A Comparison Framework for Breathing Motion Estimation Methods From 4-D Imaging

David Sarrut*, Bertrand Delhay, Pierre-Frédéric Villard, Vlad Boldea, Michael Beuve, and Patrick Clarysse

***Abstract*—Motion estimation is an important issue in radiation therapy of moving organs. In particular, motion estimates from 4-D imaging can be used to compute the distribution of an absorbed dose during the therapeutic irradiation. We propose a strategy and criteria incorporating spatiotemporal information to evaluate the accuracy of model-based methods capturing breathing motion from 4-D CT images. This evaluation relies on the identification and tracking of landmarks on the 4-D CT images by medical experts. Three different experts selected more than 500 landmarks within 4-D CT images of lungs for three patients. Landmark tracking was performed at four instants of the expiration phase. Two metrics are proposed to evaluate the tracking performance of motion-estimation models. The first metric cumulates over the four instants the errors on landmark location. The second metric integrates the error over a time interval according to an *a priori* breathing model for the landmark spatiotemporal trajectory. This latter metric better takes into account the dynamics of the motion. A second aim of this paper is to estimate the impact of considering several phases of the respiratory cycle as compared to using only the extreme phases (end-inspiration and end-expiration). The accuracy of three motion estimation models (two image registration-based methods and a biomechanical method) is compared through the proposed metrics and statistical tools. This paper points out the interest of taking into account more frames for reliably tracking the respiratory motion.**

***Index Terms*—Deformable registration, radiotherapy, thorax, validation.**

I. Introduction

ACCOUNTING for organ motion due to breathing in lung cancer radiation treatment is an important challenge [1]. Reducing uncertainties on target position should result in decreasing irradiation of healthy lung areas and should allow tumor dose escalation, potentially leading to a better outcome [2]. Several approaches are currently under investigation (breath-hold treatment, gating [3], etc.) but all require patient-specific spatiotemporal information about movements and deformations induced by breathing. Ideally, treatment planning should not rely on 3-D images only but also on a patient-specific breathing thorax model, encompassing all mechanical and functional information available. Some data can be obtained from 4-D CT imaging [4], but 4-D images alone are not sufficient and should be associated with new image analysis tools such as motion estimators and anatomical structure tracking methods [5]. They can also be used to build a "4-D model" composed of spatiotemporal trajectories of all volume elements in the thorax. Using such a model would make it possible to select the best way to manage organ motion for each patient and provide helpful information for planning real-time tracking and dose delivery.

For example, a motion margin can be defined in order to account for respiratory motion, leading to unnecessary irradiation of large volumes of normal tissues. Zhang *et al.* [6] have proposed to incorporate target motion into treatment optimization using the displacement vector fields at different breathing phases, based on patient 4-D CT images; beam targeting is optimized according to the motion. Instructing the patient to breathe following a visually displayed guiding cycle potentially allows us to spare larger volumes of normal tissue. Rietzel *et al.* [7] have delineated volumes of interest in each phase of a 4-D CT dataset and used them to determine the maximal displacement of gross tumor volume (GTV) centroids. Using B-spline deformable registrations, they have tried to quantify the impact of respiratory motion on generated dose distributions. The dose delivered to a given volume is directly related to the time of irradiation.[1] Therefore, motion of the tumors must be taken into account during the whole respiratory cycle. Brock *et al.* [8] have developed an approximation to modulate the weight of dose calculations from the exhale toward the inhale model as breathing progresses and using time weights obtained via fluoroscopy on a given population of patients. Keall *et al.* [5] have extended this concept to dynamic multi-leaf collimator[2] (DMLC)-based respiratory motion tracking. They have used deformable image registration to automatically transfer contours defined on the peak-inhale CT scan to other respiratory phase CT images. Dose distributions at each phase were then computed with phase-adapted

Manuscript received December 22, 2006; revised May 16, 2007. This work was supported in part by the French research program "ACI-Masse de données", in part by the AGIR Project, and in part by the "Région Rhônes-Alpes", France, through the EURODOC Program. *Asterisk indicates corresponding author.*

*D. Sarrut is with the Léon Bérard Cancer Center, 69373 Lyon, France, and with CREATIS-LRMN INSA, 69621 Villeurbanne cedex, France (e-mail: david.sarrut@creatis.insa-lyon.fr).

B. Delhay and P. Clarysse are with CREATIS-LRMN INSA, 69621 Villeurbanne cedex, France.

P.-F. Villard and M. Beuve are with the Université de Lyon, Lyon F-69003, France, and CNRS, Laboratoire d'InfoRmatique en Images et Systèmes d'information, Villeurbanne F–69622, France.

V. Boldea is with the Léon Bérard Cancer Center, 69373 Lyon, France, and Université de Lyon, Lyon, F-69003, France, and also with CNRS, Laboratoire d'InfoRmatique en Images et Systèmes d'information, Villeurbanne F–69622, France.

Digital Object Identifier 10.1109/TMI.2007.901006

[1]Dynamic aspect of *Intensity Modulated Radiation Therapy* (IMRT) may induce more complex behavior, but is not considered here.

[2]In radiation therapy, a multi-leaf collimator (MLC) is a device used for delimiting the radiation beam. It generally consists of two pairs of opposite jaws reshaping beams to a square or rectangular cross section. Dynamic MLC supposes that leafs can move during irradiation.

MLC-defined beam, then mapped back to a reference CT image using estimated deformation fields.

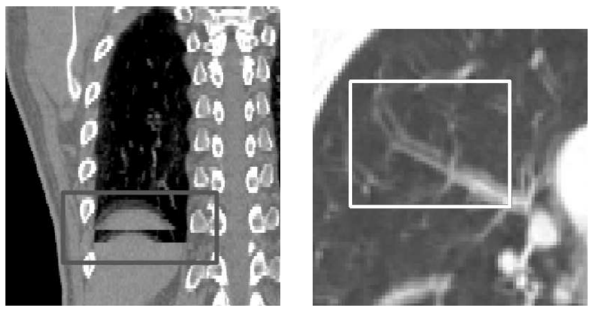
Breathing motion tracking has been a fundamental element in these recent studies and must therefore be validated. Time-related issues must also be taken into account. A major challenge in deformable motion estimation is the validation of the resulting deformation fields. Today, contrary to the rigid motion case [9]–[11], there are few evaluation standards for deformable motion estimation. A tentative evaluation framework has been proposed by Hellier *et al.* [12] which focuses on the deformable registration of the brains of different individuals. In the present paper, our goal is to quantitatively compare motion estimators by taking into account the temporal aspects of the observed motions. Our clinical motivation was related to the use of deformable motion estimators with 4-D scans to simulate radiation dose delivery inside moving and deforming organs for given irradiation configurations. We propose a framework and criteria incorporating temporal information to evaluate the accuracy of motion estimation methods for the purpose of compensating for breathing motion in 4-D CT images. The proposed framework will be illustrated with the evaluation of three different motion estimation methods in terms of accuracy.

The paper is organized as follows. Section II briefly presents the experimental 4-D CT data used in the study. Our approach for the evaluation of motion tracking methods in 4-D CT sequences is based on landmark location estimation. Section III-A and III-B explain how the landmarks have been selected and tracked by medical experts. Then, two error criteria to evaluate the accuracy of landmark location estimated by motion tracking methods are introduced. The first one is the generalization of the conventional TRE metric to the tracking in successive images (Section III-C). The second one takes into account the temporal nature of the motion and is presented in III-D. Section IV describes three motion tracking methods compared using the proposed strategy. Results are presented in Section V. Section VI compares the behavior of the motion tracking methods and discusses the respective properties of the two metrics.

II. Materials

This study considered as input data thoracic 4-D CT sequences from patients with nonsmall-cell lung cancer (NSCLC). The 4-D images were acquired according to a recent protocol, similar to the one described in [13], using a "cine" scanning protocol: multiple image acquisitions were performed along the cranio-caudal direction at a time interval greater than the average respiratory cycle. The acquisition was repeated until the prescribed volume was completely scanned. During the entire acquisition, an external respiratory signal, generated with the Real-Time Position Management (RPM) Respiratory Gating System (from Varian Medical Systems, Palo Alto, CA) was recorded. The signal was then used to sort data into respiratory phases by selecting, for each slice position and for each phase, the closest image. The resulting 4-D images were composed of ten 3-D images covering a respiratory cycle from the end of normal inspiration to the end of normal expiration.

In this paper, we focused on the *expiration part* of the respiratory cycle (six out of ten frames, including extreme phases).



(a) (b)

Fig. 1. Image artifacts. (a) Several slices were missing at given temporal phase. (b) Blurred structures inside lung due to unforeseen movement.

TABLE I
Lung Volumes (in centimeters^3 and in % of Difference Between I_E and I_I) for All Images

Images	$patient1$	$patient2$	$patient3$
I_I	5181 (100%)	3214 (100%)	3121 (100%)
I_1	5004 (50%)	2981 (49%)	2897 (47%)
I_2	4692 (44%)	2880 (27%)	2797 (24%)
I_E	4315 (0%)	2755 (0%)	2696 (0%)

The number of exploitable frames varies from one dataset to the other and the common maximum number in our series was four out of six. We thus decided to consider four images: two extreme images (denoted I_I for end-inspiration and I_E for end-expiration) and two intermediate images denoted I_1 and I_2, corresponding to intermediate lung volumes. For some reason (too rapid patient breath, inaccuracy of the external respiratory signal), selected data at a given phase and slice position may not be consistent. Hence, the corresponding 3-D images presented some misaligned slices, generally around the diaphragm (see Fig. 1). Some other (less frequent) artifacts were probably due to patient movement during scanning.

At the time when the acquisitions were performed according to this protocol, three patient datasets (referred to as patients 1, 2, and 3 in the sequel) were found exploitable. Image size was 512×512 pixels, with 88, 115, and 120 slices for patients 1, 2, and 3, respectively. In all images, pixel size was 0.97×0.97 mm^2 and slice thickness was 2.5 mm. All tumors were located in the lower part of the right lung. The tumor volume was approximately 160, 165, and 37 cm^3 in patients 1, 2 and 3, respectively.

Note also that the same phase can correspond to different lung volume percentages in different patients depending on each patient's breathing pattern. Approximated lung volumes were computed (see Table I) by automated segmentation using thresholding and morphological operations as described in [14]. Maximal displacements close to the diaphragm were estimated to: 23, 25, and 17 mm for patients 1, 2, and 3, respectively. We want to emphasize that the use of 4-D CT images is relatively new in the field of radiation therapy and that, although the technique has already been used in several clinical studies, it is still under development.

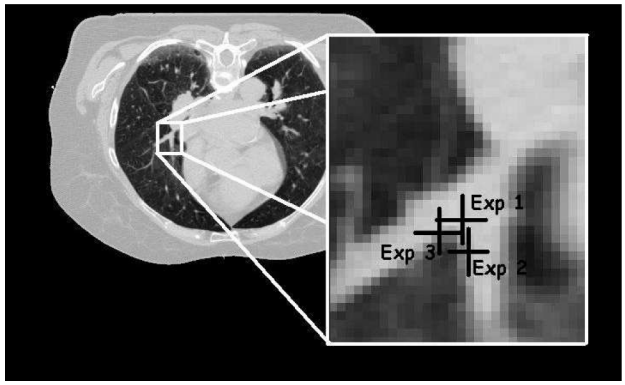


Fig. 2. Landmark selection by three experts. Mean position corresponds to pseudo-ground truth landmark. In this example, three positions lie on same slice but this was not always the case.

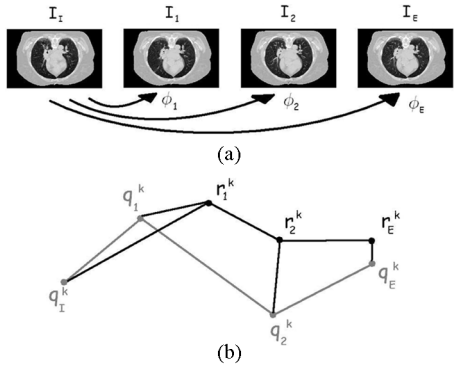


Fig. 3. (a) Transformations between all images in a sequence and end-inspiration reference image I_I are estimated. (b) Expert (q_i^k) and estimated (r_i^k) landmark definitions.

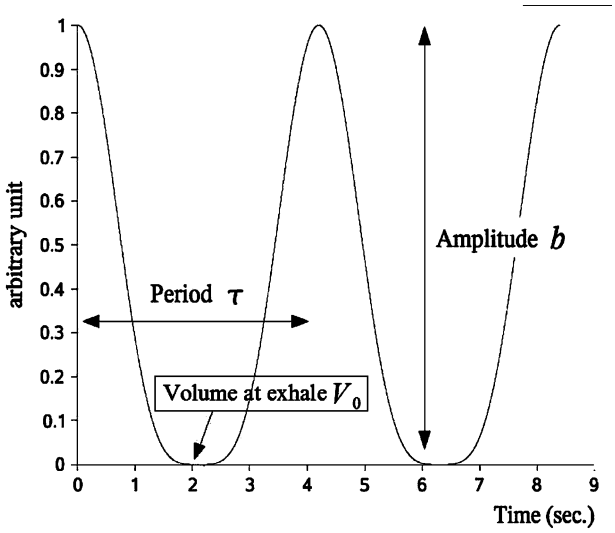


Fig. 4. Breathing cycle modeling proposed by Lujan *et. al.* [17] $(n = 2)$.

III. Methods

Criteria permitting us to evaluate and compare breathing motion estimators are proposed. They rely on the comparison of landmark locations obtained by manual reference selection (see the following) and landmark locations obtained by applying deformation fields obtained using automated motion estimation methods. Three methods (m_1, m_2, and m_3) will be described in Section IV to illustrate the proposed evaluation framework. For comparison purposes, we found it important to compare results obtained by these methods against the situation where no compensation was performed (noted m_0). Methods evaluation was based on two spatiotemporal distances between reference and observed trajectories and distance to direct straight line trajectories. Sections III-A and B describes landmark selection and tracking. Sections III-C and D present the two evaluation approaches.

A. Landmark Selection

A set of anatomical landmarks was selected and labeled inside the lungs in the reference image I_I of each of the three patients, by three medical experts. The instructions were to select salient anatomical features; each landmark should be undoubtedly identifiable and labeled with a descriptive name allowing other experts to find it. Examples of salient points are: carina, calcified nodules, culmen-lingula junction, specific branch of pulmonary arteries, apical pulmonary vein of the upper lobe, etc. (see Fig. 2). Actually, there might be some degree of statistical dependence between landmark point locations. In order to limit the impact of this dependence onto the statistical analysis, we asked the experts to select points distributed as evenly as possible all over the lungs (left/right lung, upper/lower, and central/peripheral parts of the lungs). The experts were also instructed to identify as many landmarks as possible with a minimum of 20. However, some experts extracted twice as many landmarks as the others. Up to 27 points were selected in patient 1, 41 in patient 2, and 56 in patient 3.

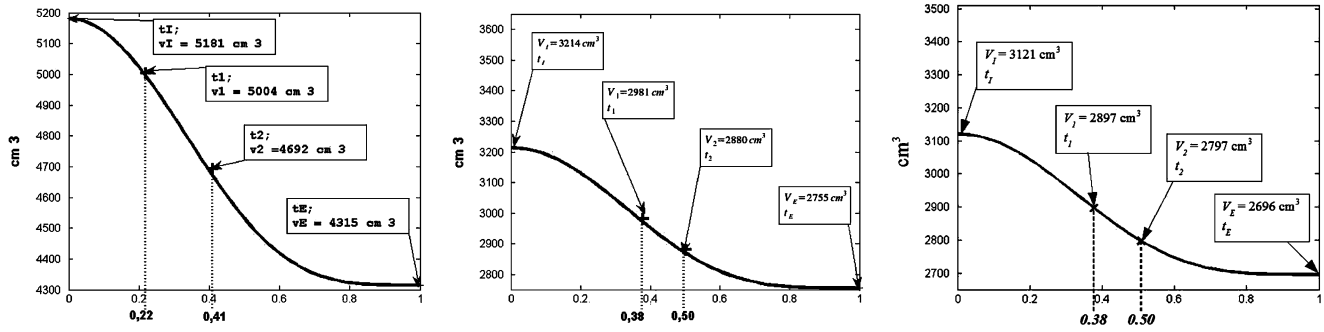
The tracking of initial landmarks across the following frames was performed by all the experts. They were not authorized to see other experts' results so as not to bias the selection. Finally, all inputs were averaged to obtain mean landmark locations. Let $\mathbf{p}_i^{e,k}$ denote the location of the kth landmark in image i (with $i \in \{I, 1, 2, E\}$, I and E corresponding to end-inspiration and end-expiration, respectively), selected by expert e. The three point locations issued from the different expert selections were averaged to define pseudo-ground truth landmarks denoted by

$$\mathbf{q}_i^k = \frac{1}{3}\sum_e \mathbf{p}_i^{e,k} \tag{1}$$

except for the reference image I_I in which $\mathbf{q}_I^k$ was the result of a unique selection. In order to estimate the inter-expert variability associated with manual identification of anatomical landmarks, we computed the standard deviation of the distances between all $\mathbf{q}_i^k$ and $\mathbf{p}_i^{e,k}$ values.

B. Landmark Tracking

Landmark motion is represented by a trajectory. A physical point at a given reference time is identified by its geometrical position: $\mathbf{x}_0 = (x_0, y_0, z_0)$. The mapping $\mathbf{x} = \boldsymbol{\phi}(\mathbf{x}_0, t)$ stands for the geometrical position of the same physical point at time t. $\boldsymbol{\phi}$ is the function which maps the physical point $\mathbf{x}_0$ from time t_0 to time t. By definition, $\boldsymbol{\phi}(\mathbf{x}_0, t_0) = \mathbf{x}_0$. The geometrical positions of the landmarks are expressed for discrete times of interest according to the previous definitions. $\mathbf{q}_I^k$ denotes the



(a) (b) (c)

Fig. 5. Lung volume curves for three patients during expiration. Values are indicated for four time points estimated by the respiratory model introduced in Section III-D1. (a) patient 1, (b) patient 2, and (c) patient 3.

position of the kth landmark in the reference image $\mathbf{I}_I$. We have the following relation:

$$\mathbf{q}_i^k = \boldsymbol{\phi}(\mathbf{q}_I^k, t_i) \tag{2}$$

where $\boldsymbol{\phi}(\mathbf{x}, t_i)$ maps every geometrical point $\mathbf{x}$ from reference time t_I to time t_i (with $i \in \{1, 2, E\}$).

In Section IV, we will introduce examples of methods allowing us to automatically estimate the displacement of the landmarks. For the three patient datasets, each method will be used to estimate the transformation $\tilde{\boldsymbol{\phi}}$ between all images in the sequence (i.e., I_1, I_2, I_E) and the reference end-inhalation image I_I (see Fig. 3). In the following, we will denote $\tilde{\boldsymbol{\phi}}_i$ the estimated function which maps image I_I to image I_i.

C. Punctual Accuracy Analysis

The first criterion to assess the accuracy of the motion estimation methods is an extension of the target registration error (TRE) proposed in [10]. Initially proposed for rigid motion, this criterion was extended to motion tracking of the sets of landmark points. Let us consider $\mathbf{q}_I^k$, the kth pseudo-ground truth landmark in the reference image I_I; its estimated geometrical location $\mathbf{r}_i^k$ in image I_i is calculated as $\mathbf{r}_i^k = \tilde{\boldsymbol{\phi}}_i(\mathbf{q}_I^k)$. The difference between pseudo-ground truth landmark positions and estimated landmark positions is illustrated in Fig. 3. The TRE for the transformation between image I_I and image I_i is defined by

$$\mathrm{TRE}_i = \frac{1}{n}\sum_{k=1}^{n}\sqrt{(\mathbf{q}_i^k - \mathbf{r}_i^k)^2} \tag{3}$$

where $n = 27$ for patient 1, $n = 41$ for patient 2, and $n = 56$ for patient 3. Error dispersion was represented using "box and whiskers plots" [15] to highlight the median and mean of each sample and its spreading and possible outliers. Bland–Altman diagrams [16] were used to compare the motion estimation results obtained with the different methods. Finally, paired student t-tests were performed to check whether two methods behaved equivalently or not, under the assumptions that the paired differences are independent and normally distributed (such assumptions were verified before applying the test).

D. Analysis of Spatiotemporal Trajectories

1) Respiratory Cycle Modeling: It is generally assumed that all the points in the volume reach their final position at the same time and that the temporal behavior along the trajectory is determined by a 1-D breathing signal. Several models of breathing cycles have been proposed in the literature. We chose the one proposed by Lujan *et al.* [17] (4) which models the dynamic breathing volume curve. It is based on a periodic but asymmetric function (more time spent at exhalation versus inhalation). In (4), V_0 is the volume at exhalation, b corresponds to the tidal volume (TV) which is the amount of air breathed in or out during normal respiration, $V_0 + b$ is the volume at inhalation, τ is the period of the breathing cycle, n is a parameter that determines the general shape (steepness or flatness) of the model, and φ is the starting phase of the breathing cycle (Fig. 4). Using the Lujan model, George *et al.* [18] have studied the correlation of respiratory motion traces between breathing cycles, based on 331 4-min respiratory traces acquired from 24 lung cancer patients. They advocated the use of $n = 2$. We followed their suggestion in the present paper. Of course, the period and magnitude of the motion due to breathing can vary, even over a short period of time. This model represents *a priori* knowledge of a conventional breathing cycle which will be incorporated into the validation procedure through the metrics introduced in the next section. Other models could also be considered. An illustration of the temporal position of each image in the test sequences according to this respiratory cycle modeling and the estimated volumes is given in Fig. 5 (see also Table I). The parameters for such a model can also be estimated using an external measurement system, such as the RPM

$$V(t) = V_0 + b\cos^{2n}\left(\frac{\pi}{\tau}t - \varphi\right). \tag{4}$$

2) Spatiotemporal Localization Error: The main drawback of the TRE metric is that it does not take into account the time spent at the main phases of a trajectory. According to the previously introduced breathing model (Section III-D1), material points move along their trajectory at variable speed (determined by the derivative of the volume curve V). In Section III-A, we mentioned that the dose deposit was mainly related to the duration of irradiation. Thus, a global and more pertinent metric

should take into account that more time is spent at the end-inspiration and end-expiration phases than between these extremes. In other words, estimation errors at an intermediate phase of the cycle should have lower weight than errors at extreme phases. This is the purpose of the following metric: for a given temporal interval $[t_a, t_b]$ of the respiratory cycle, we defined the spatiotemporal error (STE) as

$$\mathrm{STE}_{t_a,t_b}(T_1,T_2) = \frac{1}{t_b - t_a}\int_{t_a}^{t_b} \mathrm{dist}(T_1(s(t)), T_2(s(t)))dt \quad (5)$$

with *dist* the Euclidean distance and $s(t)$ the curvilinear abscissa. T_1 and T_2 denoted two trajectories. Let T be a parametric trajectory (which defines the set of the different locations of a material point during its motion) defined by

$$T[0,1] \in \mathbb{R} \to \mathbb{R}^3 \qquad s \mapsto T(s) = \begin{bmatrix} x(s) \\ y(s) \\ z(s) \end{bmatrix} \quad (6)$$

where s is the normalized curvilinear abscissa of the trajectory. This abscissa is a function of time and denotes the curve length traveled between initial time t_a and time t. The relation between time t and abscissa s is thus defined by

$$s[t_a, t_b] \in \mathbb{R} \to [0,1] \in \mathbb{R} \quad t \mapsto s(t) \quad (7)$$

where s is a strictly increasing function. The breathing cycle *a priori* model is incorporated into the STE metrics through the volume evolution function $V(t)$ defined by (4). s is thus expressed by $s(t) = (V(t) - V(t_a))/(V(t_b) - V(t_a))$. Fig. 6 illustrates the relation between respiratory cycle modeling $V(t)$ and the curvilinear abscissa. In this figure, the nonlinear relation is compared to the linear case where $s(t) = (t - t_a)/(t_b - t_a)$. Practically, a constant time step δt corresponds to a nonconstant abscissa step such as $\delta s = \delta t \cdot ds/dt$ because of the relative breathing velocity ds/dt. The relation depends on the chosen breathing model. In our case, the trajectory samples are denser at phases close to the end-inspiration and end-expiration time points than at intermediate phases.

The parametric trajectories were chosen such that the elementary displacements are approximated by linear interpolation between each pair of phases considered, and the abscissa s traverses this piecewise-linear trajectory (see Fig. 3). STE varies with the respiratory cycle model and the temporal spacing between images in the sequence (i.e., the relative position at t_1 and t_2 during expiration phase). A STE value equal to x means that, over a given portion of the cycle (from t_a to t_b), using trajectory T_1 instead of T_2 leads to x–mm shift in average. In practice, (5) was computed by approximating the integral by a sum over a set (one hundred or more) of regularly temporally spaced samples. Fig. 7 illustrates the distances between two trajectories computed with a linear and a nonlinear relation between t and s.

3) Straight-Linear and Piecewise-Linear Direct Trajectories: In order to clarify the different trajectories considered in the sequel, we introduced a set of acronyms whose meaning is illustrated in Fig. 8.

1) SRT denotes the *straight-linear reference trajectory*, that is the rectilinear trajectory obtained when directly connecting the points defined by the experts at the beginning and the end of expiration.
2) SET stands for the *straight-linear estimated trajectory*, the rectilinear trajectory obtained when directly connecting the position of a landmark extracted from the image taken at the beginning of expiration to the estimated corresponding point by a given method at the end of expiration.

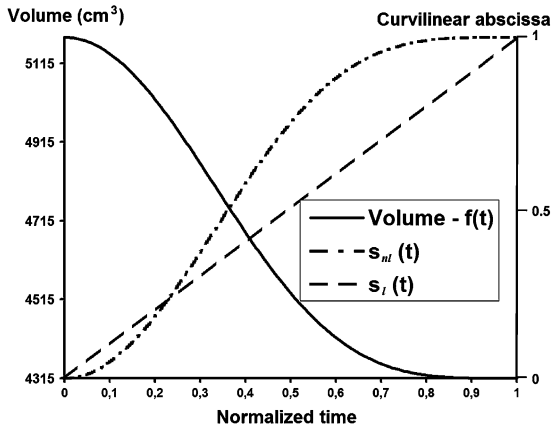


Fig. 6. Relation between global breathing cycle model and curvilinear abscissa s of trajectories. Linear case $s_l(t) = t$ is represented by dashed plot while nonlinear case s_{nl} computed from breathing cycle model (plain curve) is illustrated by dashed-dotted plot.

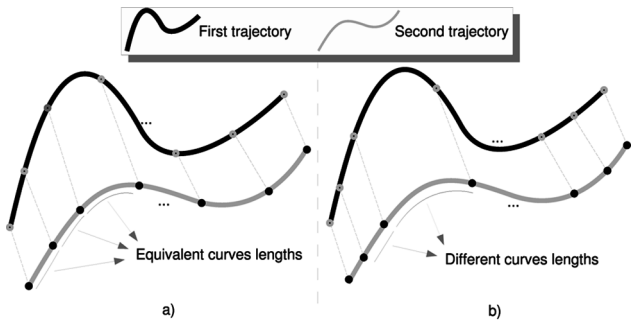


Fig. 7. (a) STE criterion with a linear and (b) nonlinear model.

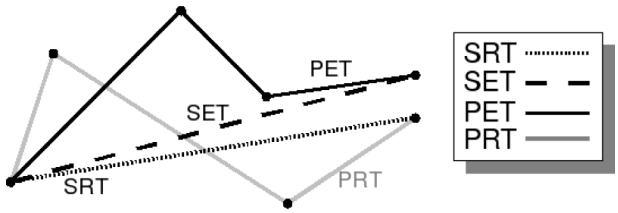


Fig. 8. Definition of straight-linear/piecewise-linear reference/estimated trajectories.

3) $\boldsymbol{PRT}$ corresponds to the *piecewise-linear reference trajectory*, piecewise linear trajectory obtained when connecting the reference points issued from all the time points.
4) $\boldsymbol{PET}$ denotes the *piecewise-linear estimated trajectory*, the piecewise linear trajectory obtained when connecting the estimated points issued from all the time points.

IV. Examples of Motion Estimators Evaluated in Our Framework

In order to illustrate the use of this evaluation framework, we selected and compared three available motion estimation methods. The literature on motion estimation methods is abundant. One approach is to seek a geometric transformation between two consecutive images in a sequence. This process is known as image registration and two of the three methods are based on this concept. Image registration algorithms are currently described as the combination of several components: a feature space, a similarity measure, a transformation model, and an optimization algorithm [19]–[21]. The goal is to find an optimal transformation that leads to maximum similarity (or minimum distance) between a reference image and a deformable floating image. Numerous methods have been proposed. Feature-based methods use landmark points [22], [23], organs contours [24], [25], or segmented surfaces to drive the transformation search. Intensity-based methods often refer to optical-flow like methods [22], [26], [27]. In this case, image similarity is defined as a statistical measure between the intensity (gray-levels) distributions of the two images, and deformable fields are the result of the optimization of a function establishing a tradeoff between image similarity and deformation smoothness. Another approach relies on biomechanical models [28]–[30] which do not explicitly use a similarity measure. Instead, they simulate organ deformation based on both physical material properties and constraints given by the initial and final states of the organs. They are usually based on the Finite-Element Method (FEM) and use physically based equations (elastic model for example) to simulate individual organ deformation (represented by triangular meshes for surface-based models or tetrahedral meshes for volume-based models). The individual material properties of each organ have to be described, with parameters such as Young's modulus and Poisson's ratio.

In order to illustrate our evaluation framework, we considered the three following motion estimation methods: m_1 is a bi-pyramidal free form deformation method, m_2 is an optimized optical flow method, and m_3 is a biomechanical method. These methods constitute rather conceptually different approaches to the problem of motion estimation. While m_1 is a parametric registration-based method, m_2 is a nonparametric one, and m_3 is based on an *a priori* physiological model of the lung dynamics. The three methods are therefore representative of different categories of motion estimation methods and good candidates to illustrate the proposed comparison techniques and metrics. The three methods are described hereafter. Let I_1 and I_2 be two images to be registered. We denote by $\mathbf{u}(\mathbf{x})$ the displacement of a point $\mathbf{x}$ and by $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$ the deformation.

A. Method m_1: Bi-Pyramidal Free Form Deformation-Based Image Registration

The nonrigid transformation is modeled using multilevel free-form deformations [31], [32]. The basic idea of the free-form deformation is to warp an object (a 3-D image in the present case) by moving an underlying set of control points distributed over a regular grid [33], [34] (the sets of control points and the landmarks defined in Section III-A are strictly uncorrelated). An interpolation function at each node of the grid is used to recover the final spatial continuous transformation. At any point $\mathbf{x}$, the deformation is computed by

$$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}' = \mathbf{x} + \sum_{j \in J} \mathbf{Q}^j B_j(\mathbf{x}) \tag{8}$$

where $J \in \mathbb{Z}^d$ defines the set of spatial parameter values, $\mathbf{Q}^j$ is a vector which contains the parameters of the transformation to be estimated (i.e., displacements of the control points), and $\mathbf{B}_j$ is a tensorial product of interpolation functions. We chose cubic B-Spline functions which are recognized to be the best choice in terms of computational efficiency, good approximation properties, and implicit smoothness (minimum curvature property) [35]. Cubic B-Spline functions have a limited support and are C^3 continuous. Thus, the influence of each control point is local and the final motion field is continuous. If we consider that the object to deform belongs to the $\mathbb{R}^d$ space and that the warping grid size is N, the transformation is thus defined by $d \times N^d$ parameters.

The algorithm relies on a bi-pyramidal formulation. In the first pyramid $\mathcal{P}_1$, a multiresolution decomposition of the original image $(I_m, I_{m-1}, \ldots, I_0)$ is stored, where each subresolution level I_{m-1} is obtained by first applying a low-pass Gaussian filter to the current image I_m, then decimating the number of pixels (or voxels). The second pyramid $\mathcal{P}_2$ allows for the multiscale decomposition of motion field $\mathbf{u}$ [35]–[37]. The final mapping function $\boldsymbol{\phi}$ belongs to the Hilbert space of finite energy deformation fields and can be approximated with a set of multilevel functions. The multilevel formulation of the transformation is described in [38]. At the coarsest level, $\boldsymbol{\phi}_0$ is defined by a few parameters. Once a deformation field has been estimated for one level l of $\mathcal{P}_2$, the next level is initialized using a projection onto the finer space. The algorithm is organized as follows: first, the transformation parameters are estimated at the coarsest image resolution and transformation level. Then, the image resolution is increased without changing any parameter of the transformation and a new estimation is performed. Afterward, the transformation level increases and previous parameters are projected onto the new finer space. These steps are repeated until the final image resolution and transformation levels are reached. The sum of squared differences (SSD) similarity criterion is used as follows:

$$\mathrm{SSD}(I_1, I_2, \boldsymbol{\phi}) = \int_{\mathbf{x} \in \Omega} (I_1(\mathbf{x}) - I_2(\boldsymbol{\phi}(\mathbf{x})))^2 \tag{9}$$

(with Ω the image overlapping domain). This criterion assumes the invariance of the material point brightness during motion

which is reasonable in our monomodal case. The optimization is achieved through a gradient descent search based on the first derivative

$$\nabla_Q \mathrm{SSD}(I_1, I_2, \boldsymbol{\phi}) = -2 \times \int_{\mathbf{x}\in\Omega} (I_1(\mathbf{x}) - I_2(\boldsymbol{\phi}(\mathbf{x}))) \times \frac{\partial I_2(\boldsymbol{\phi}(\mathbf{x}))}{\partial \mathbf{Q}} \tag{10}$$

of the SSD similarity criterion, with respect to the parameters of the current transformation level

$$\frac{\partial I_2(\phi(\mathbf{x}))}{\partial \mathbf{Q}} = \frac{\partial I_2(\phi(\mathbf{x}))}{\partial \mathbf{x}}\bigg|_{\phi(\mathbf{x})} \times \frac{\partial \phi}{\partial \mathbf{Q}}\bigg|_{\mathbf{x}}. \tag{11}$$

In (11), the terms on the right are, respectively, the gradient of I_2 at point $\mathbf{x} = \boldsymbol{\phi}(\mathbf{x})$ and the Jacobian of the transformation with respect to the parameters at point $\mathbf{x}$. At each iteration, the parameters of the current transformation level are updated according to

$$\mathbf{Q}_{i+1} = \mathbf{Q}_i + \lambda \nabla_Q \mathrm{SSD}(I_1, I_2, \boldsymbol{\phi}) \tag{12}$$

where λ is the maximum step of the gradient descent algorithm. We developed a C++ multithreaded version of the algorithm where the region of interest of the reference image was split according to the number of available processors. This considerably reduced the computing time on SMP architectures.

B. Method m_2: Optimized Optical Flow Method

The method m_2 is described in detail in [39]. It involved three main steps: 1) a preprocessing step consisting of segmenting the 3-D images into three regions labeled as air, patient, and lung; 2) *a priori* lung density modification in order to take into account the density decrease due to inhalation; 3) dense optical-flow like deformable registration.

The intensity conservation assumption implies that an image point has the same intensity in the other image but at a different location. However, lung densities are known to decrease from exhalation to inhalation according to the quantity of inhaled air. Therefore, the second step of this method aimed at artificially changing the lung density of one image in order to be closer to the intensity conservation assumption. We called this method *a priori lung density modification* (APLDM) [39]. Deformable registration was achieved by optimizing of a criterion composed of the SSD [see (9)] and a regularization measure by a steepest gradient descent algorithm. Previous works have shown that elastic and Gaussian regularizations lead to similar results for thorax CT images [39]–[41]. In this paper, we considered Gaussian regularization [42]. Gradient ∇L of the SSD criterion was expressed as proposed by Pennec *et al.* [43]

$$\nabla L(\mathbf{x}, \mathbf{u}) = \frac{I_1(\mathbf{x}) - I_2(\mathbf{x} + \mathbf{u}(\mathbf{x}))}{||\nabla I_1(\mathbf{x})||^2 + \alpha^2 (I_1(\mathbf{x}) - I_2(\mathbf{x} + \mathbf{u}(\mathbf{x})))^2} \nabla I_1(\mathbf{x}) \tag{13}$$

which limits the local displacement at each iteration according to a maximum vector displacement α. This criterion is an approximation of a second order gradient descent of the SSD [44]. The iterative process is given by

$$\mathbf{u}_{i+1}(\mathbf{x}) = G_\sigma(\mathbf{u}_i(\mathbf{x}) + \nabla L(\mathbf{x}, \mathbf{u}_i)). \tag{14}$$

$\mathbf{u}(\mathbf{x})$ denotes the displacement at point $\mathbf{x}$, $\nabla I_1(\mathbf{x})$ denotes the gradient of image I_1 at point $\mathbf{x}$, $\mathbf{u}_i$ denotes the displacement field at iteration i, and $G_\sigma(\cdot)$ denotes Gaussian kernel of variance $\sigma > 0$ (the higher the σ value the smoother the vector field). Gaussian filtering was performed using Deriche recursive Gaussian filter [45]. Images were previously resampled to an isotropic voxel's size of 2.5 mm^3.

C. Method M_3: Biomechanical Method

Various studies have analyzed organ motion with FEM methods. Some methods have been proposed to reproduce the lung behavior, such as the one by Grimal *et al.* [46] that was used to study thoracic impact injuries. In this paper, biomechanical parameters were studied in depth but breathing motion was not included into the modeling. Other methods focused on the breathing motion [47]. We recently proposed to apply, as boundary conditions, a normal displacement field to the external lung surface extracted from I_E limited by the maximal displacement field of the surface extracted from I_I. The method proposed has been detailed in [48]. It is based on a biomechanical approach and aims at physically simulating the lung behavior with laws of continuous mechanics based on physiological and anatomical studies and solved by FEM methods.

1) Model: The mechanical model was composed of: 1) a geometrical description of the lung which was discretized into small elements to constitute a mesh; 2) mechanical parameters to properly describe lung tissue behavior; and 3) boundary conditions to define the muscle actions allowing pulmonary motion. The initial state was obtained by lung surface mesh extraction from the CT images. The mesh was multilevel: 1) an external smooth mesh was obtained by a surface reconstruction method (Marching Cube) [49]; 2) this algorithm was extended to also provide an accurate tetrahedral mesh of the lung periphery [50]; and 3) bulk mesh was modeled by hexahedrons directly extracted from CT scan voxels for better convergence rate. The mechanical parameters, especially compliance, were issued from physiological measurements [51]. Compliance represents the ratio of air volume variation to the related air pressure variation. Each patient's data are linked to lung tissue elasticity, especially the Young modulus. The boundary conditions were derived from mechanical pleura action [52]. We computed the boundary conditions by imposing surface displacements. The boundaries of the lungs were modeled with a mesh extracted from CT scan image I_2 representing the deformed state. A uniform normal pressure was applied around the rib cage and around the diaphragm areas to simulate the pleural elastic recoil pressure. Adding contact condition constraints to that boundary allowed us either to block the displacement or to simulate the slipping skins. Fig. 9 illustrates these constraints. Note that in

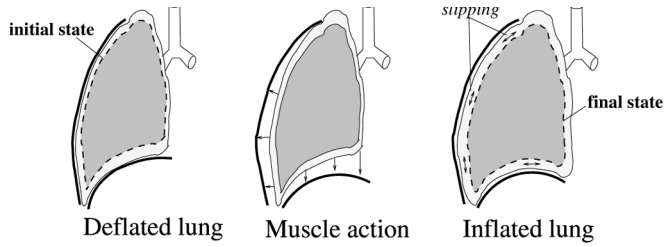


Fig. 9. Boundary conditions defined by diaphragm and rib cage actions for biomechanical model.

one dataset, the upper part of the lungs was missing. Therefore, the model could not be applied directly. To overcome this problem, the missing part of the lung apex (only about 1.5 cm) was approximated with a semi-ellipsoid. The semi-axe lengths were manually set.

2) Displacement Estimation: The solution to the problem was achieved using the FEM [53]. This numerical method consists of approaching the solution by a simple expression based on the discretization of the space into a mesh. In the present case, displacements U were estimated to minimize the residue R defined by

$$R(U) = F - K(U).U = 0 \tag{15}$$

where K is the stiffness matrix and F is the load vector. The term K expresses the rigidity of the lung. It depends both on mechanical parameters (Young modulus and Poisson's ratio) and on topological relationships between mesh nodes. The term F expresses the external forces applied to the lung, such as negative pressure. The displacement vector U represents the displacement of all the mesh nodes and allows us to estimate the displacement in the whole lung by interpolation. The space of such displacements is a subspace of functions and minimizing the residue $R(U)$ is equivalent to finding the best approximation of the solution to laws of continuous mechanics describing the behavior of a deformable solid under boundary condition stresses.

In our FEM approach, this nonlinear problem was solved using the Newton–Raphson algorithm which is an iterative method based on the computation of the gradient and the second-order gradient of $R(U)$. The displacements and strains were too large to assume that geometrical mesh changes would not influence the mechanical behavior. Therefore, we employed the iterative scheme presented in [54]. This method consists of readjusting the geometrical description at each load step in order to reevaluate $K(U)$. To account for contact conditions, we calculated algebraic distances between the nodes of the lung surface and the triangles representing the target lung surface (end-inhalation). If a distance remained positive, a negative pressure was applied to the corresponding node. When this became zero or negative, a contact between the current and the target lung surface was assumed. In this case, a restoring force was applied to ensure that the node was pulled back to the target surface. The restoring force was set as normal at the surface in order to allow surface sliding.

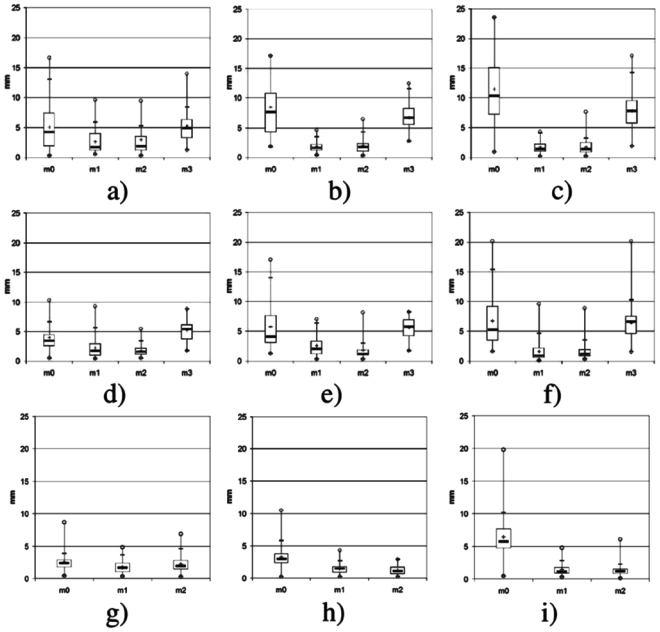


Fig. 10. Box and Whiskers plots for TRE criterion obtained for motion estimation methods. First line corresponds to patient 1, second line to patient 2 and third line to patient 3. First column corresponds to ϕ_1, second to ϕ_2, and third to ϕ_E. Each subfigure displays box and whiskers plot for four methods: m_0 is without transformation, m_1 corresponds to bi-pyramidal free-form-based image registration, m_2 to optimized optical flow method, and m_3 to biomechanical method.

Until now, we have focused on the technical aspects of the method: convergence, biomechanical parameters influence, and the interest of using a multilayer mesh. The fact that the lungs are composed of different biological tissues was not taken into account in this paper. As a consequence, mechanical properties were supposed to be uniform all over the lungs.

V. Results

The two criteria TRE and STE, introduced in (3) and (5), were used to evaluate the motion estimates obtained by the three previously described methods applied to the 4-D image sequences for the three patients presented in Section II. Box and whiskers, Bland–Altman, and Student paired t-tests analyses were derived for the two criteria. For methods m_1 and m_2, the resulting deformation field obtained between images I_I and I_1 was used as the starting deformation field for the subsequent registration (I_I to I_2), and so on. It allowed us to save some initial iterations by starting closer to the solution. Method m_1 was run on a 1.5-GHz Nonuniform Memory Access Multiprocessor SGI with 64Gb RAM, running Linux OS. The computation time for one iteration was related to the image resolution and the transformation level. For all the registrations, four image resolutions and four transformation levels were used with cubic B-Spline basis functions. The size of the regular grids were $5 \times 5 \times 5$, $7 \times 7 \times 7$, $11 \times 11 \times 11$, and $19 \times 19 \times 19$. Registration time, using ten processors, ranged from 18 min (patient 1) to 22 min (patient 2). Method m_2 was run on a 2.8-GHz PC with 1 Gb RAM running Linux OS. The computation time was about 1.5 s for one million voxels and for one iteration. Registration time

TABLE II
TRE IN MILLIMETERS FOR EACH METHOD AT EACH TIME POINT. FOR EACH CASE, FIRST VALUE CORRESPONDS TO MEAN VALUE OF TRE CRITERION. TWO VALUES IN PARENTHESES CORRESPOND TO FIRST AND THIRD QUARTILES, RESPECTIVELY

Methods	Patient 1		
	ϕ_1	ϕ_2	ϕ_E
m_0	5.1 (1.8 / 7.4)	8.4 (4.3 / 10.8)	11.4 (7.6 / 15.1)
m_1	2.7 (1.1 / 4.0)	1.8 (1.1 / 2.2)	1.7 (1.0 / 2.3)
m_2	3.0 (1.2 / 3.5)	2.0 (1.0 / 2.4)	1.7 (0.9 / 2.5)
m_3	4.5 (2.3 / 5.5)	5.9 (3.0 / 7.5)	6.5 (4.5 / 7.7)
Methods	Patient 2		
	ϕ_1	ϕ_2	ϕ_E
m_0	3.9 (2.5 / 4.5)	5.8 (3.1 / 7.7)	6.8 (3.5 / 9.2)
m_1	2.2 (1.0 / 3.0)	2.5 (1.2 / 3.4)	1.8 (0.5 / 2.2)
m_2	1.8 (1.1 / 2.2)	1.8 (1.0 / 1.9)	1.6 (1.8 / 2.0)
m_3	2.8 (2.2 / 3.5)	3.8 (2.7 / 5)	5 (3.1 / 6.5)
Methods	Patient 3		
	ϕ_1	ϕ_2	ϕ_E
m_0	2.5 (1.7 / 2.9)	3.3 (2.2 / 3.8)	6.4 (4.8 / 7.7)
m_1	1.7 (1.0 / 2.4)	1.4 (0.9 / 1.7)	1.4 (0.9 / 1.8)
m_2	2.2 (1.4 / 2.8)	1.2 (0.6 / 1.7)	1.3 (0.9 / 1.5)
m_3	–	–	–

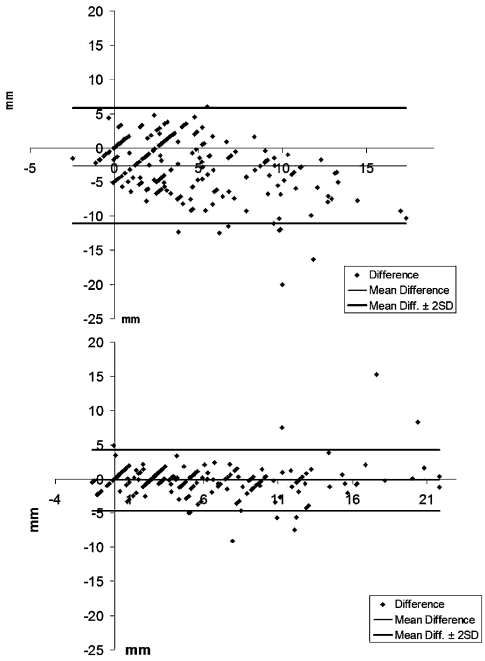


Fig. 11. Bland–Altman plots for comparing motion estimation methods (transformation ϕ_E, patient 1); (top) comparison of cranio-caudal displacements: expert's reference against m_3 estimations. (bottom) comparison of cranio-caudal displacements: expert's reference against m_1 estimations.

ranged from 5–9 min depending on the image size and the deformation to recover. Method m_3 was run on a 3.2-GHz PC. The computation time was about 2 min for a mesh composed of 7000 nodes and 20 000 elements. Method m_3 was not applied to patient 3. Indeed, as the tumor was attached to the diaphragm, the lung surface was found difficult to extract reliably.

A. First Criterion: TRE

Fig. 10 displays the Box and Whiskers plots for TRE on the three patients (one line per patient). The three columns correspond to the three transformations ϕ_1, ϕ_2, and ϕ_E, respectively. Each plot shows the statistics for the four methods (m_0 stands for "without registration," methods m_1, m_2, m_3 were described in Section IV). Table II gives the TRE statistics obtained with all the methods. Bland–Altman analysis was performed on each pair of methods. Only two representative plots are given here as other plots lead to similar behavior. Fig. 11 compares the landmark cranio-caudal coordinates given by the experts to those obtained with method m_3 [Fig. 11(a)] and method [Fig. 11(b)] for the transformation ϕ_E, in patient 1. A Bland–Altman diagram plots the differences between two methods against their mean. For each diagram, 95% of differences will lie between the two straight line limits (or, more precisely, between $d - 1.96s$ and $d + 1.96s$, where d stands for the mean difference and s for the standard deviation). Such a representation is very helpful to identify situations where the results given by two methods are truly discordant. Table III shows the Student t-test results between each pair of methods, allowing us to identify whether the TRE obtained with a method is statistically different from the TRE obtained with another method. The acceptable significance value α was set to 0.05. The p-value is a probability measure of the confidence against a null hypothesis $\mathcal{H}_0$. In the present case, hypothesis $\mathcal{H}_0$ was: "*the two methods are equivalent according to the computed* TRE metric." The lower the p-value, the more likely the difference between methods is significant.

TABLE III
STATISTICAL TESTS PERFORMED ON EACH PAIR OF METHODS (FOR ALL PATIENTS, ALL LANDMARKS, AND ALL MOTION ESTIMATION METHODS). IF p-VALUE IS GREATER THAN 0.1, DIFFERENCE IS NOT STATISTICALLY SIGNIFICANT (SYMBOL "="). IF $p < 0.1$ DIFFERENCE IS SIGNIFICANT (SYMBOL "+"), AND IF $p < 0.001$, DIFFERENCE IS HIGHLY SIGNIFICANT (SYMBOL "+ + +")

	Comparison	p-value	Is difference significant ?
Patient 1	m_0 vs m_1	<0.0001	+++
	m_0 vs m_2	<0.0001	+++
	m_0 vs m_3	0.00001	++
	m_1 vs m_2	0.42	=
	m_1 vs m_3	<0.0001	+++
	m_2 vs m_3	<0.0001	+++
	Comparison	p-value	Is difference significant ?
Patient 2	m_0 vs m_1	<0.0001	+++
	m_0 vs m_2	<0.0001	+++
	m_0 vs m_3	0.5067	=
	m_1 vs m_2	0.007	+
	m_1 vs m_3	<0.0001	+++
	m_2 vs m_3	<0.0001	+++
	Comparison	p-value	Is difference significant ?
Patient 3	m_0 vs m_1	<0.0001	+++
	m_0 vs m_2	<0.0001	+++
	m_1 vs m_2	0.34	=

B. Second Criterion: STE

Fig. 12 displays the length of the PRT according to the distance to the lung apex for patient 1. The greater displacements were observed near the diaphragm. Therefore, the magnitude of

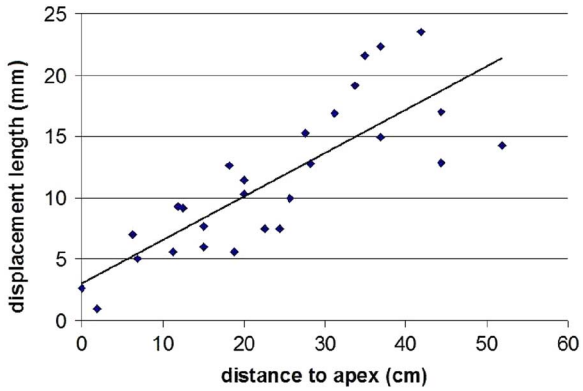


Fig. 12. Distribution of norm of displacements according to distance to lung apex for patient 1(PRT).

TABLE IV
STE METRIC (MEAN VALUE IN MILLIMETERS ± STANDARD VARIATION), FOR THREE DATASETS, BETWEEN $\boldsymbol{PRT}$s AND $\boldsymbol{PET}$s, OBTAINED WITH EACH METHOD. LAST THREE COLUMNS DEPICT MAXIMUM STE VALUES

	STE			Maximum STE		
Methods	Patient 1	Patient 2	Patient 3	Patient 1	Patient 2	Patient 3
m_0 (no motion)	6.8 ± 3.6	4.2 ± 2.7	3.2 ± 1.7	14.6	11.6	9.4
m_1 (***PRT*** vs ***PET***)	2.2 ± 1.2	1.7 ± 0.9	1.2 ± 0.4	5.1	4.4	2.6
m_2 (***PRT*** vs ***PET***)	2.3 ± 1.5	1.3 ± 0.9	1.2 ± 0.5	6.5	5.7	3.4
m_3 (***PRT*** vs ***PET***)	5.6 ± 1.9	4.3 ± 1.4	–	10.4	8.7	–

TABLE V
FOR EACH SEQUENCE, STE METRIC (MEAN VALUE IN MILLIMETERS ± STANDARD VARIATION), BETWEEN $\boldsymbol{PRT}$s AND $\boldsymbol{SRT}$s ON THE ONE HAND AND $\boldsymbol{SET}$s ON THE OTHER HAND. ALL LANDMARKS ARE TAKEN INTO ACCOUNT

	STE		
Methods	Patient 1	Patient 2	Patient 3
PRT *vs.* ***SRT***	2.9 ± 1.7	1.2 ± 0.7	0.6 ± 0.2
PRT *vs.* ***SET*** (m_1)	3.3 ± 1.8	1.7 ± 1.1	1.0 ± 0.5
PRT *vs.* ***SET*** (m_2)	3.4 ± 2.0	1.7 ± 1.2	1.0 ± 0.4
PRT *vs.* ***SET*** (m_3)	5.6 ± 2.4	3.9 ± 1.8	–

the motion to be recovered by motion estimators was variable along the thorax (from about 3 mm near the apex to 25 mm close to the diaphragm). Table IV gives the STE statistics between the $\boldsymbol{PRT}$ and $\boldsymbol{PET}$ obtained with each method in order to discuss spatiotemporal errors. The temporal sampling differed from one sequence to another as a function of the breathing cycle modeling [(4) and Fig. 5]. The importance of accounting for motion (through intermediate time points) in radiation treatment is assessed in Table V, which provides the STE statistics between the $\boldsymbol{PRT}$ and $\boldsymbol{SET}$ obtained with each registration method. In particular, these results allow us to discuss whether all the frames of the sequence are essential or if only a few of them (i.e., the two extreme phases) are needed. Table VI displays the results of the student *t*-test comparing the STE metric obtained with $\boldsymbol{SET}$ and $\boldsymbol{PET}$, respectively.

VI. DISCUSSION

The proposed framework allows the comparison, in terms of accuracy, of motion estimation methods from 4-D scans. First, the punctual accuracy of the three selected methods ($m_1, m_2,$ and m_3) was evaluated with the TRE criterion, and the behavior of the different methods was studied with the help of statistical tools (Bland–Altman, Box and Whiskers and student *t*-tests). The STE criterion introduces temporal information into the evaluation framework through a breathing model. This is to better take into account the dynamics of the organs in our context, which is of particular importance in radiotherapy of the lungs.

TABLE VI
STATISTICAL TESTS PERFORMED (FOR EACH PATIENT AND EACH POINT) TO COMPARE $\boldsymbol{SET}$s AND $\boldsymbol{PET}$s FOR EACH METHOD. IF *p*-VALUE IS GREATER THAN 0.1, DIFFERENCE IS NOT STATISTICALLY SIGNIFICANT (SYMBOL "="). IF $p < 0.1$, DIFFERENCE IS SIGNIFICANT (SYMBOL "+"), AND IF $p < 0.001$, DIFFERENCE IS VERY SIGNIFICANT (SYMBOL "+ + +")

	Comparison	p-value	Is difference significant ?
Patient 1	***SET***(m_1) vs ***PET***(m_1)	<0.0001	+ + +
	SET(m_2) vs ***PET***(m_2)	<0.0001	+ + +
	SET(m_3) vs ***PET***(m_3)	0.63	=
	Comparison	p-value	Is difference significant ?
Patient 2	***SET***(m_1) vs ***PET***(m_1)	0.95	=
	SET(m_2) vs ***PET***(m_2)	<0.0001	+ + +
	SET(m_3) vs ***PET***(m_3)	0.0039	++
	Comparison	p-value	Is difference significant ?
Patient 3	***SET***(m_1) vs ***PET***(m_1)	0.0005	+ + +
	SET(m_2) vs ***PET***(m_2)	<0.0001	+ + +

A. Method Accuracy (TRE Criterion)

Overall landmark errors (TRE, Table II) for the two intensity-based methods m_1 and m_2 (2.1 and 2.0 mm, respectively) were in agreement with the voxel size ($0.9 \times 0.9 \times 2.5$ mm^3) and the experts variability (1.2 mm). We also observed that displacements were generally slightly underestimated (mean difference of the Bland–Altman diagrams below the zero line), suggesting that the regularizations used in intensity-based methods (cubic B-splines for m_1 and Gaussian smoothing for m_2) sometimes prevent points to reach their true location. For the biomechanical method m_3, only a rather rough mesh was considered (mean hexahedron size is $24 \times 12 \times 3$ mm^3 for patient 1 and $10 \times 10 \times 10$ mm^3 for patient 2). The landmark points were defined in areas of significant intensity gradients which correspond to materially heterogeneous regions not yet included into the biomechanical model. Nevertheless, we observe in Fig. 10 that the estimated average error is approximately less than half the average mesh element size: lower than 6.5 mm for patient 1 and lower than 5 mm for patient 2. Bland–Altman diagrams (Fig. 11), revealed one specific landmark position for which the location provided by the experts was not in agreement. After discussion with the experts, this landmark was discarded from the experiments.

The TRE statistical descriptors (mean, quartiles, Fig. 10, and Table II) computed from m_1 and m_2 are similar between patient 1 and patient 2 despite the overall greater motion magnitude in patient 1 (see Table II). The slight differences observed between m_1 and m_2 may be related to the transformation model used. The nonparametric representation of method m_2 allows us to estimate deformation with a precision depending on the voxel size. For method m_1, the motion field was expressed with

a continuous model and the accuracy depended on the size of the grid and of a region of interest (ROI). For example, in patient 1 dataset, the ROI was about 200×200 pixels in the acquisition plane, corresponding to approximately one control point every ten voxels (9 mm in native image plane). The ROI was larger for patient 2 dataset (250×250 pixels) due to morphological differences between the two patients implying a distribution of one control point every 14 voxels (12.6 mm in native image plane). This might explain accuracy differences between the two sequences when using method m_1.

Methods m_1 and m_2 were found to be statistically similar but significantly different from method $m_3 (p < 0.001)$, as shown by results of the student *t*-tests summed up in Table III.

B. Trajectory Study (STE Criterion)

1) STE As Method Evaluation Criterion: The study of trajectories through the STE metric showed that the mean distance between reference $(\boldsymbol{PRT})$ and estimated piecewise-linear trajectories $(\boldsymbol{PET})$ was around 1.6 mm with methods m_1 and m_2 and around 5 mm with method m_3 (Table IV). The difference between methods m_1 and m_2 was found not statistically significant. With m_1 and m_2, the STE for patient 1 (see Table IV) was slightly higher than the mean of the three TRE (2.0 and 2.2 mm compared to 2.2 and 2.3 mm, Table II), while the STE for patients 2 and 3 was inferior to the corresponding TRE values. Even if m_1 and m_2 lead to comparable average results, the maximum STE was lower using method m_1. Whereas each time point contributes with an equal weight to the mean TRE, the STE metric introduces a variable weight according to the breathing cycle model and displacement speed along the trajectories. This implies that intermediate time points ($\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$) influence depends on their relative location in the breathing cycle (see Fig. 5). STE values are inferior to the mean TRE values for patients 2 and 3 ($1.7/1.2$ mm versus $2.2/1.5$ mm average TRE for method m_1) because the most influent time point is t_E. On the contrary, STE values are greater to the mean TRE values for patient 1 since the first intermediate time point contributes more. In conclusion, the STE metric takes into account the breathing dynamics and the acquisition time of each of the sequence frame.

The STE criterion depends on the selected breathing model. Other breathing models could be considered. The proposed framework could also be used to study the hysteresis pattern which is known to occur during breathing (different inhalation and exhalation pathways), but it would require the definition of many more landmarks. STE criteria should be well adapted to compare inhalation and exhalation trajectories and to put the focus on different parts of the breathing cycle.

2) Taking Intermediate Frames Into Account in Lung Radiotherapy Treatments: Table V illustrates, through the three sequences studied, the importance of taking into account motion in radiotherapy treatment. $\boldsymbol{PRT}$ compared to $\boldsymbol{SRT}$ represents the error committed when straight-linear trajectories are considered instead of piecewise-linear ones. This error was particularly low for patients 2 and 3 (1.2 and 0.6 mm, respectively). It suggests that, for the considered trajectories, the observed motion was almost rectilinear. Indeed, using straight-linear trajectories (one single motion estimation between end-inspiration and end-expiration images) increased the overall error for patient 1, whereas errors remain almost equivalent for patients 2 and 3 (in Tables IV and V, STE rises from 2.2 to 3.3 mm for method m_1). However, it is not clear at this stage whether such discrepancies in accuracy results between patients come from the variablity in patient organ motion or from the 4-D acquisitions. Moreover, methods behave differently: although results for m_1 and m_2 lead to similar STE, Table VI shows that the error difference between straight-linear and piecewise-linear trajectories was highly significant for method m_2 for all patients, whereas, for method m_1, it was significant for patients 1 and 3. It seems that trajectories estimated with method m_1 were more linear than those estimated with method m_2. The results concerning method m_3 called our attention to the contact conditions of the FEM. We observed afterward that this contact condition had not been properly handled. In particular, the conditions for surface contact had not been met for some nodes due to the mesh resolution, thus explaining why some differences could be observed. Inaccuracy at the contact was of the order of 5 mm, which explains why the differences were almost constant, whatever the displacement, and why the straight-linear process gave better results. Moreover, even if the method m_3 is still under development, the current evaluation study has made it possible to point out some of the problems that should be solved in the future. Overall, the benefit of incorporating additional frames for taking into account the breathing motion appears to depend on the patient. So, in the absence of *a priori* information on the patient breathing pattern, it is certainly better to dispose of more than the two extreme phase images.

Criteria were computed over the whole image domain. But it is known that lung motion is not homogeneous during the breathing cycle and that trajectories are longer and more linear near the diaphragm than near the lung apex (Fig. 12). In the future, by using the same criteria, it should be very interesting to study motion behavior in the different parts of the lungs (lower part of the lung versus upper part, tumor areas). The chronology of the landmark trajectory is globally imposed by the Lujan's model: all the landmarks are assumed to have the same temporal evolution (homogeneous behavior), but it is known that such an assumption is not rigorously true. However, if information about the breathing pattern in different lung regions would be available (for example by means of external or internal markers), it could be easily inserted into the proposed STE measure; currently $V(t)$ only depends on the time variable, it would switch to $V(\mathbf{x}, t)$ according to a spatiotemporal breathing model. Indeed, the main reason for using a breathing model is to compensate for the limited number of temporal frames. The model would be less necessary if we dispose of more temporally resolved sequences. However, there is still a tradeoff between image spatial and temporal resolutions and the acquisition costs in terms of dose delivered to the patient, the compatibility of the acquisition time with the clinical constraints, and the management of large amounts of data. The acquisitions considered in this paper take into account those constraints as they have been indeed used for patient treatment planning.

Motion validation by means of landmarks is intrinsically limited to the point location with the consequence that no information is available in between those points. Landmarks were

selected as evenly as possible all over the lungs based on visible anatomical structures. However, in homogeneous regions, no landmark could be identified and thus the quality of the estimated deformation field could not be assessed within these regions. Moreover, medical experts generally find it difficult and time consuming to select landmarks. This is the reason why the number of landmarks was limited to some tens. To our knowledge, this is one of the first times that such an evaluation is performed on the lungs with such a significant number of landmarks. More complex primitives (such as 3-D lines following vessels) would bring higher level information and thus contribute to define a better ground truth for the evaluation. This would still require patients to be evaluated by experts, which is a difficult task in 3-D.

VII. Conclusion

In this paper, we propose a strategy and criteria in order to evaluate the accuracy of motion estimators from 4-D CT sequences with a limited number of phases between end-inspiration and end-expiration. Such an evaluation is particularly crucial in radiation therapy where estimated motion can be used to estimate the distribution of the absorbed dose during the therapeutic irradiation of moving organs such as the lungs. The main contributions of this paper were the setup of test cases and of a procedure to obtain expert inputs (carefully identifying more than 500 landmarks over four phases and three patients) and the proposal of spatiotemporal criteria to evaluate the predictions of landmark displacements through the respiratory cycle. The STE criterion allows us to take into account the dynamics of the motion by introducing an *a priori* respiratory cycle model. It can be considered as a specialization of the TRE metric to the specific context of breathing motion compensation. The proposed comparison framework was illustrated by the study of three different motion estimation methods (two registration based methods, and one biomechanical model based method). The study allowed us to compare the accuracy of those methods and to highlight some of their limits. The analysis also demonstrated the interest of incorporating several frames over the respiratory cycle in view of better adapting the therapy of lung tumors to the patient. This paper has been conducted on three 4-D datasets encompassing only half the respiratory cycle. The study should be pursued by including additional datasets and extending the tracking over the entire respiratory cycle. Adding more landmarks, in particular outside the lung region, could also improve the evaluation of accuracy. Dose deposit simulations could be performed on 4-D images in order to quantify the influence of the type of motion estimators on dose distribution. Finally, a similar framework could also be used to evaluate motion tracking methods in other medical imaging contexts such as in cardiac motion analysis.

Acknowledgment

The authors would like to thank G. Sharp, S. Jiang, and N. Choi from the Massachusetts General Hospital, Boston, for providing 4-D CT datasets. The authors also wish to thank L. Claude and all other persons who spent a lot of time selecting landmarks.

References

[1] M. Goitein, "Organ and tumor motion: An overview," *Seminars Radiation Oncol.*, vol. 14, no. 1, pp. 2–9, Jan. 2004.

[2] C. Ling, E. Yorke, H. Amols, J. Mechalakos, Y. Erdi, S. Leibel, K. Rosenzweig, and A. Jackson, "Editorial: High-tech will improve radiotherapy of NSCLC: A hypothesis waiting to be validated," *Int. J. Radiation Oncol. Biol. Phys.*, vol. 60, no. 1, pp. 3–7, 2004.

[3] G. Mageras and E. Yorke, "Deep inspiration breath hold and respiratory gating strategies for reducing organ motion in radiation treatment," *Seminars Radiation Oncol.*, vol. 14, no. 1, pp. 65–75, Jan. 2004.

[4] S. Vedam, P. Keall, V. Kini, H. Mostafavi, H. Shukla, and R. Mohan, "Acquiring a four-dimensional computed tomography dataset using an external respiratory signal," *Phys. Med. Biol.*, vol. 48, no. 1, pp. 45–62, 2003.

[5] P. Keall, S. Joshi, S. Vedam, J. Siebers, V. Kini, and R. Mohan, "Four-dimensional radiotherapy planning for {DMLC}-based respiratory motion tracking," *Med. Phys.*, vol. 32, no. 4, pp. 942–951, 2005.

[6] T. Zhang, R. Jeraj, H. Keller, W. Lu., O. GH., M. TR., T. Mackie, and P. B., "Treatment plan optimization incorporating respiratory motion," *Med. Phys.*, vol. 31, no. 6, pp. 1576–86, Jun. 2004.

[7] E. Rietzel, G. Chen, N. Choi, and C. Willet, "Four-dimensional image-based treatment planning: Target volume segmentation and dose calculation in the presence of respiratory motion," *Int. J. Radiation Oncol. Biol. Phys.*, vol. 61, no. 5, pp. 1535–50, Apr. 2005.

[8] K. Brock, M. DL., T. H. RK., H. SJ., L. Dawson, and J. Balter, "Inclusion of organ deformation in dose calculations," *Med. Phys.*, vol. 30, no. 3, pp. 290–5, Mar. 2003.

[9] J. West, J. Fitzpatrick, M. Wang, B. Dawant, C. Maurer, R. Kessler, R. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. Elsen, S. Napel, T. Sumanaweera, B. Harkness, P. Hemler, D. Hill, D. Hawkes, C. Studholme, J. Maintz, M. Viergever, G. Malandain, X. Pennec, M. Noz, G. Maguire, M. Pollack, C. Pelizzari, R. Robb, D. Hanson, and R. Woods, "Comparison and evaluation of retrospective intermodality image registration techniques," *J. Comput. Assist. Tomogr.*, vol. 21, no. 4, pp. 554–566, 1997.

[10] J. Fitzpatrick and J. West, "The distribution of target registration error in rigid-body point-based registration," *IEEE Trans. Med. Imag.*, vol. 20, no. 9, pp. 917–927, Sep. 2001.

[11] N. Pauna, P. Croisille, N. Costes, A. Reilhac, T. Makela, O. Cozar, M. Janier, and P. Clarysse, "A strategy to quantitatively evaluate MRI/PET cardiac rigid registration methods using a Monte Carlo simulator," in *Proc. 2nd Int. Workshop Functional Imag. Model. Heart FIMH'03*, Lyon, France, 2003, pp. 194–204.

[12] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. L. Goualher, D. L. Collins, A. Evans, G. Malandain, N. Ayache, G. E. Christensen, and H. J. Johnson, "Retrospective evaluation of intersubject brain registration," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1120–1130, Sep. 2003.

[13] T. Pan, T. Lee, E. Rietzel, and G. Chen, "4-D-CT imaging of a volume influenced by respiratory motion on multi-slice CT," *Med. Phys.*, vol. 31, no. 2, pp. 333–340, 2004.

[14] D. Sarrut, V. Boldea, M. Ayadi, J. Badel, C. Ginestet, and S. Clippe, "Non-rigid registration method to assess reproducibility of breath-holding with ABC in lung cancer," *Int. J. Radiation Oncol. Biol. Phys.*, vol. 61, no. 2, pp. 594–607, 2005.

[15] J. W. Tukey, "Box-and-whisker plots," *Exploratory Data Anal.*, pp. 39–43, 1977.

[16] J. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, pp. 307–310, 1986.

[17] A. Lujan, L. EW, J. Balter, and R. Ten Haken, "A method for incorporating organ motion due to breathing into 3D dose calculations," *Med. Phys.*, vol. 26, no. 5, pp. 715–20, 1999.

[18] R. George, S. Vedam, T. Chung, V. Ramakrishnan, and P. J. Keall, "The application of the sinusoidal model to lung cancer patient respiratory motion," *Med. Phys.*, vol. 32, no. 9, pp. 2850–2861, Sep. 2005.

[19] W. Crum, T. Hartkens, and D. Hill, "Non-rigid image registration: Theory and practice," *Br. J. Radiol.*, vol. 77, no. 2, pp. 140–153, 2004.

[20] B. Zitova and J. Flusser, "Image registration methods: A survey," *Image Vision Comput.*, vol. 21, pp. 977–1000, 2003.

[21] C. Maurer and J. Fitzpatrick, , R. J. Maciunas, Ed., "A review of medical image registration," in *In Interactive ImageGuided Neurosurgery*. Park Ridge, IL: Amer. Assoc. Neurological Surgeons, 1993, pp. 17–44.

[22] L. Fan, C. Chen, J. Reinhardt, and E. Hoffman, "Evaluation and application of 3D lung warping and registration model using HRCT images," *SPIE Med. Imag.*, vol. 4321, pp. 234–243, 2001.

[23] B. Li, G. Christensen, E. Hoffman, G. McLennan, and J. Reinhardt, "Establishing a normative atlas of the human lung: Intersubject warping and registration of volumetric CT images," *Acad. Radiol.*, vol. 10, no. 3, pp. 255–265, 2003.

[24] D. Yan, D. Jaffray, and J. Wong, "A model to accumulate fractionated dose in a deforming organ," *Int. J. Radiation Oncol. Biol. Phys.*, vol. 44, no. 3, pp. 665–75, Jun. 1999.

[25] B. Schaly, J. Kempe, G. Bauman, J. Battista, and J. Van Dyk, "Tracking the dose distribution in radiation therapy by accounting for variable anatomy," *Phys. Med. Biol.*, vol. 49, no. 5, pp. 791–805, Mar. 2004.

[26] W. Lu, M. Chen, G. Olivera, K. Ruchala, and T. Mackie, "Fast free-form deformable registration via calculus of variations," *Phys. Med. Biol.*, vol. 49, no. 14, pp. 3067–3087, 2004.

[27] T. Sundaram and J. Gee, "Towards a model of lung biomechanics: Pulmonary kinematics via registration of serial lung images," *Med. Image Anal.*, vol. 9, no. 6, pp. 254–37, Dec. 2005.

[28] M. Birkner, D. Yan, M. Alber, J. Liang, and F. Nusslin, "Adapting inverse planning to patient and organ geometrical variation: Algorithm and implementation.," *Med. Phys.*, vol. 30, no. 10, pp. 2822–31, Oct. 2003.

[29] J. Lian, L. Xing, S. Hunjan, C. Dumoulin, J. Levin, A. Lo, R. Watkins, K. Rohling, R. Giaquinto, D. Kim, D. Spielman, and B. Daniel, "Mapping of the prostate in endorectal coil-based MRI/MRSI and CT: A deformable registration and validation study," *Med. Phys.*, vol. 31, no. 11, pp. 3087–3094, July 2003.

[30] T. Zhang, N. Orton, T. Mackie, and B. Paliwal, "Technical note: A novel boundary condition using contact elements for finite element based deformable image registration," *Med. Phys.*, vol. 31, no. 9, pp. 2412–5, Sept. 2004.

[31] B. Delhay, "Estimation spatio-temporelle de mouvement et suivi de structures déformables. Application à l'imagerie dynamique du coeur et du thorax," Ph.D. dissertation, Instit. Nat. des Sci. Appliquées de Lyon, Lyon, France, 2006.

[32] B. Delhay, P. Clarysse, C. Pera, and I. E. Magnin, "A spatio-temporal deformation model for dense motion estimation in periodic cardiac image sequences," in *Proc. Workshop MICCAI 2006: From Statistical Atlases Personalized Models: Understand. Complex Diseases Populations Individuals*, Copenhagen, Denmark, 2006, pp. 87–90.

[33] T. W. Sederberg and S. R. Parry, "Free-form deformation of solid geometric models," in *Proc. SIGGRAPH '86, Comput. Graphics 20*, Aug. 1986, vol. 4, pp. 151–159.

[34] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[35] M. J. Ledesma-Carbayo, J. Kybic, M. Desco, A. Santos, M. Shuling, P. H. ans, and M. Unser, "Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation," *IEEE Trans. Med. Imag.*, vol. 24, no. 9, pp. 1113–1126, Sep. 2005.

[36] V. Noblet, C. Heinrich, F. Heitz, and J. Armspach, "3-D deformable image registration: A topology preservation scheme based on hierarchical deformation models and interval analysis optimization," *IEEE Trans. Image Process.*, vol. 14, no. 5, pp. 553–566, May 2005.

[37] T. Rohlfing, C. Maurer, D. Bluemke, and M. Jacobs, "Volume-preserving non-rigid registration of MR breast images using free-form deformation with an incompressibility constraint," *IEEE Trans. Med. Imag.*, vol. 22, no. 6, pp. 730–741, Jun. 2003.

[38] S. Lee, G. Wolberg, and S. Y. Shin, "Scattered data interpolation with multilevel B-splines," *IEEE Trans. Visual. Comput. Graphics*, vol. 3, no. 3, pp. 228–244, Mar. 1997.

[39] D. Sarrut, V. Boldea, S. Miguet, and C. Ginestet, "Simulation of 4-D CT images from deformable registration between inhale and exhale breath-hold CT scans," *Med. Phys.*, 2006.

[40] V. Boldea, D. Sarrut, and S. Clippe, "Lung deformation estimation with non-rigid registration for radiotherapy treatment," in *Medical Image Computing and Computer-Assisted Intervention MICCAI'2003*. New York: Springer Verlag, 2003, vol. 2878, pp. 770–7.

[41] V. Boldea, D. Sarrut, and C. Carrie, "Comparison of 3D dense deformable registration methods for breath-hold reproducibility study in radiotherapy," *SPIE Med. Imag.: Visualization, Image-Guided Procedures, Display*, vol. 5747, pp. 222–230, 2005.

[42] J. Thirion, "Image matching as a diffusion process: An analogy with maxwell's demons," *Med. Image Anal.*, vol. 2, no. 3, pp. 243–260, 1998.

[43] X. Pennec, P. Cachier, and N. Ayache, , C. Taylor and A. Colschester, Eds., "Understanding the demon's algorithm: 3D non rigid registration by gradient descent," in *Medical Image Computing and Computer-Assisted Intervention MICCAI'99*. Cambridge, U.K.: Springer Verlag, 1999, vol. 1679, pp. 597–605.

[44] P. Cachier and N. Ayache, "Isotropic energies, filters and splines for vectorial regularization," *J. Math. Imag. Vision*, vol. 20, no. 3, pp. 251–265, May 2004.

[45] R. Deriche, "Recursively implementing the gaussian and its derivatives," *INRIA, Tech. Rep., 1893* Apr. 1993 [Online]. Available: http://www.inria.fr/rrrt/rr-1893.html.

[46] Q. Grimal, A. Watzky, and S. Naili, "Nonpenetrating impact on the thorax: A study of the wave propagation," *Comptes Rendus De L'Academie des Sciences*, vol. IIb, no. 329, pp. 655–662, 2001.

[47] K. Brock, M. Sharpe, L. Dawson, S. Kim, and D. Jaffray, "Accuracy of finite element model-based multi-organ deformable image registration," *Med. Phys.*, vol. 32, no. 6, pp. 1647–59, Jun. 2005.

[48] P. Villard, M. Beuve, B. Shariat, V. Baudet, and F. Jaillet, "Simulation of lung behaviour with finite elements: Influence of bio-mechanical parameters," in *MEDIVIS'05: Proc. Third Int. Conf. Med. Inf. Visualisation—BioMedical Visualisation*, Washington, DC, 2005, pp. 9–14.

[49] W. Lorensen and H. Cline, "Marching cubes: A high resolution 3D surface reconstruction algorithm," *Comput. Graphics*, vol. 21, pp. 163–169, 1987.

[50] P. Villard, M. Beuve, B. Shariat, V. Baudet, and F. Jaillet, "Lung mesh generation to simulate breathing motion with a finite element method," in *Information Visualisation*. London, U.K.: IEEE Computer Society, 2004, pp. 194–199.

[51] M. L. Moy and S. H. Loring, "Compliance," *Seminar Respiratory Critical Care Medecine*, vol. 19, no. 4, pp. 349–359, 1998.

[52] J. Humphrey, "A possible role of the pleura in lung mechanics," *J. Biomech.*, vol. 20, no. 8, pp. 773–777, 1987.

[53] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method*, 5th ed. Butterworth-Heinemann, U.K.: London, 2000.

[54] J. Simo and C. Miehe, "Associative coupled thermoplasticity at finite strains: Formulation, numerical analysis and implementation," *Comp. Meth. Appl. Mech. Eng.*, vol. 98, pp. 41–104, 1992.