The feasibility of manual parameter tuning for deformable breast MR image registration from a multi-objective optimization perspective

Kleopatra Pirpinia¹, Peter A.N. Bosman², Claudette E. Loo³, Gonneke Winter-Warnars³, Natasja N.Y. Janssen¹, Astrid N. Scholten¹, Jan-Jakob Sonke¹, Marcel van Herk⁴, Tanja Alderliesten⁵

¹ Department of Radiation Oncology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
² Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
³ Department of Radiology, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
⁴ University of Manchester, Institute of Cancer Sciences, Christie NHS Foundation Trust, Wilmslow Road, M20 4BX Manchester, UK
⁵ Department of Radiation Oncology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands
E-mail: k.pirpinia@nki.nl

Abstract. Deformable image registration is typically formulated as an optimization problem involving a linearly weighted combination of terms that correspond to objectives of interest (e.g., similarity, deformation magnitude). The weights, along with multiple other parameters, need to be manually tuned for each application, a task currently mainly addressed via trial-and-error approaches. Such approaches can only be successful if there is a sensible interplay between parameters, objectives, and desired registration outcome. This, however, is not well established. To study this interplay, we use multi-objective optimization, where multiple solutions exist that represent the optimal trade-offs between the objectives, forming a so-called Pareto front. Here, we focus on weight tuning. To study the space a user has to navigate during manual weight tuning, we randomly sample multiple linear combinations. To understand how these combinations relate to desirability of registration outcome, we associate with each outcome a mean target registration error (TRE) based on expert-defined anatomical landmarks. Further, we employ a multi-objective evolutionary algorithm that optimizes the weight combinations, yielding a Pareto front of solutions, which can be directly navigated by the user. To study how the complexity of manual weight tuning changes depending on the registration problem, we consider an easy problem, prone-to-prone breast MR image registration, and a hard problem, prone-to-supine breast MR image registration. Lastly, we investigate how guidance information as an additional objective influences the prone-to-supine registration outcome. Results show that the interplay between weights, objectives, and registration outcome makes manual weight tuning feasible for the prone-to-prone problem, but very challenging for the harder prone-to-supine problem. Here, patient-specific, multi-objective weight optimization is needed, obtaining a mean TRE of 13.6 mm
A multi-objective perspective on manual parameter tuning for DIR of breast MRI

without guidance information reduced to 7.3 mm with guidance information, but also
providing a Pareto front that exhibits an intuitively sensible interplay between weights,
objectives, and registration outcome, allowing outcome selection.

Keywords: parameter tuning, weight tuning, deformable image registration, multi-
objective optimization, MRI, breast

1. Introduction

Deformable image registration (DIR) has the potential to be a key component of
various processes in medical image analysis. It has particular potential in the field
of radiotherapy, as it can be used for dose accumulation (Schaly et al., 2004),
response assessment (Boehler et al., 2010), or treatment adaptation (Mencarelli
et al., 2014) (Foskey et al., 2005). Its use in clinical practice, however, remains limited.

For the majority of currently used registration methods, the first step when
performing a registration is to set the values of a multitude of parameters (e.g., weights
in cost function, number of image resolution levels, number of optimizer iterations),
which are often tuned by hand. This can be a laborious process, as good settings
can vary greatly depending on the type of registration problem, or even the specific
problem instance at hand. For efficient iterative trial-and-error parameter adjustment
to be feasible, a sensible relation between parameters, objectives (typically related
to similarity and deformation magnitude), and registration outcome needs to hold.
However, whether such a relation generally holds for currently widely used registration
methods is not well-established.

We study the feasibility of trial-and-error approaches for DIR problems by
investigating the relation between parameters, objectives, and desirability of registration
outcomes. We specifically focus on a subset of parameters, i.e., the weights in the
optimization function that determine the trade-offs between the objectives of interest
in DIR. We propose to use multi-objective optimization (Deb, 2001) as a weight-tuning
strategy, to navigate the space of optimal registration outcomes. In multi-objective
optimization, there is no unique optimal solution, but there are multiple equally good
solutions that represent the optimal trade-offs between the objectives.

To see how the interplay between weights, objectives, and desirability of registration
outcomes changes depending on the difficulty of the registration problem, we test the
methodology on a relatively easy problem, i.e., with limited deformations, namely prone-
to-prone breast magnetic resonance (MR) image registration of images acquired pre- and
post-radiotherapy. This can be helpful for treatment response assessment for patients
that undergo pre-operative breast irradiation. We then test the methodology on a
harder problem, prone-to-supine breast MR image registration, which is associated
with large deformations and has been shown to be very challenging for intensity-based
registration methods, e.g., (Lee et al., 2010). Registration of contrast-enhanced prone
MR imaging (MRI), currently routinely used for diagnostic purposes, to non-contrast-enhanced supine MRI, can be used for surgical planning, as it translates the diagnostic information to the patients treatment position during breast-conserving surgery. This can be particularly helpful for breast cancer patients with non-palpable tumors.

Preliminary results of this work have been reported in (Pirpinia et al., 2016a) and (Pirpinia et al., 2016b) on a limited number of test cases. In this work, we expand considerably our previous studies by increasing the number of datasets, and by including the following new elements; we investigate whether the addition of guidance information as a third objective, along with similarity and deformation magnitude, influences the quality of the registration outcome for the prone-to-supine registration problem. Further, we perform a quantitative comparison of the acquired Pareto fronts, by using two different comparison metrics. Lastly, we perform a more robust evaluation of the registration accuracy, based on the landmark annotations by two observers instead of one observer.

1.1. Related work

To address the challenge of parameter tuning, automated parameter tuning methods have been proposed for rigid image registration (Valsecchi et al., 2013). Parameter optimization (including the weights) for DIR has been investigated and applied to computed tomography (CT) lung registration, showing that there can be large variations in the optimal values of parameters even for the same type of registration problem (Dou et al., 2016). This indicates that parameter tuning is very challenging.

To specifically tackle the challenge of weight tuning in DIR, techniques for adaptively determining the weight associated with the smoothness term that regularizes the transformation during the optimization process have been proposed, (Andrews et al., 2014), (Cahill et al., 2009). The focus of such approaches was more on accommodating the absence of information (e.g., edges) in parts of the images that could guide the optimization algorithm, and were tested on MR thigh images. Further, a machine learning approach to determine the weights for DIR was also introduced (Yeo et al., 2010). This approach was validated on a specific class of problems, namely brain image registration. Moreover, all aforementioned approaches regard registration problems where little deformation is needed, in order to obtain an accurate match of the multiple anatomical structures present in the images. For such cases, optimization is typically not very challenging, making a trial-and-error approach likely viable. In contrast, in this work we focus on problems that involve (large) deformations.

2. Materials and methods

We first study the space of weights that a user has to navigate by randomly sampling sets of weight combinations and observing how those correspond to desirable registration outcomes. Further, we employ a multi-objective evolutionary algorithm
A multi-objective perspective on manual parameter tuning for DIR of breast MRI

Figure 1. Examples of source (upper row) and target (lower row) images for the prone-to-prone registration problem. The source image is acquired pre-radiotherapy, the target image is acquired post-radiotherapy. From left to right: cases $p_1$, $p_3$, $p_4$, $p_5$.

(EA), in conjunction with open-source image registration software, that finds the weight combinations that lead to multiple, equally good DIR outcomes, in the multi-objective sense. To quantify desirability of registration outcome, we associate to each solution a mean target registration error (TRE). We assume that the TRE aligns with a user’s idea of a good registration outcome, hence making a solution with a low mean TRE a preferable solution.

2.1. Datasets

2.1.1. Prone-to-prone Ten pairs of T1-weighted, non-contrast-enhanced MR images were retrospectively used. The images were acquired pre- and post-radiotherapy of breast cancer patients in prone orientation (Figure 1), in a Philips 3T scanner (Achieva Philips Medical Systems, Best, Netherlands). The patients (age $\geq$ 60 years, tumor size $\leq$ 30 mm on MR images) underwent partial breast radiotherapy between December 2012 and June 2015, prior to breast-conserving surgery (van der Leij et al., 2015). The images have a reconstructed voxel size of $0.9 \times 0.9 \times 1.2$ mm$^3$. The study for which the data was acquired was approved by the institutional review boards of the participating centers, and written informed consent was obtained prior to start of the treatment protocol.

2.1.2. Prone-to-supine Ten volunteers (age 26 - 60 years) underwent a T1-weighted, non-contrast-enhanced imaging sequence (used for routine clinical tumor diagnosis in breast cancer patients) in a Philips 3T scanner (Achieva Philips Medical Systems, Best, The Netherlands) both in prone and supine orientation during the same scan session (Figure 2). The images have a voxel size of $0.88 \times 0.88 \times 0.90$ mm$^3$. In contrast to the prone-to-prone cases, where markers were not available (due to the retrospective use of the data from a study for which marker placement was not part of the protocol), here, nine self-adhesive MR-visible fiducial markers (MM3005, IZI Medical Products
A multi-objective perspective on manual parameter tuning for DIR of breast MRI

Figure 2. Examples of source (upper row) and target (lower row) images for the prone-to-supine registration problem. The source image is in prone orientation, the target image is in supine orientation. From left to right: cases $v_1$, $v_3$, $v_7$, $v_8$. Arrows indicate marker locations.

Corporation, Baltimore, MD, USA) were attached to the breast. They were placed in a symmetrical fashion in order to achieve consistent and reproducible positioning. The markers are donut-shaped with a diameter of 15 mm and a thickness of 3.5 mm. Details about the marker configuration are provided in subsection 2.3.2. Written informed consent was obtained from the volunteers for the MR acquisition.

2.2. Multi-objective parameter tuning for DIR

DIR can naturally be viewed as a problem where multiple objectives are of interest. In DIR, not only a high degree of similarity between the target image and the transformed source image is desirable, but also an anatomically correct transformation $T$ between the source image and the target image. Another objective of interest can be to minimize the distance between points with a known correspondence so as to aid the registration by use of guidance information. In current state-of-the-art registration methods, all $m$ objectives of interest $O_i$ are linearly combined into one cost function $C$ that needs to be optimized, taking the following form:

$$C(T) = \lambda_0 O_0(T) + \lambda_1 O_1(T) + \lambda_2 O_2(T) + \ldots + \lambda_{m-1} O_{m-1}(T).$$

(1)

The weights $\lambda_i, \ i \in \{0 \ldots m-1\}$ determine the compromise between the objectives and need to be determined beforehand.

To compute such linear combinations, in this work we used elastix, an open-source image registration toolbox (Klein et al., 2010). Within elastix, we used a transformation model based on B-splines (Rueckert et al., 1999). We employed elastix to run linear combinations of two objectives: dissimilarity and deformation magnitude, as well as three objectives: dissimilarity, deformation magnitude, and guidance information.
2.2.1. Random sampling (RS) of linear combinations

We randomly sampled multiple linear weight combinations for \( m = 2, 3 \) objectives, which correspond to sets of 2 and 3 weights. The value of the weights was chosen to be sampled in \([0, 1]\).

2.2.2. Optimized linear combinations (OLC)

The preferred registration outcome encompasses the optimal trade-off between all the objectives, which, as expressed by the relative weighting of the objectives of interest, can vary greatly depending on the registration problem as well as the specific set of images to be registered. Hence, by design, the use of single-objective optimization for these problems puts the user potentially at a disadvantage. Multi-objective optimization (see Appendix A.1) focuses on such problems, where there exist multiple objectives of interest and decisions need to be taken in the presence of trade-offs between the objectives. Consequently, there exist multiple solutions that can be considered equally good, forming the Pareto front (Figure 3).

The goal is to find linear combinations that correspond to registration outcomes (i.e., after running elastix) that form a Pareto front, but also are as good as possible in objective space. To this end, we employed a multi-objective EA (Deb, 2001) to optimize the weight combinations. EAs are population-based optimization methods. Being able to simultaneously advance an entire population of solutions, and thereby approximating the Pareto front in one run, makes EAs state-of-the-art in multi-objective optimization.

The specific EA we used is an estimation-of-distribution algorithm (Lozano et al., 2006) known as iMAMaLGaM (incremental multi-objective adapted maximum likelihood Gaussian model) (Rodrigues et al., 2014). To evolve the population, in every generation 35% of the best solutions are selected based on a well-known domination ranking scheme (Deb et al., 2002) and are subsequently separated into clusters on the basis of their objective values, distributing the search bias across the Pareto front. An \( l \)-dimensional Gaussian mixture distribution is subsequently estimated (\( l \) is the number of
parameters) and then sampled to generate new candidate solutions. iMAMaLGaM has been shown to perform very well on benchmark as well as real-world problems (Bosman and Alderliesten, 2012).

For the optimization problem at hand, each candidate solution is a set of weights which is given as input to elastix, which yields a value for each objective (after elastix has converged to a solution). The values of the objectives are then passed back to iMAMaLGaM.

After testing different settings on benchmark problems, we allowed iMAMaLGaM to run for 100 generations, or equivalently use approximately 4000 evaluations per objective. The remainder of the EA-specific settings were set according to guidelines that were shown to work well on various benchmark problems (Bosman and Alderliesten, 2012).

2.3. Registration experiments and setup

For both registration problems, the images were first rigidly aligned using elastix. For DIR, default elastix parameter settings were used, excluding the number of resolution levels, the final control point grid spacing, and the number of optimizer iterations per resolution, which were determined after testing different settings on a subset of images included in this study and following elastix guidelines. For both datasets, the fixed number of iterations of the optimizer per resolution was set to 200, and the final control point grid spacing was set to $8 \times 8 \times 8$ mm. Furthermore, we used the adaptive stochastic gradient descent algorithm (Klein et al., 2008) available in elastix as the optimizer. For the prone-to-prone registration problem, we used four resolutions, whereas for the prone-to-supine problem, five resolutions.

2.3.1. Prone-to-prone  We needed to define the dissimilarity objective as well as the objective that describes the magnitude of the deformation. For the first objective we used the negative normalized correlation coefficient (NCC). To quantify the deformation magnitude, we used the bending energy penalty (Wahba, 1990).

2.3.2. Prone-to-supine  First, the positions of the external markers were annotated on all pairs of images. For each pair, a region of interest was manually defined on the target image. The region of interest included the volume of the breast to be registered up to and including the pectoral muscle. We then pre-processed the images in three different ways, resulting in three different subsets. In the first subset, we removed the MR-visible markers from the images using ImageJ (Schneider et al., 2012) (by replacing them with a uniform intensity region of gray value 0) so as to investigate the performance of the registration algorithm without any guidance information. In this case, we used two objectives, dissimilarity and deformation magnitude, as in the prone-to-prone registration experiments. In the second subset, we removed 5 out of 9 markers
from the images, in a way that results in a more sparse but still uniform distribution of markers on the breast. In the third subset, we did not remove the markers (Figure 4).

For the second and third subsets we used three objectives: dissimilarity, deformation magnitude, as defined previously, and guidance error. Guidance error is expressed as the mean Euclidean distance between the locations of the center of the external markers in the transformed source and target image that needs to be minimized (Baiker et al., 2011). The definitions of all objectives can be found in the *elastix* manual.

![Figure 4. Illustration of the marker configuration on the breast. For the first subset of experiments, all markers were removed. For the second subset of experiments, the non-filled markers were removed. For the third subset of experiments, no markers were removed.](image)

2.4. Evaluation

2.4.1. Pareto front comparison  A measure that accurately describes the quality of a Pareto front should capture several aspects, such as number of solutions on the Pareto front, proximity to the optimal Pareto front (if known), as well as diversity of the solutions; i.e., how well the solutions are distributed along the Pareto front. The hypervolume indicator (Zitzler and Thiele, 1998) (see Appendix A.2) is a unary measure that encapsulates these properties in one number. The larger the hypervolume, the higher the quality of the Pareto front. To be able to calculate the hypervolume for the RS approach, we derived Pareto fronts from the sets of the found solutions.

Another measure for Pareto front comparison is the percentage of solutions found by, e.g., algorithm A that are dominated by the solutions found by algorithm B. If this percentage is (close to) 100%, whereas the percentage of solutions of algorithm B dominated by A is small, then algorithm B can be said to be truly superior under the tested circumstances.
2.4.2. Registration accuracy  To evaluate the quality of the registration outcomes, an experienced breast radiologist annotated 8 to 12 internal anatomical landmarks in the source image and their corresponding locations in the target image (see example in Figure 5). We calculated the mean TRE as the mean Euclidean distance between the landmark locations in the target image and their locations in the transformed source image. We calculated the mean TRE after rigid registration (which was used as our baseline), as well as after DIR.

Furthermore, for the subsets of experiments where the external markers were removed, we used the removed markers (i.e., those not used as guidance information) to calculate the mean external marker-based TRE. Lastly, a second breast radiologist re-annotated the internal landmarks on the target image based on those that were annotated in the source image by the first radiologist. In this way, we could determine if the registration error falls into the range of the inter-observer variability of the landmark annotations by the two radiologists.

To compare the performance of RS and OLC on these problems, we used the Wilcoxon signed rank test (non-normal, paired data), with $p < 0.05$ as the significance level.

3. Results

3.1. Prone-to-prone

For the prone-to-prone registration problem, 86.8-96.7% of the solutions found by the RS approach are dominated by those found by OLC, with the inverse being 3.7-12.6%. Furthermore, the OLC approach obtains larger hypervolume indicator values (Table 1) for a comparable number of evaluations (approximately 4000 random samples)
The OLC obtains solutions with a larger TRE distribution (Figure 6), as it looks for solutions all along the Pareto front, including those with little to no deformation. Nonetheless, both approaches obtain solutions with a low mean TRE (and thereby preferable solutions) (Table 2). After rigid registration, mean TRE ranged from 3.8 - 7.9 mm and after DIR it was reduced to ≤ 2.9 mm for all cases and both approaches (Figure 6).

The two radiologists were in agreement about the landmark annotations, with a mean distance of approximately 1.0 mm. Furthermore, the distribution of the TRE in objective space is smooth (mean TRE decreases as deformation, described by the deformation magnitude objective, increases), for the set of solutions obtained by the RS as well as the Pareto front obtained by the OLC approach (Figure 7).

### 3.2. Prone-to-supine

Here, the algorithmic performance of the OLC approach is overall superior; the domination percentage for OLC over RS ranges from 65.6-100% (the inverse ranges from 0.0-32.4%) and hypervolume indicator values are larger (Table 1) (no markers: \( p = 0.0156 \), four markers: \( p = 0.0039 \), nine markers: \( p = 0.0039 \)). When there is no guidance information, for a comparable number of evaluations, again this did not result in large differences between the two approaches in lowest mean TRE for most cases (Figure 8). Furthermore, in 4 out of 10 cases \( (v_3, v_5, v_7, v_{10}) \) it was not possible to find acceptable registration outcomes, due to the very difficult and large breast deformation as well as some intensity inhomogeneities present in the MR images (see e.g., Figure 2b). This is reflected in the lowest mean TRE (Table 2) as well as the lowest mean external marker-based TRE (Table 3).

Including guidance in the optimization as a third objective leads to improved results for multiple cases; the overall mean TRE (Table 2) decreases from 13.6 mm to 10.4 mm...
Figure 6. Mean internal landmark-based TRE distribution of solutions found by RS (red) and OLC (blue) for prone-to-prone cases $p_1$-$p_{10}$. Mean TRE after rigid registration shown in black.

Figure 7. Solutions acquired via RS (left) and OLC (right) for prone-to-prone registration case $p_4$. Color-coded internal landmark-based mean TRE in mm.

Table 2. Average (and standard deviation) over all cases of the lowest internal landmark-based mean TRE in mm found by RS and OLC.

<table>
<thead>
<tr>
<th></th>
<th>RS</th>
<th>OLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prone-to-prone</td>
<td>2.3 (0.5)</td>
<td>2.2 (0.5)</td>
</tr>
<tr>
<td>Prone-to-supine (0 markers)</td>
<td>15.8 (11.1)</td>
<td>13.6 (8.3)</td>
</tr>
<tr>
<td>Prone-to-supine (4 markers)</td>
<td>14.7 (8.6)</td>
<td>10.4 (6.3)</td>
</tr>
<tr>
<td>Prone-to-supine (9 markers)</td>
<td>8.6 (6.2)</td>
<td>7.3 (4.4)</td>
</tr>
</tbody>
</table>
with four markers to 7.3 mm with 9 markers for the OLC approach, which performs better than the RS approach \((p < 0.05)\); registration examples found by OLC can be seen in Figure 9. Also here the observers were in agreement about the landmark annotations, with a mean distance of 1.5 mm. Further, the OLC approach finds a larger range of superior solutions for most test cases, as reflected in the distribution of mean TRE (Figure 8), see e.g., Figure 10. Here, the weights are color-coded in RGB space by straightforwardly placing \((\lambda_0, \lambda_1, \lambda_2)\) in the unit RGB cube. There is no smooth distribution of the TRE in neither parameter or objective space for the RS approach, whereas this holds for the OLC. In the case of the OLC approach, navigating the Pareto front directly, i.e., a posteriori, no longer requires the weight combinations to have an intuitive logical ordering along the Pareto front (i.e., the weight color-codings in Figures 10c,d, A2c,d, A3c,d). Instead, only the TRE distribution (the TRE color-codings in Figures 7b, 10b, A2b, A3b), which should correspond to a user’s intuition of registration desirability, should have an intuitive logical ordering. Lastly, the weight distribution as illustrated in the color-coding of Figures 10d, A2d, A3d, in combination with the solution distribution in objective space varies considerably between test cases.

![Figure 8.](image_url)

**Figure 8.** Mean internal landmark-based TRE distribution of solutions after rigid registration (black), after DIR by RS (red) and OLC (blue) for 0 markers (left), 4 markers (middle) and 9 markers (right).

4. Discussion

In this work, we studied DIR from a multi-objective perspective, to better understand the feasibility of trial-and-error approaches for parameter tuning of existing single-objective DIR software. To our knowledge, it is the first study to do so. We studied the space of trade-offs that a human has to manually navigate by means of iterative weight adjustment and how the complexity of such a task changes depending on the difficulty of the registration problem. Results suggest that for relatively simple problems, iterative weight adjustment via trial and error can be sufficient. This is indicated by the objective space shown by the RS approach on the prone-to-prone registration problem,
Figure 9. Examples of prone-to-supine registration outcomes (associated to a low mean TRE) with 9 markers, for cases v₃ (a), v₆ (b), v₈ (c) and v₉ (d). From left to right: source image, target image, transformed source image, checkerboard image of transformed source and target. Structures such as the heart are not well registered because they were not included in the region of interest.

which shows a sensible interplay between objectives and registration quality. Therefore for this particular registration problem, although the algorithmic performance of the OLC approach is superior for the majority of the cases, this does not always result in clinically relevant differences. This is not the case for the prone-to-supine registration problem with guidance information as the third objective. Here, the space shown by the RS approach does not exhibit an intuitively simple and sensible interplay neither in parameter or objective space, making trial-and-error approaches very challenging. For this hard problem, a space with a sensible interplay between parameters, objectives and registration outcomes is obtained only when a powerful optimizer such as the EA is used. Assuming that the user’s intuition about a good outcome aligns with the TRE, navigating the Pareto front obtained by the OLC should be far easier. Moreover, for
Figure 10. Solutions acquired via RS (left) and OLC (right) for prone-to-supine registration case v8 (9 markers). Upper row: color-coded internal landmark-based mean TRE in mm, lower row: color-coded weight distribution in RGB space.

Table 3. Lowest external marker-based mean TRE found by RS and OLC for prone-to-supine datasets with 4 and 0 markers. Note that for the case of 4 remaining markers, their guidance error is close to 0, since it is being explicitly minimized as a third objective, but as seen below this is not representative of registration accuracy.

<table>
<thead>
<tr>
<th>Prone-to-supine</th>
<th>4 markers</th>
<th>0 markers</th>
<th>4 markers</th>
<th>0 markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rigid</td>
<td>RS</td>
<td>OLC</td>
<td>RS</td>
<td>OLC</td>
</tr>
<tr>
<td>v1</td>
<td>16.0</td>
<td>5.1</td>
<td>5.2</td>
<td>8.6</td>
</tr>
<tr>
<td>v2</td>
<td>22.3</td>
<td>15.9</td>
<td>5.9</td>
<td>9.19</td>
</tr>
<tr>
<td>v3</td>
<td>52.3</td>
<td>23.5</td>
<td>19.4</td>
<td>25.6</td>
</tr>
<tr>
<td>v4</td>
<td>34.5</td>
<td>10.0</td>
<td>6.4</td>
<td>8.6</td>
</tr>
<tr>
<td>v5</td>
<td>48.2</td>
<td>13.2</td>
<td>14.5</td>
<td>13.1</td>
</tr>
<tr>
<td>v6</td>
<td>34.3</td>
<td>7.5</td>
<td>7.3</td>
<td>10.6</td>
</tr>
<tr>
<td>v7</td>
<td>40.1</td>
<td>15.5</td>
<td>12.1</td>
<td>23.1</td>
</tr>
<tr>
<td>v8</td>
<td>20.2</td>
<td>4.9</td>
<td>4.2</td>
<td>8.5</td>
</tr>
<tr>
<td>v9</td>
<td>28.2</td>
<td>11.9</td>
<td>4.5</td>
<td>9.2</td>
</tr>
<tr>
<td>v10</td>
<td>55.7</td>
<td>18.9</td>
<td>15.8</td>
<td>24.3</td>
</tr>
<tr>
<td>Mean</td>
<td>12.6</td>
<td>9.5</td>
<td>14.1</td>
<td>13.3</td>
</tr>
<tr>
<td>SD</td>
<td>6.0</td>
<td>5.5</td>
<td>7.2</td>
<td>6.2</td>
</tr>
</tbody>
</table>
these prone-to-supine cases the OLC approach obtains a larger range of registration outcomes that are also superior. It should however be noted that given an optimizer of a non-local nature, the results of the RS approach would potentially be equivalent to the results of the OLC approach for all cases, provided that a mechanism is used to ensure a good, even spread along the Pareto front, provided that the Pareto front is convex. Lastly, the variation in the weight and solution distribution between different test cases indicate that deriving task-optimal parameter settings may be very challenging, and hence, patient-specific tuning is needed.

Independently of the difficulty of the problem, choosing the right parameter settings is always a non-trivial task. When users need to define, e.g., the weights, they have to take into account not only the relative importance of one objective versus the other objective, but also for example the difference in the magnitude of the objectives, in order to come up with an appropriate linear combination. For these reasons, even the range from which the weights can be sampled can be a non-trivial task to determine, as was observed for the three-objective case in this article. This issue is further exemplified by additional experiments that we performed using different combinations of weights (e.g., linear combinations versus convex combinations). For the three-objective case, we restricted the search space to the unit simplex (i.e., ensuring that all weights sum up to 1), which, mathematically, describes the same optimal solutions, but led to inferior results compared to those without the simplex restriction. This is indicative that the more compact search space in the simplex is either more complex to search, or that due to implementation details, the weight combinations do not describe exactly the same solution obtained by elastix (e.g., a set of weights (0.3, 0.7) does not give the same solution as (0.03, 0.07)). We therefore additionally performed experiments with linear combinations rescaled by a constant factor, which resulted in slightly different values for the objectives. This indicates that there are also optimizer-specific details that differ per registration algorithm that the user may be unaware of and that may also affect the way in which each linear combination is being optimized.

Interestingly, the multi-objective approach used here to gain insight into the feasibility of manual parameter tuning could itself be considered as a tuning strategy, as it does not suffer from these hidden implementation details but rather exploits this information to obtain the Pareto front. It may well be better then to use less restrictive parameter ranges to allow the EA-based parameter tuner to obtain the best results, albeit potentially at the cost of a longer runtime. Further, in this context the approach can straightforwardly be used to optimize more parameters that are important for registration (e.g., number of resolutions, step size of the gradient optimizer, final control point spacing) in order to further improve the registration outcomes. Lastly, although in this work elastix was chosen as it provides the freedom to adjust parameters, in principle, the approach can be integrated with any commercially available registration framework and applied to any DIR problem.

Some of the prone-to-supine registration problems are just too hard to solve (with or without guidance information) due to the limitations of the transformation model.
as well as the intensity-based similarity metric. However, for a subset of the DIR problems with guidance information, results comparable to the current state-of-the-art in literature were obtained; namely, mean external marker-based TRE in the order of 8.4 mm found by (Han et al., 2014), which was obtained by using biomechanical modelling in combination with intensity-based registration. Biomechanical modelling methods (Hipwell et al., 2016) as well as the use of a dual-dynamic transformation model (Alderliesten et al., 2013) have been helpful in capturing the deformation for prone-to-supine registration. Recently, for four test cases, a mean internal-landmark based TRE of 3.7 mm was reported using symmetric biomechanical modelling (Eiben et al., 2016). In our study, comparable results were achieved only with the use of guidance information, which is not currently present in a clinical setting. Nonetheless, this could easily be introduced. Moreover, it is also possible that, with more insightful weight tuning for intensity-based registration, the combination of biomechanical modeling and intensity-based registration may lead to even better results.

A limitation of the approach presented in this article is the susceptibility of the selected underlying optimization method used by elastix for the registration to the inherent incapability of linear scalarization methods of capturing certain parts of a Pareto front (e.g., concave parts) (Das and Dennis, 1997) as well as to the local-search nature of the underlying optimization method. Nonetheless, potentially clinically useful solutions were obtained for both types of registration problems. Whether the best registration outcomes in terms of lowest mean TRE, however, will be actually identified and chosen as the preferred ones by an expert who will navigate the solution space of the Pareto front without having knowledge of the TRE, still needs to be investigated, as part of future work. It should be noted that the EA can be used not only as a parameter tuner, but also as the optimization component of a registration method (Alderliesten et al., 2013), which removes the limitation of not being able to capture concave parts of Pareto fronts.

Further, in some cases where the images displayed intensity inhomogeneities (inherent to the imaging sequence), the choice of NCC as a similarity metric might not be optimal, or preprocessing might be needed in the last resolution step, such as applying an unsharp mask filter (Dekker et al., 2003).

The runtime, approximately 1 day per registration case, as needed to obtain the results presented in this study, is too long for most clinical applications. It should be noted, however, that the current implementation was not optimized for speed. The long runtime can be mitigated by using, e.g., the GPU implementation of elastix (Shamonin et al., 2014). We should also note the EA has negligible overhead, resulting in (almost) the same calculation time for RS and OLC. Further, the EA can be easily parallelized, potentially significantly reducing runtime requirements. Lastly, it should be noted that the proposed approach could be a first step towards derivation of task-optimal parameter settings for at least some DIR problems, possibly removing the need for applying it for every DIR task.
5. Conclusions

In this work, we investigated the feasibility of manual parameter tuning for DIR from a multi-objective optimization perspective. We thereby obtained insight into the interplay between parameters, objectives of interest in DIR and quality of registration outcomes. By using random linear combinations as a surrogate for manual navigation of the search space by a user, we were able to give insight into why such a task can be very challenging for hard DIR problems, but feasible for easier DIR problems. By considering DIR from a multi-objective perspective, the user can navigate the space of optimized registration trade-offs in the form of a Pareto front which contains high-quality solutions, getting the most out of the registration software being used. Moreover, we found that the aforementioned interplay along this Pareto front does align with intuition in all tested cases, facilitating result selection.

Acknowledgments

This work was funded by the Dutch Cancer Society (KWF; Grant No. KWF 2012-5716).

Appendix A.1. Multi-objective optimization

In multi-objective optimization we assume to have $m$ objectives $f_i(x)$, $i \in \{0, 1, \ldots, m - 1\}$ that need to be optimized simultaneously. Without loss of generality, we assume that the goal is to minimize all objectives. Here, optimality is based on the notion of (Pareto) domination. A solution $x^1$ is said to (Pareto) dominate a solution $x^2$ (denoted $x^1 \succ x^2$) if and only if $f_i(x^1) \leq f_i(x^2)$ holds for all $i \in \{0, 1, \ldots, m - 1\}$ and $f_i(x^1) < f_i(x^2)$ holds for at least one $i \in \{0, 1, \ldots, m - 1\}$. A Pareto set of size $n$ is a set of solutions $x^j$, $j \in \{0, 1, \ldots, n - 1\}$ for which no solution dominates any other solution, i.e., there are no $j, k \in \{0, 1, \ldots, n - 1\}$ such that $x^j \succ x^k$ holds. A Pareto front corresponding to a Pareto set is the set of all $m$-dimensional objective values corresponding to the solutions, i.e., the set of all $f(x^j)$, $j \in \{0, 1, \ldots, n - 1\}$. A solution $x^1$ is said to be Pareto optimal if and only if there is no other $x^2$ such that $x^2 \succ x^1$ holds. Further, the optimal Pareto set is the set of all Pareto-optimal solutions and the optimal Pareto front is the Pareto front that corresponds to the optimal Pareto set. If the optimization problem is convex (i.e., the objective functions as well as the Pareto front are convex), then the optimization of any linear combination can potentially yield a solution on the optimal Pareto front (Das and Dennis, 1997). This, however, holds only when the optimizer solves the linear combination problem to global optimality. Furthermore, no linear combination can lead to solutions on concave parts of the Pareto front. Lastly, a uniform sampling of linear combinations does not necessarily lead to a uniformly spaced set of solutions in objective space.
Appendix A.2. Hypervolume

The hypervolume is the size of the \( n \)-dimensional space (where \( n \) is the number of objectives) contained by a set of points, bounded by one reference point, called the Nadir point (Figure A1). The Nadir point is defined as the point composed by taking the ‘worst’ (in our case, the highest) values for all objectives (as found by either RS or OLC), and multiplying by 1.1. Hypervolumes per registration case are comparable, however, hypervolumes cannot be compared between cases, since each one depends on the shape of the Pareto front of the specific registration problem. More details about the hypervolume indicator as well as how to compute it can be found in (Fonseca et al., 2006), (Auger et al., 2009).

Appendix A.3. Examples of sets of solutions and Pareto fronts

Below different examples of sets of solutions obtained by RS and Pareto fronts obtained by OLC can be found, for registrations with various degrees of success, based on the mean TRE (Figures A2, A3).

Appendix A.4. Registration parameters

Appendix A.4.1. Prone-to-prone
  (FixedInternalImagePixelType ‘float’)
  (MovingInternalImagePixelType ‘float’)
  (FixedImageDimension 3)
  (MovingImageDimension 3)
  (UseDirectionCosines ‘true’)
  (Registration ‘MultiMetricMultiResolutionRegistration’)

**Figure A1.** Visualization of the hypervolume (grey) contained by a set of points with respect to the Nadir point.
A multi-objective perspective on manual parameter tuning for DIR of breast MRI

Figure A2. Solutions acquired via RS (left) and OLC (right) for prone-to-supine registration case $v_6$ (9 markers). Upper row: color-coded internal landmark-based mean TRE in mm, lower row: color-coded weight distribution in RGB space.

(Interpolator ‘BSplineInterpolator’)  
(ResampleInterpolator ‘FinalBSplineInterpolator’)  
(Resampler ‘DefaultResampler’)  
(FixedImagePyramid ‘FixedRecursiveImagePyramid’)  
(MovingImagePyramid ‘MovingRecursiveImagePyramid’)  
(Optimizer ‘AdaptiveStochasticGradientDescent’)  
(Transform ‘BSplineTransform’)  
(Metric ‘AdvancedNormalizedCorrelation’ ‘TransformBendingEnergyPenalty’)  
(Metric0Weight x)  
(Metric1Weight y)  
(FinalGridSpacingInPhysicalUnits 8)  
(HowToCombineTransforms ‘Compose’)  
(ErodeMask ‘false’)  
(NumberOfResolutions 4)  
(MaximumNumberOfIterations 200)  
(NumberOfSpatialSamples 2048)  
(NewSamplesEveryIteration ‘true’)  
(ImageSampler ‘Random’)  
(ESplineInterpolationOrder 1)
A multi-objective perspective on manual parameter tuning for DIR of breast MRI

<table>
<thead>
<tr>
<th>Dissimilarity</th>
<th>Deformation magnitude</th>
<th>Guidance error (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.002</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>0.004</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>0.006</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>0.008</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>0.01</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>0.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.014</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure A3. Solutions acquired via RS (left) and OLC (right) for prone-to-supine registration case $v_3$ (9 markers). Upper row: color-coded internal landmark-based mean TRE in mm, lower row: color-coded weight distribution in RGB space.

(FinalBSplineInterpolationOrder 3)
(DefaultPixelValue 0)
(WriteResultImage ‘false’)

Appendix A.4.2. Prone-to-supine
(FixedInternalImagePixelType ‘float’)
(MovingInternalImagePixelType ‘float’)
(FixedImageDimension 3)
(MovingImageDimension 3)
(UseDirectionCosines ‘true’)
(Registration ‘MultiMetricMultiResolutionRegistration’)
(Interpolator ‘BSplineInterpolator’)
(ResampleInterpolator ‘FinalBSplineInterpolator’)
(Resampler ‘DefaultResampler’)
(FixedImagePyramid ‘FixedRecursiveImagePyramid’)
(MovingImagePyramid ‘MovingRecursiveImagePyramid’)
(Optimizer ‘AdaptiveStochasticGradientDescent’)
(Transform ‘BSplineTransform’).
(Metric 'AdvancedNormalizedCorrelation' 'TransformBendingEnergyPenalty' 'CorrespondingPointsEuclideanDistanceMetric')
(Metric0Weight x)
(Metric1Weight y)
(Metric2Weight z)
(FinalGridSpacingInPhysicalUnits 8)
(HowToCombineTransforms 'Compose')
(ErodeMask 'false')
(NumberOfResolutions 5)
(MaximumNumberOfIterations 200)
(NumberOfSpatialSamples 2048)
(NewSamplesEveryIteration 'true')
(ImageSampler 'Random')
(BSplineInterpolationOrder 1)
(FinalBSplineInterpolationOrder 3)
(DefaultPixelValue 0)
(WriteResultImage 'false')

References


A multi-objective perspective on manual parameter tuning for DIR of breast MRI


A multi-objective perspective on manual parameter tuning for DIR of breast MRI


