Learning Similarity Measure for Multi-Modal 3D Image Registration

Daewon Lee[1], Matthias Hofmann[1,2], Florian Steinke[1,3], Yasemin Altun[1],
Nathan D. Cahill[2,4], and Bernhard Schölkopf[1]
[1]Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany
firstname.lastname@tuebingen.mpg.de
[2]Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK
{mh,cahill}@robots.ox.ac.uk
[3]Siemens Corporate Technology, 81739 Munich, Germany
[4] Research and Innovation, Carestream Health, Inc., Rochester, NY 14618, USA

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 CT or MR, may be very different. Registration algorithms generally 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 propose learning a similarity measure in a discriminative manner such that the reference and correctly deformed floating images receive high similarity scores. To this end, we develop an algorithm derived from max-margin structured output learning, and employ the learned similarity measure within a standard rigid registration algorithm. Compared to other approaches, our method adapts to the specific registration problem at hand and exploits correlations between neighboring pixels in the reference and the floating image. Empirical evaluation on CT-MR/PET-MR rigid registration tasks demonstrates that our approach yields robust performance and outperforms the state of the art methods for multi-modal medical image registration.

1. Introduction

Many medical imaging applications require multi-modal registration, or the precise spatial alignment of images of the same person taken with different scanning devices. This is an intrinsically difficult problem, since corresponding locations in the different images show different intensities, and often there is not a one-to-one mapping between the intensities in the two images. For example in MR-CT registration, black pixels in the MR image can correspond to either bone or air tissue, which have maximally distinct CT values. Thus, one cannot simply use photo-consistency as a similarity score for this task, not even after rescaling the image intensities.

The most popular approach in multi-modal image registration maximizes the *mutual information* (MI) of the images based on their joint intensity histogram, that is, the two dimensional histogram of the pixel intensities of corresponding point pairs [11]. This approach favors similarity scores that yield the most information about the intensity distribution at one location in the floating image given the intensity at the corresponding position in the reference image. While this is a plausible and very general assumption, it also discards much useful information. Considering again the example of MR-CT registration, a black pixel in MR does not tell us that the output pixel should have a uniquely defined intensity. In fact, we know 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 when an exact registration is known, and it can be used to improve existing registration algorithms.

The first successful approaches in learning similarity functions for medical image registration have been undertaken within a generative framework [5, 4, 8, 13]. Leventon *et al*. [5] 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*. [8] use the entropic graph-based Jensen-Rényi (JR) divergence for the same minimization problem. One problem, however, with MI and all similarity measures based on sin-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

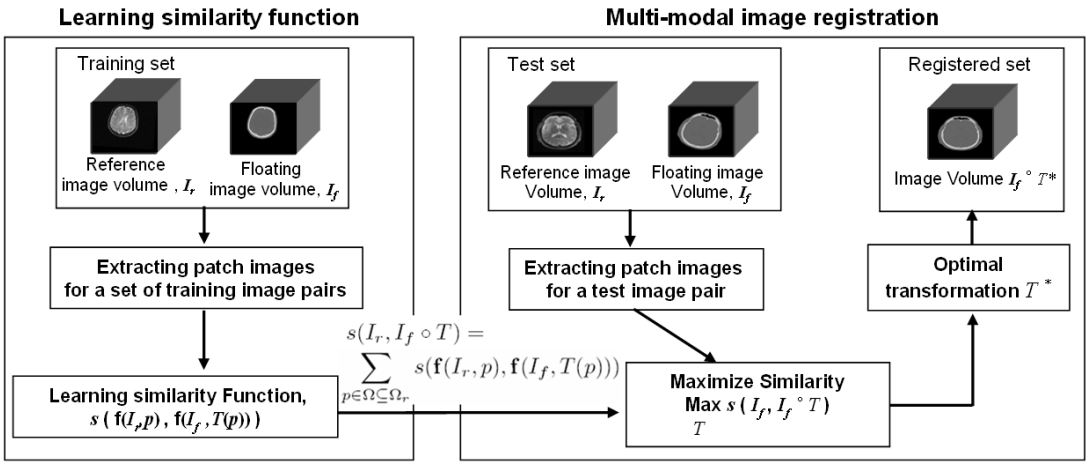

Figure 1. A flowchart of the proposed registration method using the learned similarity function.

gle pixel intensity histograms is that they give the same similarity score if the pixels in both images are randomly permuted. This is an artifact of the unrealistic modelling assumption of generative approaches that states that the intensity of the pixel in the reference image is independent of its neighbors given the corresponding pixel in the floating image. In fact, 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 in finding the corresponding pixel in the reference image than the pixel's intensity alone. Therefore, each point should be described by a whole set of features derived from the neighborhood of that point, for example, an image patch centered at that point. This can be extremely problematic for joint histogram based approaches, however, since it would require high-dimensional histograms which are generally unreliable to estimate [7]. 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.

In this paper, our goal is to develop a method that overcomes the computational and data-scarcity problems of the generative learning approaches that require joint input-output histograms over a neighborhood. To this extent, we propose a discriminative approach to learn a similarity function based on features extracted from the neighborhoods of both the reference and floating image positions. Since discriminative approaches condition on the input (reference) image, they do not impose the unrealistic conditional independence assumptions mentioned above. Furthermore, they allow the use of kernels which provide an elegant way of handling structured inputs in machine learning problems [9]. More generally, one can use *joint* kernels depending on structured input-output pairs [1], which provide an efficient way to model nonlinear dependencies between floating and reference image patches for multi-modal image registration. Joint kernels are commonly used in structured output prediction learning. In this paper, we adapt the maximum margin structured output learning method of [10] to learning similarity functions for multi-modal image registration. This formulation is preferable to other discriminative learning methods, since it provides an efficient and elegant way to incorporate joint kernel maps and cost sensitivity into the prediction problem of structured objects. We train this method with registered image patch pairs where the structured objects are the floating image patches. We define the similarity function to be the resulting discriminative function and employ it in standard registration algorithms, that search through a space of possible deformations (rigid or non-rigid) to find the deformation maximizing the learned similarity score. Our approach is outlined in Section 2.

In Sections 3 and 4, we evaluate the learned similarity function empirically. We present experiments underpinning the validity of the measure in Section 3. These experiments include visual comparisons as well as the robustness of our approach to various transformations on the training data. In Section 4 we incorporate the similarity measure in a standard rigid registration algorithm [2], allowing us to compare against MI on MR-CT and MR-PET alignment problems. We also benchmark our proposed similarity measure against some newer variants of MI that have been developed to overcome certain limitations; specifically, we compare against *normalized mutual information* (NMI), *entropy correlation coefficient* (ECC) and versions that are invariant to

the size of overlapping regions [2]. Furthermore, we benchmark against the learned joint density based measure (LJD) [5]. These analyses show that our approach outperforms the state of the art methods for multi-modal medical image registration.

2. Learning the Similarity Function

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

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,

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

Our goal is to train a similarity function s over a sample of pre-aligned image pairs such that the empirical cost Δ of misregistration, e.g. the target registration error [12], is minimized.

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

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

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. One can extend this framework further by incorporate compatibility terms $\tilde{s}$ for multiple patches of the floating image $\tilde{s}(\mathbf{f}(I_f, p), \mathbf{f}(I_f, p'))$. For example, $\tilde{s}$ can impose spatial consistency of neighboring patches after transformation. Since the standard datasets include a small number of registered full images, we restrict our attention to the formulation that ignores dependencies across floating image patches.

The optimal similarity function should give the highest score to the correctly aligned patch pairs and lower scores to the incorrectly aligned pairs. This is exactly the optimization goal of training a predictor that infers the floating image patch $\mathbf{y} = \mathbf{f}(I_f, p')$ corresponding 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 $D = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ be a sample of correctly aligned image patches. 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 $\gamma(\mathbf{x}, \mathbf{y}; w) = s(\mathbf{x}, \mathbf{y}; w) - \max_{\mathbf{y}'} s(\mathbf{x}, \mathbf{y}'; w)$ and $\mathbf{y} \neq \mathbf{y}_i$ is any structured object from the output space $\mathcal{Y}$ (the set of all patches from the floating image I_f). If we allow margin violations with linear penalties and we control the norm of w, the optimization problem can be stated as a convex program [10]

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

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; that is $\forall \mathbf{y} \in \mathcal{Y}, \mathbf{y} \neq \mathbf{y}_i$,

$$s(\mathbf{x}_i, \mathbf{y}_i; w) - s(\mathbf{x}_i, \mathbf{y}; w) \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i. \quad (4)$$

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 in 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.

Before giving further details of Algorithm 1, we compare (3) to two alternative formulations, multi-class SVM (mSVM) and a binary class SVM that discriminates the set of all $(\mathbf{x}_i, \mathbf{y}_i)$ pairs from $(\mathbf{x}_i, \mathbf{y})$ pairs for all $\mathbf{y} \neq \mathbf{y}_i$ and all i. The advantage of (3) over mSVM lies in the ability to learn across classes, which is especially important where classes (in our case floating image patches) have internal structure and the number of classes is very large. This is achieved via joint kernel maps (Section 2.2) that capture statistics within and between image modalities. It is quite possible that during test time the similarity function has to be evaluated on floating image patches that are never observed during training. The standard mSVM cannot generalize to such scenarios as opposed to the structured prediction approach. The binary classification problem stated above has been proposed in [13] and solved by a boosting algorithm. The constraints of this problem (separating

Algorithm 1 Cutting plane algorithm

1: Input: $D = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$
2: $S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$
3: Define $\mathcal{Y}_i, \forall i$ as described in Section 2.3
4: **repeat**
5: **for** $i = 1, \ldots, n$ **do**
6: $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) - s(\mathbf{x}_i, \mathbf{y}_i; w) + s(\mathbf{x}_i, \mathbf{y}; w)$
7: compute $\hat{\mathbf{y}} = \text{argmax}_{\mathbf{y} \in \mathcal{Y}_i} H(\mathbf{y})$
8: compute $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i}\}$
9: **if** $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**
10: $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
11: Solve the dual of (3) over S, $S = \cup_i S_i$
12: **end if**
13: **end for**
14: **until** no S_i has changed during iteration

all correct pairs from all incorrect pairs) are significantly harder than the constraints of (3) and are possibly infeasible. The constraints of (3) impose separation for each training instance but not across all data. Furthermore, (3) naturally incorporates the error function, Δ, into the optimization problem, which is not the case for the binary formulation. In our experiments, 0/1 loss is used for Δ function.

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 for ψ and ψ'. In each iteration of 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 $\boldsymbol{\alpha}$ 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 Algorithm 1, we have to find the most violated constraint at each iteration (see line 7 of Algorithm 1). This is computationally problematic if the whole output space $\mathcal{Y}$ is searched through exhaustively, since 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 neighborhood size 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.

In some image registration problems, the correct floating image patch for a local reference image patch can be ambiguous in the vicinity of the true positions. If the cost function Δ does not capture this ambiguity, as in the case of 0/1 loss or target registration error, it is crucial to avoid imposing margin constraints for these floating image patches, as these can lead to infeasibility of the optimization problem. To state this problem more formally, let $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}_j, \mathbf{y}_j)$ be two correctly aligned image patch pairs in D. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are very similar, i.e., $\mathbf{x}_i \approx \mathbf{x}_j$, and $\mathbf{y}_j$ is in the neighborhood of $\mathbf{y}_i$, i.e., $\mathbf{y}_j \in \mathcal{Y}_i$, then the constraints (4) require

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

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

Plugging $\mathbf{x}_i$ for $\mathbf{x}_j$ in (6) and combining with (5) yields the infeasibility problem mentioned above. We overcome this problem by explicitly removing all patches $\mathbf{y}$ such that $\mathbf{x} \approx \mathbf{x}_i$ from $\mathcal{Y}_i$ when Δ function is 0/1 loss. We define $\mathbf{x} \approx \mathbf{x}_i$ to be true if the Euclidean distance between the intensities of $\mathbf{x}$ and $\mathbf{x}_i$ is smaller than some small positive threshold 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 can yield a very large dataset that is not practical to 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, 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, during both the training and testing steps, 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

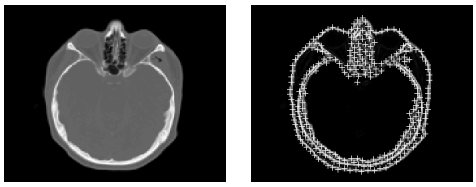

Figure 2. A CT image (left), and the locations Ω from which we extract patches for training and testing of the similarity score (right).

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 [6]. 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 simply increase the computation 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), and then we show some properties of the induced image-wise criterion (1). The description of the data sets and experimental setup for learning the similarity function is given in Sections 4.1 and 4.2, respectively.

3.1. Local similarity measure

To illustrate the learned similarity function between patches, we plot its values for a MR-CT example in Fig. 3. 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 that are 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 uniquely maximised for the correct match $(\mathbf{x}_1, \mathbf{y}_1)$. While this goal is

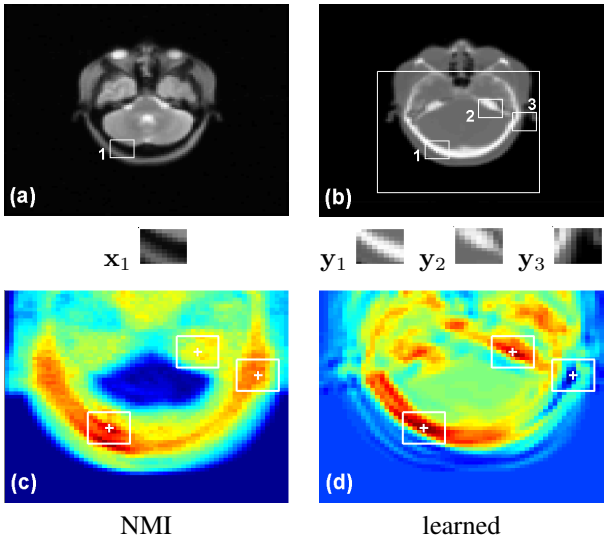

Figure 3. Comparison of local similarity values for NMI and the learned similarity measure. (a) and (b) are reference (MR) and floating (CT) image, respectively. The small rectangles below in (a) and (b) show a 2D views of the 3D patches extracted at the marked positions. (c) and (d) show the local similarity 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.

achieved by the learned similarity measure, the NMI has two maxima in the vicinity of the true match; this can easily lead to misregistration. Another shortcoming 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 counterintuitive 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. NMI was calculated both patch-wise and image-wise (Patch-wise NMI is the sum of single patch pair NMI values, similar to (1), whereas, image-wise NMI means a NMI value 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 that the learned similarity score has a unique local max-

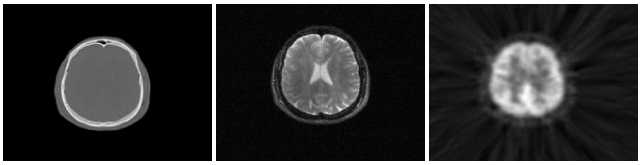

Figure 5. Axial views of RIRE brain image volumes from the patient-01 dataset. Left to right: CT, MR, and PET image.

imum at zero, which is 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 maximum for the image-wise NMI scores is obtained at positions $+1°$, $-6mm$, $+0mm$ for the rotation and translations, respectively. Thus, image-wise NMI would not yield to a precise alignment of the two datasets.

In summary, these experiments give a first, strong hint that the learned similarity score yields a more accurate and more stable criterion for registration tasks than entropy-based similarity measures. In the next section, we will further validate this claim 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 the 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 [12]. 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 different patients. Since no PET images are available for two of the seven patients, 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 physical voxel size is $0.65 \times 0.65 \times 4$ mm^3 for CT, $1.25 \times 1.25 \times 4$ mm^3 for MR, and $2.59 \times 2.59 \times 8$ mm^3 for PET images. Axial views of the CT, MR, and PET image volume of pt-01 are shown in Fig. 5.

Modality	CT to MR		PET to MR	
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.

To evaluate the accuracy of registration results for the various similarity measures, we use the *target registration error* (TRE) [12]. 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 online[1].

4.2. Experimental setup

We applied Algorithm 1 to RIRE dataset. The hyper-parameters (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 TREs for the training patient's VOIs given registrations computed with the respective parameters. The optimal parameters are reported in Table 1.

In order to obtain a fast and robust registration, we used a multi-resolution approach [11]. We resampled all images isotropically to $6 \times 6 \times 6$ mm^3 for the coarse resolution (level 2), and $3 \times 3 \times 3$ mm^3 for the finer resolution (level 1), respectively. The patch size is fixed as $30 \times 30 \times 18$ mm^3 for all resolutions and modalities. Since generalisation of s over different resolutions cannot be expected, we trained a separate similarity function for each image resolution.

The proposed learned similarity measure (Learned) is compared with several entropy-based measures: the 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 the learned joint density-based measure (LJD) [5]. Note that LJD estimates a joint density from the training image pair using the mixture of Gaussian method. For all of the comparison measures, we estimated densities (and joint densities) with histograms containing 64 (and 64x64) bins, and we accumulated samples using linear (bilinear) interpolation of histogram bins. We used the registration implementation of

[1] TRE statistics for various registration methods are available in `http://www.insight-journal.org/rire/view_results.php`

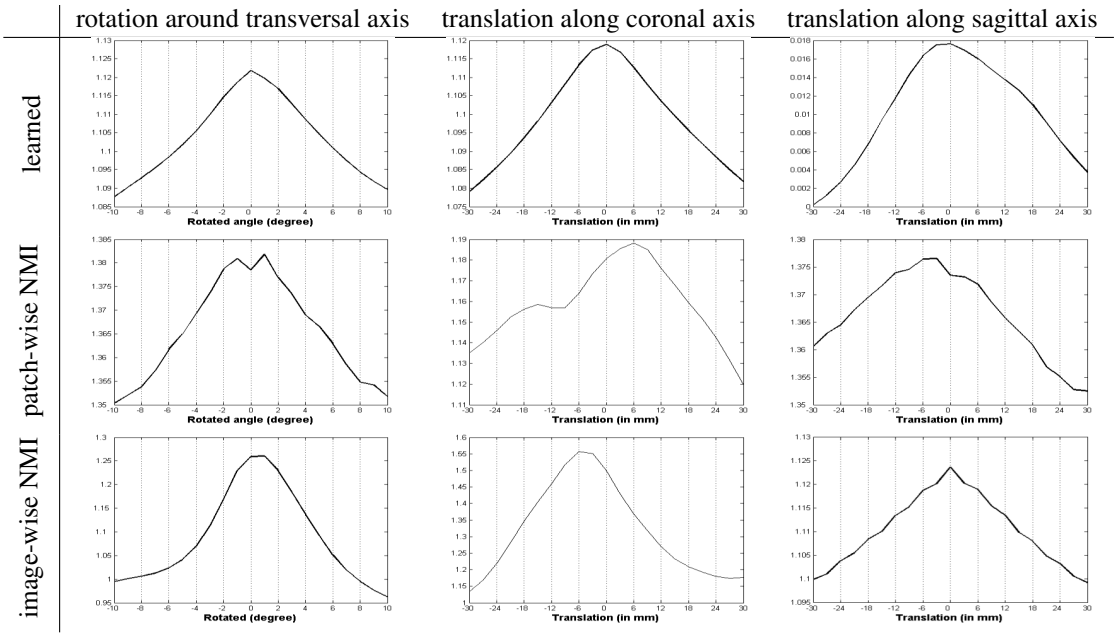

Figure 4. Image-wise similarity functions (1) for an artificial CT-MR matching task, where the CT image is transformed as in the caption of the table while the MR image is fixed. Top to bottom row represent the learned similarity function values, patch-wise NMI, and image-wise NMI, respectively.

[2] for all similarity functions.

4.3. Results

The experimental results are shown in Fig.6, with numerical values are given in Table 2. The presented values are statistics computed from the TREs of all VOIs from all test patients.

For MR-CT registrations, our learned measure outperforms all standard measures, yielding 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, 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) and shows much tighter quartile range compared to the other measures as shown in Fig.6. However, the median performance is 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 a 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 max-margin structured-output learning methods to this problem. The algorithm makes use of joint kernels for the input and the output space which provide 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. Experimental comparison on CT-MR and PET-MR registrations on brain image volumes from the RIRE project shows that our learned similarity measure outperforms other similarity measures in terms of the robustness and accuracy. In our future work, we plan to investigate the use of more sophisticated feature functions, different joint kernel maps, and different Δ functions. We also plan to apply this approach to non-rigid registration.

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	3.23	1.40
	Median	1.98	1.99	2.89	2.62	4.50	2.86	2.90	2.41	1.29
	Max	5.15	6.10	12.50	9.38	17.74	10.16	10.07	6.32	3.32
PET/MR	Mean	8.19	3.00	3.20	3.10	2.96	3.34	3.29	3.07	2.60
	Median	5.16	2.37	2.64	2.54	2.40	3.01	2.95	2.56	2.52
	Max	37.18	7.71	7.57	7.65	7.50	7.53	7.47	7.56	4.81

Table 2. Statistics of VOI TREs (in mm) across all test patients' image volumes.

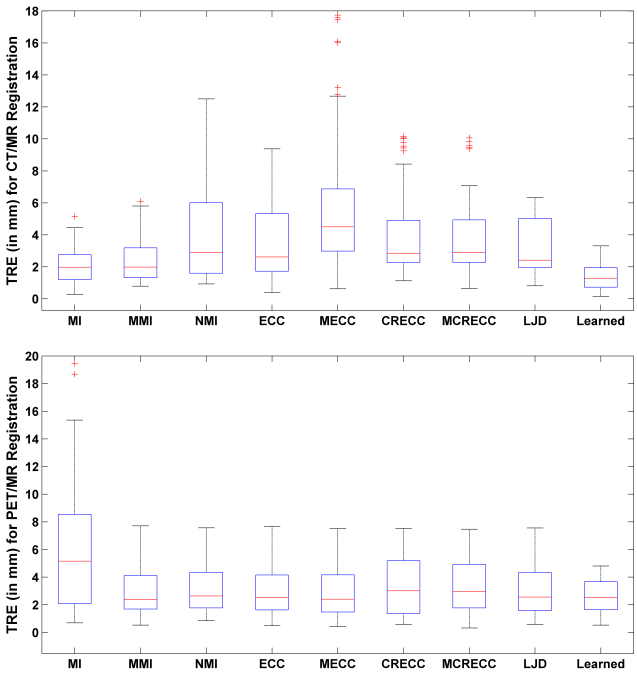

Figure 6. Box and whisker plot of the TREs for the RIRE data and different similarity measures. The box has lines at the lower quartile, median, and upper quartile values. Whiskers extend from each end of the box to the adjacent values in the data; the most extreme values within 1.5 times the interquartile 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.

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. Schölkopf, 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 multimodal image registration. In *IEEE Symposium on Biomedical Imaging*, 2006.

[4] A. C. S. Chung, R. Gan, and W. M. Wells 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] M. Leventon and W. Grimson. Multi-modal volume registration using joint intensity distribution. In *Proceedings of MICCAI*, pages 1057–1066, 1998.

[6] 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.

[7] 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.

[8] M. Sabuncu and P. Ramadge. Using spanning graphs for efficient image registration. *IEEE Transactions on Image Processing*, pages 788–797, 2008.

[9] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[10] 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.

[11] P. A. Viola, W. M. Wells III, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, pages 5–51, 1996.

[12] 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.

[13] 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.