Learning Optimization Updates for Multimodal Registration

Chapter · October 2016

DOI: 10.1007/978-3-319-46726-9_3

Benjamín Gutierrez Becker
Ludwig-Maximilians-University of Munich
16 PUBLICATIONS **28** CITATIONS

SEE PROFILE

Diana Mateus
Technische Universität München
63 PUBLICATIONS **728** CITATIONS

SEE PROFILE

Loic Peter
Technische Universität München
20 PUBLICATIONS **139** CITATIONS

SEE PROFILE

Learning Optimization Updates for Multimodal Registration

B. Gutierrez-Becker, L. Peter, D. Mateus, N. Navab

Chair for Computer Aided Medical Procedures, Technische Universität München,
Boltzmanstrasse. 3, 85748 Garching, Germany
{becker,peter,mateus,navab}@in.tum.de
http://campar.in.tum.de/

Abstract. We address the problem of multimodal image registration using a supervised learning approach. We pose the problem as a regression task where we aim to estimate the unknown transformation from the joint appearance of both fixed and moving images. Our method is based on i) context-aware features, which allow us to guide the registration using not only local structural information, but also global appearance, and ii) regression forests to map the very large feature space to transformation parameters. Our approach improves the capture range as shown in experiments on the publicly available IXI dataset. Furthermore, it also allows us to perform multimodal registration in difficult settings where other similarity metrics tend to fail, as demonstrated by the registration of Intravascular Ultrasound (IVUS) and Histology images.

Keywords: Image Registration, Machine Learning

1 Introduction

A core difficulty in multimodal registration is the lack of a general law to measure the alignment between images of the same organ acquired with different physical principles. The unknown relationship between the image intensities is in general neither linear nor bijective. Following Sotiras *et al.* [14], there have been three main approaches to address the problem: i) *information theoretic* methods [12], ii) mapping of the modalities to a *common representation* [2, 3], and iii) *learning* multimodal similarity measures [9, 10]. This paper relates to the latter category, whose main assumption is that prior knowledge (in the form of examples of aligned images) can be afforded. This extra effort can be justified both in cases where large-scale databases need to be registered or for difficult modalities for which general similarity measures do not suffice.

Up to now, the focus of learning based approaches has been on approximating multimodal similarity measures, independent of the optimization scheme used during the registration task itself. However, due to the usually complex mapping between the intensities of the two modalities, non-linear relations and ambiguities tend to shape local-optima and plateaus in the energy landscape. Therefore, the optimizer plays an important role in the success of the registration. Despite no guarantee of convergence, using gradient based methods to optimize such non-convex functions is common and may lead to good results for well behaved data. In this work we explore a combined view of the problem, where we assume a gradient-based optimizer will be used, and focus on directly inferring the motion from changes in the joint visual content of the images. We model the problem as a regression approach, where for a given pair of misaligned images the goal is to retrieve the global direction towards which the motion parameters should be updated for correct alignment. In order to ensure that the direction of the update points towards a globally optimal solution, we describe the images taking into account both their local appearance and their long-range context, by means of Haar-like features [1]. These features allow our method to rely not only on the local structure of the images but also on contextual information in a larger scale. In order to efficiency handle the resultant very high-dimensional feature space, we use regression forests [1], also known for their fast training and testing. The main contribution of our work is twofold: 1) this is the first time a regression method is used to predict registration updates in the multimodal setting; 2) the use of long-range context-aware features instead of local structural features is novel for the problem of multimodal registration. We demonstrate the advantages of our method in the difficult case of 2-D deformable registration of histological to intravascular ultrasound images (IVUS). We also perform a quantitative evaluation for the 3-D registration of T1-T2 MR images showing an advantageous increase in the capture range.

1.1 Related Work

There have been two trends in learning based methods for multimodal registration. *Generative* approaches [13], approximate the joint intensity distribution between the images to be registered and minimize the difference of a new test pair of images to the learned distribution. *Discriminative* methods, on the other hand, model the similarity learning problem as the classification of positive (aligned) and negative (misaligned) examples, typically at patch level [4, 9, 10].

Different learning strategies have been explored to approximate such patch-wise similarities, including margin-based approaches [9] and boosting [10]. In contrast to the discriminative approaches above, which aim at discerning between aligned and misaligned patches, we focus on learning a motion predictor that guides the registration process towards alignment.

There have been prior attempts of using motion prediction for monomodal tracking and registration. For instance, Jurie *et al.* [5] proposed a linear predictor for template tracking, which related the difference between the compared images to variations in template position. In the medical domain, Chou *et al.*chou20132d present an approach to learn updates of the transformation parameters in the context of 2D-3D registration. l Similarly, in [8], 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. The work presented here is, to the best of our knowledge, the first approach for motion prediction in the multimodal case. Such predictions, as shown in the experiments, save computational time and improve the capture range of the registration.

2 Method

Multimodal registration aims to find the transformation $\mathcal{W}^*$ that optimally aligns a pair of images $\mathbf{I}$ and $\mathbf{I}'$ from different modalities, with $\mathbf{I}, \mathbf{I}' : \varOmega, \varOmega' \subset \mathbb{R}^3 \to \mathbb{R}$. A common method to find $\mathcal{W}^*$ is by maximizing a similarity function $S(\mathbf{I}, \mathbf{I}')$ between the two images. In the general case, the transformation $\mathcal{W}$ is defined over the image domain and depends on a set of parameters $\mathbf{p} \in \mathbb{R}^{N_p}$, such that $\mathcal{W}_{\mathbf{p}}(\mathbf{I}')$ describes the intensity of the moving image after applying the transformation $\mathcal{W}$. In this way:

$$\mathbf{p}^* = \max_{\mathbf{p}} S(\mathbf{I}, \mathcal{W}_{\mathbf{p}}(\mathbf{I}')). \tag{1}$$

The maximization of Eq. 1 can be done either by gradient-free (usually preferred for discriminatively learned implicit similarities) or gradient-based optimization approaches. In the latter, the gradient of S is computed to iteratively estimate the parameter update, such that $\mathbf{p}_k = \mathbf{p}_{k-1} + \boldsymbol{\varDelta}_k$. In a typical steepest-ascent-like strategy, the search direction is determined in terms of the similarity gradient as $\boldsymbol{\varDelta}_k = -\frac{\partial S/\partial \mathbf{p}}{\|\partial S/\partial \mathbf{p}\|}$, which is in turn obtained based on the local approximation of this gradient. Depending on the shape of the similarity such local approximations may be poor and lead to local optima or slow convergence rates.

Here, we reformulate the multimodal registration problem as that of motion prediction, by directly learning a function $F : \mathbb{R}^{N_{\text{vox}}} \times \mathbb{R}^{N_{\text{vox}}} \to \mathbb{R}^{N_p}$ that maps the intensity variations of all voxels *vox* observed in the images to the corresponding correct motion update:

$$\widehat{\varDelta} = F(\mathbf{I}, \mathcal{W}_{\mathbf{p}}(\mathbf{I}')). \tag{2}$$

We learn F from labeled examples (images with a known misalignment), which allows us to enforce desirable properties for the optimization, namely: a parameter update pointing in the direction of the global maximum and a smooth gradient. In analogy to the steepest ascent approach, our update may be seen as a global approximation of the gradient $\frac{\partial S}{\partial \mathbf{p}}$. We explain next how to approximate F from a training set of images by means of regression.

2.1 Learning Multimodal Motion Predictors.

We choose to model the motion predictor at the image level F as the aggregation of motion predictors at the local level $f(\mathbf{x}) : \mathbb{R}^{N_{\text{vox}}} \times \mathbb{R}^{N_{\text{vox}}} \to \mathbb{R}^{N_p}$, still defined over the joint image space but relative to a point $\mathbf{x} \in \mathbb{R}^3$. We consider that the input to these local predictors are not patch intensities but rather a joint feature representation $\theta(\mathbf{x}, \mathbf{I}, \mathbf{I}') : \mathbb{R}^3 \times \mathbb{R}^{N_{\text{vox}}} \times \mathbb{R}^{N_{\text{vox}}} \to \mathbb{R}^H$. This vector, hereafter denoted $\theta(\mathbf{x})$, encodes the appearance of $\mathbf{I}$ and $\mathbf{I}'$ relative to the point $\mathbf{x}$.

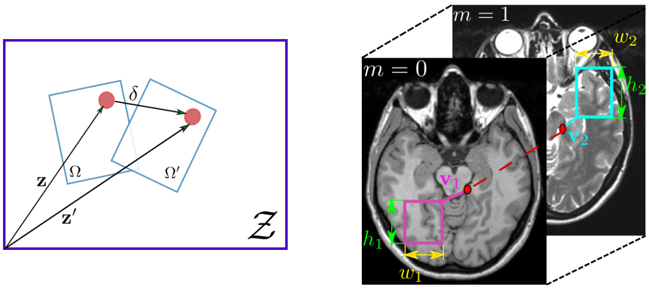

Fig. 1. Left: Learned displacement under a given transformation. Right: Long-range Haar-like features to encode local and long range context.

Given a number N_{im} of aligned multimodal images $\{\mathbf{I}_i, \mathbf{I}'_i\}_{i=1}^{N_{\text{im}}}$ our aim is then to approximate a local displacement $\boldsymbol{\delta}(\mathbf{x})$ by means of a learning-based regression approach $f(\mathbf{x}) : \theta(\mathbf{x}) \mapsto \boldsymbol{\delta}(\mathbf{x})$. Next, we describe the details of the method.

Generating a Set of Training Labels. To generate examples with known misalignment, we apply multiple known transformations $\{\mathcal{W}_j, \mathcal{W}'_j\}_{j=1}^{N_{\text{transfo}}}$ to the initially aligned images, mapping the coordinates of two originally corresponding points $\mathbf{x} \in \Omega$ and $\mathbf{x}' \in \Omega'$ to distinct locations in a common image domain $\mathbf{z}, \mathbf{z}' \in \mathcal{Z} \subset \mathbb{R}^3$ (see Fig.2) We then randomly sample points $\{\mathbf{z}_n\} \in \mathcal{Z}$ from the resultant augmented set $\{\tilde{\mathbf{I}}_{ij}, \tilde{\mathbf{I}}'_{ij}\}$. Because the applied transformations are known, we can determine the displacement $\boldsymbol{\delta}_n \in \mathbb{R}^3$ needed to find the originally corresponding point $\mathbf{z}'_n$ in the moving image, and bring it into alignment with $\mathbf{z}$, *i.e.* $\boldsymbol{\delta}_n = \mathbf{z}'_n - \mathbf{z}_n$. With this information we build the training set $\mathcal{X} = \{\theta(\mathbf{z}_n), \boldsymbol{\delta}_n\}_{n=1}^{N_{\text{points}}}$. Notice that we have chosen to use $\boldsymbol{\delta}_n$ as the regression targets instead of the transformation parameters. In this way the learning stage is independent of the motion parametrization. In fact, these displacements play the role of the similarity gradients $\frac{\partial S}{\partial \mathbf{z}}$, which can be then related to a given parametrization using the chain rule and by means of the Jacobian $J = \frac{\partial \mathbf{z}}{\partial \mathbf{p}}$ such that $\frac{\partial S}{\partial \mathbf{p}} = \frac{\partial S}{\partial \mathbf{z}} J$.

Context aware features We characterize the cross-modal appearance of each point $\mathbf{z}_n$ in the training set, by a variation of the context-aware Haar-like features [1]. These features permit to effectively capture how the joint-appearance variations in the vicinity of each point relate to different transformation parameters. The feature vector $\theta(\mathbf{z}_n)$ is a collection of H features $[\theta_1, \ldots, \theta_h, \ldots, \theta_H]^\top$; where each θ_h is computed as a simple operation on a pair of boxes located at a given offsets locations relative to point $\mathbf{z}_n$. More formally, θ_h is characterized by two boxes $\mathbf{b_1}, \mathbf{b_2}$ (*c.f.* fig.2), parametrized by their location ($\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^3$), size ($w_1, h_1, w_2, h_2, d_1, d_2 \in \mathbb{R}$), modality $m_1 = \{0, 1\}$ and an *operation* between boxes: $\{\, \overline{\mathbf{b_1}}, \overline{\mathbf{b_2}}, \overline{\mathbf{b_1}} + \overline{\mathbf{b_2}}, \overline{\mathbf{b_1}} - \overline{\mathbf{b_2}}, |\overline{\mathbf{b_1}} - \overline{\mathbf{b_2}}|, \overline{\mathbf{b_1}} > \overline{\mathbf{b_2}}\}$, where the overline denotes the mean over the box intensities. These operations are efficiently calculated with precomputed integral volumes [1]. The binary *modality* parameters m_1 and m_2 determine whether the two boxes are taken from the same modality or across modalities, thereby modeling the spatial context of each image as well as the functional relation between the two modalities. Using different offsets and box sizes enables to capture the visual context of each point without explicitly determining the scale. If we consider the combinatorial nature of the box parameters we face a very-large feature space $\mathbb{R}^H$. To deal with it, we use regression forests, which among other advantages does not require the pre-computation of features.

Regression forest Using the features described above, we characterize each point $\mathbf{z}_i$ in the training set $\mathcal{X}$ by its corresponding feature vector $\theta_i(\mathbf{z}_n)$. We then use regression forests to approximate the function $f : \theta(\mathbf{z}_n) \mapsto \boldsymbol{\delta}_n$ mapping these feature vectors to an estimation of the target displacements. We train our regression forest in a standard setting, using as splitting criteria the reduction of the covariance trace associated to the target values in a particular node. Once the forest grown, we store the Gaussian distribution (mean $\boldsymbol{\mu}_{t(l)}$ and covariance $\boldsymbol{\Sigma}_{t(l)}$) of the target displacements vectors falling in each leaf l. At test time, a new feature vector $\theta(\mathbf{z}_{\text{test}})$ is passed down through the forest. Every tree assigns an estimate of the predicted motion $\hat{\boldsymbol{\delta}}_t$ (given by the mean vector $\boldsymbol{\mu}_{t(l)}$ stored in the leaf) along with its covariance $\boldsymbol{\Sigma}_{t(l)}$. We then rank and select the $\tilde{N}_{\text{trees}}$ with the smaller values of covariance trace. The predicted displacement at point $\mathbf{z}_{\text{test}}$ is obtained as the average over the prediction of the selected trees.

2.2 Using Multimodal Motion Predictors for Registration

To register a pair of images $\mathbf{I}$ and $\mathbf{I}'$ we define a set of testing points on a grid $\{\mathbf{z}_m\}_{m=1}^{N_{\text{test}}} \in \mathcal{Z}$, extract their feature vectors $\{\theta(\mathbf{z}_m)\}_{m=1}^{N_{\text{test}}}$, and pass them through the forest to obtain the displacement estimates $\{\hat{\boldsymbol{\delta}}_m\}_{m=1}^{N_{\text{test}}}$. We then update the global transformation parameters (*c.f.* Eq. 2) by adding the contributions of each local displacement to the transformation parameters: $\boldsymbol{\Delta} = \sum_{m=1}^{N_{\text{test}}} \hat{\boldsymbol{\delta}}_m J$ where J corresponds to the Jacobian of the transformation.

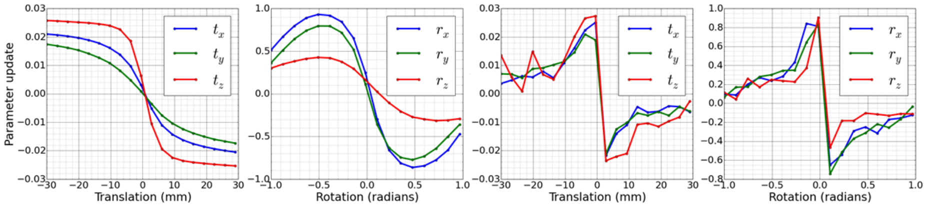

Fig. 2. Left side: Parameter updates obtained using our motion estimation method. Right side: Parameter updates obtained using the gradient of Normalized Mutual Information. Our estimated parameter updates present a smoother behavior over a larger range.

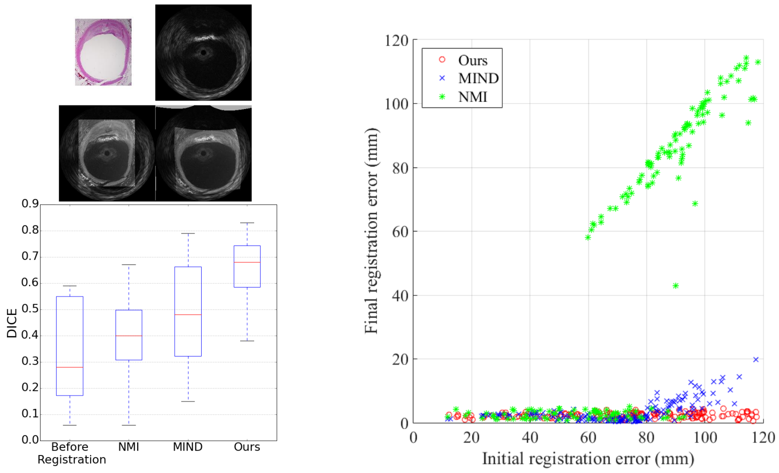

Fig. 3. Top left: Registration results on an IVUS-Histology pair. The initial unregistered images are shown as well as the overlay between the images before and after registration. Bottom left: DICE scores on the overlap between stenosis regions before and after registration. Right:Final registration error given different starting initial conditions on the T1-T2 image pairs of the IXI dataset.

3 Experiments and Results

To evaluate the performance of our method in comparison to previous registration approaches we performed two series of experiments. In the first we evaluate the performance of our method in a challenging multimodal setting: the registration of IVUS-Histology images, using the dataset from [6]. In the second, we use T1-T2 images from the IXI Dataset [1] to evaluate the capture range of our method, where we measure the registration accuracy for varying initial displacements between the fixed and moving image. This experiment shows the robustness of our method to different initial conditions.

In both cases, we compare our method to the widely used Normalized Mutual Information (NMI) [15] optimized using a gradient descent optimizer and with the Modality Independent Neighborhood Descriptor (MIND) [2] coupled with the Gauss-Newton optimization suggested by the authors.

In all the experiments we used forests consisting of 40 trees, keeping the top 10 best trees during testing. We evaluated 1000 possible splits per node and grew the trees to a maximum depth of 15, stopping earlier if not enough samples reached one of the child nodes. We limited the size of the offsets and the boxes in the feature space to half of the size of the image. To optimise the scale of these features we used the scale adaptive forest training approach presented in [11].

[1] Available at: http://brain-development.org/ixi-dataset/

3.1 IVUS-Histology Deformable Registration

In this experiment we tackled the registration between 10 Intravascular Ultrasound images (IVUS) and histological slices. We used the method in [6] [7] to obtain the initial set of aligned images needed for training. For evaluation we performed deformable registrations using our method and we compare to MI and MIND by measuring the overlap (DICE) of segmented stenosis regions both in IVUS and the histology images. For all methods we use the same 3rd-order b-spline parametrization with 5 nodes per dimension.

During training we split the dataset in 2 groups of 5 images and perform cross validation. The final registration results are shown in fig. 3. This dataset is particularly challenging because the underlying assumptions of most similarity metrics, like local structural similarities or relationships between statistics on the intensities of the images, are not strong enough. The methods we used for comparison therefore presented high registration errors for the IVUS-Histology pairs. Our supervised approach, on the other hand, was capable to register the images thanks to prior knowledge and the non-local context of each point.

3.2 Capture Range

To test the capture range, we take a set of 10 prealigned T1-T2 image pairs from the IXI dataset splitting them in 2 groups of 5 images for cross validation. For each image pair we apply a rigid transformation to one of the images and then we find the transformation that brings it back into alignment. The applied transformations were in the range of ±100 mm for translations along each axis and $\pm\pi/2$ radians for rotations. We repeat this procedure 20 times per image with different transformations for a total of 200 registration evaluations.

The results of this experiment can be seen in Fig. 3. Each point in the plot corresponds to the registration of a pair of images. We can clearly observe that our method presents a larger capture range than the metrics we compared with. Note that there is a breaking point where MIND and MI start to fail, as these metrics tend to underperform when the overlap between images is small and no local structure can be used to evaluate the metrics reliably. Our method on the other hand, was able to register the images even when they had no overlap, thanks to the prior knowledge and the use of context aware features which together to pull the optimizer in the right direction.Additionally, our method was able to converge in a smaller number of iterations (5) compared to NMI ($\sim$ 250 gradient ascent iterations) and MIND (16 iterations). In terms of computational time our method performed each registration in an average of $\sim$10 secs compared to $\sim$200 secs for NMI and $\sim$35 secs for MIND. The faster convergence can be attributed to the smoothness of our parameter updates in comparison to the updates estimated using the derivative of NMI (see Fig. 2). In this way, we are entitled to use a more aggressive step size without a decrease on the final registration error and makes it less dependent on the initial misalignment between images.

4 Conclusions

We present a novel approach to the problem of multimodal registration, which combines supervised regression with simple gradient-based optimizers. Supervised regression let us infer motion from changes in the visual appearance of the images to be registered. In this way, it is no longer necessary to rely on prior assumptions about local appearance correlations. Although our method requires the use of aligned images for training, we have observed that the required amount of training images to achieve good results is reasonably small (not more than 5 images in each case). Building datasets with aligned multimodal images requires additional effort, but this extra effort can be justified in cases where other metrics are not sufficient or when a large dataset of similar images has to be registered. For more common scenarios (such as multimodal MR registration), our method produces registrations with comparable accuracy to other similarities but with faster convergence and a larger capture range.

References

1. A Criminisi, J Shotton, and S Bucciarelli. Decision forests with long-range spatial context for organ localization in ct volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 69–80. Citeseer, 2009.
2. M. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. Schnabel. MIND: Modality independent neighbourhood descriptor for multimodal deformable registration. *Med. Image Analysis*, 2012.
3. M. P. Heinrich, M. Jenkinson, B. Papiez, M. Brady, and J. Schnabel. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *MICCAI*, pages 187–194, 2013.
4. J. Jiang, S. Zheng, A. Toga, and Zhuowen T. Learning based coarse-to-fine image registration. In *CVPR*, pages 1–7, June 2008.
5. F. Jurie and M. Dhome. Hyperplane approximation for template matching. *TPAMI*, 24(7):996–1000, July 2002.
6. A. Katouzian, A. Karamalis, J. Lisauskas, A. Eslami, and N. Navab. Ivus-histology image registration. In *Biomedical Image Registration*. Springer, 2012.
7. Amin Katouzian, Shashidhar Sathyanarayana, Wenguang Li, Tom Thomas, and Stéphane G Carlier. Challenges in tissue characterization from backscattered intravascular ultrasound signals. In *Medical Imaging*, pages 65130O–65130O. International Society for Optics and Photonics, 2007.
8. M. Kim, G. Wu, P. T. Yap, and D. Shen. A general fast registration framework by learning deformation appearance correlation. *IEEE Transactions on Image Processing*, 21(4):1823–1833, April 2012.
9. D. Lee, M. Hofmann, F. Steinke, Y. Altun, N.D. Cahill, and B. Scholkopf. Learning similarity measure for multi-modal 3d image registration. In *CVPR*, 2009.
10. F. Michel, M. Bronstein, A. Bronstein, and N. Paragios. Boosted metric learning for 3d multi-modal deformable registration. In *ISBI*, pages 1209–1214. IEEE, 2011.
11. L. Peter, O. Pauly, P. Chatelain, D. Mateus, and N. Navab. Scale-adaptive forest training via an efficient feature sampling scheme. In *MICCAI*. Springer, 2015.
12. J. Pluim, J. Maintz, and M. Viergever. f-information measures in medical image registration. *TMI*, 23(12):1508–1516, 2004.

13. M.R. Sabuncu and P. Ramadge. Using spanning graphs for efficient image registration. *TMI*, 17(5):788–797, May 2008.
14. A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *TMI*, 32(7):1153–1190, July 2013.
15. P. Viola and W.M.I.I.I. Wells. Alignment by maximization of mutual information. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 16–23, Jun 1995.