Assessment of Rigid Registration Quality Measures in Ultrasound-Guided Radiotherapy

Roozbeh Shams, Yiming Xiao, François Hébert, Matthew Abramowitz, Rupert Brooks and Hassan Rivaz*

***Abstract*—Image guidance has become the standard of care for patient positioning in radiotherapy, where image registration is often a critical step to help manage patient motion. However, in practice, verification of registration quality is often adversely affected by difficulty in manual inspection of 3D images and time constraint, thus affecting the therapeutic outcome. Therefore, we proposed to employ both bootstrapping and the supervised learning methods of linear discriminant analysis and random forest to help robustly assess registration quality in ultrasound-guided radiotherapy. We validated both approaches using phantom and real clinical ultrasound images, and showed that both performed well for the task. While learning-based techniques offer better accuracy and shorter evaluation time, bootstrapping requires no prior training and has a higher sensitivity.**

***Index Terms*—Radiotherapy, Image registration, Quality management, Motion management, Bootstrapping, Supervised learning.**

I. Introduction

ACCURATELY targeting the pathological loci during radiotherapy is crucial to ensure the treatment outcomes. However, patient motions limit the precision with which radiation can be applied, resulting in less effective treatment plans. In modern radiotherapy, image guidance is used to align and update the patient's anatomy with the treatment isocenter, proving better target coverage and in some cases reducing dose to surrounding healthy tissues. Such alignment (i.e., patient positioning) can be achieved through widely used image registration algorithms based on a number of techniques, including external surface motion, implanted markers, X-ray imaging, and ultrasound imaging[1], [2], [3]. Compared with X-ray imaging, ultrasound is non-ionizing and provides good soft tissue contrast in real time [4], and thus it has become a popular imaging modality to track patient motions.

Radiotherapy frequently involves the delivery of radiation dose in multiple sessions, known as fractions. Two types of patient motions can occur, including interfraction motion (i.e., on each day of treatment, as the patient is positioned for that day), and intrafraction motion (i.e., short term during radiation delivery). Interfraction positioning affects the entire treatment fraction. Although it must be completed reasonably quickly, more time is available for calculation and review. However, intrafraction positioning, or monitoring, must be completed in near real time to be of use. An operator often has to rapidly verify the positioning quality during the entire duration of the treatment, which is challenging due to time limitations and 3D nature of the images. To help ensure the quality of patient positioning and mitigate the workload of the operator, who may not offer consistent quality assurance, a robust automatic method for assessing image registration quality is needed.

Most of the previous work in quality evaluation can be broadly categorized into Bayesian and supervised learning methods [5]. Typically in the former, a Bayesian framework for the registration problem is proposed and a posterior distribution over the model parameters is calculated. Next, using the posterior, a measure of uncertainty is given. For instance, Risholm et al. [6] proposed a Bayesian non-rigid registration framework using Boltzmann's distribution for the prior and likelihood and Markov Chain Monte Carlo (MCMC) to estimate the most likely deformation and the uncertainty associated with it. Janoos et al. [7] proposed a similar framework which is used for image registration for multi-modal images. In [8], the authors introduce ways to summarize the uncertainty of an elastic registration framework which they proposed in [9]. Simpson et al. [10] try to solve the problem of choosing the regularization coefficient with a Bayesian approach which also can estimate the uncertainty in the form of a covariance matrix. As shown in [11], the uncertainty can also be used to construct a filter to smooth the areas with higher uncertainty.

A supervised classification method was first introduced by Wu and Samant [12] for automatic detection of unsuccessful registrations during radiotherapy. The authors used one feature (e.g. mutual information or cross correlation) as an input for the classifier and the classifier itself, used a threshold calculated based on the training data to classify different registrations. Wu and Murphy [13] then improved their previous work by extracting more features and also using a neural network as the classifier. Muenzing et al. [14] did a comprehensive study of different features and classifiers that could be of use for the task at hand and evaluated their method on lung CT images. Finally, Sokooti et al. [15] constructed a random regression forest to estimate the registration error of chest CT scans and also classify based on the estimated error.

An advantage of learning methods over the Bayesian approach is the lower computational complexity for classifying the registrations at runtime. This advantage makes these methods suitable for real-time applications. However, to train such a classifier, an appropriately sized training set is needed

* indicates corresponding author.

R. Shams, Y. Xiao, R. Brooks and H. Rivaz are with the Department of Electrical and Computer Engineering and the PERFORM Centre, Concordia University, Montreal, QC, Canada.

F. Hébert is with Elekta Ltd, Montreal, QC, Canada.

M. Abramowitz is with the University of Miami Hospital, Miami, FL, USA.

R. Brooks was with Elekta Ltd, Montreal, QC, Canada. He is now with Ossimtech, Montreal QC Canada

Emails: ro_shams@encs.concordia.ca and yiming.xiao@concordia.ca and francois.hebert@elekta.com and rupert@rupertbrooks.ca and hrivaz@ece.concordia.ca

Manuscript received –; revised –.

and acquiring such training set is not always feasible. An unsupervised method may prove to be useful in such cases.

In [16], the author has taken a frequentist approach to measuring uncertainty using bootstrapping. It is assumed that the input images are realizations of random processes. Given several realizations of the input images, the registration method could be run on these images and the uncertainty can be calculated based on the results of these registrations. Since only one realization of the random variable, the image at hand, is available, bootstrapping is used to simulate different realizations. This method does not require any training which makes it an attractive candidate for registration quality evaluation that can be readily applied to different ultrasound systems and even other modalities. We will therefore propose a technique for assessing the quality of ultrasound registration using bootstrapping, and validate it using phantom and *in-vivo* data.

In this paper, we propose to use bootstrapping and supervised learning methods for assessing the quality of rigid ultrasound image registration in the context of ultrasound-guided radiotherapy. More specifically for supervised learning methods, we employed Linear Discriminant Analysis (LDA) [17] and Random Forest (RF) [18] to classify the registration quality. All methods were compared using both phantom and *in-vivo* data for intrafractional prostate motion management. In this work, we have made three major novel contributions to the field. First, to the best of our knowledge, this is the first work that introduces automatic registration assessment techniques for ultrasound-guided radiotherapy, and more generally for registration of ultrasound images. Second, in the context of machine learning techniques, we introduced new features due to the unique characteristics of ultrasound images. Lastly, we compared the performance of bootstrap and machine learning techniques for the application, which has not been reported previously. Given that ultrasound has numerous applications in image-guided applications, this work can be further extended and utilized in several other applications. The article is organized as follows. In the next section the methodology is explained. In Section III, the results are presented and are discussed in Section IV. The conclusions are provided in the final section.

II. Methods

A. Registration

Assume $f, g : \mathbb{R}^m \rightarrow \mathbb{R}^n$ to be the fixed and moving images. Also, let $\Omega \in \mathbb{R}^m$ be a set of points from the domain of f. We aim to find a transformation, $T(x, \theta) : \mathbb{R}^m \rightarrow \mathbb{R}^m$, with $\theta \in \Theta \subseteq \mathbb{R}^d$, such that $f(x)$ corresponds to $g(T(x, \theta))$. To calculate $\hat{\theta}$, the transform parameters, a cost function, $J(\theta)$, is constructed and $\hat{\theta}$ is estimated by:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, J(\theta) \tag{1}$$

$$J(\theta) = D(f(\mathrm{x}), g(T(\mathrm{x}, \theta))), \tag{2}$$

where $D(.)$ is the dissimilarity function.

Both f and g can be considered outcomes of random processes and therefore $J(\theta)$ is a random process and $\hat{\theta}$ is a random variable.

In order to evaluate the registration results, it is necessary to measure how close these results are to the true value. A popular approach is to use mean Target Registration Error (mTRE) [19], [20], [21]. Since a rigid transformation model is used in our work, mTRE is calculated on 4 or 6 points for 2D and 3D data respectively. We define the distance between two transforms, T_1 and T_2, as follows. Let $\{P_i\}$ be a set of N points in the fixed image near the center of the transformation, C, and the center itself. The points are selected by moving r millimeters away from C in each cardinal direction; therefore 4 points in the 2D images and 6 in 3D volumes are chosen. The distance is then defined as:

$$d(T_1, T_2) = \frac{1}{N} \sum_{i=1}^{N} ||T(P_i, \theta_1), T(P_i, \theta_2)|| \tag{3}$$

In other words, the distance between two transformations is the mean distance of the corresponding transformed points. Calculating the distance as explained, instead of doing so on a grid, reduces the computational complexity of the evaluation while keeping the evaluation valid because of the rigidness of the transform. Also by using the 4 or 6 point distance measurement method, the comparison between ROIs with different sizes will be equivalent.

Before we present the supervised learning and bootstrapping techniques, it is important to clearly state what is called a "successful" registration or "poor" registration. During registration, the optimizer either converges to an optimum or not. If it diverges, the result is a poor registration. If it converges, but converges to a local optima which is far from the true parameters, the result is again a poor registration. A successful registration is one that the optimizer converges to the correct optimum.

B. Data Preparation

To validate the registration assessment methods, a great number of both cases of poor and successful registrations were needed. Both phantom and *in vivo* patient data were utilized for validation. The following procedure was used to obtain poor and successful registrations. First, a reference registration was carried out to be used as the true registration. For the phantom data, this registration was known *a priori* with a robotic system. For the 3D patient data, as each session represented a tracking sequence, each sequence was made into a video showing anterior-posterior (A-P) and superior-inferior (S-I) cuts through the center of the original prostate position. These videos were visually inspected by experts experienced in prostate radiotherapy to ensure that the reference registration was of high quality. To further evaluate the automatic registration quality for the ground truths, we selected 2 image pairs from each of the 7 treatment sessions for the patient, and asked an expert to manually align the image pairs based on visual inspection. In addition, for each image pair, 10 pairs of homologous anatomical landmarks were selected, and the mean target registration error (mTRE) was obtained with

these landmarks for both manual and automatic registrations. The mTREs (mean±sd) from 14 image pairs for manual and automatic registrations are 1.91±0.83 mm and 1.86±1.09 mm, respectively. The difference in the results obtained from the two approaches is not statistically significant based on a Wilcoxon signed-rank test ($p = 0.358$).

Next, the true transform parameters were moved in the parameters space in a random direction and the registration was restarted from that point. If the result of the new registration was within a determined distance defined by Eq. 3 (i.e., the smallest resolution of the images), the registration result was regarded as "good", and the initial parameters are moved further away from the true registration result. This was repeated until the new registration either diverges or converges to another point far from the true result, hence generating a bad registration. Instead of changing the registration parameters with equal step sizes along a direction in the registration parameter space, the parameter steps for each new starting point was defined as an increase of 2 mm by Eq. 3. This way, the interpretation was more intuitive as the metric was the same as the measurement of the image resolution. Furthermore, these incremented registrations were selectively inspected by clinical experts who are experienced in prostate radiotherapy. As such, a set of successful and poor registrations were generated from an initial limited set of inspected, good registrations. This procedure is depicted in Fig. 1, and instances of good and failed registration results for the patient data are demonstrated in Fig. 2. Here, the moving image was moved in the parameter space until a failed registration occur while the fixed image was kept the same. The successful registration visibly improved the alignment of the walls of the bladder and prostate.

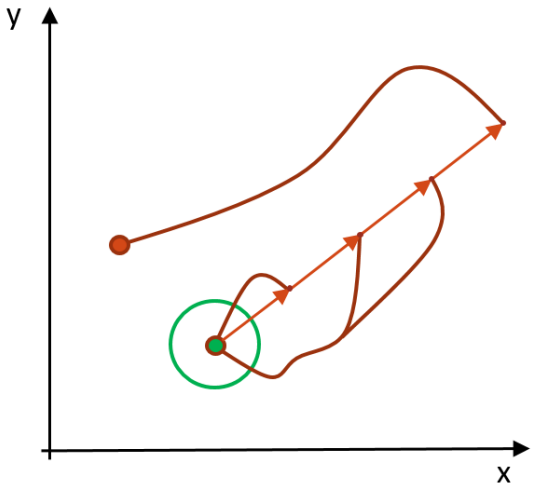


Fig. 1: Generating poor and successful registrations. The green dot shows parameters of the correct registration. Each arrow shows the start of a new registration process. In this schematic example, three registrations converge to the correct result, and one converges to an incorrect result (red dot). The green circle shows the area wherein the registration is still considered successful.

C. Supervised Learning Methods

There are numerous classifiers available in the literature; we chose two for our experiments: LDA [17], a simple classifier, and RF [18] as a state-of-the-art classifier. As with other supervised learning methods, this requires feature extraction, training and validation.

1) Feature Extraction: There may be a trade-off between calculation time and discriminative value of a feature. The ideal feature would cost no additional calculation. We selected a subset of 10 features from a pool of features for training and classification. This selection was done based on feature importances (Gini importance [18]) resulting from an RF classifier using all the features.

The registration is implemented by optimizing the negative Normalized Cross-Correlation(NCC) between a selected set of pixels in the reference and target images. The resulting optimal NCC can be used as a criterion for distinguishing between successful and poor registrations. This measure costs no additional computation, as we are already computing it.

Let $f_i = f(\mathrm{x}_i)$ and $g_i = g(T(\mathrm{x}_i, \hat{\theta})$ represent the fixed and moving image intensities where $\{\mathrm{x}_i\}$ is the set of points used to calculate the NCC and $T(x, \theta) : \mathbb{R}^m \rightarrow \mathbb{R}^m$ is the transformation. N is the number of pixels (i.e. the number of points in $\{\mathrm{x}_i\}$. The NCC can be calculated in a single pass over the image using:

$$NCC = \frac{S_{fm} - S_f \cdot S_g/N}{C_f \cdot C_g} \tag{4}$$

where

$$S_f = \sum_i f_i \qquad S_g = \sum_i g_i$$

$$S_{ff} = \sum_i f_i f_i \qquad S_{fg} = \sum_i f_i g_i$$

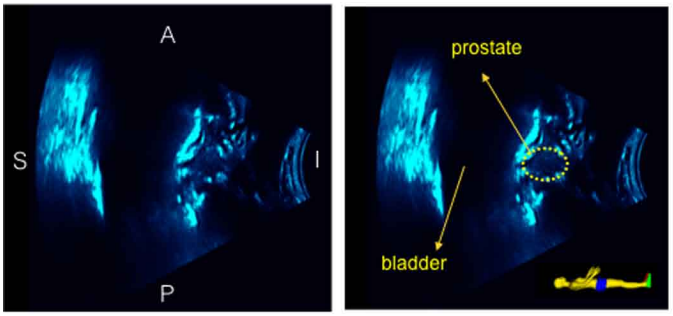
$$S_{gg} = \sum_i g_i g_i$$

These sums can be accumulated during a loop over the pixels. Note that S_{ff} , S_f and N are not necessarily constant, as some of the reference pixels may map outside of the moving image and will therefore be excluded from the calculation. From these, we can compute the contrast of each image, C_x, using:

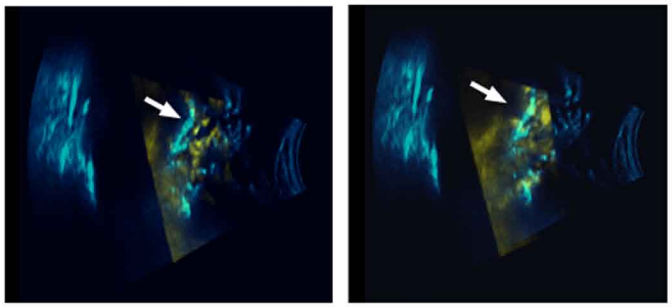
$$C_x = \sqrt{S_{xx} - S_x \cdot S_x/N} \tag{5}$$

and the NCC using Equation 4. An advantage of calculating the NCC this way is that each part can be used as a feature. It was conceivable that one or more of these measures were more distinctive than correlation alone. There is no additional cost to these measures as they are already computed.

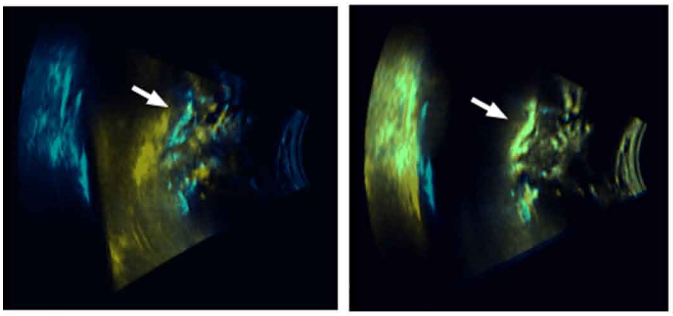
The Distinctiveness of Optimum (DO) [22] was used together with Mirror Symmetry (MS) in [13]. It is an average descriptor of the shape of the dissimilarity function around the found solution. It requires $2U$ evaluations of the dissimilarity measure over the registration cost function J with respect to each registration parameter in the positive and negative directions. Here, U is the number of registration parameters,



(a) Fixed image and the annotation



(b) An example of a bad registration



(c) An example of a good registration

Fig. 2: Demonstration of good and bad registration results for the patient US data. The fixed image (cyan) and moving images (yellow) are overlaid to show the quality of registration. (a) shows the fixed image along with the anatomical annotation and the orientation of the image with respect to the patient. (b) shows a failed registration (left: before registration; right: after registration). (c) shows a case of successful registration (left: before registration; right: after registration). Here, the white arrows point to the wall of the bladder.

and $U = 6$ for rigid registration. Therefore, the DO metric is defined as:

$$DO(s) = \frac{1}{2sU}\sum_{u}\Big(\frac{J(\hat{\theta} - se_u) + J(\hat{\theta} + se_u)}{2} - J(\hat{\theta})\Big) \quad (6)$$

where s is the step size, $\hat{\theta}$ are the optimal parameters, J is the cost function to be optimized for registration and e_u is a unit parameter vector in direction u.

The Mirror Symmetry (MS) [13], [23] is a measure of the evenness of the shape of the similarity function around the found solution. Letting

$$\bar{J}_u = \frac{J(\hat{\theta} - se_u) + J(\hat{\theta} + se_u)}{2}, \quad (7)$$

MS can be calculated as:

$$MS(s) = \frac{1}{2P}\sum_{u}\Big(\frac{(J(\hat{\theta} - se_u) - \bar{J}_u)^2 + (J(\hat{\theta} + se_u) - \bar{J}_u)^2}{(J(\hat{\theta}) - \bar{J}_u)^2}\Big). \quad (8)$$

It can be generated from the same samples as the distinctiveness of optimum.

An indication of a good registration is that the correlation score at one step size away in any direction from the found location is significantly worse. Therefore, we also include individual cost evaluations as features. For the convenience of annotation in the later sections, we name these evaluations as

$$\{DissimProbes[2k], DissimProbes[2k+1]\}$$

where

$$DissimProbes[2k] = J(\hat{\theta} + se_u)$$

$$DissimProbes[2k+1] = J(\hat{\theta} - se_u).$$

and

$$k = 0, 1, ..., U-1.$$

With $k = \{0, 1, 2, 3, 4, 5\}$ corresponding to the probing of J in the direction of each of the transformation parameters (3 rotations and 3 translations), we obtained the evaluations as $DissimProbes[0]$ to $DissimProbes[12]$. Here, $DissimProbes[.]$ is short for "Dissimilarity Probes".

In a successful registration, it is expected for all the pixels in the ROI to be registered equally well. To quantify this, the ROI is divided into orthants and the correlation score is calculated for each. In a poor registration, the correlation score varies between the orthants. Therefore, several measures of quality can be considered regarding this. The individual orthant scores (OrthantScores), the maximum and minimum score (MaxOrthantScores and MinOrthantScores) and finally, the score difference between the maximum and minimum (OrthantScoreRange).

Instead of treating the two intensity distributions as identical, we can instead examine the joint distribution of intensity. The Mutual Information (MI) of this joint intensity distribution is a commonly used similarity measure. We did not use it to construct the cost function because our work is focused on mono-modal registration. In addition, MI costs more to compute than the above, as it involves keeping track of a joint distribution.

If the images are correctly aligned, it is reasonable to presume that the corresponding set of pixels in the fixed and the moving images have similar intensity distribution. The Kullback-Leibler divergence [24] can be used to quantify the difference between the two distribution functions and therefore is used as a feature.

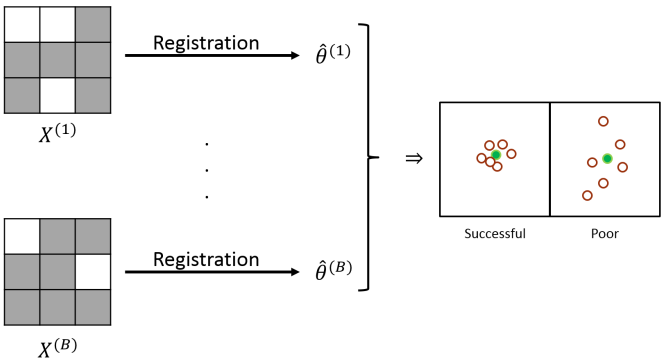


Fig. 3: An overview of bootstrapping for registration evaluation. $X^{(b)}$ shows different multisets, and $\hat{\theta}^{(b)}$ denotes the results of the registration using each multiset. The grayed out pixels on the left show selected pixels for registration (possibly more than once). The green dots show the correct registration parameters, and the red dots represent registration results (i.e. $\hat{\theta}^{(b)}$). In a poor registration, $\hat{\theta}^{(b)}$ are expected to be more dispersed than a successful case.

2) Training and validation: As mentioned before, we used LDA and RF to classify the registrations. Here, half of the total data were used as a testing set, and the other half was used to train the classifiers through a 4-fold cross-validation process (training set vs. validation set ratio = 3:1). The machine learning algorithms and validations were implemented in scikit-learn package, version 0.17 [25].

D. Bootstrapping

Bootstrap resampling is a technique that can be used to estimate the properties of an estimator, such as mean, variance, etc. [26]. Assume a random variable X with N i.i.d samples $X = \{x_1, \ldots, x_n\}$ drawn from it. A bootstrap resample, $X^{(b)}$, is a multiset constructed by selecting N points from X with replacement. This is repeated B times, thus leading to B multisets: $X^{(b)}, b = 1, \ldots, N$.

Assume a statistic on X, ϑ, and its estimator $\hat{\vartheta} \approx \varphi(X)$. Our goal is to measure the reliability of this estimator. This can be done by finding the estimates of ϑ based on each bootstrap: $\hat{\vartheta}^{(b)} = \varphi(X^{(b)})$. These bootstrap values can be used to form a non parametric distribution on the estimates which can be used to express a measure of reliability, such as the covariance matrix.

E. Bootstrapping for registration evaluation

Image registration can be thought of as an estimator of the transformation parameters, $\boldsymbol{\theta}$. Therefore we can use bootstrapping to measure the reliability of this estimator similar to what was explained in the previous section [16]. In our case, we use the result of bootstrapping to classify the registration as reliable and unreliable.

To this end, it is needed to solve B registration problems, based on each bootstrap:

$$J^{(b)}(\boldsymbol{\theta}) = D(f(\mathrm{x}), g(T(\mathrm{x}, \theta))), \mathrm{x} \in \Omega^{(b)} \tag{9}$$

$$\hat{\theta}^{(b)} = \underset{\theta}{\operatorname{argmin}}\, J^{(b)}(\theta) \tag{10}$$

where $\Omega^{(b)}$ is a bootstrap resample. From here, the measure of dispersion on $\hat{\theta}^{(b)}$ can be calculated.

$$\bar{d}_B = \frac{1}{B} \sum_{b=1}^{B} d(T(\mathrm{x}, \theta^{(b)}), T(\mathrm{x}, \bar{\theta}_B)) \tag{11}$$

where $\bar{\theta}_B$ is the mean of parameters resulting from bootstraps: $\operatorname{mean} \theta^{(b)}$. In order to exclude outliers from the calculations, the trimmed mean [16] is used: the furthest 10% of the results from the mean bootstrap result is taken out and the mean is recalculated accordingly. If this dispersion measure is higher than a threshold τ, then the registration is poor and if not, successful. In other words, if we are not able to estimate the registration parameters with sufficient confidence through the sampling process, then a single registration is likely to provide a bad image alignment that is far off the optimum. Note that for each bootstrap sampling, only a portion of the pixels/voxels were randomly selected for registration. To facilitate easier interpretation of the dispersion measurement, instead of measuring the metric in the registration parameter space, we employed Eq. 3 to evaluate the transform distance. This way, the mean transform distance is in the same spacing unit as the images or volumes, the threshold can be set based on the resolution of the data and according to what accuracy is needed.

Figure 3 shows a general overview of the bootstrapping scheme for classifying registrations and Algorithm 1 describes an in depth implementation.

Algorithm 1 Bootstrap resampling for image registration quality evaluation

for b=1 to B **do**
  $S \leftarrow$ empty multiset
  for i=1 to N **do**
    $\Omega^{(b)} \leftarrow \Omega^{(b)} \cup \{\mathrm{x_k}\}$; $k \sim_{i.i.d} \{1, \ldots, N\}$
  end for
  $J^{(b)}(\boldsymbol{\theta})$=$D(f(\mathrm{x}), g'(\mathrm{x})), \mathrm{x} \in \Omega^{(b)}$
  $\hat{\theta}^{(b)} \leftarrow \underset{\theta}{\operatorname{argmin}}\, J^{(b)}(\theta)$
end for
Calculate $\bar{d}_B$ from $\{\hat{\theta}^{(b)}; b = 1 \ldots N\}$
if $\bar{d}_B > \tau$ **then**
  poor registration
else
  successful registration
end if

F. Experimental setup

We compare the two approaches using experiment data and patient data. For the experimental data, 2D images were acquired with a Clarity (Anticosti Research Version, Elekta AB, Stockholm, Sweden) monitoring system with a linear ultrasound probe. The patient data was collected with the Clarity system (Version 3.0) with a wobbler probe, providing

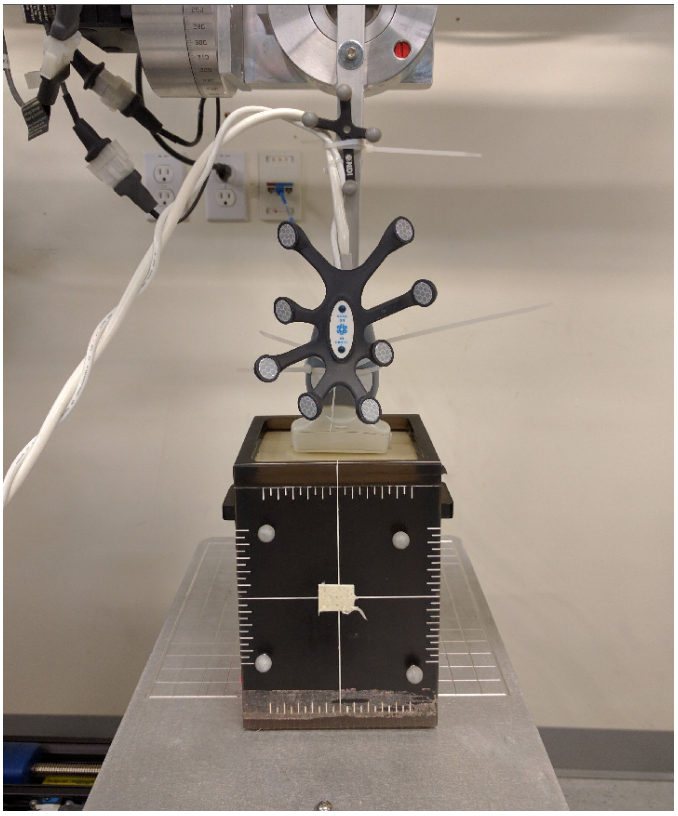


Fig. 4: The experiment setup. The phantom, ultrasound probe, the tracked markers and the robot can be seen.

a sequence of 2D images in a sweeping pattern and thus forming a 3D volume. The 2D phantom data were collected using translational motions by a robotic arm. With better controlled ground truths, this approach is ideal for preliminarily testing the proposed techniques. Then, we further validated the methods with 3D patient data under full rigid body motions in order to reveal their performance for potential real clinical applications.

G. Phantom study

We imaged a Clarity QC phantom (Elekta AB), with the ultrasound probe attached to a Cartesian gantry robot (Velmex, Inc. Bloomfield, NY, USA) to control the probe movements. A laser level was used to set the orientation of the probe so that a) the image plane and the motion plane of the probe would be parallel and b) the probe would be perpendicular to the surface of the phantom. The former is to minimize any movement not on the image plane and the latter, to assert only translation in one dimension of the images. The probe was also tracked with a Polaris Spectra optical tracking system (Northern Digital Inc, Waterloo, ON, Canada); this was used so that the true probe translation would be available (within the precision of the tracker). Moreover, the tracking information was used to ensure correct movement of the probe. Figure 4 shows the experimental setup.

The following procedure was used to acquire the images. First between 15 to 20 frames were captured. The probe was moved then in the lateral direction for 10, 15, or 20 mm. After each translation, another 15 to 20 frames were captured and the two image sets, from before and after the translation, were registered.

This was repeated for 8 different runs. Between runs, the amount of probe movement, the settings of the ultrasound machine, the part of the phantom which was imaged and the medium (gel or water) were changed to produce a wide range of images with different qualities. To have more variety in registrations, image sets from different runs were also registered. These image sets were chosen so that they would be images from the same structure in the phantom, with the same orientation of the probe. The difference between them being the settings of the ultrasound machine. Good and bad registrations were generated with the procedure described in Section II.B. As a result, 1688 sets of registrations were used for supervised learning methods, and 3376 sets for testing bootstrapping. The good vs. bad ratio is about 4:1.

For the supervised learning methods, features were extracted and different classifiers were trained and evaluated. Bootstrapping was also carried out, with 20 bootstrap resamples and $\tau = 0.14$ mm. τ was chosen based on the pixel size, which was 0.14 mm in this experiment.

H. Patient Trials

For the experiment, ultrasound data were collected from one patient acquired during a previously scheduled and planned radiotherapy treatment session for the prostate [27]. The data were acquired using the same scanner mentioned in the previous section, and included 7 separate treatment sessions to help increase the variability among the images. Imaging in each session lasted about 4-10 min. The patient images were acquired in the context of an Institutional Review Board (IRB) approved clinical study, and were not used to make clinical decisions. The patients were undergoing radiation treatment and as such had bladders comfortably full and rectums empty, increasing internal patient anatomy uniformity with radiation planning CT studies. During each session, the patient was positioned supine, legs akimbo, with the probe imaging via the perineum. In this scan position, the prostate can be imaged between the pelvic bones. The probe position was adjusted to obtain a good image of the prostate with, and fixed in place. The patient was instructed not to voluntarily move during the procedure. The probe continuously sweeps the image plane, forming a continuously updated 3D dataset, and a total of 2193 images were acquired from all the sessions. Intrafractional target tracking is performed by registering the current 3D dataset to the first reference dataset and the quality of registration was visually inspected by a clinical expert.Using the same procedure, we generated a set of good and bad registrations, and used them to compare the learning methods against bootstrapping. For bootstrapping, 43328 registrations were used, and for supervised learning, 21664 were used. The ratio between good and bad registrations is about 4:1. Since the resolution of the volumes were different from that of the experiment images, the threshold, τ was set to 0.4 mm, which is the axial voxel resolution of the volumes.

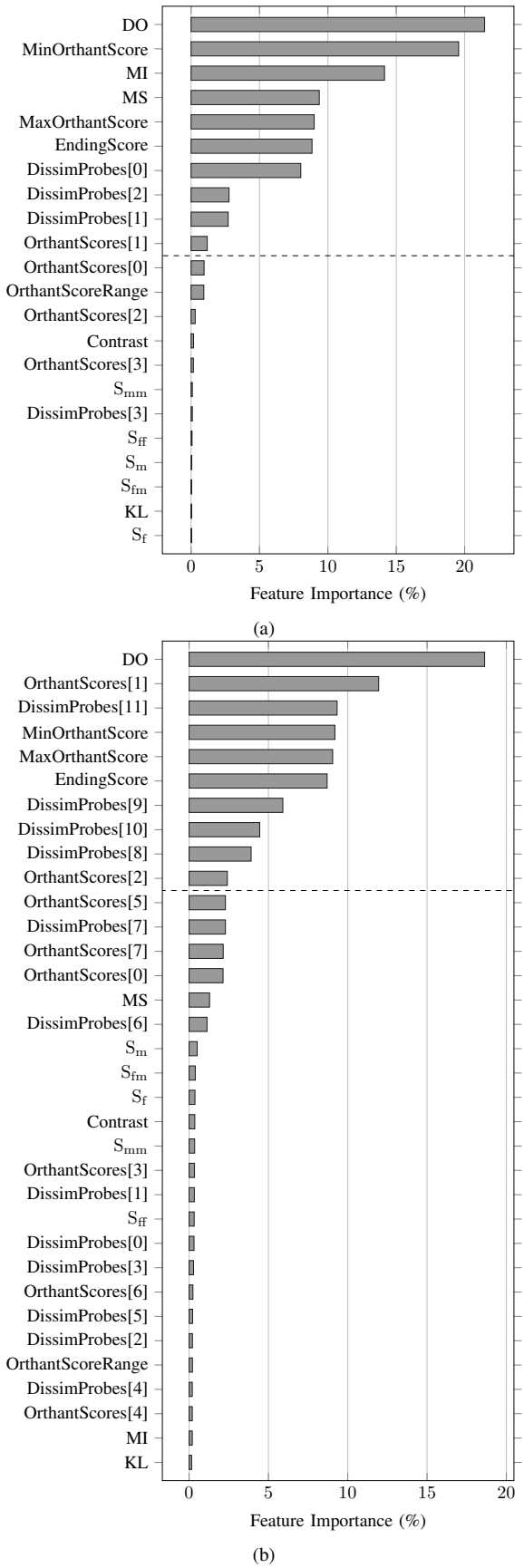


Fig. 5: Feature importance according to RF. (a) shows the feature importance for the 2D experimental data and (b) for that of the patient data. The features above the dashed lines are those chosen for classification.

Registration was performed using the Insight Toolkit (ITK) [28] parametric registration framework by optimizing the Normalized Cross Correlation (NCC) computed over the selected pixels using gradient descent (specifically the regular step gradient descent optimizer in ITK was used) and linear interpolation. In the 3D case, the images are collected on a fan-shaped geometry. To avoid unnecessary interpolation errors the images were directly registered as originally sampled (i.e., in the fan shape) using the methodology described in [29]. For the experiment, 35% of the pixels in the ROI were randomly selected to build the cost function. Also, a translation transformation model was used since the movement was in one direction. For the 3D data, 20,000 voxels were used to construct the cost function. As for the transformation model, one was used to accommodate the non-rectilinear image frames which is the result of using a wobbler probe [29]. The supervised learning methods were implemented in Python using the scikit-learn library [25].

III. Results

A. Feature selection

The feature importances for the 2D and 3D data can be seen in Figures 5(a) and 5(b) respectively, and the features shown in the figures above the dashed lines are selected to be used for the classifiers. Although DO was ranked as the most prominent feature for both cases classifying good and bad registrations, the rest of the features differ.

B. Registration evaluation

The results for registration evaluation using bootstrapping and machine learning methods are shown in Table I for both 2D phantom and 3D patient data. In our experiments, supervised learning methods outperformed the bootstrapping method in terms of accuracy. An advantage of using bootstrapping is that it does not require training data for classifying at the expense of lower classification accuracy and higher computation complexity. Another advantage of the bootstrapping method is the higher sensitivity of 99.92% compared to 96.15% and 96.95% for LDA and RF respectively for the patient data. This makes bootstrapping a reliable method for ensuring the registration is yielding correct results. To further demonstrate the performance of the techniques, their receiver operating characteristics (ROC) curves for assessing the registrations of patient data are shown in Fig. 6.

IV. Discussions

For supervised learning techniques, we have explored existing features and explored new ones for classifying good and bad registrations. For the 2D and 3D data, the selected features differ greatly, and this is likely a result of the differences in image dimensions, imaging contents, as well as the degrees of freedoms in registration. From Table I, we have observed a superior registration assessment quality when using machine learning approaches than bootstrapping. Besides the inherent power of machine learning techniques, the phenomena may

	2D			3D		
	BOOT	LDA	RF	BOOT	LDA	RF
ACC	86.54%	99.35%	99.76%	87.45%	96.97%	97.73%
MCC	66.58%	97.63%	99.14%	75.14%	92.42%	94.29%
TP	16.19%	15.94%	16.47%	26.94%	25.92%	26.14%
TN	70.35%	83.41%	83.29%	60.51%	71.05%	71.59%
FP	12.80%	0.06%	0.18%	12.53%	1.99%	1.45%
FN	0.65%	0.59%	0.06%	0.02%	1.04%	0.82%
N	3376	1688	1688	43328	21664	21664
Sensitivity	96.12%	96.42%	99.64%	99.92%	96.15%	96.95%
Specificity	84.60%	99.93%	99.79%	82.84%	97.28%	98.02%

TABLE I: ACC and MCC are accuracy and Mathews Correlation Coefficient (MCC) [30]; TP,TN,FP and FN are percentages of true positive, true negative, false positive and false negative relative to the total number of samples, N. Sensitivity (true positive rate) and Specificity (true negative rate) of the classifiers are also included

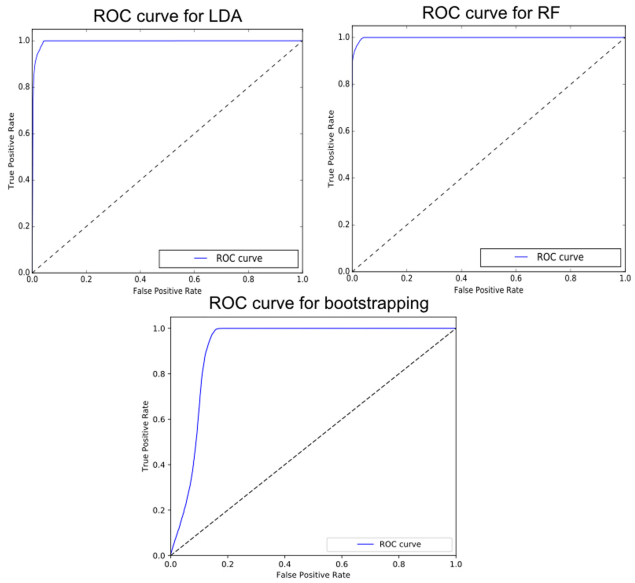


Fig. 6: ROC curves of different registration assessment methods for the 3D patient data.

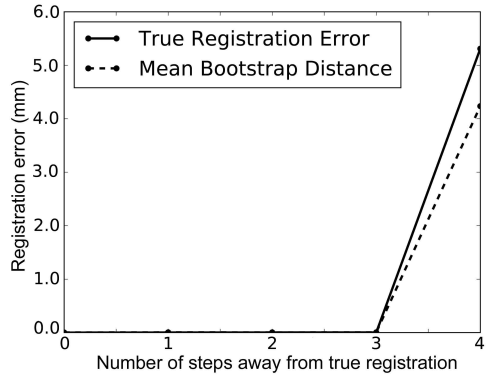
be partially explained by the employed data that lacked high variability. For the 2D data, the images were collected from the same phantom. Although for the 3D data, the images were obtained from seven different treatment sessions, the variabilities due to diseases and individual anatomy of the data are relatively low. Thus, we postulate that this limitation may cause the evaluation results to be slightly better. However, this does not mean that the classifiers were over-fitted in the scope of the data. We will examine our methods based on multiple patients' data in the future. In order to translate the proposed machine learning methods to clinical applications, data will be gathered from more patient cases and also from human volunteers, and the classifiers will be retrained to improve their generalizability. In the case of volunteers, a wider range of imaging settings and patient motions can be explored, as there is no risk of affecting the patient's treatment. Therefore, it is not required to retrain the system for every new patient.

Aside from the distances from the ground truth image alignment, there can be other factors, such as image noise, that can influence the registration quality. In our experiments, we have attempted to incorporate variability in image quality in both phantom and patient data through varying image settings and obtaining images from different treatment sessions, thus incorporating the factors in the experiments implicitly. Explicitly analyzing the effect of individual factors, which require meticulous control during data acquisition and are very difficult to isolate for patient data, is out of the scope of this work. However, the potential of the proposed techniques has been demonstrated with real clinical data.

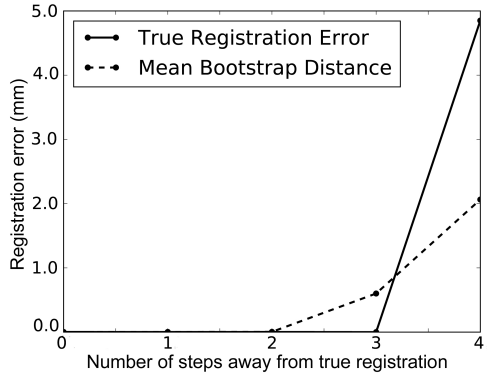
Both approaches mentioned in the paper have advantages and disadvantages and therefore both can be viable choices depending on the application. The supervised learning approach has higher classification accuracy and is faster. More specifically, for each registration to be assessed, the bootstrapping takes around 18 seconds while the machine learning methods take less than 1 second. Note that the algorithms were implemented on a Window7 desktop computer with a 16GB RAM and an Intel core i7-4770@3.40GHz processor. The bootstrap method does not require training data, is less accurate, and has a higher computational cost. As a result, supervised learning methods are a better fit for intrafraction motion management, where speed and accuracy are critical. Nonetheless, bootstrapping can still be considered for this application. Since the calculation of each bootstrap result is independent of the other, it is possible to run the registrations in parallel and reduce the runtime. Moreover, bootstrapping is also applicable in cases where the timing requirements for registration validation are not as strict as the target tracking itself, whereby it can be calculated independently from the registration. Bootstrapping is a more natural fit for interfraction registration wherein the algorithm does not need to run in real time. Also, as the patient is positioned, if a reliable registration can not be performed, the registration setup could be modified until the registration can be reliably carried out. Due to the variation of the images from each day, acquiring the interfraction training data for supervised learning methods that generalize well would be a challenging task, which makes bootstrapping a better choice for this application.

A contributing factor in bootstrapping having lower accuracy is the relatively higher FP rate. Upon further investigation, we realized these FPs occur at steps close to where the registration fails when the initial parameter is far from the true registration. However, due to the randomness involved (the pixel selection), the registration result is successful. An example of this is given in Figure 7. In Figure 7(a), a case of TP can be seen. The initial registration parameter is far away from the true registration parameters (at step 4) and therefore the result is a poor registration. By inspecting at the mean bootstrap distance, the same can be deduced. In Figure 7(b), the registration is poor at step 4, however, judging by the bootstrap result, it seems that the registration was poor from step 3 onwards. Although the registration at step 3 was successful, it was not reliable and had a good chance of failing. Using bootstrapping enables us to detect these cases. From a

technical perspective it adds to the number of false positives, but from a practical perspective it ensures that the registration result is reliable.



(a)



(b)

Fig. 7: Figures showing the registration error as the initial parameters are moved further away from the true result. The x-axis shows the number of steps away from the true registration parameters. (a) and (b) were generated from different bootstrap sampling of the same image pair.

It should be acknowledged that bootstrapping cannot detect the bias as briefly mentioned in [16]. For instance if the minimum of the cost function is at a distance from the true parameters, caused by the interpolation for example, bootstrapping would fail to detect the bias. In this work, we have intentionally neglected the bias. As shown in [16] bias of image registration algorithms is often quite small, and in this work, we adopted the same assumption as in [16].

We focused on rigid registration to follow the intrafraction motion of the prostate for two main reasons. First, in most current radiotherapy workflows in the clinic, only 3 dimensional (translational), or in certain cases, 6 dimensional (3 translations & 3 rotations) patient positioning is possible. Adaptation of the dose delivery plan to account for deformed anatomy remains an open research problem. Second, for the *in vivo* data collected, no rectal probe or balloon is used. Under these conditions, while some nonlinear deformation may occur, the majority of the motion can be described by a combination of rotations and translations [31]. Deformations of the prostate during treatment are considered relatively small with respect to the margins in use. To help select features for supervised learning methods, we employed RF, which has been employed previously in image feature selection in medical image analysis [32], [33] and offered satisfactory results. While there are also many other techniques for feature selection, a comprehensive comparison is out of the scope of this article and will be studied in a future work.

V. Conclusion

In this work, we proposed to use bootstrapping and supervised learning methods (i.e., LDA and RF) to assess ultrasound registration quality. By using both phantom and real clinical data, the two categories of methods were evaluated and compared against each other. While both bootstrapping and supervised learning methods demonstrate good performance, the latter showed better accuracy. In addition, we explored existing features and devised new features that are essential given the unique characteristics of ultrasound images to robustly evaluate the registration quality using machine learning methods. To the best of our knowledge, it is the first time that automatic registration assessment techniques are proposed for ultrasound imaging, which is widely used in image-guided procedures.

Acknowledgment

This work was funded by Natural Science and Engineering Research Council of Canada (NSERC) Engage grant, NSERC Discovery Grant RGPIN-2015-04136, and a grant from the Richard and Edith Strauss Canada Foundation. R. Brooks and H. Hébert were with Elekta Ltd. when this research was carried out. The authors would like to thank anonymous reviewers for constructive comments, and Dr. Martin Lachaine for helpful discussions.

References

[1] T. O'Shea, J. Bamber, D. Fontanarosa, S. van der Meer, F. Verhaegen, and E. Harris, "Review of ultrasound image guidance in external beam radiotherapy part ii: intra-fraction motion management and novel applications," *Physics in medicine and biology*, vol. 61, no. 8, p. R90, 2016.

[2] D. A. Jaffray, "Image-guided radiotherapy: from current concept to future perspectives," *Nature Reviews Clinical Oncology*, vol. 9, no. 12, pp. 688–699, 2012.

[3] J. De Los Santos, R. Popple, N. Agazaryan, J. E. Bayouth, J.-P. Bissonnette, M. K. Bucci, S. Dieterich, L. Dong, K. M. Forster, D. Indelicato *et al.*, "Image guided radiation therapy (IGRT) technologies for radiation therapy localization and delivery," *Int J Radiat Oncol Biol Phys*, vol. 87, no. 1, pp. 33–45, 2013.

[4] M. Lachaine and T. Falco, "Intrafractional prostate motion management with the clarity autoscan system," *Med. Phys. Int*, vol. 1, no. 1, pp. 72–80, 2013.

[5] R. Datteri, Y. Liu, P.-F. D'Haese, and B. Dawant, "Validation of a Non-Rigid Registration Error Detection Algorithm using Clinical MRI Brain Data." *IEEE transactions on medical imaging*, vol. 34, no. 1, pp. 1–11, 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25095252

[6] P. Risholm, F. Janoos, I. Norton, A. J. Golby, and W. M. Wells, "Bayesian characterization of uncertainty in intra-subject non-rigid registration," *Medical Image Analysis*, vol. 17, no. 5, pp. 538–555, 2013.

[7] F. Janoos, P. Risholm, and W. Wells, "Bayesian characterization of uncertainty in multi-modal image registration," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7359 LNCS, no. Iii, pp. 50–59, 2012.

[8] P. Risholm, S. Pieper, E. Samset, and W. M. Wells, "Summarizing and visualizing uncertainty in non-rigid registration," *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, vol. 6362 LNCS, no. PART 2, pp. 554–561, 2010.

[9] P. Risholm, E. Samset, and W. Wells, "Bayesian estimation of deformation and elastic parameters in non-rigid registration," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6204 LNCS, pp. 104–115, 2010.

[10] I. J. A. Simpson, J. A. Schnabel, A. R. Groves, J. L. R. Andersson, and M. W. Woolrich, "Probabilistic inference of regularisation in non-rigid registration," *NeuroImage*, vol. 59, no. 3, pp. 2438–2451, 2012.

[11] I. J. Simpson, M. Woolrich, A. R. Groves, and J. A. Schnabel, "Longitudinal brain MRI analysis with uncertain registration," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*. Springer, 2011, pp. 647–654.

[12] J. Wu and S. S. Samant, "Novel image registration quality evaluator (RQE) with an implementation for automated patient positioning in cranial radiation therapy," *Medical Physics*, vol. 34, no. 6, pp. 2099–2112, 2007.

[13] J. Wu and M. J. Murphy, "A neural network based 3D/3D image registration quality evaluator for the head-and-neck patient setup in the absence of a ground truth," *Medical Physics*, vol. 37, no. 11, pp. 5756–5764, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21158287

[14] S. E. A. Muenzing, B. van Ginneken, K. Murphy, and J. P. W. Pluim, "Supervised quality assessment of medical image registration: Application to intra-patient CT lung registration," *Medical Image Analysis*, vol. 16, no. 8, pp. 1521–1531, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.media.2012.06.010

[15] H. Sokooti, G. Saygili, B. Glocker, B. P. Lelieveldt, and M. Staring, "Accuracy estimation for medical image registration using regression forests," pp. 107–115, 2016.

[16] J. Kybic, "Bootstrap resampling for image registration uncertainty estimation without ground truth," *Image Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 64–73, 2010.

[17] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] P. Jannin, J. M. Fitzpatrick, D. J. Hawkes, X. Pennec, R. Shahidi, and M. W. Vannier, "Validation of medical image processing in image-guided therapy," 2002.

[20] E. B. van de Kraats, G. P. Penney, D. Tomazevic, T. Van Walsum, and W. J. Niessen, "Standardized evaluation methodology for 2-d-3-d registration," *IEEE Transactions on Medical Imaging*, vol. 24, no. 9, pp. 1177–1189, 2005.

[21] H. Rivaz, S. J.-S. Chen, and D. L. Collins, "Automatic deformable mr-ultrasound registration for image-guided neurosurgery," *IEEE transactions on medical imaging*, vol. 34, no. 2, pp. 366–380, 2015.

[22] D. Skerl, B. Likar, and F. Pernus, "A protocol for evaluation of similarity measures for rigid registration," *IEEE Transactions on Medical Imaging*, vol. 25, no. 6, pp. 779–791, 2006.

[23] H. Zabrodsky, S. Peleg, and D. Avnir, "A measure of symmetry based on shape similarity," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1992, pp. 703–706.

[24] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] M. R. Chernick, *Bootstrap methods: A guide for practitioners and researchers*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, 2011, vol. 619.

[27] M. Abramowitz, E. Bossart, L. Martin, R. Brooks, F. Lathuiliere, F. Laura, A. Iskanian, and A. Pollack, "Noninvasive real-time prostate tracking using a transperineal ultrasound: A clinical trial comparison to rf transponders with visual confirmation," *International Journal of Radiation Oncology Biology Physics*, vol. 87, no. 2, p. S682, 2013.

[28] T. S. Yoo, M. J. Ackerman, W. E. Lorensen, W. Schroeder, V. Chalana, S. Aylward, D. Metaxas, and R. Whitaker, "Engineering and algorithm design for an image processing api: a technical report on itk-the insight toolkit," *Studies in health technology and informatics*, pp. 586–592, 2002.

[29] R. Brooks, "Intrafraction prostate motion correction using a non-rectilinear image frame," in *International Workshop on Prostate Cancer Imaging*. Springer, 2011, pp. 57–59.

[30] P. Baldi, S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[31] M. J. Ghilezan, D. A. Jaffray, J. H. Siewerdsen, M. Van Herk, A. Shetty, M. B. Sharpe, S. Zafar Jafri, F. A. Vicini, R. C. Matter, D. S. Brabbins, and A. A. Martinez, "Prostate gland motion assessed with cine-magnetic resonance imaging (cine-MRI)," *Int J Radiat Oncol Biol Phys*, vol. 62, no. 2, pp. 406–17, 2005. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/15890582

[32] F. Aalamifar, H. Rivaz, J. J. Cerrolaza, J. Jago, N. Safdar, E. M. Boctor, and M. Linguraru, "Classification of kidney and liver tissue using ultrasound backscatter data," *Medical Imaging 2015: Ultrasonic Imaging and Tomography*, vol. 9419, 2015.

[33] N. Uniyal, H. Eskandari, P. Abolmaesumi, S. Sojoudi, P. Gordon, L. Warren, R. N. Rohling, S. E. Salcudean, and M. Moradi, "Ultrasound rf time series for classification of breast lesions," *IEEE Trans Med Imaging*, vol. 34, no. 2, pp. 652–61, 2015. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25350925