Medical Image Analysis 000 (2017) 1–16

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# Guiding multimodal registration with learned optimization updates

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

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

ARTICLE INFO

*Article history:*
Received 22 February 2017
Revised 1 May 2017
Accepted 3 May 2017
Available online xxx

*Keywords:*
Image registration
Machine learning
Multimodal registration
Motion estimation

ABSTRACT

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



## 1. Introduction

Multimodal image registration is a fundamental task in medical image analysis, consisting in the alignment of two images of a given anatomical location acquired with different modalities. Multimodal registration is an important tool in clinical diagnosis, image-guided interventions, medical augmented reality, as well as in the validation of new imaging modalities (Markelj et al., 2012; Navab et al., 2016). In all these applications multimodal registration plays the key role of bringing and presenting complementary information in a spatially consistent way. In addition to the challenges of the monomodal case, multimodal registration has to deal with the potentially large appearance differences that result from each modality's acquisition principles. As the relation between the intensities from the two modalities is unknown and often neither linear nor bijective, an open question is the definition of a general energy function capable of relating the two modalities and guiding a multi-modal registration algorithm.

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

Unfortunately, explicitly defining a general and well-behaved energy function that models the unknown intensity relationship between the two modalities is not straightforward. Current multi-modal similarity standards based on information theory (Pluim et al., 2004), structural information (Heinrich et al., 2013; Wachinger and Navab, 2012) or metric learning (Michel et al., 2011; Simonovsky et al., 2016) rely on the strong assumption that the same structures are visible in both modalities (Fig. 3). In the latter case, such similarities do not have an analytical gradient nor guarantee the desired properties for an optimization energy. Therefore, their gradient-based optimization calls for local gradient

* Corresponding author at: Computer Aided Medical Procedures (CAMP), Technische Universität München, Boltzmanstr. 3 85748, Garching, Germany.
*E-mail addresses:* gutierrez.becker@tum.de (B. Gutierrez-Becker), mateus@in.tum.de (D. Mateus), peter@in.tum.de (L. Peter), navab@in.tum.de (N. Navab).

http://dx.doi.org/10.1016/j.media.2017.05.002
1361-8415/© 2017 Published by Elsevier B.V.

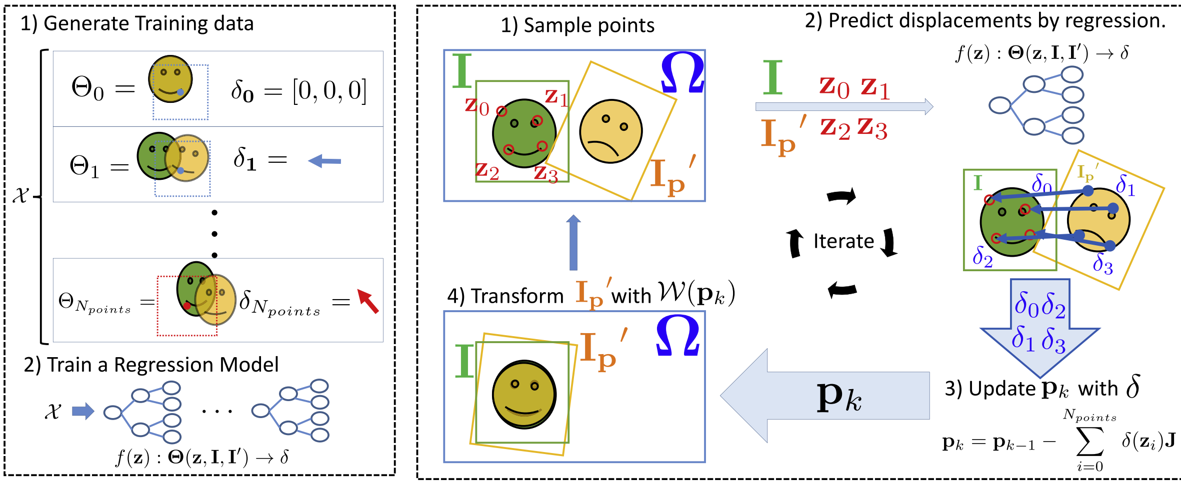


**Fig. 1.** Method Overview. **Training stage (left)**: A set of aligned multimodal images is used to generate a training set of images with known transformations. From this training set we train an ensemble of trees mapping the joint appearance of the images to displacement vectors. **Testing stage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

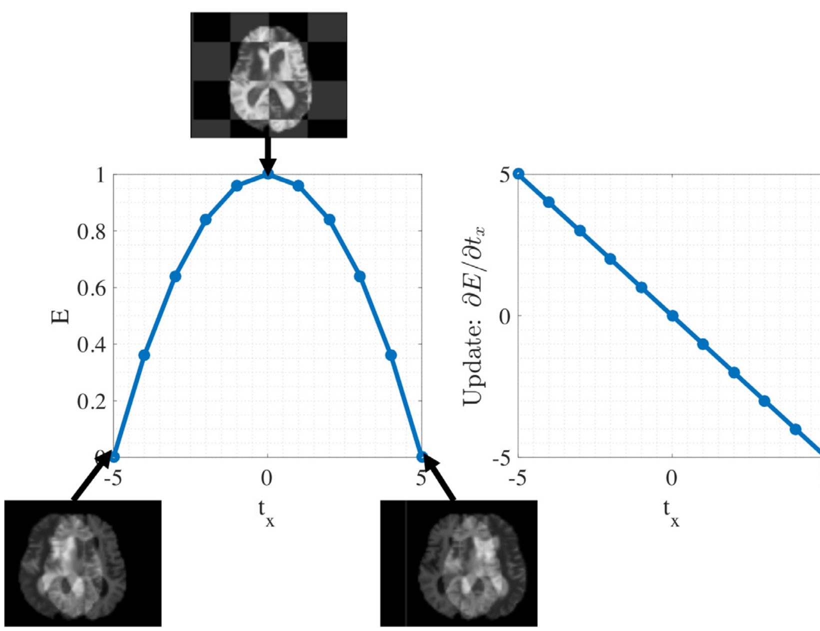


**Fig. 2.** Exemplary energy function *E*. **Left**: Continuous, convex and smooth behavior of *E* w.r.t. a transformation parameter. **Right**: Parameter update obtained by obtaining the derivative of the energy function with respect to a transformation parameter.

approximations or gradient free methods, which require advanced updates rules and an increased number of evaluations of the similarity metric.

In this work, we design a multimodal energy function that: i) is general, since it can create models capturing complex relationships between a wide range of modality pairs by using a small set of aligned examples, ii) can model such relationships based on global *and* local appearance, iii) can be easily optimized using a gradient-based method, and iv) that adapts to different transformation parameterizations. We model multimodal registration as a supervised regression problem, where given a pair of misaligned images we predict updates of the transformation parameters towards the correct alignment (*c.f* Fig. 1).

The joint appearance of the images is represented via a multimodal version of the Haar-like features (Criminisi et al., 2009) extracted from a sampling grid, which allows describing both the local and global-range context of each point. The regression task is formalized with gradient boosted trees, capable of handling the very high-dimensional Haar-like feature space, as well as of accurately approximating the transformation updates.

Our work is to the best of our knowledge, the first approach aiming at learning functions that map multimodal appearance to motion predictions, and showing how to effectively integrate them into a simple optimization scheme.

This paper is based on our previous work (Gutiérrez-Becker et al., 2016) but includes several extensions. First, we modify the method in order to predict not only the optimal update direction, but also the magnitude of the update vector in each iteration of the gradient-based optimizer. Second, we replace the regression model from random forest to gradient boosted trees (Friedman, 2001). We show how these two modifications lead to faster convergence times during testing as well as to an accurate registration. In addition, we include an evaluation of the improved properties of our method in terms of convergence and its training requirements using the IXI dataset. To demonstrate the generality of our method, we also extended our experiments to the publicly avail-