A method to detect landmark pairs accurately between intra-patient volumetric medical images

Deshan Yang[a)]
Department of Radiation Oncology, Washington University in Saint Louis, Saint Louis, MO, USA

Miao Zhang
Department of Physics and Astronomy, University of Missouri, Columbia, MO, USA

Xiao Chang, Yabo Fu, Shi Liu, Harold H. Li, and Sasa Mutic
Department of Radiation Oncology, Washington University in Saint Louis, Saint Louis, MO, USA

Ye Duan
Department of Computer Science & IT, University of Missouri, Columbia, MO, USA

(Received 11 April 2017; revised 14 June 2017; accepted for publication 14 August 2017; published 13 September 2017)

Purposes: An image processing procedure was developed in this study to detect large quantity of landmark pairs accurately in pairs of volumetric medical images. The detected landmark pairs can be used to evaluate of deformable image registration (DIR) methods quantitatively.
Methods: Landmark detection and pair matching were implemented in a Gaussian pyramid multi-resolution scheme. A 3D scale-invariant feature transform (SIFT) feature detection method and a 3D Harris–Laplacian corner detection method were employed to detect feature points, i.e., landmarks. A novel feature matching algorithm, Multi-Resolution Inverse-Consistent Guided Matching or MRICGM, was developed to allow accurate feature pairs matching. MRICGM performs feature matching using guidance by the feature pairs detected at the lower resolution stage and the higher confidence feature pairs already detected at the same resolution stage, while enforces inverse consistency.
Results: The proposed feature detection and feature pair matching algorithms were optimized to process 3D CT and MRI images. They were successfully applied between the inter-phase abdomen 4DCT images of three patients, between the original and the re-scanned radiation therapy simulation CT images of two head-neck patients, and between inter-fractional treatment MRIs of two patients. The proposed procedure was able to successfully detect and match over 6300 feature pairs on average. The automatically detected landmark pairs were manually verified and the mismatched pairs were rejected. The automatic feature matching accuracy before manual error rejection was 99.4%. Performance of MRICGM was also evaluated using seven digital phantom datasets with known ground truth of tissue deformation. On average, 11855 feature pairs were detected per digital phantom dataset with TRE = 0.77 ± 0.72 mm.
Conclusion: A procedure was developed in this study to detect large number of landmark pairs accurately between two volumetric medical images. It allows a semi-automatic way to generate the ground truth landmark datasets that allow quantitatively evaluation of DIR algorithms for radiation therapy applications. [https://doi.org/10.1002/mp.12526]

Key words: deformable image registration, feature matching, image feature detection, medical image processing, radiation therapy

1. INTRODUCTION

Deformable image registration[1–3] (DIR) is a key enabling technology for many important advanced radiotherapy techniques (e.g., adaptive radiotherapy[4–6]) and critical clinical tasks (e.g., target definition,[7,8] automatic segmentation,[9–12] motion estimation,[13–16] dose accumulation[17–20] and treatment response evaluation[5,6,21,22]). The current DIR algorithms compute the tissue deformation by minimizing one or multiple of (a) the image intensity difference, (b) irregularities of deformation, and (c) the boundary mismatches of delineated structures.[4] DIR accuracy, which is the correspondence of matching points between two images under DIR, is often inadequate and largely dependent on the operator, DIR algorithm, implementation, and image quality. It is frequent that DIR workflow (e.g., pre-DIR rigid registration, uses of region-of-interest) and parameters need to be manually adjusted repetitively on a trial-and-error basis before a "reasonable" result (visually acceptable but the absolute accuracy cannot be quantitatively assessed) can be attained. It is critical to evaluate DIR algorithms quantitatively using the benchmark datasets before the accuracies could be assessed

and the DIR parameters could be understood and optimized by the operators.

Many DIR evaluation methods have been previously published in the literature. Existing methods can be roughly grouped into three categories. (a) Analyzing image intensity and deformation vector field (DVF) to assess DIR accuracy indirectly. Image intensity based metrics include the sum of squared intensity difference (SSD), mutual information (MI) and cross-correlation (CC).[23,24] DVF based metrics include Jacobian, stress, inverse consistency, transverse consistency, smoothness, divergence, unbalanced stress energy.[25–33] Because none of these metrics is directly related to DIR accuracy, methods based on these metrics cannot give a quantitative assessment of DIR errors (in term of mm), would fail in many cases, and cannot be trusted to support clinical decision.[23] (b) Using digital phantom with artificially generated deformation, or physical phantoms with known deformation.[25,28,34–39] These methods could test the absolute DIR accuracy, but only on these phantom images. The measured DIR errors for any DIR algorithm are not generalizable to patient images because the patient deformation is not known, much more irregular and complex[23,34] than the artificially generated and usually very smooth deformation. (c) Using manually selected landmarks and manually delineated structure contours to compute TRE (target registration error), contour displacement, or volume overlapping ratio.[30,31,40–44]

To compute Target Registration Error (TRE) on manually selected landmarks is commonly known as the only trustable way to test DIR accuracy on a patient image dataset. However, the manual landmark selection process is very labor intensive if a relatively large number of landmarks are to be selected. To authors' knowledge, the lung DIR evaluation framework[42] by Castillo et al. is the only published work in which a total 6762 landmarks were manually registered over the set of five thorax 4DCT image datasets. It took the authors a total of 60 h. Manual selecting landmarks can be subjective and biased. For example, human observers select the most visually apparent landmark points, e.g., the bifurcation points, in a lung CT image. However, such bifurcation points usually have very strong image gradient and therefore DIR algorithms would be naturally more accurate at such points than at image voxels with minimal image gradient. TRE evaluated at these biasedly selected landmarks might be underestimated.

A few methods have been proposed to detect the landmarks automatically in CT image pairs. Veckress et al. applied SIFT (Scale Invariant Feature Transform) detection and matching methods and detected on average 64 landmarks per lung.[45] On average four landmarks (or 6%) were rejected after manual verification. Paganelli et al. reported a method using SIFT feature detection and matching methods to verify the results of DIR.[46] Between the radiation therapy simulation CT and cone-beam CT for head-neck cancer patients, on average 50 to 250 landmarks were detected per case. The average accuracy of landmark pair matching was about 96.5%, and about 50% of the detected landmark pairs were for bony structures. Mazur et al. reported a method to detect SIFT features in 2D cine MRIs and compute tissue motion by tracking the local groups of SIFT features.[47] Murphy et al. constructed DIR benchmark datasets of 47 pairs of temporal thoracic CT scans, 100 landmark pairs per case, using a customized automatic landmark detection method following by automatic block-matching based landmark matching and manual confirmation.[48] Authors thoroughly described and evaluated the manual confirmation process, however, they did not report the accuracy of the automatic feature matching prior to manual confirmation.[48]

Current SIFT feature detection and matching methods are not mature enough to be a viable solution for automatic verification of DIR results on arbitrary patient dataset due to two correlated problems: (a) inability to detect larger number of features and (b) inability to accomplish higher feature pair matching accuracy. A density of 64 landmarks per lung,[45] as 1 to 2 landmarks per axial slice, is inadequate to assess DIR accuracy in whole lung. A feature detection and matching procedure with a matching accuracy of 94%,[45] which seems to be very high, cannot be fully trusted as an automatic DIR verification procedure.

In authors' observation, these two problems are actually rooted in a single problem—inaccuracy in feature pair matching. In fact, it is very simple to use a lower threshold value in the SIFT detection to detect tens of thousands of SIFT features in a standard size CT volume. The standard SIFT feature pair matching method, based on the feature descriptor similarity measurement (to be described in Section 2.F.1), is however not optimized to accurately match the feature pairs among such larger number of features. Inspired by this observation and motivated to accurately detect large number of feature pairs (i.e., landmark pairs), we have developed (a) a novel procedure to perform feature detection and feature pair matching in the multiple-resolution pyramid scheme, and more importantly, (b) a novel inverse-consistent and guided feature matching method. The proposed procedures were successfully applied in multiple different clinical cases—between the inhale and exhale phases of abdominal 4DCT images, between the original and rescanned radiation therapy simulation CT for head-neck patients, and between MRI obtained at different treatment fractions. The proposed procedures were able to detect thousands of feature pairs with pair matching accuracy >99% in each case, and noteworthy large number of feature pairs in the low contrast soft tissue regions without potential bias associated with manual landmark selection. The generated feature pairs, i.e., landmarks, after a quick manual verification and outlier rejection, are useful as ground-truth landmark dataset for evaluation of any DIR algorithms. With further development, the proposed method is potentially useful as a fully automated tool for verification of DIR on arbitrary patient image datasets.

2. MATERIAL AND METHODS

2.A. A multiple-resolution feature detection and matching workflow

The workflow for the proposed feature pair detection/matching procedure is shown in Fig. 1. This workflow is

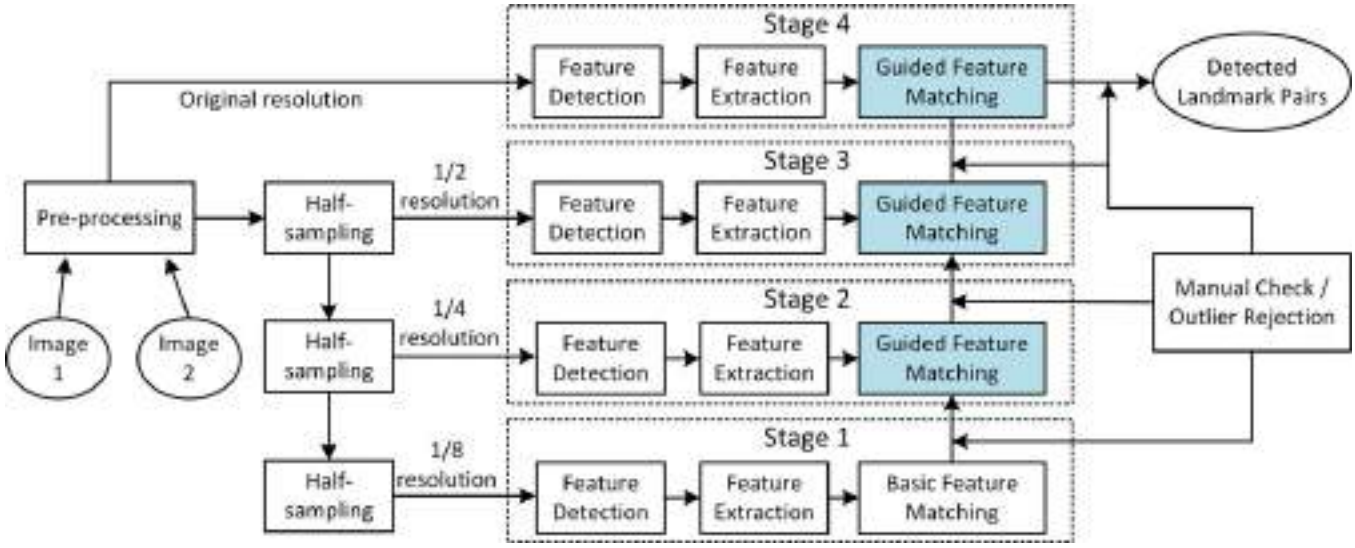

FIG. 1. The proposed feature detection and matching workflow. Image features are detected and matched at 4 stages in the order of 1/8, 1/4, 1/2, and full resolutions. Guided feature matching, indicated by the shaded blocks, is employed at the higher resolution stages (stages 3, 2, and 1). The dashed lines indicate the optional components or procedures. [Color figure can be viewed at wileyonlinelibrary.com]

optimized for two primary goals—to detect as many as possible feature pairs and to have high accuracy (e.g., >99%) in feature matching. To accomplish these two goals, we have made two important unique and novel design choices.

First, image features are detected and matched in the multiple-resolution pyramid scheme. Our hypothesis is that both feature detection and feature pair matching at a lower resolution stage are more robust and accurate than the features detection and matching steps at a higher stage, even though the locations of the detected features might be less precise. The local details in the images, that would compromise the feature detection and matching at a higher resolution stage, would be reduced at a lower stage. Empirically determined in preliminary experiments, four stages were used in this study. The image resolutions at the stages 1 to 4 are corresponding to 1/8, 1/4, 1/2 and 100% of the original image resolution respectively. The images applied at the stage 3 are half-sampled from the stage 4 images using a Gaussian pyramid filter,[49] and further half-sampled for stages 2 and 1. Image feature detection and matching are sequentially performed starting at the lowest stage, e.g., the stage 1. Matching results at a lower stage are used to guide pair matching at the stage one level higher. The workflow is designed to support generic feature detectors and feature descriptors. In this work, a combination of SIFT and Harris–Laplacian corner detectors were employed to detect more feature points.[46,50] The SIFT 3D feature descriptor was employed.[46]

Second, the novel multi-resolution inverse-consistent guided feature matching algorithm, or MRICGM, was developed to allow accurate feature matching. With the detail to be further explained in 2.6, MRICGM is used in the workflow in two ways. The matching results from a lower resolution stage are used to guide feature matching at the higher resolution stage. The higher confidence matching results at a given stage, obtained using a tighter threshold, are used to guide additional iterations of feature matching using looser thresholds to allow more feature pairs to be matched and to retain the accuracy.

2.B. Materials

Three types and seven cases of patient image pairs were included in this study as listed in Table I. These images were all acquired during the standard patient cancer treatments at authors' institution. With an IRB approval, they were retrospectively obtained from the clinical treatment systems and de-identified. These three types were selected because tissue motion is significant between two images in each pair, the motion magnitude would generally pose challenges to most DIR algorithms and no DIR benchmark datasets of these types has been previously published. As multiple thorax 4DCT DIR ground-truth datasets were previously published and publically available, thorax 4DCT datasets were therefore of relatively lower priority and currently not included in this study.

The 4DCT images in the cases 1 to 3 were of patients with abdominal cancer. Full sets of 4DCT images were acquired using the respiratory phase re-binning method and reconstructed in 10 respiration phases. Only the end-of-inhale and the end-of-exhale phases were included in this study because the tissue motion is the most significant between these two phases. Case 4 was of a head-neck cancer patient's two treatment plan CTs of 20 days apart. In the second CT, a lower neck support was used and patient's neck was less curved. Case 5 was of another head-neck cancer patient. The first CT image was the radiation treatment plan CT and the second image was the diagnostic CT with the neck surrounded by a soft pillow. The original simulation CT and the rescan simulation CT were included in this study. The cases 6 and 7 were of patients with pelvic cancer who received the MRI guided

TABLE I. List of patient image datasets utilized in this study.

Types	Case #	Image information
Abdominal 4DCT	1, 2, 3	In-plane resolution = 0.9866–1.2 mm, Slice thickness = 3 mm
Head-neck CT	4, 5	In-plane resolution = 0.9866–1.2 mm, Slice thickness = 3 mm
Pelvis fractional MRI	6, 7	In-plane resolution = 1.5 mm, Slice thickness = 1.5 mm

adaptive radiation therapy[51] on an MR guided RT (MRIdian, Viewray, Oakwood Village, OH, USA) treatment machine.

2.C. Preprocessing

For each CT image, the skin surface was detected by thresholding the image at HU = −700 on each 2D axial slice and filling the holes with a morphologic flood fill filter. All voxels outside the skin mask was set to 0 to avoid feature detection on the noise voxels outside the body. The images were then cropped in the axial plane to remove the CT couch table and excessive empty space outside the skin mask. The cropped images were denoised using a 3D bilateral filter[52] with the distance sigma = 1 and intensity sigma = 20.

Similar preprocessing steps were applied for each MR image. A thresholding value = 20 was used in skin surface detection. The distance sigma = 1 and the intensity sigma = 5 were used in the 3D bilateral denoising filter.

2.D. Feature detection

To detect more features in a patient image pair, we employed a 3D SIFT feature detector[46] and a 3D Harris–Laplacian corner detector[50] in this study. Our preliminary results had shown that these two feature point detection methods worked complementarily and detected different features.

2.D.1. SIFT (Scale Invariant Feature Transform) feature detection

A SIFT feature detector detects local extrema of the difference of Gaussian (DOG) scale space of the 3D image.[53,54] The Gaussian scale space is created by applying Gaussian smoothing repetitively on the 3D image and arranging the series of 3D smoothing results into a 4D space. DOG is calculated as the differential of the ordered smoothing results in the 4th dimension.

Cheung et al. generalized SIFT from 2D to N-dimensions and showed potential applications in medical images.[54] However their implementation in Linux[55] performed poorly on all our datasets, detecting less than 40 and mostly bad matches per case. We have therefore implemented our own version of 3D SIFT method with the following important improvements. (a) Any points outside the skin mask are rejected. (b) The point positions and scales are iteratively refined in the 4D DOG space based on the 4D image gradient according to the method previously published by Otero.[56] (c) The points on the edges and in the low contrast regions are rejected based on the Eq. (1), i.e., point feature strength, measured using the structure tensor matrix per point.[57] (d) The duplicated points, that are the points within 1 voxel distance from the kept points, are removed. Parameters of 3D SIFT algorithm used in this study are provided in Table II. Readers could find additional information about SIFT algorithm in the original papers.[46,54]

2.D.2. Harris–Laplacian corner point detection

Corner points in a 3D image are the points having significant intensity gradient in all three cardinal directions. Harris algorithm detects corner points by identifying the local minima in the map of ratios between the trace and the determinant of the structure tensor matrix at every voxel, and the local minima points shall satisfy

$$\frac{Trace^3(M)}{Det(M)} < \frac{(1+2r)^3}{r^2} \tag{1}$$

where M is the structure tensor matrix (at every voxel) and r is an empirical threshold.

To determine the scales (i.e., sizes in voxels) of the corner points, a series of smoothed images are created by applying Gaussian smoothing repetitively. A corner point and its scale (i.e., on which smoothed image) are selected only if the Laplacian (the second order spatial derivative) of the point at the scale is greater than the Laplacians of the point at the adjacent scales. Parameters of the Harris–Laplacian corner detection algorithm used in this study are provided in Table II. Readers could find additional information of Harris–Laplacian corner detection algorithm in the original paper by Lindeberg.[50]

2.E. Feature description

Feature description is the process to extract and process voxel information in the neighborhood of a feature point, and to construct the feature descriptor. The SIFT descriptor was used in this study because it is robust to affine transformation and changes in illumination, and had been successfully applied in many previous medical image processing studies.[45–47] Other feature descriptors, including SURF[58] and BRISK,[59] were also implemented and experimented in this study but were not chosen because of lower matching accuracy and greater inverse-consistent rejection rates (to be described in Section 2.F.6).

2.F. Multi-resolution, inverse-consistent guided matching (MRICGM)

2.F.1. The common feature matching method and the challenges

Feature matching is the process to identify the matching between two sets of image features that have been independently detected in two images. The regular method is, for each feature detected in one image, to exclusively search the

TABLE II. Parameters and values used in this study.

Method	Parameters	Values
SIFT feature detection	Threshold	0.01
	Gaussian scale space length	6
	Gaussian kernel σ	[1, 1.15, 1.32, 1.52, 1.74, 2]
	Number of Octaves	1
	Number of iterations in point location refinement	25
	Duplication removal distance threshold	1 voxel
Harris–Laplacian corner detection	Threshold parameter r	10
	Number of Gaussian smoothing steps	5
	Gaussian kernel σ	[1, 1.26, 1.59, 2, 2.52]
SIFT feature descriptor	Number of histogram bins per quadrant	20
	Number of quadrant in 3D	8
MRICGM (multi-resolution inverse-consistent guided matching)	Number of multi-resolution stages	4
	Maximal number of guidance feature pairs	10
	Range to select guidance feature pairs	15 mm × (5—stage number)
	Range to search for matching points	15 mm × (5—stage number)
	Number of iterations per stage	2 (for stages 1 to 3) and 5–7 for stage 4
	Descriptor matching threshold t_1 for stages 1 to 3	0.2, 0.3 (2 iterations)
	Guidance matching threshold t_2 for stages 1 to 3	0.2, 0.3 (2 iterations)
	Descriptor matching threshold t_1 for stage 4	0.2, 0.3, 0.4, 0.5 and 0.5 (5 iterations)
	Guidance matching threshold t_2 for stage 4	0.2, 0.2, 0.2, 0.2 and 0.3 (5 iterations)
	Ratio between the 1st and 2nd best overall confidences	1.11
	Minimal matching confidence for a matched feature pair to be used for guidance	0.95
Standard feature matching method	Range to search for matching points	20 mm
	Ratio between the 1st and 2nd best overall confidences	1.11
	Descriptor matching threshold t_1	Adjustable[a]

[a]The descriptor matching threshold is adjustable for the standard feature matching method to control the total number of detected feature pairs.

best matched one from the all available features detected in the second image. The *descriptor matching confidence* C_D between two features, i.e., a feature p in the set #1 and the feature q in the set #2, is computed as the dot product of two corresponding feature descriptor vectors:

$$C_D(p,q) = D(p) \cdot D(q) \quad (2)$$

where $D(p)$ and $D(q)$ are the feature descriptors for p and q. C_D ranges between 0 (complete mismatch, perpendicular to each other) and 1 (complete match).

Feature matching is not trivial task considering that (a) the quantities of features detected in each 3D medical image are very large, often in tens of thousands, (b) feature detection is neither stable nor repeatable between two images due to image noise, tissue motion, postural and anatomical changes, and (c) large amount of features are very similar to each other and cannot be differentiated solely based on their feature descriptors.

Many improvements could be introduced to the regular feature matching procedure. Two images can be rigidly registered so that corresponding features are closer to each other. In addition, matching feature searching could be limited to a maximal distance therefore to reduce the total computation cost, assuming the magnitude of tissue motion between these two images is limited to certain maximal distance, e.g., 30 mm. Ambiguous matchings, for which the second best match is as good as its best match, could be discarded. A RANSAC (Random Sample Consensus) method[60] or its varieties could be applied after the matching procedure to reject the outliers.

However, as authors had experimented in preliminary studies, feature matching even with the improvements is insufficient for medical images. As observed in initial experiments, the feature matching accuracy was <95% if the number of feature pair matchings was limited to 200 (by controlling the feature matching thresholds), and <90% and <60% if the numbers of matching were extended to 1000 and 5000, respectively. Such observed accuracies were similar to previous published results.[45,46]

2.F.2. Obtain the high confidence matchings using the multi-resolution and multi-step scheme

To improve the feature matching accuracy, the novel method—multi-resolution inverse-consistent guided matching, or MRICGM was developed in this study. The general idea is that the high confidence matchings can be detected first and then be used to guide the detection of additional matchings.

In MRICGM, the high confidence matchings are obtained in two ways. First, feature detection and matching are implemented in a multi-resolution scheme, as shown in Fig. 1. Four resolution stages are used in this study. The stages 1 to 4 are corresponding to the 100%, 1/2, 1/4, and 1/8 of the original resolution. Gaussian pyramid half-sampling filter[49] was applied. This design choice was based on our observations that image features in the lowered resolution of the images are in much less quantity but each feature is more stable because the most image details are blurred out by the half-sampling process. The stable feature could then be matched with high confidence. Second, at a given stage, feature

matches are implemented in multiple iterations. Higher confidence feature matchings are detected first by using tighter threshold. The detected higher confidence feature pairs are then applied to guide the matching of additional feature pairs in the following iterations with sequentially loosing thresholds.

2.F.3. Feature matching with guidance

The guided matching procedure is illustrated in Fig. 2. Note that the illustration is in 2D for the purpose of simplicity but the actual implementation is in 3D. The goal is to identify the matching point for the point P in the first image (I_1) among the available points in the second image (I_2), provided there is a good match in I_2. The points A to I in I_2 are the available points in the close proximity of P', which is the position of P mirrored to I_2 based on the image registration between these two images. The points 1 to 7 in I_1 and $1'$ to $7'$ in I_2 are the already matched feature pairs, obtained either at the lower resolution stage or in the previous iteration at the same resolution stage, with good matching confidence. One shall note that (a) the point A is slightly off from P' in I_2 and (b) the positions of the points 1 to 7 in I_1 and the points $1'$ to $7'$ in I_2 are also slightly off.

For a point, e.g., the point P in I_1, its *guidance vector* V is defined as:

$$V' = [(X_P - X_1), (X_P - X_2), \cdots, (X_P - X_N)] \tag{3}$$

$$V = V'/|V'| \tag{4}$$

where $X_P, X_1, \cdots, X_N$ are the position vectors for the point P and the guidance points 1 to N in the same image of P, V' is essentially the concatenation of N differential vectors between P and its guidance points, and V is V' normalized to the unit length. In the example shown in Fig. 2, $N = 5$ and the guidance points for P are the points 1 to 5 in I_1.

Similarly, for each available point in the other image, i.e., the points A to I in I_2, the corresponding guidance vectors can be computed for the point, e.g., the point A, and the corresponding guidance points, i.e., the points $1'$ to $5'$.

With the guidance vectors, the *guidance confidence* C_G between a point p in I_1 and the point q in I_2 is computed as the dot product between two guidance vectors:

$$C_G(p, q) = V(p) \cdot V(q) \tag{5}$$

where $V(p)$ and $V(q)$ are the guidance vectors for p and q. C_G ranges between -1 (complete opposite) and 1 (complete match).

2.F.4. Select the better guidance feature pairs

To search for the best matching point in the second image for a point in the first image, e.g., the point P, only the points in the second image within a radius from the mirrored position P' are checked. This is to ensure the computation efficiency. For the same reason, only the already matched feature pairs within a radius from the point P in I_1 are used for guidance. The search and guidance range are shown as a single circle in dashed line in Fig. 2. The radius of the range is configurable in this study. Considering the multi-resolution stages 1 to 4 (the stage 4 is the full resolution), the default value for a stage was 15 mm $\times$ (5—stage number). In addition, the searching radius and the guidance radius do not have to be the same.

Within the guidance radius, there could be many matched feature pairs that might be used for guidance. To ensure the computation efficiency, we select the maximal number of guidance feature pairs as 10. To ensure the quality of guidance, only the feature pairs with matching confidence >0.95 are used for guidance. If there are less than 10 matched and qualified feature pairs available in the guidance range, all of them will be used. If there are more than 10 pairs, the 10 pairs of the best matching confidence will be used. If there is no

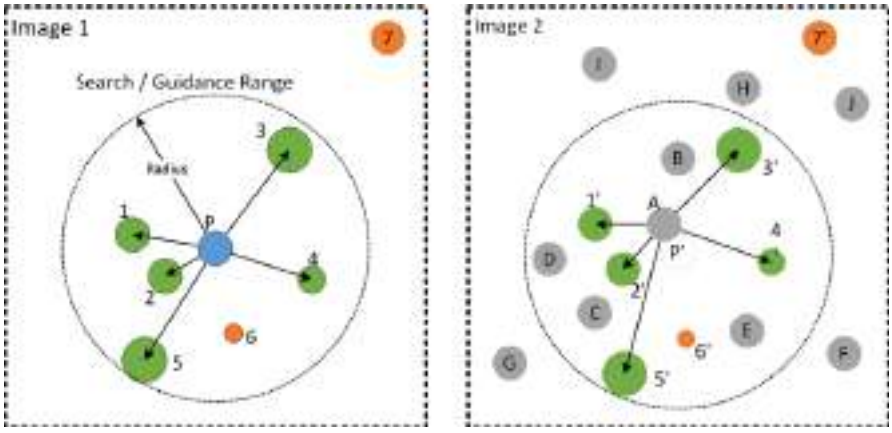

FIG. 2. Demonstration of guided feature matching in 2D. The point P in I_1 is successfully matched to the point A in I_2 under the guidance of the already matched feature pairs 1 to 5. P' is the position of P mirrored to I_2 based on the image registration between these two images. The sizes of the matched feature pairs 1 to 7 suggest the corresponding matching confidence for each feature pair. The points 1 to 5 and 1' to 5' are the previously matched feature pairs to be used for guidance. The points 6 and 6' are the previously matched pairs not to be used for guidance due to their distances or their insufficient matching confidences. The points A to J in Image 2 are the available features to be matched to the point P in Image 1. [Color figure can be viewed at wileyonlinelibrary.com]

matched pair available in the guidance range for a point to find matching, feature matching will be performed in the regular way, i.e., to find the best match based on feature descriptors, without guidance.

2.F.5. Integrate feature descriptor matching with guidance

For the point P in I_1, to find the best matching point in I_2 within the searching range of the mirrored position P' under the guidance of feature pairs 1 to 5, the overall matching confidence is computed for all the candidates. The combined confidence for a point p in I_1 and the point q in I_2 is defined as:

$$C(p,q) = (1-\alpha)C_D(p,q) + \alpha C_G(p,q) \quad (6)$$

where $\alpha \in [0,1]$ is an arbitrary ratio to weight contributions linearly between the descriptor matching confidence and the guidance confidence. In this study, $\alpha = 0.4$ is used, selected empirically according to the preliminary results.

After the confidence values are computed for all the matching candidates, i.e., the points A to E in I_2, the best matching is identified as:

- The candidate with the best combined matching confidence
- The descriptor matching confidence is greater than the descriptor matching threshold t_1
- The guidance matching confidence is greater than the guidance matching threshold t_2
- If there are more than one candidates, the ratio of the combined confidences of the best and the second best candidate is greater than a threshold t_3

2.F.6. Enforce inverse consistency

Feature matching between two images could be performed bi-directionally. Previous studies showed that inverse-consistent matching is generally more accurate than unidirectional matching.[61] We have therefore adopted a simple inverse-consistent matching strategy. For each feature in image 1, e.g., P in I_1, its best matching will be identified in image 2. If a match is indeed detected in image 2, e.g., A in I_2, its best matching in image 1 will be sequentially detected with the matching procedure. The feature pair P and A is confirmed if each of two feature points is the corresponding best matching. Otherwise, the matching is considered as ambiguous and therefore rejected.

2.G. The pseudo code

The pseudo code for the entire feature detection and matching procedure is shown in Fig. 3. The feature pairs detected at the highest resolution stage are the final results. The feature pairs detected at lower resolution stages are utilized only for guidance but are not included in the final results because the feature points do not have enough locational accuracy. It shall be noted that two general feature matching improvement methods—limiting the searching distance and rejecting the ambiguous matchings, are already incorporated in MRICGM. However, feature matching with guidance in MRICGM has eliminated the need of using RANSAC to reject outliers.

2.H. Implementation details

Most algorithms applied in this study were implemented in MATLAB 2016a (Mathworks, Natick, MA, USA) except that SIFT feature descriptor generation and 3D bilateral filter were implemented in C++. Parameters used in these algorithms are listed in Table II. A stand-alone tool with graphic user interface was developed using MATLAB to allow visualization and manual verification of the detected feature pairs on top of the images.

2.I. Evaluation using digital phantom datasets with known ground truth

2.I.1. Digital phantom generation

To evaluate the performance of the proposed multi-resolution inverse-consistent guided matching (MRICGM) method quantitatively, digital phantom datasets with ground truth landmarks were generated in the following four steps for each patient dataset listed in Table I:

(1) The deformation vector field (DVF) was computed between the image pair of the dataset using the Horn-Schunck deformable image registration (DIR) algorithm[62] in the MATLAB deformable image registration toolkit DIRART.[4] The first image in the pair is denoted as the moving image in the DIR computation. The second image is the target image. Applied parameters were 4 multi-resolution stages, in-iteration smoothing setting = 3, between-pass smoothing setting = 3, the number of passes = 9 for each stage, and the number of iterations = 20 for every pass and every stage.
(2) The first image of the digital phantom, denoted as I_{DM}, was computed using DVF_50%, i.e., DVF at 50% magnitude.
(3) The inversion of DVF_50% was computed using DIRART according to the DVF inversion method by Ashburner[63] and was denoted as IDVF_50%.
(4) The second image of the digital phantom, denoted as I_M, was computed using the IDVF_50%. In this way, I_M and I_{DM} were both computed by deforming the moving image by 50% DVF in the reverse and forward directions, respectively. This is to ensure that the magnitude of deformation between I_M and I_{DM} is equal to 100% and the smoothness due to interpolation in I_M and I_{DM} is similar. It is important to obtain similar smoothness in these two images to ensure the image feature processing results are not biased to either image.

- Pre-processing
- Build the multi-resolution Gaussian pyramid of 4 stages
- For each stage
  - Detect the SIFT features using the 3D SIFT algorithm
  - Detect the corner features using the 3D Harris-Laplacian algorithm
  - Compute the SIFT descriptors for the combined SIFT and corner features
- At stage 1 (the lowest resolution stage)
  - Use non-guidance feature matching procedure to detect the feature pairs, using tight threshold values t_1, t_2 and t_3
  - Apply MRICGM and use the feature detected in the last step to detect additional feature pairs, using less tight threshold values t_1, t_2 and t_3
- For each other stage in the order of stages 2, 3 and 4
  - Apply MRICGM and use the feature detected in the lower resolution stage to detect new feature pairs, using tight threshold values t_1, t_2 and t_3
  - Apply MRICGM in multiple iteration, and use the feature detected in the last iteration to detect additional feature pairs, using less tight threshold values t_1, t_2 and t_3

FIG. 3. Pseudo code for the proposed detection and feature matching procedure.

Each digital phantom dataset contains the two images I_M, I_{DM} and the known voxel mapping between I_M and I_{DM} (DVF computed by DIR). The DVF that was used to compute I_{DM}, was computed using DIR, and was more realistic than a DVF synthesized an analytical equation. Using a computed DVF to form digital phantom datasets has been previously reported by multiple studies.[25] The DIR results for the patient case #3 are shown in Fig. 4. From the difference images before and after DIR, one can see that DIR performed well. The deformation was mainly inside the ventral cavity. There was minimal deformation on and outside the rib cage, and on the vertebrae.

2.I.2. Evaluation using digital phantoms

I_M and I_{DM} of each digital phantom dataset, both computed from the respective moving image, were applied in the feature detection and matching procedures. For any feature point detected in either image, the ground truth corresponding coordinates in the other image could be precisely computed using the known voxel mapping between I_M and I_{DM}.

Feature matching accuracies of the three methods were measured using the generated seven digital phantom datasets. The three methods are the regular method (finding the best matching feature by testing the dot product of two feature descriptors), the regular method plus inverse consistency (to run the regular method in both matching directions and to reject the inconsistent results), or simply denoted as the IC method, and the proposed multi-resolution inverse-consistent guided matching (MRICGM) method. The parameters and values used by the regular method are listed in Table II. The IC method uses the same parameters as the standard method.

2.J. Evaluation using patient datasets

The proposed method was also evaluated using the datasets from seven patients listed in Table I. We implemented a software tool in MATLAB that includes a graphic user interface (GUI). It enables users to visualize detected landmark pairs in three orthogonal views side by side (see Fig. 6). This tool allows users to verify the landmark pairing and flag the false matches. The detected landmark pairs for each patient dataset were manually and independently verified by three observers using the software tool. As thousands of landmark pairs were detected in each case, the following procedure was adopted to speed up the manual evaluation process: (a) Detected landmark pairs were sorted according to the overall matching confidence, from low to high. (b) Landmark pairs were manually verified one by one starting from the landmark pairs at the lowest confidence. A landmark pair was flagged as a false match if the landmark position in the second image was off by more than one voxel in any orthogonal direction. (c) If there is no error for the last 20 pairs, the remaining landmark pairs were to be checked at an interval of 5, then 10, then 20. (d) If a new error is found when one of every 5, 10, or 20 landmark pairs is checked, the interval will be reduced. This process allows the low confidence landmark pairs to be checked carefully one by one and the high confidence pairs to be checked quickly. At the end, a feature pair was considered a false match if it was flagged by any of the three observers.

3. RESULTS

3.A. Digital phantom results

The performance evaluation results using seven digital phantom datasets are provided in Table III. There was a total of 82 987 pairs of features detected or 11 855 on average per dataset. Numbers of feature pairs detected by the regular and IC methods were within 0.5% respectively by manual configuring the Descriptor matching threshold t_1 (see Table II). TREs were computed per feature pair against the ground truth DVFs. Analyzed using student t-tests, TREs on feature pairs detected by MRICGM were significantly smaller than TREs obtained by the other two matching methods ($P \ll 0.001$).

Figure 5(a) shows the comparison of histograms of all TRE values measured by the regular, IC, and MRICGM methods on the seven digital phantom datasets. It is evident that MRICGM outperformed the other two methods, for example, only 1.35% of feature pairs obtained by MRICGM have TRE > 3 mm, and 0.52% have TREs > 4 mm. In

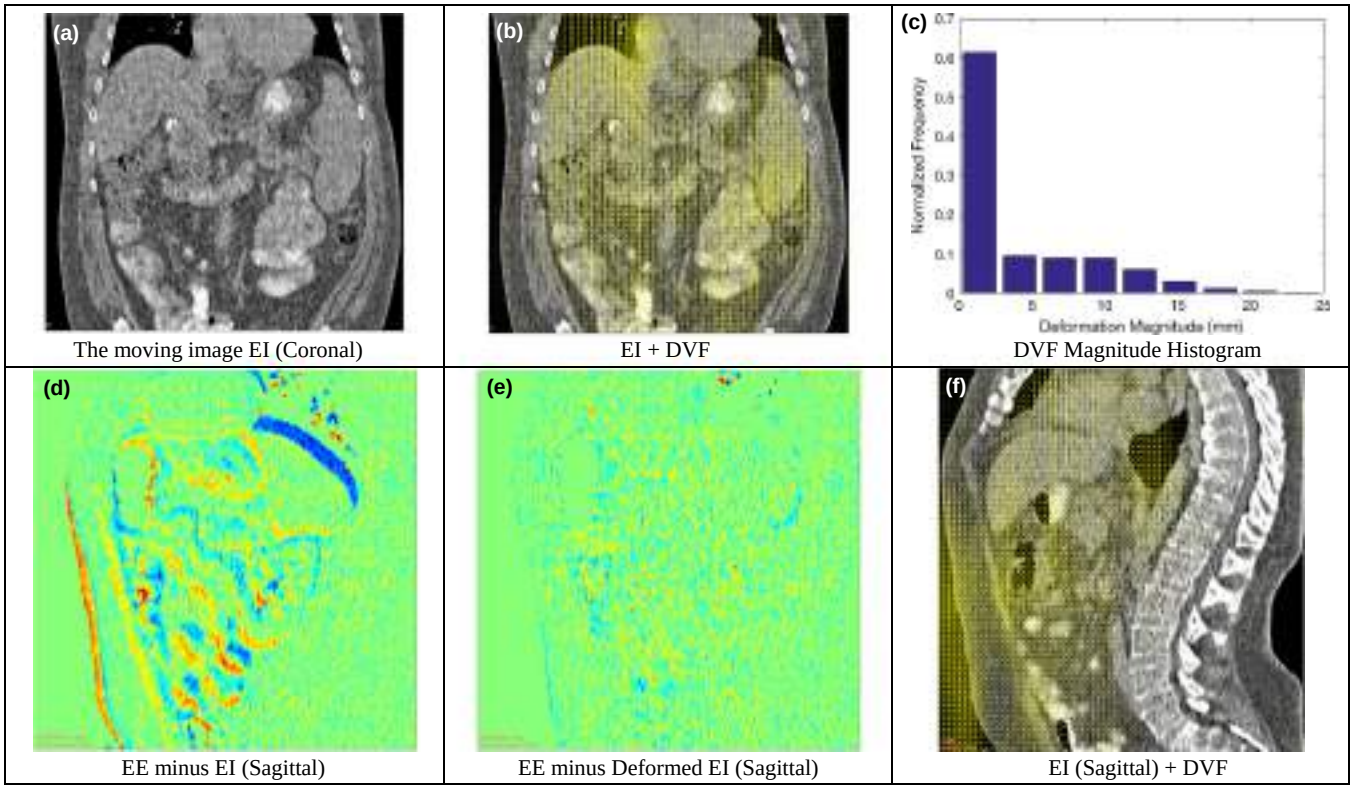

FIG. 4. DVF was computed between the EI and EE phases, and then applied to deform the EI phase to form the ground truth digital phantom images. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE III. Summary of feature matching results on seven digital phantoms.

Case #	Image type	# of Feature pairs detected[a]	Ground truth DVF[b] (mm)	TRE (MRICGM) (mm)	TRE (regular) (mm)	TRE (IC[c]) (mm)
1	4DCT	21472	1.67 ± 2.08	0.65 ± 0.60	1.10 ± 2.95	0.83 ± 1.73
2	4DCT	15180	4.13 ± 4.71	0.78 ± 0.72	1.61 ± 3.89	1.16 ± 2.70
3	4DCT	28203	4.2 ± 4.55	0.74 ± 0.68	1.35 ± 3.49	0.96 ± 2.16
4	H/N CT	2038	2.98 ± 1.94	1.29 ± 1.18	2.28 ± 3.30	1.75 ± 2.56
5	H/N CT	910	1.95 ± 1.52	1.19 ± 1.69	1.66 ± 2.58	1.30 ± 2.06
6	Pelvis MRI	4648	5.96 ± 2.50	1.05 ± 1.02	2.33 ± 4.17	1.92 ± 3.23
7	Pelvis MRI	10545	1.25 ± 1.21	0.85 ± 0.79	1.85 ± 3.59	1.33 ± 2.32
Average		11855	3.20 ± 3.30	0.77 ± 0.72	1.48 ± 3.46	1.09 ± 2.23

TRE, Target Registration Error.
[a]The listed numbers of feature pairs were detected by MRICGM. Numbers of feature pairs detected by the regular and IC methods were within 0.5%, respectively.
[b]The ground truth DVF (positional displacement) between the matched feature pairs.
[c]IC = regular + inverse consistency.

comparison, the respective values are 8.33% and 6.89% for the regular method, and 4.17% and 3.49% for the IC method.

Figure 5(b) shows the comparison of TRE values for all 38821 SIFT feature pairs and 53 992 Harris corner pairs measured by MRICGM on seven digital phantom datasets. It is interesting to see that there were more corner features detected than SIFT features. The average positional accuracy with SIFT features (TREs = 0.67 ± 0.74 mm) are greater than that with the corner features (TREs = 0.85 ± 0.68 mm), probably due to the iterative position refinement step (described in Section 2.D.1) which we have implemented in our 3D SIFT detection algorithm. This step allows sub-voxel positional accuracy with SIFT features. In comparison, the positions of the corner feature detected by the Harris–Laplacian method are only on the whole voxels. An additional note is that fewer percentages of Harris corners have larger TREs, e.g., TREs > 3 mm. This suggests Harris corners are more robust, even though the positions being less accurate, than SIFT features. These observations justify our motivation to use the combination of SIFT feature detection and Harris–Laplacian corner detection methods to detect more features.

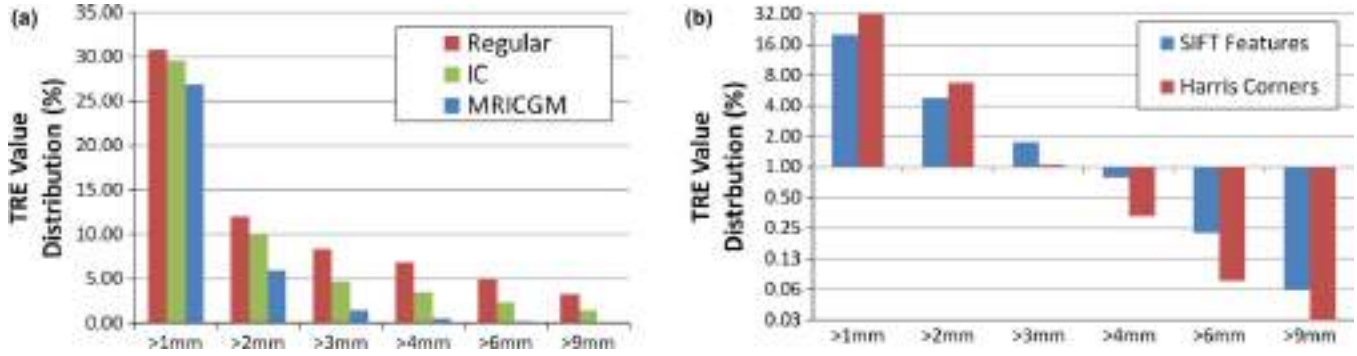

FIG. 5. (a) Comparison of distribution of all TRE values measured on seven digital phantom datasets. By having lower percentages of TREs, MRICGM outperformed both the regular and the IC methods. (b) A comparison of TREs distribution between 38 821 SIFT and 53 992 Harris corner feature pairs, detected by the MRICGM method on seven digital phantom datasets, shows SIFT features have greater positional accurate and Harris corners have fewer greater TREs. [Color figure can be viewed at wileyonlinelibrary.com]

3.B. Patient image results

Figure 6 shows the feature detection and matching results on the dataset #3—the end-of-inhale and end-of-exhale phases of patient's abdominal 4DCT image. There were 12050 feature pairs detected at the original resolution stage on this dataset. The maximal respiratory motion magnitude, measured by the distance between the matched feature pairs was 14 mm. One can see that feature pairs of a good quantity are located in the organs and other soft tissues inside the ventral cavity. These feature pairs would be very useful as landmarks to estimate the organ respiratory motion and to verify the deformable image registration on the organs and soft tissues.

Figure 7 demonstrates the detected feature pairs in the dataset #3. One can see that feature pairs were roughly evenly detected cross the entire CT volume. Feature pairs in relatively higher density were detected along the spinal column because the stable bones with complicated shapes have greater amount of image features.

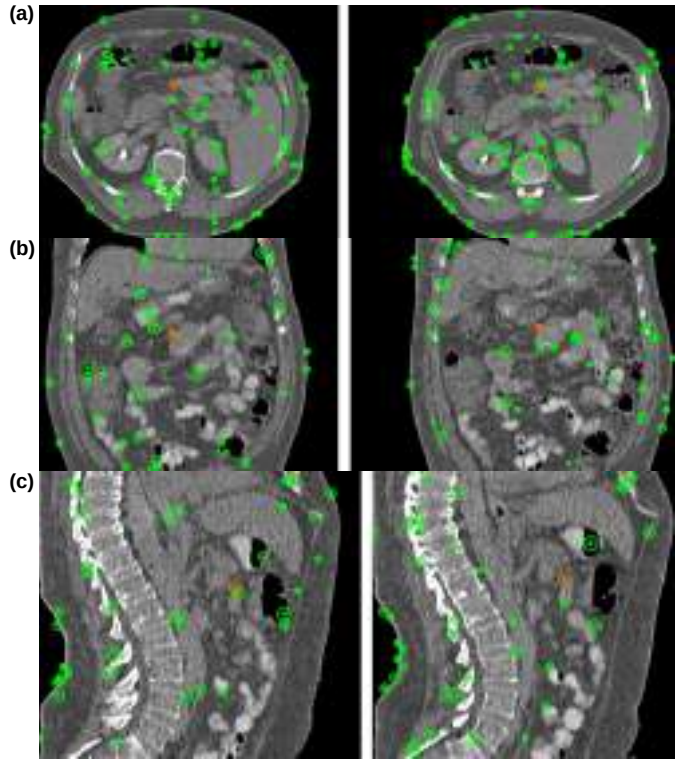

FIG. 6. The detected feature pairs on the dataset #3—abdominal 4DCT. There are 12 050 feature pairs detected, relatively distributed uniformly. (a), (b), and (c) are the axial, coronal and sagittal views. In each view, the end-of-inhale phase is on the left and the end-of-exhale phase is on the right. The displayed slices from two images are centered on the focused feature point. The other feature points are on the same image slices of each image but are not necessary the same points between two images because the corresponding point of a pair on the other images could be off the slice. [Color figure can be viewed at wileyonlinelibrary.com]

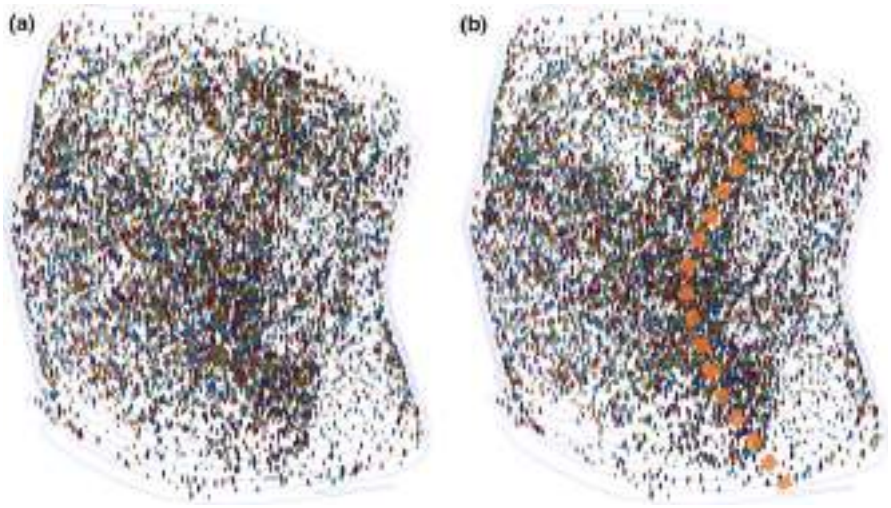

FIG. 7. Examples of the detected 12 050 feature pairs between the two CT images of the abdominal 4D-CT dataset #3, rendered in 3D (azimuth angle = 45 degree to the patient's left, and elevation angle = 30 degree). Features are relatively uniformly distributed. (a) The end-of-inhale phase. (b) The end-of-exhale phase. The skin surface is shown in contours. The sizes of the ellipsoids are corresponding to the scales of the feature points, computed by the feature point detection algorithms. The spinal column in (b) is approximately suggested by the added-on dashed-curve. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE IV. Statistics of the detected feature pairs.

Matching method	Soft tissues (%)	Soft tissue-air (%)	Soft tissue-bone (%)
Regular	50.79	11.30	36.25
Regular + inverse consistency (IC)	55.56	8.04	34.27
MRICGM	56.43	7.45	34.66

Table IV lists the statistics of the types of the detected feature pairs in the dataset #3. One shall note that MRICGM was able to detect relatively more feature pairs on soft tissues. This could be useful for clinical applications that DIR accuracy on soft tissues is important.

The overall feature matching results on the seven patient datasets are provided in Table V. As described in Section 2.J, the detected landmark pairs for each patient dataset were manually evaluated by three observers. One can see that thousands of landmark pairs were successfully detected for each dataset and the feature matching accuracy at the full resolution stage are overall greater than 99%. Matching accuracies in the lower resolution stages 1 to 3 were essentially 100% for all patients. Fewer landmark pairs detected on the two head-neck scan/rescan CT cases because the physical volumes of patient bodies, i.e., necks and upper shoulders, in these two datasets were much smaller than other cases.

3.C. Computational performance

Table VI lists the computation time on each step measured with the dataset #3 on a Windows-7 64bit PC with an Intel Cord i7 960 CPU at 3.20 GHz and 20 GB RAM. Image dimension is 369 × 512 × 123. As one can see, the feature detection and feature descriptor generation steps are the slowest. Fortunately, these two steps are suitable for GPU acceleration. Feature matching with MRICGM was also slow, mostly because that it ran seven iterations at the highest resolution stage. In each iteration, MRICGM was roughly as fast as the IC (regular + inverse consistency) method, which was

TABLE V. Summary of results on seven patient datasets.

	Number of detected landmark pairs and matching accuracy (%)			
Dataset	Stage 1 (1/8 resolution)	Stage 2 (1/4 resolution)	Stage 3 (1/2 resolution)	Stage 4 (Full resolution)
1. 4DCT	102/99.5%	435/99.6%	1636/>99.9%	6302/99.7%
2. 4DCT	72/100%	414/99.9%	2114/>99.9%	6980/99.7%
3. 4DCT	87/100%	570/100%	2113/>99.9%	12050/99.9%
4. H/N CT	32/100%	194/100%	951/100%	3126/99.9%
5. H/N CT	5/100%	17/100%	70/99.5%	1115/98.3%
6. Pelvis MRI	28/100%	181/98.3%	818/99.3%	8934/98.8%
7. Pelvis MRI	9/100%	59/100%	465/99.9%	6020/99.4%
Average accuracy	99.9%	99.7%	99.8%	99.4%

TABLE VI. Computation time in each step with all values. The listed computation time for the regular method and the IC method is for comparison purpose only. All values are in seconds except the percentage values.

	Feature detection[a,c]					Other feature matching methods	
Preprocessing	SIFT[a,c]	Harris–Laplacian[a,c]	Descriptor generation[a,c]	Feature matching with MRICGM[a,b]	Total	Regular[d]	IC[d]
16.0	77	208	680	125	1106	10.6	22.9
1.53%	6.96%	18.8%	61.5%	11.3%			

IC = Inverse Consistency.
[a]With multi-resolution.
[b]With multiple iterations at each image resolution.
[c]For both images.
[d]Only at the highest resolution stage.

as ~50% as fast as the regular method because that computation needs to be run in both directions.

4. DISCUSSION

Accurate detection of large quantity of landmarks in patient image datasets in this study is accomplished by multiple designed choices. While it is important to use both SIFT and Harris–Laplacian algorithms together to improve the total number of detected features, the multi-resolution inverse-consistent guided matching method, or MRICGM, is the most important contribution in this study. The use of multiple iterations at each resolution stage, starting from the tight thresholds and stopping at looser thresholds, was an important component in MRICGM.

While the current results are relatively satisfactory, many steps of the proposed methods could be further improved, or further optimized for certain types of applications or image modalities. Image preprocessing would probably require different settings for cone-beam CT image. Multi-resolution can be done differently, e.g., to half-sample the image in the axial only at certain stage, if the image voxels are anisotropic, e.g., in $1 \times 1 \times 3$ mm^3. Feature matching thresholds might need to be loosened to allow more feature pair detection for certain image modalities, e.g., cone-beam CT, and might affect the feature matching accuracy. The needed numbers of detected feature pairs in a patient image dataset would likely be application-dependent. As shown in Figs. 6 and 7, one can expect that 12 050 feature pairs are probably adequate to generally verify the DVF in this patient abdominal CT. However, if the goal is to precisely verify the DVF locally in a certain region to measure the tumor response to treatment, e.g., around the right kidney, more feature pairs in the region-of-interest would be desirable.

The proposed image processing methods are useful, as a set of tools, to generate landmark pairs in patient CT or MR images. Although the measured feature matching accuracies are very good, i.e., >99% in most cases, the generated landmarks should be manually verified before they can be confidently used as benchmarks for DIR verification. Therefore, the proposed methods are probably not yet ready as an automatic DIR verification tool, at least not for online applications, e.g., online plan adaptation.

Computational speed is currently a concern. GPU acceleration is clearly a potential solution. In fact, the slowest steps, i.e., the feature detection and feature description generation steps, are very suitable for GPU acceleration. To speed up the feature detection, it might also be possible to combine the SIFT and Harris–Laplacian algorithms because they share a few common computations, e.g., repetitive Gaussian smoothing. The computational efficiency of MRICGM can also be further improved. MRICGM is currently applied in multiple iterations at the highest resolution stage. Many temporary results could be shared between the iterations and therefore many repetitive computations could be avoided.

Our near-term future works would be (a) to implement the proposed methods in GPU, (b) to apply the proposed methods on additional image modalities, e.g., cone-beam CT, and (c) to apply the methods on additional anatomical sites, e.g., thorax 4DCT, inter-fractional helical CTs of HDR brachytherapy patients, (d) to further optimize the methods for additional applications, e.g., between CT images of different patients, (e) to optimize parameters in the feature detection and feature matching methods per tissue types, and (f) to organize the current and additional results and make them available to the community.

5. CONCLUSION

A procedure was developed in this study to automatically and accurately detect large number of landmark pairs in CT and MRI image pairs. It allowed semi-automatic ways to quickly generate the ground-truth datasets for evaluation of DIR algorithms.

ACKNOWLEDGMENT

The project described was partially supported by the AHRQ (Agency for Healthcare Research and Quality) grant number 1 R01 HS022888-01 and its contents are solely the responsibility of the authors and do not necessarily represent

the official views of the Agency for Healthcare Research and Quality.

[a]Author to whom correspondence should be addressed. Electronic mail: yangdeshan@wustl.edu.

REFERENCES

1. Kadoya N. Use of deformable image registration for radiotherapy applications. *J Radiol Radiat Ther.* 2014;2:1–7.
2. Crum WR, Hartkens T, Hill D. Non-rigid image registration: theory and practice. *Br J Radiol*. 2014;77(Suppl 2):S140–S153.
3. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imaging*. 2013;32:1153–1190.
4. Yang D, Brame S, El Naqa I, et al. DIRART – a software suite for deformable image registration and adaptive radiotherapy research. *Med Phys*. 2011;38:67–77.
5. Yan D, Vicini F, Wong J, Martinez A. Adaptive radiation therapy. *Phys Med Biol*. 1997;42:123–132.
6. Castadot P, Lee JA, Geets X, Grégoire V. Adaptive radiotherapy of head and neck cancer. *Semin Radiat Oncol*. 2010;20:84–93.
7. Awan M, Kalpathy-Cramer J, Gunn GB, et al. Prospective assessment of an atlas-based intervention combined with real-time software feedback in contouring lymph node levels and organs-at-risk in the head and neck: quantitative assessment of conformance to expert delineation. *Pract Radiat Oncol*. 2013;3:186–193.
8. Reed VK, Woodward WA, Zhang L, et al. Automatic segmentation of whole breast using atlas approach and deformable image registration. *Int J Radiat Oncol Biol Phys*. 2009;73:1493–1500.
9. Faggiano E, Fiorino C, Scalco E, et al. An automatic contour propagation method to follow parotid gland deformation during head-and-neck cancer tomotherapy. *Phys Med Biol*. 2011;56:775.
10. Sharp G, Fritscher KD, Pekar V, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys*. 2014;41:050902.
11. Thornqvist S, Petersen JBB, Høyer M, Bentzen LN, Muren LP. Propagation of target and organ at risk contours in radiotherapy of prostate cancer using deformable image registration. *Acta Oncol*. 2010;49:1023–1032.
12. Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. *Radiother Oncol*. 2011;98:373–377.
13. Li X, Wang X, Li Y, Zhang X. A 4D IMRT planning method using deformable image registration to improve normal tissue sparing with contemporary delivery techniques. *Radiat Oncol*. 2011;6:83–83.
14. Wang J, Gu X. High-quality four-dimensional cone-beam CT by deforming prior images. *Phys Med Biol*. 2013;58:231.
15. Wu G, Wang Q, Lian J, Shen D. Reconstruction of 4D-CT from a single free-breathing 3D-CT by spatial-temporal image registration. *Inf Process Med Imaging*. 2011;22:686–698.
16. Yang D, Lu W, Low DA, Deasy JO, Hope AJ, Naqa IE. 4D-CT motion estimation using deformable image registration and 5D respiratory motion modeling. *Med Phys*. 2008;35:4577–4590.
17. Cherpak A, Serban M, Seuntjens J, Cygler JE. 4D dose-position verification in radiation therapy using the RADPOS system in a deformable lung phantom. *Med Phys*. 2011;38:179–187.
18. Niu CJ, Foltz WD, Velec M, Moseley JL, Al-Mayah A, Brock KK. A novel technique to enable experimental validation of deformable dose accumulation. *Med Phys*. 2012;39:765–776.
19. Yeo UJ, Taylor ML, Supple JR, et al. Is it sensible to ''deform'' dose? 3D experimental validation of dose-warping. *Med Phys*. 2012;39:5065–5072.
20. Hasan Y, Kim L, Wloch J, et al. Comparison of planned versus actual dose delivered for external beam accelerated partial breast irradiation using cone-beam CT and deformable registration. *Int J Radiat Oncol Biol Phys*. 2011;80:1473–1476.
21. Castadot P, Geets X, Lee JA, Christian N, Grégoire V. Assessment by a deformable registration method of the volumetric and positional changes of target volumes and organs at risk in pharyngo-laryngeal tumors treated with concomitant chemo-radiation. *Radiother Oncol*. 2010;95:209–217.
22. Mencarelli A, van Kranen SR, Hamming-Vrieze O, et al. Deformable image registration for adaptive radiation therapy of head and neck cancer: accuracy and precision in the presence of tumor changes. *Int J Radiat Oncol Biol Phys*. 2014;90:680–687.
23. Li S, Carri G-H, Lu M, et al. Voxel-based statistical analysis of uncertainties associated with deformable image registration. *Phys Med Biol*. 2013;58:6481.
24. Fallone BG, Rivest DR, Riauka TA, Murtha AD. Assessment of a commercially available automatic deformable registration system. *J Appl Clin Med Phys*. 2010;11:3175.
25. Yang D, Li H, Low DA, Deasy JO, El Naqa I. A fast inverse consistent deformable image registration method based on symmetric optical flow computation. *Phys Med Biol*. 2008;53:6143–6165.
26. Leow A, Huang S-C, Geng A, et al. Inverse consistent mapping in 3D deformable image registration: its construction and statistical properties. In: Christensen GE, Sonka M, eds. *Information Processing in Medical Imaging*, vol. 3565. Berlin/Heidelberg: Springer; 2005:493–503.
27. Varadhan R, Karangelis G, Krishnan K, Hui S. A framework for deformable image registration validation in radiotherapy clinical applications. *J Appl Clin Med Phys*. 2013;14:192–213.
28. Zhong H, Kim J, Chetty IJ. Analysis of deformable image registration accuracy using computational modeling. *Med Phys*. 2010;37:970–979.
29. Mencarelli A, Beek SV, Kranen SV, Rasch C, Herk MV, Sonke J-J. Validation of deformable registration in head and neck cancer using analysis of variance. *Med Phys*. 2012;39:6879–6884.
30. Kim J, Kumar S, Liu C, et al. A novel approach for establishing benchmark CBCT/CT deformable image registrations in prostate cancer radiation therapy. *Int J Radiat Oncol Biol Phys*. 2013;87:S713.
31. Shunshan L, Carri G-H, Mei L, et al. Voxel-based statistical analysis of uncertainties associated with deformable image registration. *Phys Med Biol*. 2013;58:6481.
32. Bender ET, Tome WA. The utilization of consistency metrics for error analysis in deformable image registration. *Phys Med Biol*. 2009;54:5561.
33. Schreibmann E, Pantalone P, Waller A, Fox T. A measure to evaluate deformable registration fields in clinical settings. *J Appl Clin Med Phys*. 2012;13:126–139.
34. Yeo UJ, Supple JR, Taylor ML, Smith R, Kron T, Franich RD. Performance of 12 DIR algorithms in low-contrast regions for mass and density conserving deformation. *Med Phys*. 2013;40:101701.
35. Pukala J, Meeks SL, Staton RJ, Bova FJ, Mañon RR, Langen KM. A virtual phantom library for the quantification of deformable image registration uncertainties in patients with cancers of the head and neck. *Med Phys*. 2013;40:111703.
36. Cheung Y, Sawant A. An externally and internally deformable, programmable lung motion phantom. *Med Phys*. 2015;42:2585–2593.
37. Nie K, Chuang C, Kirby N, Braunstein S, Pouliot J. Site-specific deformable imaging registration algorithm selection using patient-based simulated deformations. *Med Phys*. 2013;40:041911.
38. Stanley N, Glide-Hurst C, Kim J, et al. Using patient-specific phantoms to evaluate deformable image registration algorithms for adaptive radiation therapy. *J Appl Clin Med Phys*. 2013;14:4363–4363.
39. Castillo R, Castillo E, Fuentes D, et al. A reference dataset for deformable image registration spatial accuracy evaluation using the COPDgene study archive. *Phys Med Biol*. 2013;58:2861.
40. Latifi K, Zhang G, Stawicki M, van Elmpt W, Dekker A, Forster K. Validation of three deformable image registration algorithms for the thorax. *J Appl Clin Med Phys*. 2013;14:19–30.
41. Ramadaan I, Peick K, Hamilton D, et al. Validation of Varian's SmartAdapt(R) deformable image registration algorithm for clinical application. *Radiation Oncology.* 2015;10:73.
42. Castillo R, Castillo E, Guerra R, et al. A framework for evaluation of deformable image registration spatial accuracy using large landmark point sets. *Phys Med Biol*. 2009;54:1849.
43. Nithiananthan S, Brock KK, Daly MJ, Chan H, Irish JC, Siewerdsen JH. Demons deformable registration for CBCT-guided procedures in the head and neck: convergence and accuracy. *Med Phys*. 2009;36:4755–4764.
44. Hoffmann C, Krause S, Stoiber EM, et al. Accuracy quantification of a deformable image registration tool applied in a clinical setting. *J Appl Clin Med Phys*. 2014;15:237–245.

45. Vickress J, Battista J, Barnett R, Morgan J, Yartsev S. Automatic landmark generation for deformable image registration evaluation for 4D CT images of lung. *Phys Med Biol*. 2016;61:7236.
46. Paganelli C, Peroni M, Riboldi M, et al. Scale invariant feature transform in adaptive radiation therapy: a tool for deformable image registration assessment and re-planning indication. *Phys Med Biol.* 2013;58:287.
47. Mazur TR, Fischer-Valuck BW, Wang Y, Yang D, Mutic S, Li HH. SIFT-based dense pixel tracking on 0.35 T cine-MR images acquired during image-guided radiation therapy with application to gating optimization. *Med Phys*. 2016;43:279–293.
48. Murphy K, van Ginneken B, Klein S, et al. Semi-automatic construction of reference standards for evaluation of image registration. *Med Image Anal*. 2011;15:71–84.
49. Burt P, Adelson E. The Laplacian pyramid as a compact image code. *IEEE Trans Commun*. 1983;31:532–540.
50. Lindeberg T. Scale selection properties of generalized scale-space interest point detectors. *J Math Imaging Vis*. 2013;46:177–210.
51. Laurence EC, Roy BT, Joshua P, Robert C, Lee C. Automatic online adaptive radiation therapy techniques for targets with significant shape change: a feasibility study. *Phys Med Biol*. 2006;51:2493.
52. Tomasi C, Manduchi R, Presented at the IEEE Sixth International Conference on Computer Vision; 1998.
53. Lowe DG, Presented at the Proceedings of the Seventh IEEE International Conference on Computer Vision.; 1999.
54. Cheung W, Hamarneh G, Presented at the 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2007); 2007.
55. Klein A, Andersson J, Ardekani BA, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*. 2009;46:786–802.
56. Rey Otero I, Delbracio M, Anatomy of the SIFT Method. *Image Processing on Line*. 2014; 4: 370–396.
57. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vision*. 2004;60:91–110.
58. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-Up Robust Features (SURF). *Comput Vis Image Underst*. 2008;110:346–359.
59. Leutenegger S, Chli M, Siegwart RY, Presented at the 2011 International Conference on Computer Vision; 2011.
60. Fischler MA, Bolles RC. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun ACM*. 1981;24:381–395.
61. Morago B, Bui G, Duan Y. An ensemble approach to image matching using contextual features. *IEEE Trans Image Process*. 2015;24:4474–4487.
62. Horn BKP, Schunck BG. Determining optical flow. *Artif Intell*. 1981;17: 185–203.
63. Ashburner J, Andersson JL, Friston KJ. Image registration using a symmetric prior–in three dimensions. *Hum Brain Mapp*. 2000;9: 212–225.