

Learning the Similarity Measure for Multi-Modal 3D Image Registration

Daewon Lee¹, Matthias Hofmann^{1,2}, Florian Steinke¹, Yasemin Altun¹,
Nathan D. Cahill², and Bernhard Schölkopf¹

¹Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

firstname.lastname@tuebingen.mpg.de

²Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK

{mh,cahill}@robots.ox.ac.uk

Abstract

Multi-modal image registration is a challenging problem in medical imaging. The goal is to align anatomically identical structures, however, their appearance in images acquired with different imaging devices, such as for example CT or MR, may be very different. Registration algorithms generally try to deform one image, the floating image, such that it matches with a second, the reference image, by maximizing some similarity score between the deformed and the reference image. Instead of using a universal, but a priori fixed similarity criterion such as mutual information, we here try to learn a similarity measure that is optimally adapted to a given task. To this end, we use an algorithm derived from max-margin structured output learning, and the learned similarity measure is then built into a standard rigid registration algorithm. Compared to other approaches, our learning framework adapts to the specific registration problem at hand, it can exploit correlations between neighboring pixels in the reference and the floating image, and we demonstrate its superior and robust performance on difficult CT-MR/PET-MR rigid registration tasks.

1. Introduction

Many medical imaging applications require the precise spatial alignment of images of the same person taken with different scanning devices, so-called multi-modal image registration. This is an intrinsically difficult problem, since corresponding locations in the different images show different intensities, and often there is no one-to-one mapping between the intensities in one image and the intensities in the other image. For example in MR-CT registration, black pixels in the MR can correspond to either bone or air tissue, which have maximally distinct CT values. One thus cannot simply use photo-consistency as a similarity score for this task, not even after rescaling the image intensities.

Instead, the most commonly used approach maximizes the *mutual information* (MI) based on the joint intensity histogram, that is, the two dimensional histogram of the pixel intensities of corresponding point pairs [13]. The idea is that the deformation should be such that knowing the intensity at one position in the floating image should tell us as much as possible about the intensity distribution at the corresponding location in the reference image. While this is a plausible and very general assumption, it also discards much useful information. Considering again the example MR-CT, a black pixel in MR does not only tell us that the output pixel should have one, well-defined intensity, but we know more precisely that the output pixel intensity in CT should be either one, in the case of bone, or zero, in the case of air. Such knowledge can be learned from cases where an exact registration is known, and it can be used to improve existing registration algorithms.

First successful approaches in that direction have been undertaken by [7, 4, 10, 15]. For example, Leventon *et al.* [7] propose to estimate the underlying joint intensity distribution from registered example image pairs, and then to employ a maximum likelihood (ML) approach to define the alignment measure for new image pairs. Chung *et al.* [4] minimize the Kullback-Leibler (KL) divergence between the learned joint intensity distribution and the joint distribution of the new images. Similarly, Sabuncu *et al.* [10] use the entropic graph-based Jensen-Rényi (JR) divergence for the same minimization problem. Note, however, that MI and all similarity measures based on single pixel intensity histograms would not change value if the pixels in both images were randomly permuted. Yet, if considered in conjunction with its neighbors, each pixel carries much more helpful information than its pixel intensity alone. For example, observing that a pixel in the floating image is part of a boundary between two different tissue types will be much more informative to find the corresponding pixel in the reference image than the pixel's intensity alone. Therefore, each point should be described by a whole set of fea-

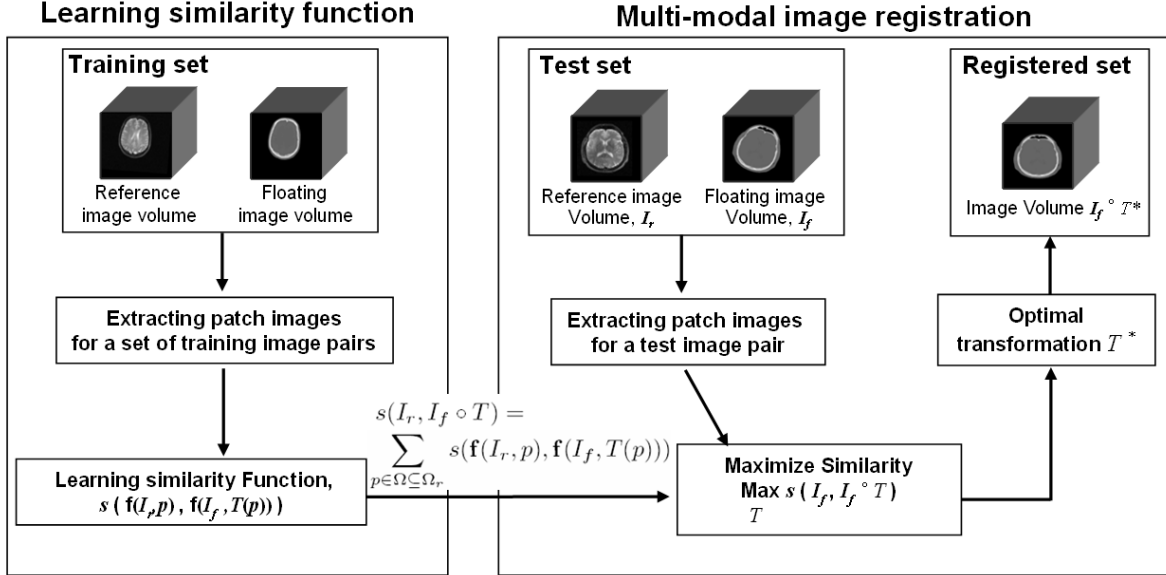


Figure 1. A flowchart of the proposed registration method using the learned similarity function.

tures derived from the neighborhood of that point, for example, an image patch centered at that point. This, however, is extremely problematic for histogram based approaches, since it would require high-dimensional histograms, which are generally unreliable to estimate [9]. To some degree these problems can be overcome by partitioning the feature space and considering only histograms of the resulting class labels [3]. However, if for computational reasons only few partitions are used, we lose again much information. Furthermore, a partitioning optimal for representing either the input or the output distribution of patches will not necessarily represent the joint distribution well. We are therefore aiming at an approach to exploit local image structure without the need of histograms.

Bringing above two ideas together, the goal of this paper is to learn a similarity function that quantifies for each position in the reference image how well it would fit another position in the floating image, based on features extracted from the neighborhood of that point in the reference and the floating image, respectively. Assuming that we can learn such a function from one or several examples where a good registration is known, we can then include it into standard registration algorithms, that search through a space of possible deformations, rigid or non-rigid, for the element maximizing the thus defined similarity score.

In Section 2 we will show how learning ideas from maximum margin structured output learning [12] can be exploited for the efficient learning of the similarity function for multi-modal registration. Part of the success of kernel methods is due to the fact that kernels were seen to provide an elegant way of handling structured inputs in ma-

chine learning problems [11]. It was subsequently noticed, however, that structured output prediction can also be done, by defining kernels on the outputs of a prediction problem. More generally, one can use joint kernels depending on inputs and outputs [1]. In Section 3 and Section 4, we experimentally evaluate the learned similarity function. We first show experiments underpinning the plausibility of the measure itself in Section 3. Then in Section 4 we incorporate the similarity measure in a standard rigid registration algorithm [2], allowing us to compare against MI on a difficult task of MR-CT and MR-PET alignment. We also benchmark our proposed similarity measure against some newer variants of MI, that have been developed to overcome certain limitations of MI. Specifically, we also compare against *normalized mutual information* (NMI), *entropy correlation coefficient* (ECC) and versions thereof that are invariant to the size of overlapping regions [2].

2. Learning the Similarity Function

In this section we describe the steps involved in learning the similarity measure for multi-modal image registration. The whole methodology, which includes learning the similarity measure from a pre-registered set of images and applying the learned measure for registering new images, is sketched in Fig. 1.

2.1. Max margin structured prediction

In multi-modal image registration, we are interested in the task of inferring a spatial transformation $T : \Omega_r \rightarrow \Omega_f$ for a reference image $I_r : \Omega_r \rightarrow \mathbb{R}$ and its corresponding

floating image $I_f : \Omega_f \rightarrow \mathbb{R}$ image, where $\Omega_r, \Omega_f \subset \mathbb{R}^d$ are the feasible position sets of the reference and floating images respectively. Given a similarity function s that quantifies the compatibility of aligned reference/floating image pairs, the optimal transformation of I_r, I_f is found by maximizing the similarity over all possible transformations,

$$T^* = \operatorname{argmax}_{T \in \mathcal{T}} s(I_r, I_f \circ T).$$

Our goal in this paper is to train a similarity function s over a sample of pre-aligned image pairs such that the empirical cost of misregistration, e.g. the target registration error [14], is minimized.

We assume that the similarity of two images decomposes into the similarities of local regions,

$$s(I_r, I_f \circ T) = \sum_{p \in \Omega_r \subseteq \Omega_f} s(\mathbf{f}(I_r, p), \mathbf{f}(I_f, T(p))), \quad (1)$$

where \mathbf{f} extracts a vectorial description of the local surroundings of point p from the given image. In this paper we focus on rectangular image patches centered at p , but other feature representations would be possible as well. Note that whereas it is possible to incorporate compatibility terms for multiple patches of the floating image $s(\mathbf{f}(I_f, p), \mathbf{f}(I_f, p'))$ (for example in order to impose spatial consistency of neighboring patches after transformation), in this paper we restrict our attention to this formulation that ignores such long-range dependencies across patches.

The optimal similarity function should give the highest score to the correctly aligned patch pairs and lower score to the others. This is exactly the optimization goal of training a predictor h that infers the floating image patch $\mathbf{y} = \mathbf{f}(I_f, p')$ that corresponds to a given reference image patch $\mathbf{x} = \mathbf{f}(I_r, p)$ by maximizing $s(\mathbf{x}, \mathbf{y})$ over all the possible patches of I_f . Let D be a sample of correctly aligned image patches $D = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$. We train s such that the empirical cost of predicting an incorrect floating image patch for a reference image patch is minimized.

We restrict the space of s to linear functions over some feature representation ϕ that is defined on the joint input-output space,

$$s(\mathbf{x}, \mathbf{y}; w) = \langle w, \phi(\mathbf{x}, \mathbf{y}) \rangle. \quad (2)$$

We train the similarity function s by maximizing the minimum margin of the sample with respect to w , where the margin is defined as $\gamma(\mathbf{x}, \mathbf{y}; w) = s(\mathbf{x}, \mathbf{y}; w) - \max_{\mathbf{y}' \neq \mathbf{y}_i} s(\mathbf{x}, \mathbf{y}'; w)$ and $\mathbf{y} \neq \mathbf{y}_i$ is any from the output space \mathcal{Y} , that is, the set of all patches from the floating image I_f . Allowing margin violations with linear penalties and controlling the norm of w , the optimization problem can be

Algorithm 1 Cutting plane algorithm for solving Eq. (3)

```

1: Input:  $D = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ 
2:  $S_i \leftarrow \emptyset$  for all  $i = 1, \dots, n$ 
3: repeat
4:   for  $i = 1, \dots, n$  do
5:      $H(\mathbf{y}) = 1 - s(\mathbf{x}_i, \mathbf{y}_i; w) + s(\mathbf{x}_i, \mathbf{y}; w)$ 
6:     compute  $\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}_i} H(\mathbf{y})$ 
7:     compute  $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i} H(\mathbf{y})\}$ 
8:     if  $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$  then
9:        $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$ 
10:    Solve the dual of (3) over  $S, S = \cup_i S_i$ 
11:    end if
12:  end for
13: until no  $S_i$  has changed during iteration

```

stated as a convex program [12]

$$\begin{aligned} \min_{w, \xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & s(\mathbf{x}_i, \mathbf{y}_i; w) - \max_{\mathbf{y} \in \mathcal{Y}, \mathbf{y} \neq \mathbf{y}_i} s(\mathbf{x}_i, \mathbf{y}; w) \geq 1 - \xi_i, \forall i \\ & \xi_i \geq 0, \forall i. \end{aligned} \quad (3)$$

Note that due to the non-linearity of the constraints, this is not a quadratic program (QP), but it can be converted into a QP by replacing each margin constraint with a set of constraints,

$$s(\mathbf{x}_i, \mathbf{y}_i; w) - s(\mathbf{x}_i, \mathbf{y}; w) \geq 1 - \xi_i, \forall \mathbf{y} \in \mathcal{Y}, \mathbf{y} \neq \mathbf{y}_i. \quad (4)$$

The important point here is that the number of constraints for each \mathbf{x} is the number of patches in the floating image. For efficient optimization of this objective function, we use a *cutting plane* approach, where at each iteration the most violated constraint is included to a set of active constraints S_i for each training instance \mathbf{x}_i , and the quadratic program is optimized over the set of active constraints $S = \cup_i S_i$. An algorithmic overview is given in Algorithm 1, some further details are explained in the following.

2.2. Joint Kernel Map

The choice of feature representation $\phi(\mathbf{x}, \mathbf{y})$ should reflect the correlations between the components of the input and output variables. We consider feature maps that are implicitly induced by a kernel function defined over the joint input-output space via

$$\begin{aligned} k((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) &= \langle \phi(\mathbf{x}, \mathbf{y}), \phi(\mathbf{x}', \mathbf{y}') \rangle \\ &= \psi(\mathbf{x}, \mathbf{x}') \cdot \psi'(\mathbf{y}, \mathbf{y}') \end{aligned}$$

where ψ and ψ' denote the inner product kernel in input and output space, respectively. In our experiments, we use

Gaussian kernels both for ψ and ψ' . In each iteration of the cutting plane algorithm, Algorithm 1, we solve the dual problem of (3) via kernel functions and obtain the learned similarity function $s(\mathbf{x}, \mathbf{y}; w)$ expressed in terms of the dual variables α as

$$\begin{aligned} s(\mathbf{x}, \mathbf{y}; w) &= \langle w, \phi(\mathbf{x}, \mathbf{y}) \rangle \\ &= \sum_{i=1}^n \sum_{\bar{\mathbf{y}} \in S_i} \alpha_{i\bar{\mathbf{y}}} \psi(\mathbf{x}_i, \mathbf{x}) \cdot (\psi'(\mathbf{y}_i, \mathbf{y}) - \psi'(\bar{\mathbf{y}}, \mathbf{y})). \end{aligned}$$

2.3. Output Space $\mathcal{Y}_i \subseteq \mathcal{Y}$

When iteratively optimizing problem (3) using the aforementioned cutting plane algorithm, we have to find the most violated constraint at each iteration (see line 6 of Algorithm 1). This is computationally problematic if the whole output space \mathcal{Y} is searched through exhaustively, since the cardinality of \mathcal{Y} , the number of patches in the floating image, may be large. A practical approach to solve this problem is to search only over a reduced set $\mathcal{Y}_i \subset \mathcal{Y}$ in the i -th iteration, where \mathcal{Y}_i includes neighboring patches of training output patch \mathbf{y}_i in the floating image. By assuming that the size of neighborhood is larger than the maximum shift of the center point of patch \mathbf{y}_i for the optimal transformation T^* , our restricted set \mathcal{Y}_i remains plausible for registration purposes.

Furthermore, we exclude additional patches that lead to contradictions with training patches from \mathcal{Y}_i . Let \mathbf{x} be the patch in the reference image corresponding to patch $\mathbf{y} \in \mathcal{Y}_i$ in the floating image of an aligned training image pair. If there exists $\mathbf{y} \in \mathcal{Y}_i$ such that $\mathbf{x} \approx \mathbf{x}_i$, then the constraints (4) require

$$s(\mathbf{x}_i, \mathbf{y}_i; w) > s(\mathbf{x}_i, \mathbf{y}; w), \quad (5)$$

$$s(\mathbf{x}, \mathbf{y}; w) > s(\mathbf{x}, \mathbf{y}_i; w). \quad (6)$$

However, smoothness of s implies $s(\mathbf{x}_i, \mathbf{y}_i; w) < s(\mathbf{x}_i, \mathbf{y}; w)$ by putting \mathbf{x}_i instead of \mathbf{x} in (6). Thus, this leads to a contradiction with (5). We avoid such contradicting training patches by excluding all patches \mathbf{y} such that $\mathbf{x} \approx \mathbf{x}_i$ from \mathcal{Y}_i . We implement $\mathbf{x} \approx \mathbf{x}_i$ to mean that the Euclidean distance between their intensities of the patches is less than some sufficiently small value.

2.4. Selecting training and test image patches

For training our similarity measure for multi-modal registration, we need a training set D of well-aligned image patches. Although one could simply extract patches from all positions of a set of registered training image pairs I_r, I_f , this would yield a set too large to practically work with. Moreover, we can only expect the patch-wise similarity measure to be informative in regions that have some image contrast and are not uniformly coloured. Otherwise,

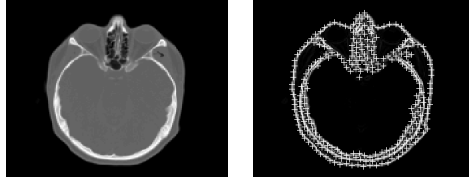


Figure 2. A CT image (left), and the locations Ω from which we extract patches for training and testing of the similarity score (right).

we can always shift the patches relative to each other without changing the similarity score $s(\mathbf{f}(I_r, p), \mathbf{f}(I_f, T(p)))$. Thus, also at test time, that is for computing the similarity score between two new images via (1), our similarity score is only needed for a subset of the image space.

We therefore define a restricted region $\Omega \subseteq \Omega_r$ from which we extract both the training and the test set patches. We focus on regions with high contrast, that is,

$$\Omega := \{p \mid \|\nabla I_r(p)\| > \theta, p \in \Omega_r\},$$

where $\nabla I_r(p)$ denotes the norm of the image gradient at p and θ is a threshold parameter. This selection of the training and test set implies that our similarity score s and the resulting registration algorithms will mainly focus on aligning anatomical boundaries. This is a plausible assumption in cases where images from different modalities fundamentally depict the same anatomical structures, as in CT and MR images [8, 5]. When other modalities such as PET or SPECT are used, this assumption may be violated; however, a learned similarity measure may still be useful as long as one modality contains structural information.

Moreover, using patches that are very close together in Ω_r will always yield the same similarity scores. Such patch pairs thus contribute neither to the training nor to the testing step. Instead, they result in a useless increase of calculation time. We therefore constraint the positions in Ω to also have at least some minimal distance from each other. The resulting positions Ω for one example image pair are shown in Fig. 2.

3. Validating the learned similarity measure separately from multi-modal registration

In this section, we first examine the learned local similarity measure between patches (2), then we show some properties of the induced image-wise criterion (1). The description of the used data sets and experimental setup for learning the similarity function is given in Section 4.1 and Section 4.2, respectively.

3.1. Local similarity measure

To evaluate the learned similarity function between patches we show its values for a MR-CT example in Fig. 3. More precisely, we show a reference MR image of a human head in a), and the corresponding CT image in b). To show the validity of the learning function s as a local similarity measure, we pick one position \mathbf{x}_1 in the reference MR image, and compute the scores $s(\mathbf{x}_1, \mathbf{y})$ for all patches \mathbf{y} of the CT image (within a box of the size of the maximal expected shift). The results are color-coded in d). While MI and NMI are typically computed between two whole images, they can also be computed between two local image patches. In order to compare NMI against our proposed local approach, we show such similarity scores for patch-wise NMI in c).

A good similarity measure should be maximised for the correct match $(\mathbf{x}_1, \mathbf{y}_1)$. This should also be the only maximum of the similarity landscape. While these goals are well-achieved by the learned similarity measure, the NMI has two maxima in close neighborhood of the true match, which can easily cause registration errors.

Another short-coming of NMI can be seen by examining the matches $(\mathbf{x}_1, \mathbf{y}_2)$ and $(\mathbf{x}_1, \mathbf{y}_3)$ more closely. \mathbf{y}_2 looks relatively similar to \mathbf{y}_1 , whereas the appearance of \mathbf{y}_3 is quite different to \mathbf{y}_1 . Nevertheless, NMI gives a better score to the pair $(\mathbf{x}_1, \mathbf{y}_3)$ than to $(\mathbf{x}_1, \mathbf{y}_2)$. This counter-intuitive behaviour is not seen for our learning based approach, which may thus be more easily interpretable.

3.2. Image-wise similarity measure

To evaluate the combined image-wise similarity function (1), we conducted the following synthetic experiments. We took a correctly aligned CT-MR image pair, and translated and rotated the CT image in different directions while the MR image was kept fixed. For each CT-MR image pair resulting from such a transformation we computed our learned similarity score of (1) as well as NMI. Here, NMI is calculated both patch-wise and image-wise. Patch-wise NMI is the sum of single patch pair NMI values, similar to (1). On the other hand, image-wise NMI means a NMI values between the whole image pair. The results are shown in Fig. 4.

Note that the obtained graphs for the learned similarity score are smoother than those for patch-wise NMI. They also show only a unique local maximum, whose position is the identity, the true transformation. In contrast, the patch-wise NMI curves show multiple local maxima, which will lead any local optimization algorithm to stop short of the true global maximum. On the other hand, image-wise NMI shows comparable smooth graphs, but the global minimum for the image-wise NMI scores is obtained at positions $+1^\circ$, $-6mm$, $+0mm$ for the rotation and translations, respectively. Thus, image-wise NMI would not yield to a precise

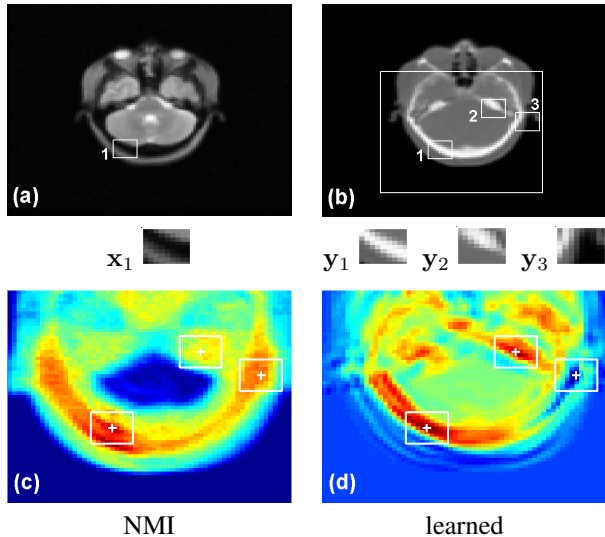


Figure 3. Comparison of local similarity values for NMI and the learned similarity measure. (a) and (b) are reference (MR) and floating (CT) image, respectively. The small rectangles below in (a) and (b) show a 2D views of the 3D patches extracted at the marked positions. (c) and (d) show the local similarity values of NMI and the learned similarity measure for all pairs of \mathbf{x}_1 and a patch \mathbf{y} within the rectangle marked in (b). Red codes for high similarity, blue for low values.

alignment of the two datasets.

In summary, these experiments give a first, strong hint that the learned similarity score yields a more accurate as well as more stable criterion for registration tasks. In the next section, we will validate these claims on a set of real multi-modal registration problems.

4. Validating the learned similarity measure within multi-modal registration

In order to evaluate the effectiveness of the learned similarity function in real applications, we conducted rigid registration experiments on a set of clinical brain image volumes comparing the performance of our learned similarity measure with state-of-the-art alternatives.

4.1. RIRE dataset

In our experiments, we used CT, MR-T2, and PET image volumes from the Retrospective Image Registration Evaluation (RIRE) Project [14]. Note that all the MR-T2 images have been rectified by the RIRE project. The dataset consists of a training set for learning a similarity function and a test set for evaluating the registration performance. The RIRE project provides the ground truth transformation for one patient (pt-00), which we use for building a pair of pre-aligned training images. The test images are from seven

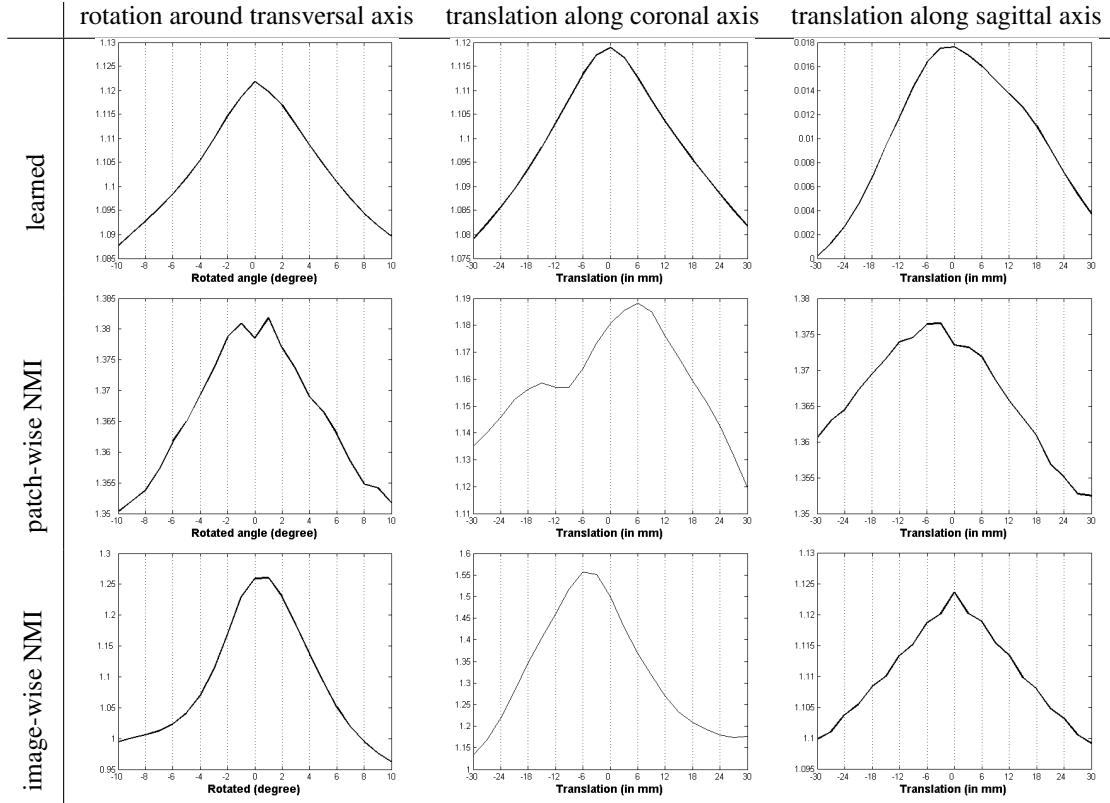


Figure 4. Image-wise similarity functions (1) for an artificial CT-MR matching task, where the CT image is transformed as in the caption of the table while the MR image is fixed. Top to bottom row represent the learned similarity function values, patch-wise NMI, and image-wise NMI, respectively.

different patients. Since for two of the seven patients no PET images are available, we only use five patients’ images for PET to MR registration (pt-01, pt-02, pt-05, pt-06, and pt-07), but all seven for CT to MR registration. The voxel size is $0.65 \times 0.65 \times 4 \text{ mm}^3$ for CT, $1.25 \times 1.25 \times 4 \text{ mm}^3$ for MR, and $2.59 \times 2.59 \times 8 \text{ mm}^3$ for PET images. Axial views of the CT, MR, and PET image volume of pt-01 are shown in Fig. 5.

To evaluate the accuracy of registration results for the various similarity measures, we use the *target registration error* (TRE) [14]. For each patient, the RIRE project has defined a set of volume of interests (VOIs) which are anatomically meaningful. TRE is the Euclidean distance between the VOI center in the reference image and its corresponding location in the deformed floating image. To obtain the TREs, we submit our transformation for each test image pair to the RIRE website, which computes the TREs and posts them on a webpage¹.

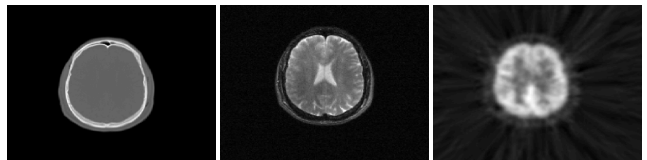


Figure 5. Axial views of RIRE brain image volumes from the patient-01 dataset. Left to right: CT, MR (T2-Rectified), and PET image.

4.2. Experimental setup

In order to learn the similarity function s , we first computed the position set Ω from the reference images (MR) of the training image pair, see Section 2.4. Then we extracted training patches centered in Ω from both of the reference and the floating training image, and learned the similarity score as described in Section 2. The width of the Gaussian kernel (σ) and threshold (θ) for the magnitude of image gradients were selected by 5-fold cross validation among all combinations of values on a finite grid. The candidate parameter values were evaluated via their average TRE for the training patient’s VOIs given registrations computed with

¹The TRE statistics for the various registration methods are available in the following webpage as open access.: http://www.insight-journal.org/rire/view_results.php

| Modality | CT to MR-T2 | | PET to MR-T2 | |
|------------|-------------|---------|--------------|---------|
| parameter | level 2 | level 1 | level 2 | level 1 |
| σ | 2.0 | 4.0 | 2.0 | 4.0 |
| θ | 0.2 | 0.2 | 0.1 | 0.2 |
| $ \Omega $ | 262 | 306 | 263 | 276 |

Table 1. Learning parameters as determined via cross-validation. $|\Omega|$ denotes the number of training patch pairs.

the respective parameters. The optimal parameters are reported in Table 1.

In order to obtain a fast and robust registration we use a multi-resolution approach [13]. We resampled all images isotropically to $6 \times 6 \times 6 \text{ mm}^3$ for the coarse resolution (level 2), and $3 \times 3 \times 3 \text{ mm}^3$ for the finer one (level 1), respectively. The patch size is fixed as $18 \times 18 \times 18 \text{ mm}^3$ for all resolutions and modalities. For each image resolution, we trained a separate similarity function, since generalisation of s over different resolutions cannot be expected.

The proposed learned similarity measure (Learned) is compared with several entropy-based measures; mutual information (MI), normalized MI (NMI), entropy correlation coefficient (ECC), cumulative residual entropy correlation coefficient (CRECC), their modified overlap invariant measures (MMI, MECC, MCRECC) [2], and learned joint density-based measure (LJD) [4]. For these measures, we set the bin size to 64 and used linear interpolation.

Since our experiments are supposed to compare different similarity functions with each other, optimally independently of a specific implementation of the registration, we used the same implementation code [2] for all measures except for LJD.

4.3. Results

The experimental results are shown in Fig.6, their numerical values are given in Table 2. The presented values are statistics computed from the TREs of all VOIs from all test patients. Note that statistics for LJD measure are only available for CT-MR registrations on 5 patients through the RIRE webpage. Thus, to compare the results on the same test patient set, we reported the values separately with () in Table 2.

For MR-CT registrations, our learned measure outperforms all standard measures. The proposed one has the lowest mean and median TRE among all measures. A MR-CT registration can be judged successful if the TRE value is smaller than 4 mm, which is the largest voxel dimension of the respective image pairs; otherwise, it should be considered a misregistration [4]. In Table 2, one can see that for the learned measure the maximum TRE is smaller than 4 mm, implying that all of VOIs of the test patients were successfully registered. On the other hand, the maximum TREs for the other similarity measures are all larger than 4 mm,

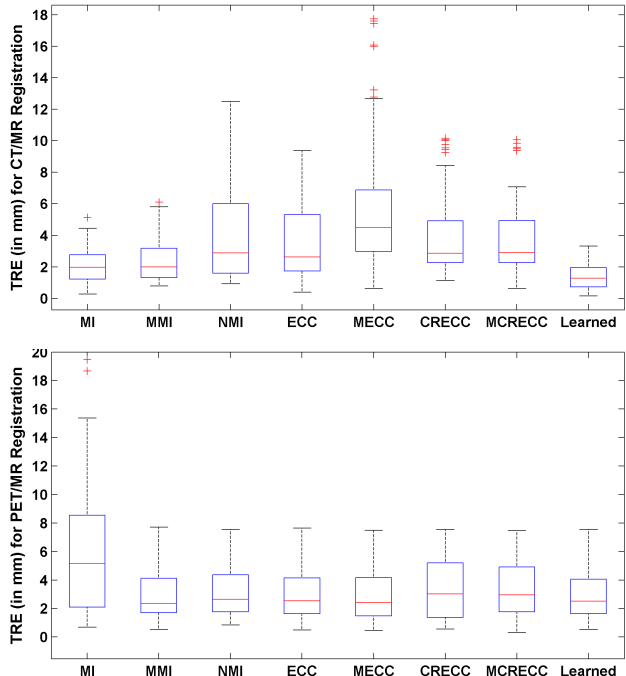


Figure 6. Box and whisker plot of the TREs for the RIRE data and different similarity measures. The box has lines at the lower quartile, median, and upper quartile values. Whiskers extend from each end of the box to the adjacent values in the data; the most extreme values within 1.5 times the interquartile range from the ends of the box. Outliers are data with values beyond the ends of the whiskers. Outliers are displayed with a red + sign.

which means they failed to register some VOIs successfully.

Concerning PET-MR registration, our proposed similarity measure also leads to registrations for which the worst case TRE is still smaller than 8mm, the maximum voxel dimension of the PET images. However, the mean and median performance are not significantly better than the other measures except MI which performs much worse. This might be due to the low resolution and the high noise levels in the PET images, which renders the PET patches much less informative.

5. Conclusions and Future Works

In this paper, we have shown a method to learn a similarity measure for multi-modal 3D image registration. In contrast to universal similarity measures such as mutual information, our learned score can be adapted optimally for each given task. Furthermore, the new method also allows to exploit structural information contained in neighbourhoods around a voxel of interest. These two effects are achieved through applying a modified version of recent max-margin structured-output learning methods to this problem. The algorithm makes use of joint kernels for the input and the

| Modality | Statistics | MI | MMI | NMI | ECC | MECC | CRECC | MCRECC | LJD | Learned |
|----------|------------|-------|------|-------|------|-------|-------|--------|--------|-------------|
| CT/MR | Mean | 2.08 | 2.47 | 4.51 | 3.48 | 6.17 | 3.97 | 3.87 | (1.85) | 1.40 (1.40) |
| | Median | 1.98 | 1.99 | 2.89 | 2.62 | 4.50 | 2.86 | 2.90 | (1.55) | 1.29 (1.29) |
| | Max | 5.15 | 6.10 | 12.50 | 9.38 | 17.74 | 10.16 | 10.07 | (6.83) | 3.32 (3.30) |
| PET/MR | Mean | 8.19 | 3.00 | 3.20 | 3.10 | 2.96 | 3.34 | 3.29 | - | 2.60 |
| | Median | 5.16 | 2.37 | 2.64 | 2.54 | 2.40 | 3.01 | 2.95 | - | 2.52 |
| | Max | 37.18 | 7.71 | 7.57 | 7.65 | 7.50 | 7.53 | 7.47 | - | 4.81 |

Table 2. Statistics of VOI TREs (in mm) across all test patients’ image volumes. Values within () denote statistics over 5 patients.

output space which are an efficient way of capturing the statistics within and between the respective image modalities to be registered, through the implicit use of an infinite-dimensional feature space representation. To compare the performance between the proposed measure and the standard entropy-based measures, we conducted CT-MR and PET-MR registrations on brain image volumes from RIRE project. For CT-MR registrations, the proposed learned measure outperforms all other measures in terms of the robustness and accuracy. For PET-MR, although no distinct differences were found between the learned measure and the standard ones, it shows comparable performance.

Several issues of the proposed method can be improved on or should be investigated further. For example, additional features could be tested towards obtaining even more robust and accurate registration results [6]. Also, we will develop a nonrigid registration method based on the learned measure in the future.

Acknowledgements

The images and the standard transformation(s) were provided as part of the project “Retrospective Image Registration Evaluation”, National Institutes of Health, No. 8R01EB002124-03, PI J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN.

References

- [1] G. H. BakIr, T. Hofmann, B. Schlkopf, A. J. Smola, B. Taskar, and S. V. Vishwanathan. *Predicting Structured Data*. Advances in neural information processing systems. MIT Press, Cambridge, MA, USA, 09 2007.
- [2] N. D. Cahill, J. A. Schnabel, J. A. Noble, and D. J. Hawkes. Revisiting overlap invariance in medical image alignment. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, June 2008.
- [3] N. D. Cahill, C. M. Williams, S. Chen, L. A. Ray, and M. M. Goodgame. Incorporating spatial information into entropy estimates to improve umtimodal image registration. In *IEEE Symposium on Biomedical Imaging*, 2006.
- [4] A. C. S. Chung, R. Gan, and W. M. W. III. Robust multi-modal image registration based on prior joint intensity distributions and minimization of kullback-leibler distance. *HKUST CSE Technical Report, HKUST-CS07-01*, 2007.
- [5] E. Haber and J. Modersitzki. Intensity gradient based registration and fusion of multi-modal images. *Lecture Notes in Computer Science (MICCAI 2006)*, pages 726–733, 2006.
- [6] A. Kelman, M. Sofka, and C. V. Stewart. Keypoint descriptors for matching across multiple image modalities and non-linear intensity variations. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.
- [7] M. Leventon and W. Grimson. Multi-modal volume registration using joint intensity distribution. In *Proceedings of MICCAI*, pages 1057–1066, 1998.
- [8] J. Pluim, J. B. A. Maintz, and M. A. Viergever. Image registration by maximization of combined mutual information and gradient information. *IEEE Transactions on Medical Imaging*, (8):809–814, 2000.
- [9] D. Rueckert, M. J. Clarkson, D. L. G. Hill, and D. J. Hawkes. Non-rigid registration using higher-order mutual information. In *Medical Imaging: Image Processing*, K. M. Hanson, Ed. Bellingham, WA: SPIE, pages 438–447, 2000.
- [10] M. Sabuncu and P. Ramadge. Using spanning graphs for efficient image registration. *IEEE Transactions on Image Processing*, pages 788–797, 2008.
- [11] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
- [12] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, pages 1453–1484, 2005.
- [13] P. A. Viola, W. M. W. III, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, pages 5–51, 1996.
- [14] J. West, J. Fitzpatrick, M. Wang, B. Dawant, C. Maurer, R. Kessler, and R. Maciunas. Comparison and evaluation of retrospective intermodality image registration techniques. In *Proceedings of the SPIE Conference on Medical Imaging*, 1996.
- [15] S. K. Zhou, J. Shao, B. Georgescu, and D. Comaniciu. Boostmotion: Boosting a discriminative similarity function for motion estimation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.