Fig. 2, for a given step size s , the DO is given by

$$DO(s) = \frac{1}{2sN} \sum_{i=1}^N (\bar{f}_i - f_0), \quad (1)$$

where $\bar{f}_i = (f_{i1} + f_{i2})/2$ is the average of the two sampled function values along the axis i . N , which equals 3 in our application, is the number of the sampled axes. f_0 , f_{i1} , and f_{i2} are the function values of the point x_0 , $x_0 - s$, and $x_0 + s$, respectively, on the axis i .

II.B.1.b. MS. Our MS definition was based on Zabrodsky's continuous symmetry measure.¹⁶ As shown in Fig. 2, for a given step size s , the MS is defined as

$$MS(s) = \frac{1}{2N} \sum_{i=1}^N \frac{(f_{i1} - \bar{f}_i)^2 + (f_{i2} - \bar{f}_i)^2}{(f_0 - \bar{f}_i)^2}, \quad (2)$$

where all the symbols are defined the same as in Eq. (1). The difference between our definition and Zabrodsky's definition is that in our definition, the term inside the summation is weighted by the inverse of the square of the depth of the

TABLE I. Range of random disturbances to the initial transformation parameters and the registration success rates when generating training and test data. Abbreviations: MI=matural information; MSD=mean-squared difference.

Similarity measure	Image content	Range of random disturbances		Success rate
		Translational (mm)	Rotational (deg)	
MI	Entire image	± 20	± 8	0.553
	Bony landmark	± 10	± 5	0.127
MSD	Entire image	± 60	± 30	0.617
	Bony landmark	± 30	± 15	0.736

minimum, i.e., $1/(f_0 - \bar{f}_i)^2$. This would remove the dependence of the MS on the depth of the valley, which is desirable since the depth of the valley has already been described by the DO. It is worth noting that the denominator in Eq. (2) may become zero under two trivial cases. One is for a flat local surface, in which $f_0 = f_{i1} = f_{i2}$. The other is the case when $f_{i1} < f_0$ or $f_{i2} < f_0$. In this case, f_0 is not a global minimum. So in both cases, the solution can be easily classified as an unsuccessful registration.

The DOs computed for the small step size (1 mm) and the large step size (5 mm) were denoted by DO(1) and DO(5), respectively. The similar notations were used for MS(1) and MS(5). Above four features were computed for each registration solution.

II.B.2. Determination of true categories

To determine the true category of a registration solution, first we must quantify the registration error. Then a user-defined error threshold can be used to determine whether a particular solution is successful. Thus the “gold standard” registration solutions for each pair of registrations must be established in the first place. Because we do not have an independent and more accurate approach to carry out the registration (e.g., using fiducial markers), the gold standard solutions were taken as the best solutions over 50 registration attempts for each registration problem. For registrations using the whole image content, the initial transformation parameters of multiple registration attempts were randomly distributed within the range of ± 5 mm and $\pm 5^\circ$ in the parametrical space. Because the capture ranges of the similarity measure functions when using only bone landmarks are much smaller, the randomly distributed initial transformation parameters were centered at the best solution given by registration using the whole image content and the magnitudes of disturbances were kept the same. The solution that had the lowest objective function value was considered as the gold standard solution. (In the case that the MI was used as the objective function, we took the negated MI so we were still minimizing the function.) The registered images given by the gold standard solutions were visually verified using an in-house visualization software.

Once the gold standard solutions were established, a new round of 50 registration repetitions were carried out for each registration problem in order to generate the training and test

data set for the RQE. This resulted in a total of 1200 registration solutions (i.e., 6 patients $\times$ 4 registration pairs per patient $\times$ 50 repetitions per registration pair) that could be used to train and test the RQE. In this new round of registrations, the initial transformation parameters were generated by applying a larger magnitude of random disturbances to the gold standard parameters. The ranges of the random disturbances were empirically determined in balancing the number of successful registrations and the number of unsuccessful registrations. Because of the differences of the capture ranges for various combinations of the similarity measures and the image contents, the ranges of random disturbances were separately determined for each registration strategy. The ranges of random disturbances we used in the training and the test data generation and their corresponding registration success rates were given in Table I. For example, MSD+entire image had the largest capture range, so the range of the disturbances was the largest. Note that the range of large random starting positions that we used for the registrations will result in a much larger registration failure rate compared to normal clinical registration practice. However, this is just what we wanted in order to obtain a reasonable number of unsuccessful registrations to train our network without running an extremely large number of registrations.

The registration errors for these new solutions were quantified by comparing with the gold standard solutions. We defined the registration error as the mean voxel displacement (MVD) within the ROI that brings the registered image given by the current registration solution to the “perfectly” registered image given by the gold standard solution. As described by Skerl and colleagues,⁶ this MVD is equivalent to the Euclidean distance from the point of the current solution to the point of the gold standard solution in the normalized parametrical space. For a given center of transform, the units of six transformation parameters were normalized so that the unit changes of each parameter produced the same MVD. Now we can easily calculate the MVD of an individual registration solution from its transformation parameters and a user-defined threshold can be used to classify all solutions with their MVD larger than this threshold as unsuccessful registrations. Empirically, we set the MVD as 1.5 mm throughout this study. In this study, we observed that except for a few exceptions, the registrations were either converged at the solutions within 1 mm from the global optimum or

FIG. 3. Schematic plot of the structure of the neural network used as the registration quality evaluator.

trapped at the local minima at least 5 mm away. A small change of the MVD threshold would not affect the performance of the RQE significantly.

II.C. Training and validation of the neural network

A two-layer feed-forward neural network was trained on the above generated data. The network was implemented using the MATLAB Neural Network Toolbox (The MathWorks, Inc., Natick, MA). Hyperbolic tangent sigmoid transfer functions were used in both the hidden and the output layers. To avoid the problem of overfitting, an effort has been taken to keep the network structurally simple. This was carried out by starting from having 20 neurons in the hidden layer and gradually reducing the number of neurons. We found out that the RQE still has very good performance when the number of neurons in the hidden layer was reduced to two. Thus this network structure, shown in Fig. 3, was chosen and used throughout our study.

The mean sum of squares of the network errors was used as the network performance function. The scaled conjugate gradient algorithm¹⁷ was used as the network optimization method. This method is based on the conjugate gradient method but was designed to avoid the time-consuming line search by combining with the model-trust region approach.

A common problem in neural network construction is overtraining, in which the weights of the network are too closely optimized to particular details of the training data set, but perform poorly on a new data set that is not used in training. To overcome the problem of overtraining and improve generalization, the early stopping technique was adopted, in which the available data were divided into three subsets. First, the data generated from each individual patient were chosen in turns as the test data set. Then the rest of the data from four other patients were randomly divided into the training data set (80%) and the validation data set (20%). The training set was used for computing the gradient and

updating the network weights and biases. After each training iteration, the network was tested on the validation set. The error on the validation set was monitored during the training process. Initially the error decreased, but at some later stage it started increasing. This was the point at which the network began to adapt specifically to the idiosyncrasies of the specific training set, thus adversely affecting the generalization performance of the network. When the validation error increased for a specified number of iterations, the training was stopped and the weights and biases at the minimum of the validation error were returned. The test data had no effect on training and they provided an independent measure of network performance during and after training. Thus it provided us with a valid method to evaluate how the network was generalized to new patients. Four criteria, sensitivity, specificity, positive predicative value (PPV), and negative predicative value (NPV), were used to evaluate the performance of the classifier. The definitions of these criteria were given in Table II.

III. RESULTS

The scatter plots of all data that we used in training, validation, and testing were shown in Figs. 4(a)–4(d). In all figures, the data points were plotted over two features: DO(1) and DO(5). The data computed using MI were shown in the figures on the top, whereas the data using MSD were shown in the figure on the bottom. The data calculated using the full intensity range are shown in the figures on the left, whereas figures on the right are based on registrations using bony landmarks. As shown in these figures, the unsuccessful registrations (plotted as crosses) can be separated from the successful ones (plotted as circles) with relatively less difficulty when MI was used as the similarity measure function. The same conclusion can be drawn when the other two features [MS(1) and MS(5)] were used, as indicated in the scatter plots shown in Figs. 5(a)–5(d). Because of the large dynamic ranges of these two features, the data were plotted in the logarithmic scales. As shown in Figs. 5(a) and 5(b), the data computed using MI show good clustering, indicating good feature performance.

To better understand how and why the proposed geometrical features can distinguish between the global minimum and local minima, the cost functions were probed along the x, y, and z axes from the point-of-solutions in the parametrical space. As shown in Figs. 6(a)–6(d), the negated MI values were plotted against the shifts from the selected solution

TABLE II. Definitions of the criteria that were used to evaluate the performance of a classifier. Abbreviations: TP=true positive; FP=false positive; FN=false negative; TN=true negative; PPV=positive predictive value; NPV=negative predictive value.

		True category		
		Unsuccessful	Successful	
Classifier output	Unsuccessful	TP	FP	PPV=TP/(TP+FP)
	Successful	FN	TN	NPV=TN/(FN+TN)
		Sensitivity=TP/(TP+FN) Specificity=TN/(FP+TN)

FIG. 4. Scatter plots of all the data over two features: DO(1) and DO(5). The networks were constructed using (a) mutual information and the whole image content; (b) mutual information and bony landmarks; (c) mean-squared difference and the whole image content; and (d) mean-squared difference and bony landmarks.

points. The sampled points used in feature calculations are shown as the square marks in the plots. The calculated four feature values are also shown besides the curves. The center points in Figs. 6(a) and 6(b) are the gold standard solutions and the cost function curves show good symmetry and have nice deep valleys. These geometrical properties have been

FIG. 5. Same as Fig. 4, except that the data were plotted over two other features: MS(1) and MS(5).

FIG. 6. Cost function (i.e., the negated mutual information computed using the whole image content) profiles probed along the x, y, and z axes from the selected point-of-solutions. The profiles centered on the successful solutions from two different patients are shown in (a) and (b). The profiles centered on the local minimum solutions are shown in (c) and (d).

indicated by their small MS values and large DO values. The two figures on the bottom show failed solutions that were trapped in local minima. In particular, Fig. 6(d) shows a typical local minimum when the optimization process is drifting away from the global minimum. The cost function near these solutions lost symmetry and distinctiveness, which are indi-

FIG. 7. An example of the misregistration caused by local minimum trapping. The target FBCT images were shown on the left panels and the registered CBCT images on the right. The top panels give the axial views and the sagittal views show on the bottom.

TABLE III. Classifier performance when the mutual information and the whole image content were used. Abbreviations: See Table II.

Test data set	Sensitivity	Specificity	PPV	NPV
Patient 1	1.000	1.000	1.000	1.000
Patient 2	1.000	0.993	0.983	1.000
Patient 3	0.960	1.000	1.000	0.994
Patient 4	0.993	1.000	1.000	0.985
Patient 5	1.000	1.000	1.000	1.000
Patient 6	1.000	1.000	1.000	1.000

cated by their large MS values and small DO values. Figure 6(c) presents one of the toughest cases in our study, in which the similarity function curve looks very similar to those for the global minima. However, there are still substantial differences that are indicated by the decreased DO values and the increased MS values. These differences had been used by the RQE to successfully identify those local minima in our study. The cause of this particular local minimum was investigated and the registered images generated by the transformation parameters associated with this local minimum were shown in Figs. 7(a)–7(d). The target FBCT images were shown on the left panels and the registered CBCT images on the right. The top panels gave the axial views and the sagittal views were shown on the bottom. The CBCT image appears to be fairly aligned. However, a close inspection revealed that the CBCT image had been shifted one vertebra down. In other words, the C6 in the FBCT image was registered to the C5 in the CBCT image. This type of registration failure can be very difficult to identify by visual inspection.

The performances of the classifiers constructed using the MI and the whole image content, the MI and bony landmarks, the MSD and the whole image content, and the MSD and bony landmarks, are shown in Tables III–VI, respectively. As described earlier, the test data set computed from a selected patient was not used in training. Thus, the performance of the classifier given in these tables virtually predicted how the classifier would perform on the registration solutions from new patients. For a perfect classifier that gives no classification errors, the sensitivity, the specificity, the PPV, and the NPV should all equal to 1. As shown in Tables III and IV, the classifiers constructed using the MI have very good performances. The sensitivity and the specificity range from 0.96 to 1.0 and from 0.993 to 1.0, respectively. On the other hand, as shown in Tables V and VI, the

TABLE IV. Classifier performance when the mutual information and bony landmarks were used. Abbreviations: See Table II.

Test data set	Sensitivity	Specificity	PPV	NPV
Patient 1	1.000	1.000	1.000	1.000
Patient 2	1.000	1.000	1.000	1.000
Patient 3	0.989	1.000	1.000	0.909
Patient 4	1.000	1.000	1.000	1.000
Patient 5	1.000	1.000	1.000	1.000
Patient 6	1.000	1.000	1.000	1.000

TABLE V. Classifier performance when the mean-squared difference and the whole image content were used. Abbreviations: See Table II.

Test data set	Sensitivity	Specificity	PPV	NPV
Patient 1	1.000	1.000	1.000	1.000
Patient 2	0.937	1.000	1.000	0.960
Patient 3	0.966	1.000	1.000	0.994
Patient 4	1.000	1.000	1.000	1.000
Patient 5	0.981	1.000	1.000	0.993
Patient 6	0.266	1.000	1.000	0.455

RQE constructed using the MSD did not perform well in a few cases. The sensitivity of the RQE for patient 6 was only 0.266 when the MSD and the whole image content were used. When the MSD and the bony landmark were used, the sensitivities were low for patients 4–6. This indicates the geometrical features proposed in this paper are not reliable when the MSD is used as the similarity measure function for the head-and-neck registration applications.

IV. DISCUSSION

IV.A. Training data generation method

In the RQE proposed by Wu and Samant¹⁵ for 2D/2D image registrations, the training data were generated by randomly shifting the transformation parameters from the gold standard solution in the parametrical space and computed the corresponding similarity function values. This method has the advantage of requiring much less calculation time since no registrations are required. However, the simulated training data are not the typical registration solutions that may be encountered in real registration applications. There is no guarantee that the simulated point in the parametrical space would be a local minimum and an optimizer would present it as a solution. Owing to this fact, we used the real registration solutions as the training data in this study. Because a RQE classifier only needs to be trained once for a particular registration application, the increased training data preparation time is not a concern. For our applications, using an Intel Core 2 Extreme 3 GHz PC, the data preparation time was about 2 h when the MI was used as the similarity measure and about 3 h when the MSD was used. The classifier training time required just a few seconds.

TABLE VI. Classifier performance when the mean-squared difference and bony landmarks were used. Abbreviations: See Table II.

Test data set	Sensitivity	Specificity	PPV	NPV
Patient 1	1.000	1.000	1.000	1.000
Patient 2	1.000	1.000	1.000	1.000
Patient 3	0.903	1.000	1.000	0.983
Patient 4	0.464	1.000	1.000	0.920
Patient 5	0.864	0.925	0.851	0.932
Patient 6	0.669	1.000	1.000	0.664

IV.B. Geometrical feature calculation

To calculate four geometrical features, the similarity functions values at 12 sampling points need to be computed. This is translated to about 10%–20% increase in the overall registration time. For applications where the speed of the registration is of great concern, the number of sampling points may be reduced (e.g., sampling using only a small step size or a large step size, or sampling only along the x, y, or z axis). Due to the high correlations among a number of quantities, the performance of the classifier may not be greatly affected. If necessary, a future study may also be carried out to investigate the impact of the sampling step size on the classifier performance.

IV.C. Precision vs accuracy

The proposed RQE classifier used with MI provided us with a reliable way to tell whether a registration solution converged to the global minimum. However, it reveals nothing about how far away this global minimum deviates from the true alignment location. The difference is the systematic error of the similarity measure that can be estimated by conducting a similarity measure study and/or a forward/backward registration method.¹⁴ The systematic error can be affected by the imaging system calibration error, the imaging system stability, the changes of the region-of-interest, the deformation of the imaging content, and the presence of foreign objects. Fortunately, the systematic error is usually small (typically <2 mm) under the controlled clinic environment for daily patient setup.

IV.D. Clinical registration scheme

The RQE proposed in the paper can be readily incorporated into a clinical image registration system for patient setup. For example, a multiresolution registration strategy can be used. In the coarse resolution, a less accurate but more robust similarity measure, such as the MSD, can be used. Once the optimization process is close to a global solution, a more accurate similarity measure, such as the MI, can be used instead to drive the optimization. When the registration is terminated, the similarity measure function is sampled and the DO and the MS are computed. The RQE, which has been constructed offline using the historical image database, is used to evaluate the registration solution. If the registration is successful, the setup adjustment is applied to the patient if necessary and then start with the daily treatment. Otherwise, the registration can automatically start over using a different set of initial transform parameters (e.g., adding a randomly generated vector). If it still does not pass after a number of tries, a different ROI may need to be defined or the registration has to be carried out manually. The registration solutions and their verified categories can be saved and used retrospectively to improve the RQE or update the RQE adaptively.

IV.E. Generality of the RQE method

The application of the RQE on the head-and-neck patient is probably the toughest case compared to rigid registrations on other anatomical sites. The misregistrations caused by the alignments of wrong vertebrae generate very deep local minima and have the local geometrical properties of the similarity function very similarity to those for the global optima. Even so, the RQE constructed using MI was still able to identify those local minima with very few classification errors.

These deep local minima are very unique for the registration problems involve only a few vertebrae and are unlikely to happen on other anatomical sites. For other applications, the classification job would be relatively easier and the performance of the RQE is expected to be even better. The concept of the RQE can be readily applied to rigid 2D/3D registrations as well.

V. CONCLUSIONS

Our patient study has demonstrated the proposed neural network based RQE had very good performance when used with the MI in identifying unsuccessful 3D/3D registrations for daily patient setup. The classifier had very good generality and required only to be trained once for each implementation. When the RQE is incorporated with an automated 3D/3D image registration system, it can improve the robustness of the system while contributing only a small fractional increase of overall processing time.

ACKNOWLEDGMENTS

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).

⁵D. 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).

⁶D. Skerl, B. Likar, and F. Pernus, "A protocol for evaluation of similarity measures for rigid registration," *IEEE Trans. Med. Imaging* **25**, 779–791 (2006).

⁷F. Maes, D. Vandermeulen, and P. Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," *Med. Image Anal.* **3**, 373–386 (1999).

⁸P. Thevenaz and M. Unser, "Optimization of mutual information for mul-

- tiresolution image registration," *IEEE Trans. Image Process.* **9**, 2083–2099 (2000).
- ⁹M. S. Weinhaus, "Treatment verification using a computer workstation," *Int. J. Radiat. Oncol., Biol., Phys.* **19**, 1549–1554 (1990).
- ¹⁰P. M. Evans, J. Q. Gildersleve, E. J. Morton, W. Swindell, R. Coles, M. Ferraro, C. Rawlings, Z. R. Xiao, and J. Dyer, "Image comparison techniques for use with megavoltage imaging systems," *Br. J. Radiol.* **65**, 701–709 (1992).
- ¹¹M. L. Graham, A. Y. Cheng, L. Y. Geer, W. R. Binns, M. W. Vannier, and J. W. Wong, "A method to analyze 2-dimensional daily radiotherapy portal images from an on-line fiber-optic imaging system," *Int. J. Radiat. Oncol., Biol., Phys.* **20**, 613–619 (1991).
- ¹²J. Wu and S. S. Samant, "Registration quality evaluator: Application to automated patient setup verification in radiotherapy," in *Medical Imaging 2004: Image Processing*, edited by M. Sonka and J. M. Fitzpatrick (SPIE, San Diego, 2004), Vol. 5370, pp. 137–142.
- ¹³D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellyn, and W. Eubank, "Nonrigid multimodality image registration," in *Medical Imaging 2001: Image Processing*, edited by M. Sonka and K. M. Hanson (SPIE, San Diego, 2001), Vol. 4322, pp. 1609–1620.
- ¹⁴J. Wu and M. J. Murphy, "Assessing the intrinsic precision of 3D/3D rigid image registration results for patient setup in the absence of a ground truth," *Med. Phys.* **37**, 2501–2508 (2010).
- ¹⁵J. Wu and S. S. Samant, "Novel image registration quality evaluator (RQE) with an implementation for automated patient positioning in cranial radiation therapy," *Med. Phys.* **34**, 2099–2112 (2007).
- ¹⁶H. Zabrodsky, S. Peleg, and D. Avnir, "A measure of symmetry based on shape similarity," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '92)*, Cham-paign, IL, 1992, pp. 703–706.
- ¹⁷M. F. Moller, "A scaled conjugate-gradient algorithm for fast supervised learning," *Neural Networks* **6**, 525–533 (1993).