Guiding Multimodal Registration with Learned Optimization Updates

Benjamin Gutierrez-Becker[a,c,*], Diana Mateus[a], Loic Peter[a], Nassir Navab[a,b],

[a] *Computer Aided Medical Procedures (CAMP), Technische Universität München. Boltzmanstr. 3 85748, Garching, Germany*
[b] *Computer Aided Medical Procedures (CAMP), Johns Hopkins University, USA.*
[c] *Department of Child and Adolescent Psychiatry, Psychosomatic and Psychotherapy. Ludwig-Maximilian-University Waltherstr. 23. Munich, Germany.*

Abstract

In this paper, we address the multimodal registration problem from a novel perspective, aiming to predict the transformation aligning images directly from their visual appearance. We formulate the prediction as a supervised regression task, with joint image descriptors as input and the output are the parameters of the transformation that guide the moving image towards alignment. We model the joint local appearance with *context aware descriptors* that capture both local and global cues simultaneously in the two modalities, while the regression function is based on the gradient boosted trees method capable of handling the very large contextual feature space. The good properties of our predictions allow us to couple them with a simple gradient-based optimization for the final registration. Our approach can be applied to any transformation parametrization as well as a broad range of modality pairs. Our method learns the relationship between the intensity distributions of a pair of modalities by using prior knowledge in the form of a small training set of aligned image pairs (in the order of 1 to 5 in our experiments). We demonstrate the flexibility and generality of our method by evaluating its performance on a variety of multimodal imaging pairs obtained from two publicly available datasets, RIRE (brain MR, CT and PET) and IXI (brain MR). We also show results for the very challenging deformable registration of Intravascular Ultrasound and Histology images. In these experiments, our approach has a larger capture range when compared to other state-of-the-art methods, while improving registration accuracy in complex cases.

Keywords: Image Registration, Machine Learning, Multimodal Registration, Motion Estimation

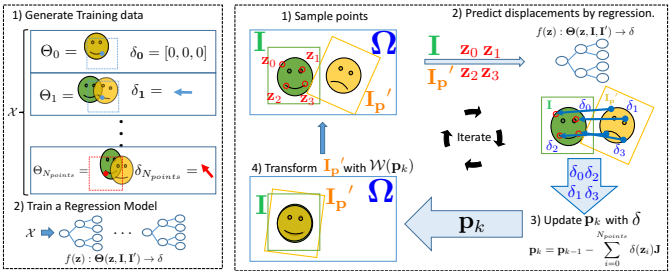

Figure 1: Method Overview. **Training stage (left)**: A set of aligned multimodal images is used to generate a training set of images with known transformations. From this training set we train an ensemble of trees mapping the joint appearance of the images to displacement vectors. **Testing stage (right)**: We register a pair of multimodal images by predicting with our trained ensemble the required displacements δ for alignment at different locations $\mathbf{z}$. The predicted displacements are then used to devise the updates of the transformation parameters to be applied to the moving image. The procedure is repeated until convergence is achieved.

*Corresponding author
Email addresses: gutierrez.becker@tum.de (Benjamin Gutierrez-Becker), mateus@in.tum.de (Diana Mateus), peter@in.tum.de (Loic Peter), navab@in.tum.de (Nassir Navab)

1. Introduction

Multimodal image registration is a fundamental task in medical image analysis, consisting in the alignment of two images of a given anatomical location acquired with different modalities. Multimodal registration is an important tool in clinical diagnosis, image-guided interventions, medical augmented reality, as well as in the validation of new imaging modalities [23, 26]. In all these applications multimodal registration plays the key role of bringing and presenting complementary information in a spatially consistent way. In addition to the challenges of the monomodal case, multimodal registration has to deal with the potentially large appearance differences that result from each modality's acquisition principles. As the relation between the intensities from the two modalities is unknown and often neither linear nor bijective, an open question is the definition of a general energy function capable of relating the two modalities and guiding a multi-modal registration algorithm.

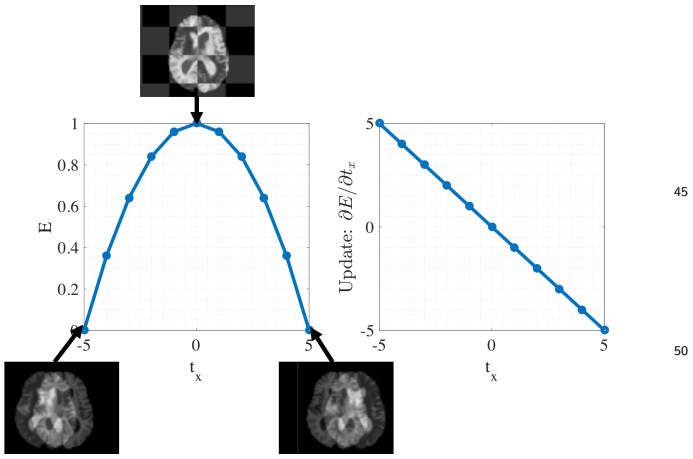

Figure 2: Exemplary energy function E. **Left**: Continuous, convex and smooth behavior of E w.r.t. a transformation parameter. **Right**: Parameter update obtained by obtaining the derivative of the energy function with respect to a transformation parameter.

For instance, one common approach is to define similarity energy functions that map the appearance of both images to a scalar value (Fig. 2. left). Once the function is defined, the optimal spatial transformation between the images is computed maximizing the similarity. Under well-behaved energies (convex, smooth, *etc.*), the optimal transformation can be reached with simple gradient-based optimization algorithms, which compute iterative updates based on the energy gradient with respect to the transformation parameters (Fig. 2. right).

Unfortunately, explicitly defining a general and well-behaved energy function that models the unknown intensity relationship between the two modalities is not straightforward. Current multi-modal similarity standards based on information theory [30], structural information [13, 37] or metric learning [25, 34] rely on the strong assumption that the same structures are visible in both modalities (Fig. 3). In the latter case, such similarities do not have an analytical gradient nor guarantee the desired properties for an optimization energy. Therefore, their gradient-based optimization calls for local gradient approximations or gradient free methods, which require advanced updates rules and an increased number of evaluations of the similarity metric.

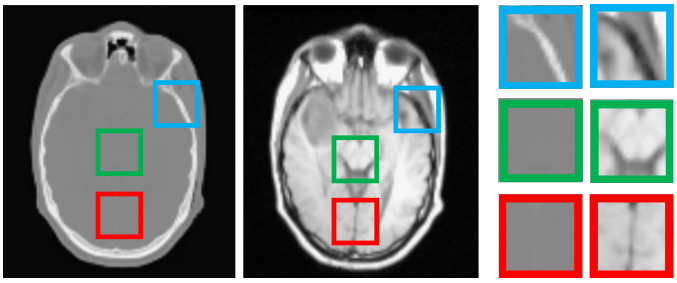

Figure 3: Corresponding CT (**left**) and MR-T1 (**middle**) images of the brain obtained from the RIRE dataset. The highlighted regions are corresponding areas between both images (**right**). Some multimodal similarity metrics rely on structural similarities between images obtained using different modalities, like the ones inside the blue boxes. However in many cases structures which are clearly visible in one imaging modality correspond to regions with homogeneous voxel values in the other modality (red and green boxes).

In this work, we design a multimodal energy function that: i) is general, since it can create models capturing complex relationships between a wide range of modality pairs by using a small set of aligned examples, ii) can model such relationships based on global *and* local appearance, iii) can be easily optimized using a gradient-based method, and iv) that adapts to different transformation parameterizations. We model multimodal registration as a supervised regression problem, where given a pair of misaligned images we predict updates of the transformation parameters towards the correct alignment (*c.f.* Fig. 1). The joint appearance of the images is represented via a multi-modal version of the Haar-like features [7] extracted from a sampling grid, which allows describing both the local and global-range context of each point. The regression task is formalized with gradient boosted trees, capable of handling the very high-dimensional Haar-like feature space, as well as of accurately approximating the transformation updates. Our work is to the best of our knowledge, the first approach aiming at learning functions that map multimodal appearance to motion predictions, and showing how to effectively integrate them into a simple optimization scheme.

This paper is based on our previous work [11] but includes several extensions. First, we modify the method in order to predict not only the optimal update direction, but also the magnitude of the update vector in each iteration of the gradient-based optimizer. Second, we replace the regression model from random forest to gradient boosted trees [9]. We show how these two modifications lead to faster convergence times during testing as well as to an accurate registration. In addition, we include an evalua-

tion of the improved properties of our method in terms of convergence and its training requirements using the IXI dataset. To demonstrate the generality of our method, we also extended our experiments to the publicly available and widely used RIRE dataset [40] for the rigid registration of three additional modalities: Computed Tomography (CT), Magnetic Resonance (MR) as well as Positron Emission Tomography (PET). We performed quantitative comparisons on the convergence of the proposed energy with a baseline [24] and a state-of-the-art method [13].

2. Related Work

Borrowing the classification of Sotiras *et al.* [35], previous approaches to multimodal-registration fall in one among three categories. The first category comprises the *information-theoretic (IT)* methods like mutual information [21] and its variations [39, 24], which are arguably the most widely used methods given their simple implementation and their effectiveness to register different modalities [31]. Assuming that a global mapping between the intensities of the two modalities exists, such methods look for the transformation that maximizes the information of the intensity distributions. However, they are typically non-convex and suffer from the discrete approximations of the densities. Furthermore, IT methods suffer from a limited capture range and thus require a good initial transformation in order to converge.

A second family of approaches seeks to reduce the multimodal registration problem to a monomodal one. This can be done by synthesizing one modality from the other [38, 5] or by building an intermediate representation common to the two modalities [37, 13, 12]. Learning has also been used for both synthesis [36] or to build intermediate representations [28]. These methods have been shown to achieve lower registration errors compared to information theoretic approaches in a variety of applications [35]. However, they are usually designed to register a specific pair of modalities or rely on strong structural similarities between the modalities to be registered.

The third category corresponds to *similarity learning* approaches that leverage on *a priori* information in the form of a training set of aligned examples.

Among these, *Generative* approaches approximate the joint intensity distribution of the images and minimize the difference of a new test pair of images to the learned distribution [33], possibly in a Bayesian Framework [43]. *Discriminative* methods, on the other hand, model the similarity learning problem as the classification of positive (aligned) and negative (misaligned) examples, discrimination typically done at the patch level [15, 20, 25]. Different strategies have been explored to approximate such similarities, including margin-based approaches [20], boosting [25] and most recently, deep learning [34, 3]. In contrast to the discriminative approaches above which aim at discerning between aligned and misaligned patches, we focus on regressing a motion predictor that guides the registration process towards alignment.

In the Computer Vision community, prior work has used motion predictions for monomodal tracking and pose-estimation. Jurie *et al.* [16] proposed a linear predictor for template tracking, which relates the difference between the compared images to variations in template position. Dollar *et al.* [8] introduced a cascaded regression approach to learn a mapping from image features to object pose parameters. The cascaded approach reduces the parameter error progressively by means of an ensemble of boosted regressors (ferns) that re-computes the features at each iteration. Xiong *et al.* [41] provides a generalization of the cascaded method of Dollar *et al.* **to solving non-linear least squared problems** via a supervised descent method in the context of face alignment. In practice, [41] implements the supervised descent approach as a sequence of linear regressors that link the differences in appearance (SIFT descriptors) to the distance between landmarks. In this paper, we formulate the multi-modal registration problem in terms of a quadratic alignment error between the two images. We optimize this function iteratively using a gradient-based scheme. Similar to [41], we learn to predict the parameter updates at each iteration, although **with gradient-boosting trees** instead of linear regressors in order to be able to handle the higher dimensionality and larger complexity of the multi-modal task. This means that a boosted sequence of prediction takes place at each iteration of the gradient optimization approach. Such two level regression approach also bears some similarities to the work of Cao *et al.* [1], who use two levels of gradient boosting regression together with feature selection and sparse coding to regress the whole facial shape in a non-parametric manner.

In the context of registration of medical images, Chou *et al.* [4] presented an approach for learning updates of the transformation parameters in the context of 2D-3D monomodal registration. Similarly, in [19], Kim *et al.* proposed the prediction of a deformation field for registration initialization, achieved by modeling the statistical correlation between image appearances and deformation fields with Support Vector Regression. Hu *et al.* [14] proposed a regression model which can predict a deformation field given changes of appearance on monomodal images of the fetal brain. Similarly to Hu *et al.*, our work uses motion prediction for registration but does it for the multimodal case. To the best of our knowledge, our work is the first approach aiming at predicting motion for the registration of medical multimodal images.

From a higher-level perspective, our work is also related to contemporary methods combining learning motion predictions with optimization methods, such as the approach of Ghesu *et al.* [10] to predict the next search direction towards an anatomical landmark based on reinforcement learning or the approach by Yang *et al.* [42] using a deep patch-wise network to predict mono-modal image deformations.

3. Background: gradient-based optimization

The simplest form of optimization for a smooth and unconstrained continuous energy is an iterative gradient-based search [27]. Starting at iteration $k = 0$ and from an initial estimate $\mathbf{p}_k$, gradient-based optimization algorithms follow the next steps:

(i) Convergence test: if conditions satisfied stop and use $\mathbf{p}_k$ as the solution.

(ii) Compute the search direction vector $\mathbf{\Delta}_k \in \mathbb{R}^{N_p}$.

(iii) Compute the step length, a positive scalar α_k such that $E(\mathbf{p}_k - \alpha_k \mathbf{\Delta}_k) < E(\mathbf{p}_k)$.

(iv) Update $k = k + 1$ and $\mathbf{p}_k = \mathbf{p}_{k-1} - \alpha \mathbf{\Delta}_k$, and go back to step (i).

The various algorithms mainly differ in the way to compute the search direction $\mathbf{\Delta}_k$ and the step size α_k. Usually, the update direction is set as the gradient of the energy function $\mathbf{\Delta}_k = \frac{\partial E}{\partial \mathbf{p}}$ at the current point $\mathbf{p}_k$ [1]. The step size may be considered as a hyper-parameter or estimated with the help of heuristics or approximate optimizations. One such approximation known as Newton approach, takes into account the energy's second derivatives (the Hessian $\mathbf{H}(E)$) to determine the update:

$$\mathbf{p}_k = \mathbf{p}_{k-1} - \mathbf{H}(E)^{-1} \frac{\partial E}{\partial \mathbf{p}}. \tag{1}$$

Newton's improved convergence rates comes at the price of an increased computational cost, as the estimation of the inverse of the Hessian can be expensive and ill-conditioned, in particular, for high dimensional problems. Computing the Hessian can be avoided by using quasi-Newton methods which approximate the Hessian matrix using an update rule leading to a faster computation. However such approximations can require a higher number of iterations when compared to the full Newton method if the approximation of the Hessian is not accurate.

4. Method

Multimodal registration is the problem of finding the optimal transformation $\mathcal{W}(\mathbf{p})$ that brings into alignment a fixed image $\mathbf{I_f} : \Omega_f \subset \mathbb{R}^3 \rightarrow \mathbb{R}$ and a moving image $\mathbf{I_m} : \Omega_m \subset \mathbb{R}^3 \rightarrow \mathbb{R}$, each of a different modality. Let the transformation be described by a vector of parameters $\mathbf{p} \in \mathbb{R}^{N_p}$. Then, the problem is formalized as that of finding the optimal displacement vector $\mathbf{p}^*$ such that:

$$\mathbf{p}^* = \arg\max_{\mathbf{p}} E(\mathbf{I}, \mathbf{I}'_{\mathbf{p}}), \tag{2}$$

where $\mathbf{I}'_{\mathbf{p}}$ stands for the moving image transformed to a joint domain $\Omega \subset \mathbb{R}^3$ by $\mathcal{W}(\mathbf{p})$, $\mathbf{I}$ is the fixed image also resampled in Ω, and E is an energy measuring the similarity between $\mathbf{I}$ and $\mathbf{I}'_{\mathbf{p}}$.

In this work, we describe a multimodal energy E compatible with simple gradient-based optimization algorithms. The resultant updates, including search direction and step-size are effectively learned from a training set of aligned images. The problem is modeled as a supervised regression task. During the training phase, we learn to predict the search direction and step size given the local joint appearance of the two images. During the test phase, we aggregate local predictions towards a global parameter update. An overview of the method is presented in Fig. 1.

4.1. An optimization-aware energy for registration.

Without loss of generality[2], we consider the transformation between the two multi-modal images as a discrete deformation field anchored to the elements of a set of control points $\{\mathbf{z}_i\}_{i=1}^{N_{\text{samples}}}$ on a joint domain Ω (see Fig 4). Formally, the deformation field is described by parameters $\mathbf{p} = \left[\overrightarrow{\delta}(\mathbf{z}_1), \ldots, \overrightarrow{\delta}(\mathbf{z}_i), \ldots, \overrightarrow{\delta}(\mathbf{z}_{N_{\text{samples}}})\right]^\top$, where each displacement $\overrightarrow{\delta}(\mathbf{z}_i)$ is a vector in $\mathbb{R}^3$ and the number of parameters equates that of the "control" points times three, *i.e.* $N_p = N_{\text{samples}} \times 3$.

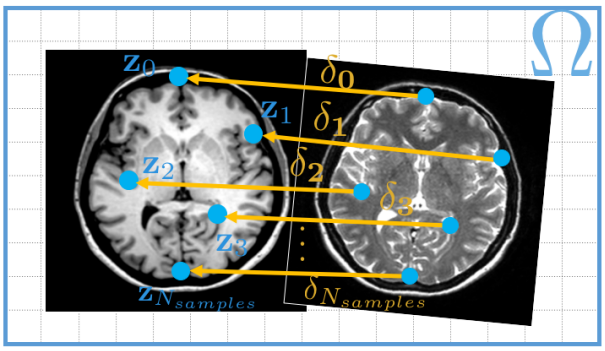

Figure 4: The transformation $\mathcal{W}$ can be defined by the local displacements $\overrightarrow{\delta}(\mathbf{z}_i)$ at the grid positions $\mathbf{z}$. These local displacement vectors $\overrightarrow{\delta}(\mathbf{z}_i)$ point towards corresponding locations in the fixed and moving image.

The displacement field connects anatomical corresponding points $\mathbf{z}_i \sim \mathbf{z}'_i$ in the two images such that $\mathbf{z}'_i = \mathbf{z}_i + \overrightarrow{\delta}(\mathbf{z}_i)$, with $\{\mathbf{z}'_i\}_{i=1}^{N_{\text{samples}}}$. We then define the registration energy function as the sum of distances between the corresponding points $\sum_{\mathbf{z}_i \in \Omega} ||\mathbf{z}_i - \mathbf{z}'_i||^2$, or equivalently as the L2-norm of the displacement field:

$$E(\mathbf{I}, \mathbf{I}'_{\mathbf{p}}) = \frac{1}{2} \sum_{\mathbf{z}_i \in \Omega} ||\overrightarrow{\delta}(\mathbf{z}_i)||^2 \tag{3}$$

The energy in equation 3 is convex, has a smooth gradient, and leads to gradient-based parameter updates pointing

[1] This is the update direction to find the minimum of the energy function. The same formulation can be used to maximize an energy function by using the update rule $\mathbf{p}_k + \alpha_k \mathbf{\Delta}_k$ minimizing E.

[2] In Sec. 4.2.4 we explain how to generalize the method to other parameterizations.

towards the global minimum, and thus favors fast convergence. The global minimum is located at the transformation for which all the displacement updates $\overrightarrow{\delta}(\mathbf{z}_i) = 0$, which corresponds to a perfect alignment.

We can easily compute the energy gradient $\frac{\partial E}{\partial \mathbf{p}_i} = \frac{\partial E}{\partial \overrightarrow{\delta}} = \overrightarrow{\delta}(\mathbf{z}_i)$ as well as the Hessian given by the 3×3 identity matrix $\mathbf{H}(\mathbf{z}) = \mathbb{I}_3$. Leading to a Newton-like update (Eq. 1):

$$\mathbf{p}_k = \mathbf{p}_{k-1} - \sum_{\mathbf{z}_i \in \Omega} \overrightarrow{\delta}(\mathbf{z}_i) \tag{4}$$

Our definition of E is so far based on the assumption that correspondences $\mathbf{z}_i \sim \mathbf{z}'_i$ are given. In the real registration setting, correspondences are unknown. However, instead of addressing the correspondence problem and explicitly defining E, we focus on predicting directly from the images the displacement field $\{\overrightarrow{\delta}(\mathbf{z}_i)\}$. We interpret this field in the context of a gradient-based optimization, as the next search direction and step-size of the parameter updates towards alignment. Such a formulation is independent of the image intensities of each modality and their relationship, while allowing for an iterative refinement of the predictions. Furthermore and as we show later, given an appropriate predictor, the behavior of the updates will be close to that of an ideal energy function.

4.2. Learning Multimodal Motion Predictors

In equation 3, we model the registration energy as the squared sum of local offsets $\overrightarrow{\delta}$ between corresponding points in both images. In practice, since for a new pair of images these offsets are unknown, we estimate them by learning a regression function $f(\mathbf{z}) : \Theta(\mathbf{z}, \mathbf{I}, \mathbf{I}'_\mathbf{p}) \mapsto \overrightarrow{\delta}(\mathbf{z})$. The input to f is a feature vector $\Theta(\mathbf{z}, \mathbf{I}, \mathbf{I}'_\mathbf{p})$ describing the joint appearance of the point $\mathbf{z}$ in both modalities. Hereafter, we denote it $\Theta(\mathbf{z})$ for simplicity. In the following subsections we describe in details the different steps of our method:

(i) Creating a training dataset $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_{\text{points}}}$ from multi-modal images under known misalignments (sec. 4.2.1).

(ii) Defining a descriptor for the joint appearance features $\Theta(\mathbf{z})$ (sec. 4.2.2).

(iii) Modeling and fitting the regression function $f(\mathbf{z}) : \Theta(\mathbf{z}) \mapsto \overrightarrow{\delta}(\mathbf{z})$ (sec. 4.2.3).

(iv) Generalizing the motion predictions to other transformation parameterizations (sec. 4.2.4).

(v) Using predicted parameter updates to solve the multi-modal registration problem during test time (sec. 4.2.5).

4.2.1. Generating the Training Set

We assume we are given prior knowledge about the relationship between the intensity distributions of the two modalities in the form of aligned image pairs. To generate the training set $\mathcal{X}$, we apply multiple known transformations $\{\mathcal{W}_j, \mathcal{W}'_j\}_{j=1}^{N_{\text{transfo}}}$ to the aligned images, mapping the coordinates of two originally superposed points $\mathbf{x} \in \Omega_f$ and $\mathbf{x}' \in \Omega_m$ to distinct locations in a common image domain $\mathbf{z}, \mathbf{z}' \in \Omega \subset \mathbb{R}^3$ (see Fig. 5).

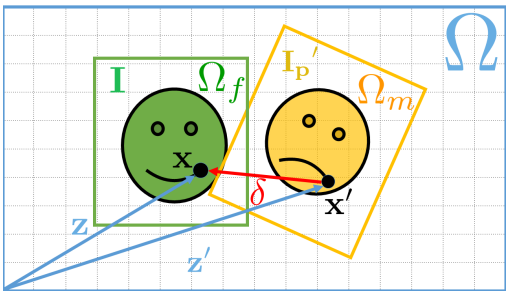

Figure 5: Generating a displacement field. The vector $\vec{\delta}$ relates corresponding points $\mathbf{x} \sim \mathbf{x}'$ after they have been moved to locations $\mathbf{z} \sim \mathbf{z}'$ by applying transformations $\{\mathcal{W}_j, \mathcal{W}'_j\}$. During training, vector $\vec{\delta}$ becomes the regression target required at location $\mathbf{z}$ to bring $\mathbf{I}'_\mathbf{p}$ into alignment with $\mathbf{I}$.

Because the applied transformations are known, we can determine the ground truth displacement $\vec{\delta}_n \in \mathbb{R}^3$ needed to find the originally corresponding point $\mathbf{z}'_n$ in the moving image, and bring it into alignment with $\mathbf{z}$, *i.e.* $\vec{\delta}_n = \mathbf{z}'_n - \mathbf{z}_n$. With this information and sampling N_{points} from the transformed images, we build a training set consisting of pairs of feature vectors Θ and their corresponding offset vector $\vec{\delta}$, *i.e.* $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_{\text{points}}}$. The process is illustrated in Fig. 6.

4.2.2. Describing joint appearance with context-aware multimodal features.

We characterize the joint appearance of a pair of images around a location $\mathbf{z}$ by means of a feature vector $\Theta(\mathbf{z}) \in \mathbb{R}^H$. We model $\Theta(\mathbf{z})$ with a multi-modal adaptation to the context-aware Haar-like features [7]. We use such rich high-dimensional descriptors to be able to encode the very large input space consisting of the joint local appearance of all image regions under all considered transformations.

The feature descriptor $\Theta(\mathbf{z})$ is built as a collection of H scalar features $[\theta_1, \ldots, \theta_h, \ldots, \theta_H]^\top$, where each θ_h is computed as an operation on a pair of boxes located at given offset locations relative to the point $\mathbf{z}$. More formally, θ_h is characterized by two boxes $\mathbf{b_1}, \mathbf{b_2}$ (*c.f.* Fig.7-left), parametrized by:

- Their *relative position* and *size* ($\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^3, w_1, h_1, w_2, h_2, d_1 \mathbb{R}$) (*c.f.* Fig. 7, top right). The position and size of the boxes are allowed to range from a couple of pixels to half of the size of the image. Using small boxes close to the sample location $\mathbf{z}$ allows the feature vector to accurately describe the local joint appearance around the point. Larger boxes and further positions instead capture the global context, which is important to perform registration when little or no overlap between images exist or when ambiguities can not be resolved using local appearance.

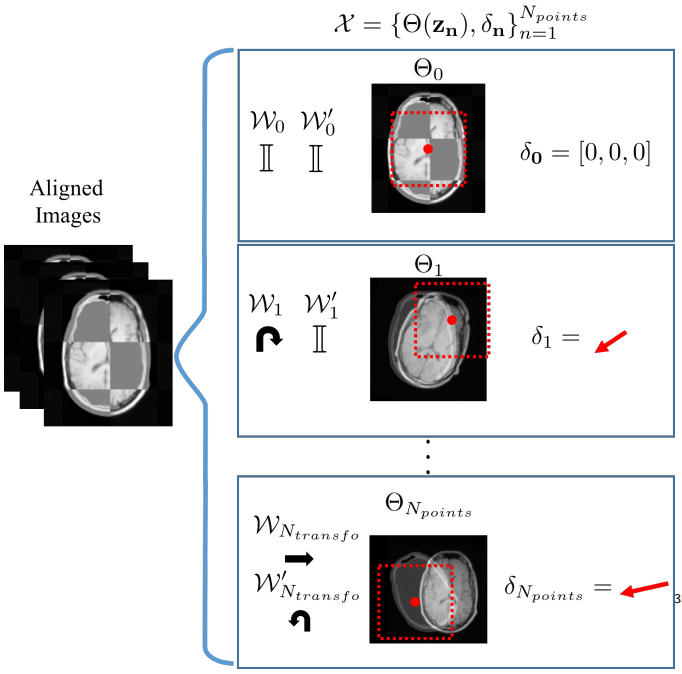

Figure 6: Generating the training dataset. Pairs of aligned images are generated and arbitrary known transformations are applied to them. The region surrounding a point, here depicted with doted lines, is characterized using the feature vector Θ described in section 4.2.2. To each feature vector we assign the displacement $\vec{\delta}$ required to bring the image at location $\mathbf{z}$ into alignment. The training set $\mathcal{X}$ is built from the collection of features and their corresponding displacements, *i.e.* $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_{\text{points}}}$.

- The *modality* where the box operates $m = \{0, 1\}$ (*c.f.* Fig. 7-middle-right). If m has the same value for both boxes we can capture the spatial context of each point within an image. If the value of m is different for each box, the feature is able to capture the functional relation across modalities.

- An operation $\mathit{0}$ between boxes taken from the set : $\mathcal{D} = \{\ \overline{\mathbf{b_1}}, \overline{\mathbf{b_2}}, \overline{\mathbf{b_1}} + \overline{\mathbf{b_2}}, \overline{\mathbf{b_1}} - \overline{\mathbf{b_2}}, |\overline{\mathbf{b_1}} - \overline{\mathbf{b_2}}|, \overline{\mathbf{b_1}} > \overline{\mathbf{b_2}}\}$, where the overline denotes the mean over the box intensities.

Considering the combinatorial nature of the parameters above, we face an infinite-dimensional feature space $\mathbb{R}^H$, which could be inefficient for learning. However, such high-dimensional feature spaces can be naturally handled by ensemble trees with axis-aligned splits, which enable individual features θ_h to be computed *on the fly* during training instead of precomputing the full vectors $\Theta(\mathbf{z})$. In addition, feature calculation is sped up by using precomputed integral volumes.

4.2.3. Displacement prediction with ensemble methods

In this subsection, we explain how to predict the offsets $\overrightarrow{\delta}(\mathbf{z}) \in \mathbb{R}^3$ from the features $\Theta(\mathbf{z})$ by approximating a function $f(\mathbf{z}) : \Theta(\mathbf{z}) \mapsto \overrightarrow{\delta}(\mathbf{z})$. We model $f(\mathbf{z})$ with an

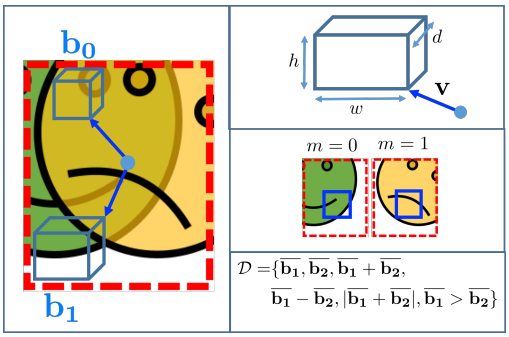

Figure 7: Context-aware features. Each element of the feature vector $\Theta(\mathbf{z})$ is constructed by obtaining a pair of boxes at different positions relative to a location $\mathbf{z}$ (left). These boxes are described by its position and size (right, top), the modality that they describe (right, middle) and an operation between boxes (right, bottom).

ensemble of regression trees, given their ability to handle high-dimensional feature spaces. A regression tree is a binary tree consisting of a set of nodes and leaves [6]. Each internal node splits the feature space into two parts according to an axis-aligned test function $g(\Theta(\mathbf{z}), h, T)$, where θ_h designates one of the dimensions of the feature vector $\Theta(\mathbf{z})$ and $T \in \mathbb{R}$ is a threshold. Given a subset of training samples $S \subset \mathcal{X}$ arriving to a given node, the split function creates a partition $S = \{S_L, S_R\}$, where S_L corresponds to the set $\theta_h < T$ and conversely, S_R to the set of features for which $\theta_h > T$. Finally, nodes without children are called leaves, and in the case of regression trees store a continuous value, *i.e.* $\overrightarrow{\delta}(\mathbf{z}) \in \mathbb{R}^3$.

During *training*, a set of labeled examples $\mathcal{X} = \{\Theta(\mathbf{z}_n), \vec{\delta}_n\}_{n=1}^{N_{\text{points}}}$ is passed through all of the trees, and the parameters of the node splitting functions h, T are optimized to minimize the prediction error. The criteria used to determine the best split parameters is the minimization of the sample covariance:

$$\theta_h^*, T^* = \arg\min_{\theta_h, T} \text{trace}\left(\mathbf{\Sigma}_{S_L}\right) + \text{trace}\left(\mathbf{\Sigma}_{S_R}\right), \qquad (5)$$

where $\mathbf{\Sigma}_{|S_{L,R}|}$ stands for the covariance matrix of the training offsets $\{\vec{\delta}_k\}_{k=1}^{|S_{L,R}|}$ of the features falling in each subset. **Computing the trace instead of the full covariance matrix allows a faster computation of the splitting criteria.**

To preserve the generalization benefits of randomized splits over the forest the parameters are usually obtained through randomized node optimization. However, given the high dimensionality of $\Theta(\mathbf{z})$, we opt instead for the automatic scale selection strategy proposed by Peter *et al.* [29], which enables us to optimize the value for the box parameters responsible for the position and scale. This choice has a positive impact on the performance of our method.

During *testing*, the prediction of the displacement $\vec{\delta}$ at a given location $\mathbf{z}$ is computed by passing the feature

vector through the ensemble, and summing the individual tree predictions. The prediction at each node is performed independently, without explicitly taking into account the spatial position of each grid point.

We considered two approaches for ensemble tree regression. The first approach is a regression forest (RF), as presented in our previous work [11], where the predictions of the individual trees are combined through a simple average $f(\mathbf{z}) = \sum_{t=1}^{N_{\text{trees}}} \frac{1}{N_{\text{trees}}} \mathcal{T}_t(\mathbf{z})$. Here, each tree is independent of each other, allowing for their parallelization during both training and prediction.

The second regression approach is based on Gradient Boosted Trees (GBT), introduced by Friedman *et al.* [9]. GBT has shown to have lower prediction errors when compared to general random forests when tuned correctly in a variety of scenarios [2]. The predictor f in GBT is a weighted sum of functions:

$$f(\mathbf{z}) = \sum_{t=1}^{N_{\text{trees}}} \beta \mathcal{T}_t(\mathbf{z}), \tag{6}$$

where each $\mathcal{T}_t$ corresponds to a regression tree and β is a scalar weighting each regression tree. However, Different to regression forests, where the training of each tree is independent, in GBT the function $f(\mathbf{z})$ is built sequentially as:

$$f_t(\mathbf{z}) = f_{t-1}(\mathbf{z}) + \beta_t \mathcal{T}_t(\mathbf{z}). \tag{7}$$

At each stage t, a tree in GBT $\mathcal{T}_t$ minimizes the squared loss between the currently predicted displacement and the ground truth $||f_{t-1}(\mathbf{z}_n) - \vec{\delta}_n||^2$, instead of trying to recover the $\vec{\delta}_n$ directly. Apart from the change in the target value, each regression tree is trained as before finding the splits that reduce the sample covariance trace (*c.f.* Eq. 5). Even though GBT requires a sequential training and therefore individual trees can not be trained in parallel, the sequential aggregation allows for shallower trees when compared to the forests, leading also to comparable training times with lower prediction errors.

4.2.4. Generalizing to arbitrary transformations

Notice that so far we have chosen $\vec{\delta}_n$ as the regression targets instead of the transformation parameters. This choice is compatible with having the transformation parametrized as a displacement field. However, we now show that by the simple chain rule of derivatives, the results of equation 4 can be generalized to other types of transformation while keeping the learning stage independent of the parametrization.

Indeed, using the chain rule the gradient of the energy may be split as $\frac{\partial E}{\partial \mathbf{p}} = \frac{\partial E}{\partial \overrightarrow{\delta}} \frac{\partial \overrightarrow{\delta}}{\partial \mathbf{p}}$, where $\frac{\partial E}{\partial \overrightarrow{\delta}}$ are the spatial derivatives and $\frac{\partial \overrightarrow{\delta}}{\partial \mathbf{p}}$ corresponds to a Jacobian relating the displacement to the transformation parameters which we denote hereafter $\mathbf{J}(\mathbf{z})$ for simplicity. The Jacobian is only dependent on the chosen parametrization and therefore does not change during the optimization. This means that we only require computing $\frac{\partial E}{\partial \overrightarrow{\delta}}$ at each iteration in order to retrieve the update direction. In the same way the Hessian of E will be computed as :

$$\mathbf{H} = \frac{\partial^2 E}{\partial \overrightarrow{\delta}^2} \frac{\partial^2 \overrightarrow{\delta}}{\partial \mathbf{p}^2} \tag{8}$$

4.2.5. Using Multimodal Motion Predictors for Registration

Once the regression function $f(\mathbf{z}) : \Theta(\mathbf{z}) \mapsto \overrightarrow{\delta}(\mathbf{z})$ is trained, we use it to perform multimodal registration on a pair of previously unseen images $\mathbf{I_f}$ and $\mathbf{I_m}$. We follow a standard gradient-based optimization (*c.f.* Sec. 3), where we calculate the search direction vector $\mathbf{\Delta}$ and the optimal step size α at every iteration k. The iterative procedure is illustrated in Fig. 1. First, a set of testing points $\{\mathbf{z}_m\}_{m=1}^{N_{\text{test}}} \in \Omega$ is randomly sampled from the fixed image. We then extract the feature vectors for the point set $\{\Theta(\mathbf{z}_m)\}_{m=1}^{N_{\text{test}}}$ and pass them through the tree ensemble. The output of the ensemble are the predicted local displacement estimates $\{\hat{\vec{\delta}}_m\}_{m=1}^{N_{\text{test}}}$. We then compute the global update (*c.f.* Eq. 4) by adding the contribution of each local displacement to the transformation parameters $\hat{\mathbf{\Delta}} = \sum_{m=1}^{N_{\text{test}}} \hat{\vec{\delta}}_m J(z)$. Finally, a convergence test is performed and if necessary the procedure is repeated. In our case we stop the optimization when the difference of the energy function between iterations E falls below a threshold ϵ.

We have seen in section 4.1 that with perfect displacement predictions, the Hessian estimate of the step length for our method is $\alpha = 1$. However, as we expect the predictions to have some error we reduce the step size by a factor λ, which we empirically evaluate.

5. Experiments and Results

To evaluate the performance of our method under different scenarios we perform three series of experiments on different multimodal datasets.

Our *first* experiments rely on pairs of multi-modal MR images of the brain from the IXI dataset[3]. The focus is on studying the amount of data required for training the regression of the displacement field as well as on demonstrating the fast convergence of our registration approach.

The *second series of experiments* evaluates the performance of our method both in terms of registration accuracy and capture range for a variety of imaging modality pairs. To this end, we performed rigid registration on the publicly available RIRE dataset[4] [40], consisting of images of adult brains obtained using different MR protocols as well as CT and PET. We evaluated our algorithm using *all* the modality pairs available in the RIRE database, which

[3] http://brain-development.org/ixi-dataset/
[4] http://www.insight-journal.org/rire/

includes: CT-T1, CT-T2, CT-PD, PET-T1, PET-T2 and PET-PD pairs, showing the generality of our approach.

In the *third experiment*, we use our method for the deformable registration of two complex modalities: Intravascular Ultrasound images and histological slices [18] (see Fig. 16). This dataset is particularly challenging, first, because the images are noisy and have acquisition artifacts, and second, because the underlying assumptions of most similarity metrics, like local structural similarities between statistics on the intensities of the images, are not valid. During these experiments, we do a comparative evaluation of our method with respect to two other similarity metrics, namely Normalized Mutual Information (NMI) and the Self-Similarity Context descriptor (SSC) [13]. We show that our learning based approach improves the results of multimodal registration in terms of accuracy and capture range. We also provide a detailed analysis of the properties of our method in terms of smoothness of the optimization updates and fast convergence. Finally, for all experiments, we also compare the behavior of our initial regression using random forest (LOU) and the new one based on gradient boosted trees (LOU2), where LOU stands for Learning Optimization Updates.

5.1. Implementation details

Our registration framework was implemented using the Insight Segmentation and Registration Toolkit (ITK) [5]. For all our experiments we performed optimization using a simple gradient descent optimizer and the same parameterizations were used for all methods. In the case of NMI we used the Mattes Mutual Information Metric included on the ITK framework and in the case of SSC we adapted the implementation provided by the authors to our framework. In all cases the control points to evaluate the similarity metrics were sampled randomly, taking approximately a proportion of 0.1 of the total voxels in the image. All metrics were evaluated using the same number of control points to ensure a fair comparison. Interpolation between control points was performed with a b-spline interpolation.The size of the boxes and offsets for the Haar-like features was limited to a maximum of half the size of the image in each dimension.

Images were processed by performing histogram matching to a reference image. This was done in order to reduce the amount of possible intensity variations observed during training and testing.

5.2. Evaluation on the IXI dataset: convergence and amount of training data

Our first experimental setup is based on the IXI dataset, which contains T1, T2 and PD-weighted images of the brain from healthy subjects. We perform two different types of experiments. First, we evaluate the registration accuracy of our method given different training dataset sizes (*c.f.* Sec. 5.2.1). Second, we study the convergence of our algorithm in terms of number of iterations required for convergence as well as for different step sizes α (*c.f.* Sec. 5.2.2). For the purpose of these experiments, we extract a dataset consisting of pairs of corresponding T1-T2 images from 10 subjects. We pre-processed the images with skull-stripping and performed histogram matching to a reference image. We carefully selected pairs of images with little or no misalignment between the T1-T2 images. We further removed any residual alignment error by aligning manually placed landmarks in both images.

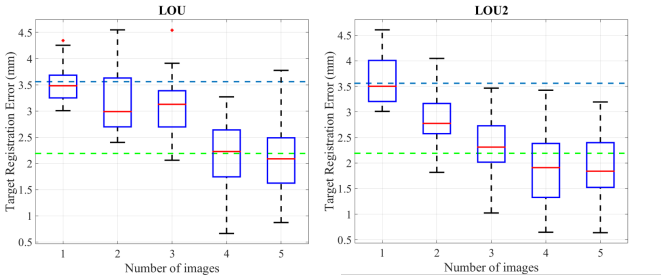

Figure 8: Comparison of registration error for models trained using different dataset sizes for LOU and LOU2. The dotted blue line indicates the median registration error for registration in the same images using Normalized Mutual Information **and the green line the median registration error for registration using SSC.**

5.2.1. Dataset size

Here, we evaluate the number of aligned images required to build a regression model capable of performing accurate registrations.

Training: We split our dataset into two groups: 5 image pairs for training and 5 for testing. We then train 5 different regression models, each with an increasing number of training images. To each image pair, we apply a random transformation, sampled from a uniform distribution in the range of $\pm size$ for translations and $\pm 1 rad$ for rotations, where $size$ corresponds to the size of the image. In total 1250 image pairs are generated for each modality pair and 10% of their voxels are taken at random for training.

Testing: We perform rigid registration using the 5 different regression models on the 5 images left out for testing. In order to assess the robustness of our algorithm to different initializations, we perform 30 registrations per image pair, each one at a different initial position for the moving image in a range between $\pm size$ for translations and $\pm 1 rad$ for rotations. We evaluate models created using both our previously presented method using random forests (LOU) and our new approach based on gradient boosted trees (LOU2). The results are shown in the box plots in Fig. 8. For reference, we perform registration on the same set of images using NMI and SSC as a similarity metric and using the same simple gradient descent optimizer and we plot the median registration error as a dotted line.

[5] https://itk.org/

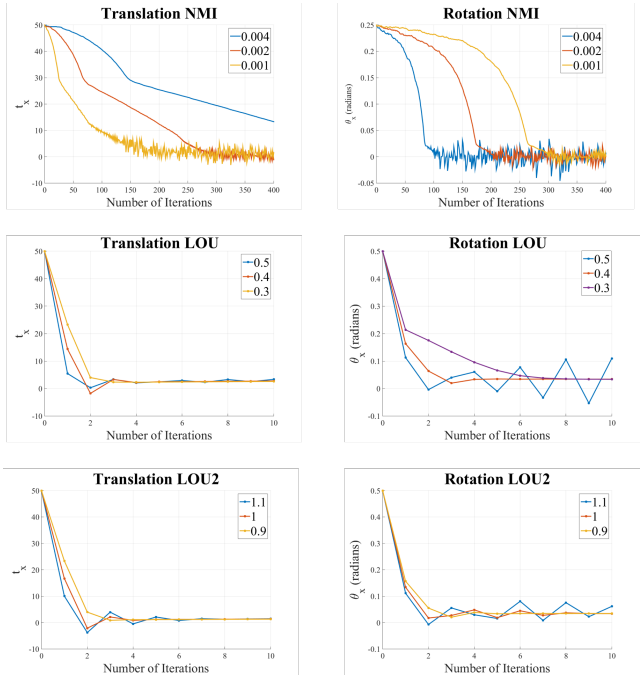

Figure 9: Behavior of a gradient descent optimizer using different strategies to calculate parameter updates. (**Top**) Gradient of NMI. (**Middle**) Updates calculated using LOU. (**Bottom**) Updates from LOU2. Each plot line corresponds to a different step size α. Our methods have a faster convergence (∼10 iterations) and a smoother behavior when compared to NMI.

The box plots show that our method is able to accurately register the test image pairs under a large range of initializations. The median registration error was comparable to the error obtained using NMI and SSC, even when the number of training images was reduced to a single pair of aligned images. Including additional images into the training dataset helps our regression model to reduce the final registration error even further. We also observe that LOU2 produces slightly lower registration errors when compared to LOU, and that, LOU2 is able to reduce the registration error using a lower number of training images. This can be attributed to the lower prediction error obtained using Gradient Boosted Trees as an ensemble technique when compared to Random Forests.

5.2.2. Convergence

For our second experiment with IXI, we take the models trained on 5 images (last box plot in Fig. 8-right) and perform rigid registration for a test pair of images with an arbitrary initial transformation. For clarity, we illustrate the behaviour of the updates independently for different parameters. To this end, we perform two separate experiments. First, we initialize the moving image with a 50mm translation offset in the axial direction. In the second experiment, we rotate the moving image around the axial axis by 0.5 radians for LOU and LOU2, but only by 0.25 radians for NMI given its smaller capture range. We show in Fig. 9 the evolution of the error over the iterations in each case and for both our methods and NMI using the same gradient-based optimizer.

In order to assess the influence of the step size α for each method, we present three curves with different step sizes. The step size shown in the red corresponds to the step size that presented the lowest final error after a line search in the parameter α.

We can draw some interesting observations from the results in Fig. 9. The path across the energy function for LOU and LOU2 is smoother and reaches a transformation close the global optimum in just a few iterations. In general LOU and LOU2 were able to find an accurate solution after no more than 10 iterations compared to the hundreds required using NMI.

The longer optimization time can be explained by the noisy approximations of the gradient of NMI. This noisy gradient forces the optimization algorithm to use a very small step size in order to ensure that the optimization converges to a solution close to the global optimum. For this reason, the convergence times for LOU (∼10 seconds) and LOU2 (∼5 seconds) were an order of magnitude faster when compared to NMI (∼100 seconds).

5.3. Evaluation on the public dataset (RIRE)

In order to demonstrate the flexibility and generality of our multi-modal registration dataset, we train an independent regression model for each available modality pair in the publicly available RIRE dataset (CT-T1, CT-T2, CT-PD, PET-T1, PET-T2 and PET-PD). Only the single pre-aligned image pairs provided in the RIRE dataset are used for training. We report the average Target Registration Error (TRE) as obtained from the online RIRE evaluation platform and compare the results for both our LOU and LOU2 methods with respect to Normalized Mutual Information (NMI) and the Self-Similarity Context descriptor (SSC) [13].

Training: We follow a similar procedure as for the IXI experiments. We again generate transformations using translations ranging from $\pm size$ and rotations from $\pm 1 rad$. However, this time we consider the raw images without skull stripping. The only pre-processing step is histogram matching between the test images and the training image in order to account for differences in the dynamic range of the images.

Testing: For each testing image pair (in total 33 image pairs) in the dataset we perform rigid registration 30 times, each starting from a different initial misalignment of the moving image. This initial transformation is sampled at random from a uniform distribution in the range of $\pm 0.5 * size$ for translation in each of the axis and rotations of $\pm$ 0.5 radians.

Results of our evaluation are shown in Figure 11. The box plots indicate the final registration error after convergence for the four compared methods and all combinations of image modality pairs. In the case of **CT-MR**, we observe that the final median registration error is comparable across the different methods. When the initialization

is close to the optimum solution, SSC, LOU and LOU2 lead to comparable low registration error. However, NMI and SSC result more often in higher registration errors when the initial transformation is large. As such transformations are not covered by capture range of the algorithm, the optimizer converges to a local optimum. We performed Mann-Whitney U statistical tests [22] between each pair of methods for all modality pairs. Almost all of these tests resulted on a significant difference between methods ($p < 0.05$) with the exception of the test between LOU and LOU2 on the CT-PD data where the null hypothesis was not rejected.

LOU and LOU2 are more robust to the initial alignment between the images and converged to a low registration error for a broader range of initial transformations.In the case of registering **PET- MR** images, the registration error of SSC is higher, which can be attributed to the poor structural information in PET images. LOU and LOU2 result in lower errors for all the PET experiments . Among our two methods, LOU2 has a broader capture range resulting in lower registration errors .

In order to asses the accuracy of each one of the methods for a standard initialization **we also performed registration for each pair of images with the initial position given by the RIRE database. Our results are shown in Table 1 and are also available online on the RIRE website using the ids shown in Table 2. We observe that given a good initialization all methods presented a similar final accuracy** . The main advantage of our method in this dataset lies therefore the increased capture range as observed in Figure 11 and the fastest convergence times due to the reduced number of required iterations for convergence. Qualitative results of these experiment are also shown on Figure 10 **where a pair** of images from the RIRE database are shown before and after registration using LOU2

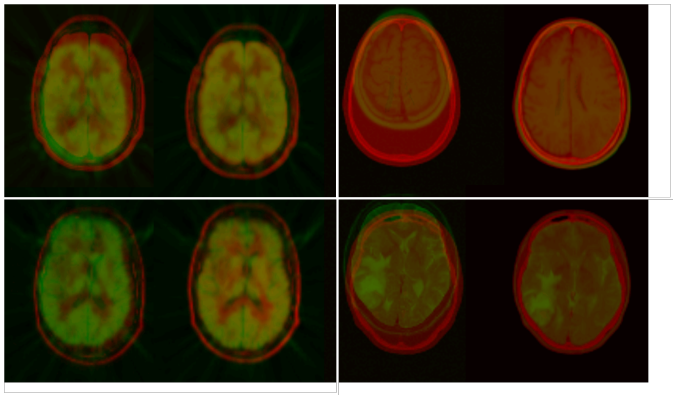

Figure 10: Exemplary results of registered images of the RIRE database using our method (LOU2). In each box the image of the left corresponds to the initial position given by the RIRE database and the image of the right corresponds to the image after registration. Top left: PET-PD; Bottom Left: PET-T2; Top left: CT-T2; Bottom left: CT-T1.

Table 1: Median TRE obtained after registering the image pairs of the RIRE dataset without initialization.

	NMI	SSC	LOU	LOU2
CT-T1	1.06 ± 1.16	2.06 ± 1.17	0.95 ± 1.30	0.89 ± 0.90
CT-T2	4.07 ± 7.74	5.53 ± 11.67	3.80 ± 4.17	3.07 ± 3.62
CT-PD	1.07 ± 12.46	1.58 ± 1.43	3.49 ± 1.91	3.07 ± 1.89
PET-T1	4.10 ± 1.83	8.11 ± 3.21	5.22 ± 3.48	5.29 ± 3.70
PET-T2	3.72 ± 2.66	7.29 ± 7.75	3.27 ± 2.91	4.17 ± 1.91
PET-PD	4.07 ± 2.11	2.62 ± 2.81	3.46 ± 2.39	2.98 ± 2.24

Table 2: Experiment ID in the RIRE database for the experiments in Table 1.

	NMI	SSC	LOU	LOU2
CT-T1	185750	185753	185752	185751
CT-T2	185754	185755	185808	185756
CT-PD	185762	185763	185764	185759
PET-T1	185769	185771	185804	185777
PET-T2	185773	185772	185784	185785
PET-PD	185767	185766	185801	185800

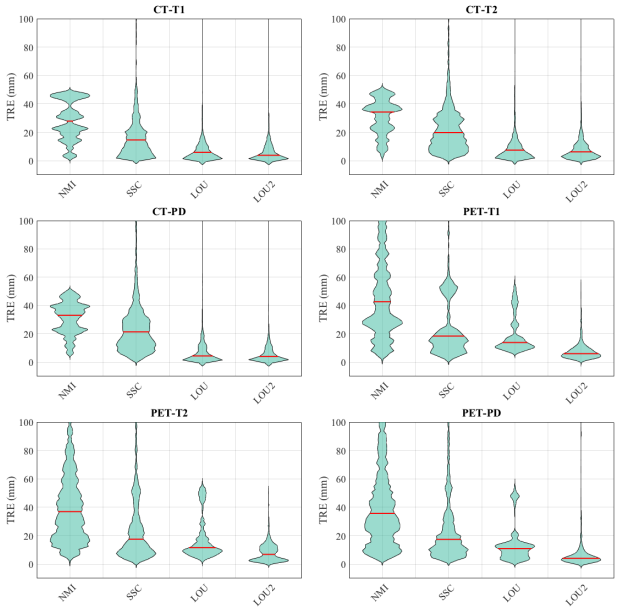

Figure 11: Final registration error for the RIRE dataset. The width of the violin plot represents the distribution of the TRE, the red line indicates the median and the black line the mean. Results are shown for registration on PET - MR image pairs and CT- MR image pairs. The plot summarizes the results for all patients from 30 different initializations.

One of our driving hypothesis is that the prior knowledge used in our learning-based approach should serve to increase the capture range for multimodal registration. To demonstrate this hypothesis is verified, we compute our predicted updates for different initial misalignments and compare them with the gradient-based updates of NMI. The update plots are shown in Fig. 12 for one pair of PET-

MR images and in Fig. 13 for one CT-MR pair.

For NMI, the updates based on the similarity gradient tend to be smooth for the range of parameters close to the optimal transformation, but noisy when the transformation parameters are far from the optimal alignment. This behavior causes the NMI gradient-based registration algorithm to fail when the initialization is far from the optimal solution. Furthermore, the optimal step size for NMI is small[6], leading the gradient-ascent algorithm to converge in a larger number of iterations when compared to our method.

For the learning-based methods, LOU and LOU2 , the optimization updates generated by our metric are smoother for all modality pairs. Additionally, in the case of LOU2, the predicted step size conforms to the ideal optimal step (*c.f.* Fig. 2) for a wide capture range. The fact that the update is proportional to the distance to the optimal solution, allows the gradient-based algorithm to converge with the fewest iterations. The differences among the updates of different are most notable for the PET-MR pair (Fig. 12), most probably given the lack of structural similarity between the modalities. Similar behavior was observed for all converged instances of the algorithm given different image pairs and initializations.

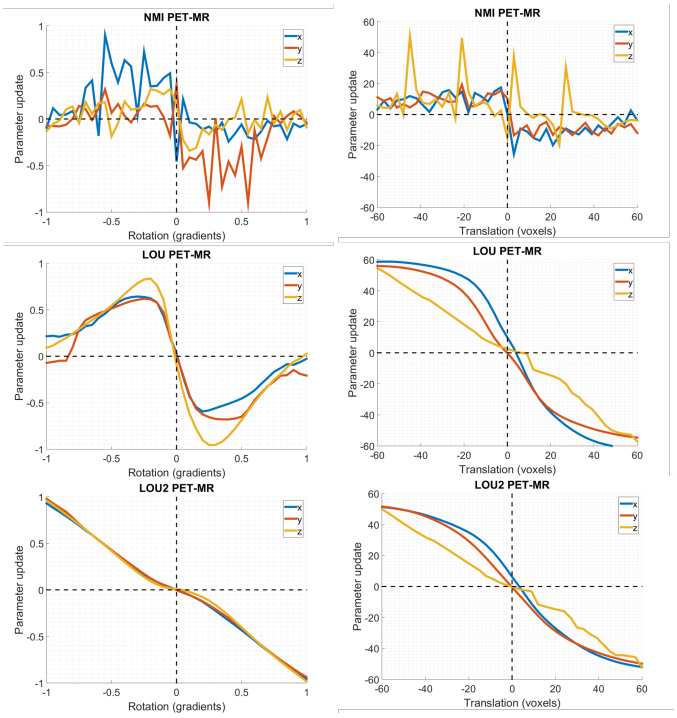

Figure 12: Comparison of estimated parameter updates using different methods for **PET-MR** pairs: NMI, updates calculated using the gradient of NMI with respect to transformation parameters; LOU, updates calculated using our approach presented on [11]; LOU2: updates calculated using the presented method. The updates with our method are smoother and the estimated step size is close to the optimal.

[6]Found through line search seeking to maximize the capture range of NMI while keeping a comparable error to our method

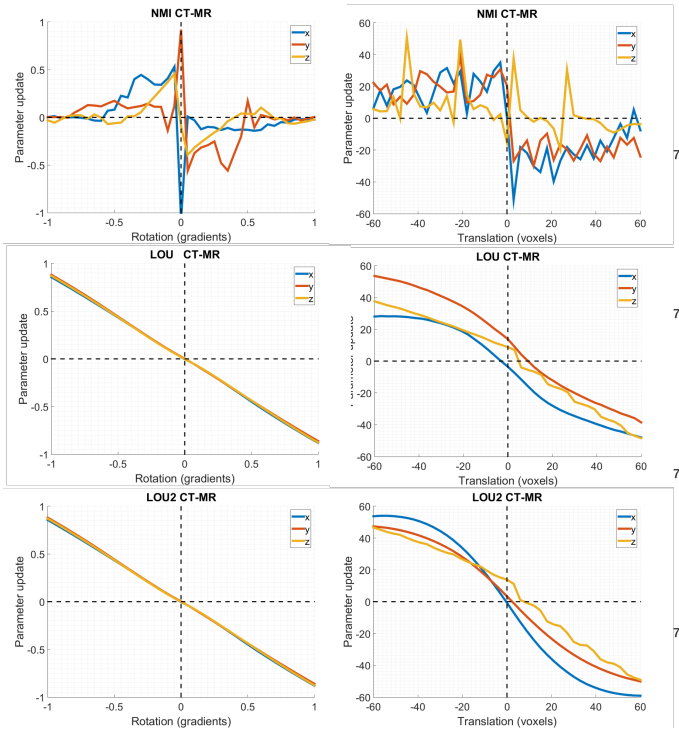

Figure 13: Comparison of estimated parameter updates using different methods for **CT-MR** pairs: NMI, updates calculated using the gradient of NMI with respect to transformation parameters; LOU, updates calculated using our approach presented on [11]; LOU2: updates calculated using the presented method. The updates with our method are smoother and the estimated step size is close to the optimal.

5.3.1. Feature relevance

We analyze which features are more relevant for the registration task by analyzing how many times each feature type is selected in the training process. In figure 14 we observe the frequency at which different types of features were selected at different trees in a gradient boosting ensemble trained for the registration of CT and T1 images of the brain. The histograms of the first column correspond to the first trained tree of the ensemble, while the second and third columns correspond to the 30th and 100th tree respectively.

We can observe that in general features with short offsets and small boxes are favored by the ensemble of regression trees. However, features corresponding to long range appearance are still useful and are considered by the trees. We observed that in general, early trees tended to select a broader range of scales, while trees corresponding to later stages of the boosted ensemble selected mostly short range features. This behavior occurs because long range features are useful to perform a rough initialization for images with large initial misalignments but are less useful for the posterior fine alignment of the images. The first few trees of the ensemble are therefore able to perform a rough alignment of the images and later trees added to the ensemble reduce the final registration error.

In the third row of Figure 14 we also show the proportion of times that trees select boxes from either the fixed or the moving image, or both images simultaneously. By observing the histograms we can conclude that our method extracts information within a single image to determine the relative position of each control point in the image and simultaneously obtains information from both images to determine the relationship of the intensities of both modalities. Interestingly while the first trees tend to select all features in an even distribution, further trees tend to rely more on features that take both modalities into account at the same time.

In the bottom row of Figure 14 we show which operations between boxes are selected. All operations seem to have equal importance on the first trained trees with the exception of the binary operation between boxes. The low importance of binary operations between boxes can be explained by the high importance of the relationship between the intensity values of both modalities. Later trees tend to prefer operations between boxes instead of operations using a single box. This is related to the previous observation that trees trained on the later stages of the ensemble require more information on the local intensity relationship between images in order to reduce registration error.

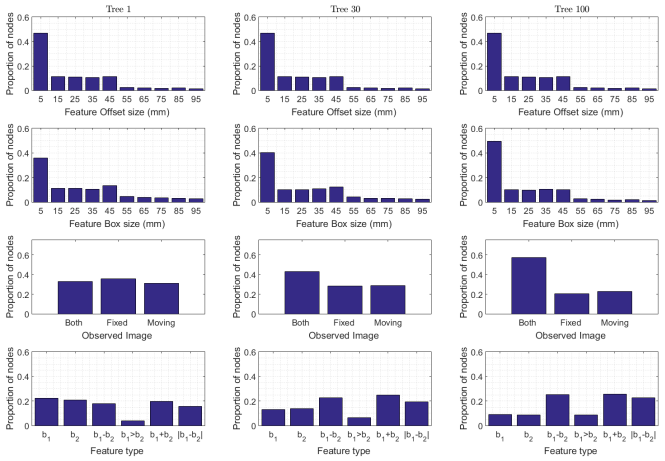

Figure 14: Histograms showing the distribution of the selected features during the training of a boosting ensemble for CT-T1 registration. The top row corresponds to the size of the offsets, the second row to the size of the boxes, the third row to the image where the features are extracted from and the fourth row to the operation between boxes.

5.4. IVUS-Histology Deformable Registration

For our third set of experiments, we perform deformable multi-modal registration in a dataset of IVUS and histology image pairs. Registration of IVUS and histology pairs is important for the characterisation of artherosclerotic tissues [17]. To build the training set we align the IVUS and histological slices using the method presented in [17],

which is based on the alignment of manually-segmented structures. We then generate 100 transformations of each aligned pair by deforming the images using a B-spline with random parameters in a range between ±20 for the B-spline coefficients , leading to a training set of 5 initially aligned images and 500 transformed images. We train our model following the same settings as for the RIRE experiments, but this time using a mask around the region actually containing tissue in the histological image. Please note that the synthetic transformations have only been used for training purposes, but testing was performed on pairs of histology and IVUS images withouth any additional transformations applied to them.

We quantitatively compare our approach against other methods by measuring the overlap (DICE) of segmented stenosis regions both in IVUS and the histology images. Even though using overlap measures is not the ideal measure to assess registration accuracy, it is still reliable for distinguishing between reasonable from inaccurate registrations [32]. For testing, we use again gradient-based optimization and we parametrize the transformation with a 3rd-order B-spline with 5 nodes per dimension distributed uniformly along the image. We do a 2-fold cross-validation evaluation with the 10 image pairs. The DICE scores after registration are shown in Fig. 15. Here, NMI and SSC present, in general, lower overlap measures when compared to our two supervised methods. Reasons for the comparably lower scores are the complex relationship between the intensities of both modalities which is difficult to capture by the joint histogram of mutual information, and the lack of structure which can be leveraged on by SSC. Our supervised approaches, on the other hand, result in much larger DICE values, indicating more accurate registration. After performing a Mann-Whitney U test between our method, SSC and NMI, our approach proved to yield a statistical significant improvement in the registration error ($p < 0.05$) . The median registration error was similar between our two approaches, but the Gradient Boosted trees of LOU2 reduces the maximum registration error. Visual examples of the experiment are illustrated in Figure 16, where we show overlays of IVUS and histology pairs before and after registration as well as the generated deformation fields.

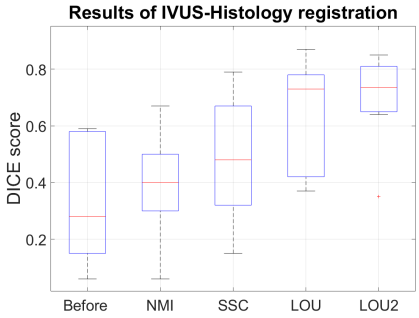

Figure 15: Results of the multi-modal deformable registration of IVUS-Histology images. The registration success of the different methods is measured by means of the DICE score, indicating the overlap between the IVUS and Histology tissue masks after registration. The boxplot shows the results of a 2-fold cross-validation experiment on ten images.

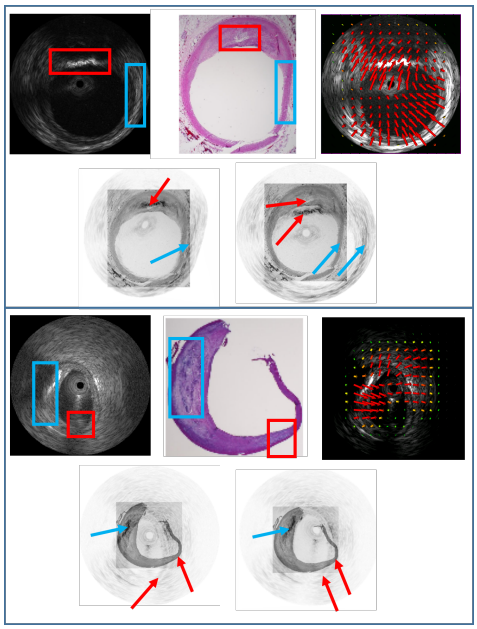

Figure 16: Results of the IVUS-Histology deformable registration experiment. On the top box the example with the highest DICE score after registration using LOU2 (DICE = 0.81), and in the bottom box an example with a low DICE score after registration (DICE=0.65). On top of each box the IVUS and histology images prior to registration and the deformation field obtained after the first iteration of LOU2. The red and blue boxes show corresponding regions in both images. In the bottom left part of the box, an overlay of the images before registration is shown. The arrows point towards the same regions in the boxes on the top. On the bottom right of each box, we show the overlay of the registered images using our method (LOU2), with the arrows indicating how the previously mismatching regions overlap after registration.

5.5. Amount of training data

Similar to the experiments performed on the IXI dataset, we evaluated the differences on the performance of our trained models depending on the amount of aligned images used to generate the training set. As the IVUS-Histology dataset presents a bigger variation both in terms of appearance and deformations, we expect that increasing the dataset size has a bigger impact in registration accuracy when compared to previous experiments. We therefore evaluated 5 different models, each one trained using different number of aligned images ranging from 1 to 5. In Figure 17 we can observe the results of this evaluation. Similar to our previous experiments, increasing the number of images improved the registration accuracy in the IVUS-Histology dataset. We can observe that our method is able to perform a more accurate registration when compared to NMI and SSC even after a single training image is included. However

adding extra images on the training set allowed our method both to reduce the number of registration outliers as well as to reduce the final registration error.

Figure 17: Results of the multi-modal deformable registration of IVUS-Histology images using LOU2 as a function of the number of aligned images used for training. We can observe an increase of registration accuracy. The blue dotted line represents the median DICE score obtained after registration using NMI and the red line the DICE score after registration with SSC.

6. Conclusions

In this work, we have a presented a novel approach to solving the multimodal registration problem based on a supervised regression and a gradient-based optimizer. Different to prior methods based on similarity design or learning, we directly target the prediction of the optimizer updates. To this end, we first show how the updates are related to the displacement field aligning the two images. We then demonstrate that using a training set of image pairs under known misalignments it is possible to train a regressor predicting the displacement fields from changes in the joint visual appearance of the images. Finally, we described how the predicted displacements can be generalized to other transformation parameterizations, and how the transformation updates can be inscribed within a simple gradient-based optimizer.

In the experimental evaluation, we have shown the flexibility and generality of our method to work on scenarios with very different modality pairs. Our method achieves comparable registration accuracy for several modality pairs were other methods have proven to be successful (for example, CT to MR). However, we have also shown that the same method is able to accurately register difficult pairs of modalities, such as IVUS to histology, for which other multimodal registration methods tend to fail (IVUS to Histology). Indeed, the supervised regression allows our method to deal with modalities displaying very different appearances and with weak structural similarities.

Our method requires sets of aligned images to train our supervised method. Nevertheless, we have observed that LOU and LOU2 are able to perform reliable registrations even when the training set size is very small. For example, the low reported errors in the RIRE experiments resulted from our model being trained on a single pair of aligned images transformed 100 times. The extra effort of generating such training sets can be justified in large scale studies, or when the ability of our method to perform registration on complex modality pairs is required (*i.e.* when other metrics fail). Adding extra pairs of aligned images to the training set enhances the accuracy of our method over other multimodal registration approaches, but it is not a requirement if focusing on our method's large capture range for comparable registration errors.

We have also shown that our method can easily be adapted to work with different parametrizations. By modeling our transformation using a displacement field we were able to easily integrate both a rigid registration parametrization and a deformable b-spline parametrization. Although this has not been throughly explored on this work, an additional advantage to parametrize our transformation as a displacement field is that a spatial regularization term could be easily applied to the displacement fields estimated by our regression model. Such a regularization term could prove important for the success of our method in scenarios where larger deformations are expected. At this point, it is important to mention that generating training sets for a highly deformable registration setting is not trivial and is an issue that has not yet been throughly explored in the supervised learning of similarity metrics. Generating training sets for deformable registration that are both realistic and extensive is an area to be addressed in order to extend our method to other scenarios and is an interesting area for future research.

The experiments have also shown that our method has an increased capture range and a faster convergence than the compared approaches. This is the result of modeling our metric as a motion prediction problem which takes the optimization into account. In the case of rigid registration, our method was able to converge in a maximum of 10 iterations, while 50 iterations were enough for an accurate deformable registration of the IVUS-Histology database. In both cases, the registration was successful even when the initial transformations were far from the optimal solution.

Our experiments mainly focused on registration on imaging settings on which acquisition protocols can be controlled and remain fairly homogeneous. However an open challenge still to be addressed by our method and other learning based approaches is that of highly variable environments, such as deformable registration of multimodal images in an intraoperative setting or registration of US images acquired at arbitrary positions and acquisition angles. The main reasons we have not yet

tackled this challenge are the requirement to generate ground truth data which can be used to train our regression models and the difficulty of modeling the large difference in appareance which occur in an intra operative scenario. However our experiments so far have shown that our method is able to handle multiple modalities as well as different parametrization, which encourage us to further explore solutions which can tackle more challenging cases.

In the future, we plan to test our method in other scenarios where prior knowledge is required to improve registration accuracy, for instance, for the registration of intra-operative 2D ultrasound images to pre-operative MR for surgical navigation. We also believe the approach can contribute to mono-modal and volume-to-slice registration problems.

7. Acknowledgements

Benjamin Gutierrez thanks CONACYT and the DAAD for their financial support. We thank Amin Katouzian and Andrew Laine for allowing us to use the IVUS-Histology dataset. We also thank Mattias Heinrich for kindly providing the implementation of SSC.

8. References

[1] Cao, X., Wei, Y., Wen, F., Sun, J., 2014. Face alignment by explicit shape regression. IJCV 107 (2), 177–190.

[2] Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on Machine learning. ACM, pp. 161–168.

[3] Cheng, X., Zhang, L., Zheng, Y., 2016. Deep similarity learning for multimodal medical images. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 1–5.

[4] Chou, C.-R., Frederick, B., Mageras, G., Chang, S., Pizer, S., 2013. 2d/3d image registration using regression learning. Computer Vision and Image Understanding 117 (9), 1095–1106.

[5] Coupé, P., Hellier, P., Morandi, X., Barillot, C., 2012. 3d rigid registration of intraoperative ultrasound and preoperative mr brain images based on hyperechogenic structures. Journal of Biomedical Imaging 2012, 1.

[6] Criminisi, A., Shotton, J. (Eds.), 2013. Decision Forests for Computer Vision and Medical Image Analysis. Advances in Computer Vision and Pattern Recognition. Springer London, pp. 245–260.

[7] Criminisi, A., Shotton, J., Bucciarelli, S., 2009. Decision forests with long-range spatial context for organ localization in ct volumes. In: MICCAI. Citeseer, pp. 69–80.

[8] Dollár, P., Welinder, P., Perona, P., 2010. Cascaded pose regression. In: CVPR. IEEE, pp. 1078–1085.

[9] Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232.

[10] Ghesu, F. C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D., 2016. An artificial agent for anatomical landmark detection in medical images. In: Ourselin, S., Joskowicz, L., Sabuncu, M. R., Unal, G., Wells, W. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III. Springer International Publishing, Cham, pp. 229–237.

[11] Gutiérrez-Becker, B., Mateus, D., Peter, L., Navab, N., 2016. Learning optimization updates for multimodal registration. In: MICCAI. Springer, pp. 19–27.

[12] Heinrich, M., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F. V., Brady, M., Schnabel, J., 2012. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. Med. Image Analysis.

[13] Heinrich, M. P., Jenkinson, M., Papiez, B., Brady, M., Schnabel, J., 2013. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: MICCAI. pp. 187–194.

[14] Hu, S., Wei, L., Gao, Y., Guo, Y., Wu, G., Shen, D., 2016. Learning-based deformable image registration for infant mr images in the first year of life. Medical Physics 44, 158,170.

[15] Jiang, J., Zheng, S., Toga, A., T., Z., June 2008. Learning based coarse-to-fine image registration. In: CVPR. pp. 1–7.

[16] Jurie, F., Dhome, M., Jul. 2002. Hyperplane approximation for template matching. TPAMI 24 (7), 996–1000.

[17] Katouzian, A., Karamalis, A., Lisauskas, J., Eslami, A., Navab, N., 2012. Ivus-histology image registration. In: International Workshop on Biomedical Image Registration. Springer, pp. 141–149.

[18] Katouzian, A., Sathyanarayana, S., Li, W., Thomas, T., Carlier, S. G., 2007. Challenges in tissue characterization from backscattered intravascular ultrasound signals. In: Medical Imaging. International Society for Optics and Photonics, pp. 65130O–65130O.

[19] Kim, M., Wu, G., Yap, P. T., Shen, D., April 2012. A general fast registration framework by learning deformation appearance correlation. IEEE Transactions on Image Processing 21 (4), 1823–1833.

[20] Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N., Scholkopf, B., 2009. Learning similarity measure for multimodal 3d image registration. In: CVPR.

[21] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P., 1997. Multimodality image registration by maximization of mutual information. IEEE transactions on medical imaging 16 (2), 187–198.

[22] Mann, H. B., Whitney, D. R., 1947. On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, 50–60.

[23] Markelj, P., Tomaževič, D., Likar, B., Pernuš, F., 2012. A review of 3d/2d registration methods for image-guided interventions. Medical Image Analysis 16 (3), 642–661.

[24] Mattes, D., Haynor, D. R., Vesselle, H., Lewellyn, T. K., Eubank, W., 2001. Nonrigid multimodality image registration. In: Medical Imaging 2001. International Society for Optics and Photonics, pp. 1609–1620.

[25] Michel, F., Bronstein, M., Bronstein, A., Paragios, N., 2011. Boosted metric learning for 3d multi-modal deformable registration. In: ISBI. IEEE, pp. 1209–1214.

[26] Navab, N., Hennersperger, C., Frisch, B., Fürst, B., 2016. Personalized, relevance-based multimodal robotic imaging and augmented reality for computer assisted interventions. Medical Image Analysis 33, 64–71.

[27] Nocedal, J., Wright, S. J., 2006. Numerical Optimization, 2nd Edition. Springer, New York.

[28] Oktay, O., Schuh, A., Rajchl, M., Keraudren, K., Gomez, A., Heinrich, M. P., Penney, G., Rueckert, D., 2015. Structured decision forests for multi-modal ultrasound image registration. In: MICCAI. Springer, pp. 363–371.

[29] Peter, L., Pauly, O., Chatelain, P., Mateus, D., Navab, N., 2015. Scale-adaptive forest training via an efficient feature sampling scheme. In: MICCAI. Springer.

[30] Pluim, J., Maintz, J., Viergever, M., 2004. f-information measures in medical image registration. TMI 23 (12), 1508–1516.

[31] Pluim, J. P., Maintz, J. A., Viergever, M. A., 2003. Mutual-information-based registration of medical images: a survey. TMI 22 (8), 986–1004.

[32] Rohlfing, T., 2012. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE transactions on medical imaging 31 (2), 153–163.

[33] Sabuncu, M., Ramadge, P., May 2008. Using spanning graphs for efficient image registration. TMI 17 (5), 788–797.

[34] Simonovsky, M., Gutierrez-Becker, B., Mateus, D., Navab, N., Komodakis, N., 2016. A deep metric for multimodal registration. In: MICCAI. pp. 10–18.

[35] Sotiras, A., Davatzikos, C., Paragios, N., July 2013. Deformable medical image registration: A survey. TMI 32 (7), 1153–1190.

[36] Van Nguyen, H., Zhou, K., Vemulapalli, R., 2015. Cross-domain synthesis of medical images using efficient location-sensitive deep network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 677–684.

[37] Wachinger, C., Navab, N., 2012. Entropy and laplacian images: Structural representations for multi-modal registration. Med. Image Analysis 16 (1), 1 – 17.

[38] Wein, W., Khamene, A., Clevert, D.-A., Kutter, O., Navab, N., 2007. Simulation and fully automatic multimodal registration of medical ultrasound. In: MICCAI. Springer, pp. 136–143.

[39] Wells, W. M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Medical image analysis 1 (1), 35–51.

[40] West, J., Fitzpatrick, J. M., Wang, M. Y., Dawant, B. M., Maurer Jr, C. R., Kessler, R. M., Maciunas, R. J., Barillot, C., Lemoine, D., Collignon, A., et al., 1997. Comparison and evaluation of retrospective intermodality brain image registration techniques. Journal of computer assisted tomography 21 (4), 554–568.

[41] Xiong, X., De la Torre, F., 2013. Supervised descent method and its applications to face alignment. In: CVPR. pp. 532–539.

[42] Yang, X., Kwitt, R., Niethammer, M., 2016. Fast predictive image registration. In: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J. M. R. S., Belagiannis, V., Papa, J. P., Nascimento, J. C., Loog, M., Lu, Z., Cardoso, J. S., Cornebise, J. (Eds.), Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings. Springer International Publishing, Cham, pp. 48–57.
URL http://dx.doi.org/10.1007/978-3-319-46976-8_6

[43] Zöllei, L., III, W. M. W., 2006. Multi-modal image registration using dirichlet-encoded prior information. In: Biomedical Image Registration, Third International Workshop, WBIR 2006, Utrecht, The Netherlands, July 9-11, 2006, Proceedings. pp. 34–42.
URL http://dx.doi.org/10.1007/11784012_5