Learning Optimization Updates for Multimodal Registration

Benjamín Gutiérrez-Becker(✉), Diana Mateus, Loïc Peter, and Nassir Navab

Computer Aided Medical Procedures, Technische Universität München,
Munich, Germany
gutierrez.becker@tum.de, {mateus,peter,navab}@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, whose goal is to estimate the unknown geometric transformation from the joint appearance of the fixed and moving images. Our method is based on (i) context-aware features, which allow us to guide the registration using not only local, but also global structural information, and (ii) regression forests to map the very large contextual feature space to transformation parameters. Our approach improves the capture range, as we demonstrate on the publicly available IXI dataset. Furthermore, it can also handle difficult settings where other similarity metrics tend to fail; for instance, we show results on the deformable registration of Intravascular Ultrasound (IVUS) and Histology images.

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.* [15], there have been three main approaches to address the problem: (i) *information theoretic* methods [13], (ii) mapping of the modalities to a *common representation* [3,4], and (iii) *learning* multimodal similarity measures [10,11]. 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 when the two modalities are so different that general multi-modal 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-linearities and ambiguities tend to shape local-optima and plateaus in the energy landscape. Thereby, the optimizer plays an important role in the success of the registration. In this work we explore a combined view of the problem, where we take the optimizer into account. In particular, we restrict ourselves to gradient-based methods, and focus on directly inferring the motion parameters from changes in the joint visual

S. Ourselin et al. (Eds.): MICCAI 2016, Part III, LNCS 9902, pp. 19–27, 2016.
DOI: 10.1007/978-3-319-46726-9_3

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 [2]. In order to efficiency handle the resultant very high-dimensional feature space, we use regression forests [2], 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 [14], 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 [5,10,11]. Different learning strategies have been explored to approximate such patch-wise similarities, including margin-based approaches [10] and boosting [11]. 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.* [6] 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.* [1] present an approach to learn updates of the transformation parameters in the context of 2D-3D registration. Similarly, in [9], 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.

2 Method

Multimodal registration aims to find the transformation $\mathcal{W}(\mathbf{p})$ that optimally aligns two images of different modalities, namely, a fixed image $\mathbf{I} : \Omega \subset \mathbb{R}^3 \to \mathbb{R}$

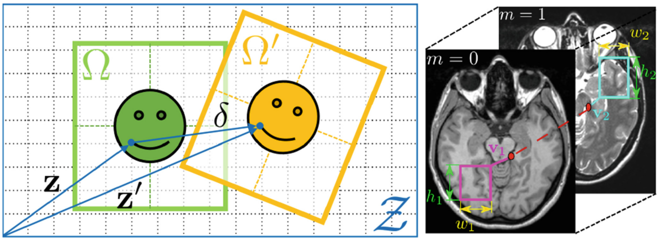

Fig. 1. Left: Learned displacement under a given transformation. Right: long-range Haar-like features to encode local and long range context.

and a moving image $\mathbf{I}' : \Omega' \subset \mathbb{R}^3 \rightarrow \mathbb{R}$. A common method to find the optimal parameters $\mathbf{p} \in \mathbb{R}^{N_p}$ that define $\mathcal{W}$ is by maximizing a similarity function $S(\mathbf{I}, \mathbf{I}')$ between the two images. Denoting by $\mathbf{I}'_{\mathbf{p}}$ the moving image resampled in the fix domain Ω according to parameters $\mathbf{p}$, we have:

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

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 $\Delta_k \in \mathbb{R}^{N_p}$, such that $\mathbf{p}_k = \mathbf{p}_{k-1} + \Delta_k$, where k is the iteration index. In a typical steepest-ascent-like strategy, the update direction is determined in terms of the similarity gradient as $\Delta_k = -\frac{\partial S/\partial \mathbf{p}}{\|\partial S/\partial \mathbf{p}\|}$, which is in turn obtained based on the *local* approximation of this gradient. Depending on the 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 learning a motion predictor, *i.e.*a function that directly maps the intensity configuration of the two images in the fixed space to the corresponding motion update:

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

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 local motion predictors f. We consider that the input to these local predictors are not patch intensities but rather a joint feature representation

$\Theta(\mathbf{z}, \mathbf{I}, \mathbf{I}'_{\mathbf{p}}) \in \mathbb{R}^H$, which describes the local appearance of $\mathbf{I}$ and $\mathbf{I}'_{\mathbf{p}}$ relative to a point $\mathbf{z} \in \mathbb{R}^3$. Hereafter, we denote the feature vector $\Theta(\mathbf{z})$ for simplicity. Given a number N_{im} of aligned multimodal images $\{\mathbf{I}_i, \mathbf{I}'_i\}_{i=1}^{N_{\mathrm{im}}}$ our aim is to approximate a function $f(\mathbf{z}) : \Theta(\mathbf{z}) \mapsto \boldsymbol{\delta}$ capable of predicting a local displacement $\boldsymbol{\delta} \in \mathbb{R}^3$ towards alignment. The approximation of f is done by means of a learning-based regression approach. In the following, we describe the details of our method.

Generating a Set of Training Labels. To generate examples with known misalignment, we apply multiple known transformations $\{\mathcal{W}_j, \mathcal{W}'_j\}_{j=1}^{N_{\mathrm{transfo}}}$ 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. 1). Because the applied transformations are known, we can determine the ground truth 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 $\mathcal{X} = \{\Theta(\mathbf{z}_n), \boldsymbol{\delta}_n\}_{n=1}^{N_{\mathrm{points}}}$. 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 $\frac{\partial S}{\partial \mathbf{p}} = \frac{\partial S}{\partial \mathbf{z}} J$, by means of the Jacobian $J = \frac{\partial \mathbf{z}}{\partial \mathbf{p}}$.

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 [2]. These features 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 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. 1), 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 [2]. 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 capturing 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 do 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

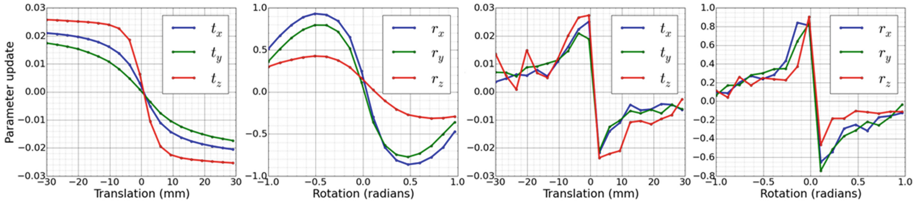

Fig. 2. Left side: parameter updates obtained using our motion estimation method. Right side: parameter updates obtained using the gradient of normalized mutual information. Our estimated parameter updates are smoother over a larger range.

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 has 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 local displacement estimates $\{\hat{\boldsymbol{\delta}}_m\}_{m=1}^{N_{\text{test}}}$. We then compute the global update (*c.f.* Eq. 2) by adding the contributions of each local displacement to the transformation parameters: $\hat{\Delta} = \sum_{m=1}^{N_{\text{test}}} \hat{\boldsymbol{\delta}}_m J$ where J corresponds to the Jacobian of the transformation.

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 [7]. 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.

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

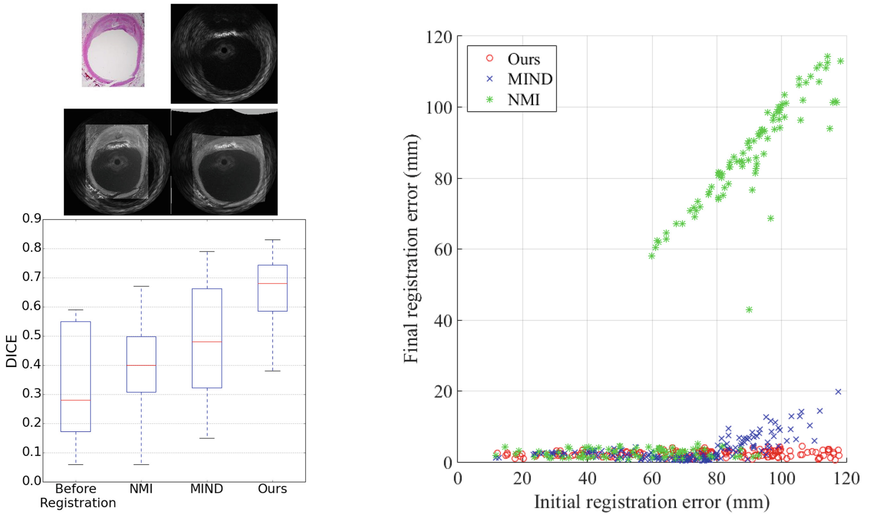

Fig. 3. Top left: registration results on an IVUS-Histology pair. The initial unregistered images are shown as well as the overlay between the images before and after registration. Bottom left: DICE scores on the overlap between stenosis regions before and after registration. Right: final registration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

In both cases, we compare our method to the widely used Normalized Mutual Information (NMI) [16] optimized using a gradient descent optimizer and with the Modality Independent Neighborhood Descriptor (MIND) [3] 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 image size. To optimise the scale of these features we used the scale adaptive forest training approach presented in [12].

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 [7] [8] 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 verified. 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 s compared to $\sim$200 s for NMI and $\sim$35 s 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 depend less 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.

Acknowledgements. Benjamin Gutierrez thanks CONACYT and the DAAD for their financial support.

References

1. Chou, C.-R., Frederick, B., Mageras, G., Chang, S., Pizer, S.: 2D/3D image registration using regression learning. Comput. Vis. Image Underst. **117**(9), 1095–1106 (2013)
2. Criminisi, A., Shotton, J., Bucciarelli, S.: Decision forests with long-range spatial context for organ localization in CT volumes. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 69–80. Citeseer (2009)
3. Heinrich, M., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.: MIND: Modality independent neighbourhood descriptor for multimodal deformable registration. Med. Image Anal. **16**, 1423–1435 (2012)
4. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 187–194. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40811-3_24
5. Jiang, J., Zheng, S., Toga, A., Tu, Z.: Learning based coarse-to-fine image registration. In: CVPR, pp. 1–7, June 2008
6. Jurie, F., Dhome, M.: Hyperplane approximation for template matching. TPAMI **24**(7), 996–1000 (2002)
7. Katouzian, A., Karamalis, A., Lisauskas, J., Eslami, A., Navab, N.: IVUS-histology image registration. In: Dawant, B.M., Christensen, G.E., Fitzpatrick, J.M., Rueckert, D. (eds.) WBIR 2012. LNCS, vol. 7359, pp. 141–149. Springer, Heidelberg (2012). doi:10.1007/978-3-642-31340-0_15
8. Katouzian, A., Sathyanarayana, S., Li, W., Thomas, T., Carlier, S.: Challenges in tissue characterization from backscattered intravascular ultrasound signals. In: Medical Imaging, p. 6513. International Society for Optics and Photonics (2007)
9. Kim, M., Wu, G., Yap, P.T., Shen, D.: A general fast registration framework by learning deformation appearance correlation. IEEE Trans. Image Process. **21**(4), 1823–1833 (2012)
10. Lee, D., Hofmann, M., Steinke, F., Altun, Y., Cahill, N.D., Scholkopf, B.: Learning similarity measure for multi-modal 3D image registration. In: CVPR (2009)
11. Michel, F., Bronstein, M., Bronstein, A., Paragios, N.: Boosted metric learning for 3D multi-modal deformable registration. In: ISBI, pp. 1209–1214. IEEE (2011)
12. Peter, L., Pauly, O., Chatelain, P., Mateus, D., Navab, N.: Scale-adaptive forest training via an efficient feature sampling scheme. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 637–644. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24553-9_78
13. Pluim, J., Maintz, J., Viergever, M.: f-information measures in medical image registration. TMI **23**(12), 1508–1516 (2004)

14. Sabuncu, M.R., Ramadge, P.: Using spanning graphs for efficient image registration. TMI **17**(5), 788–797 (2008)
15. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: a survey. TMI **32**(7), 1153–1190 (2013)
16. Viola, P., Wells III, W.M.: Alignment by maximization of mutual information. In: IEEE International Conference on Computer Vision (ICCV), pp. 16–23, June 1995