Simultaneous Registration of Multiple Images: Similarity Metrics and Efficient Optimization

Christian Wachinger, *Member, IEEE,* and Nassir Navab, *Member, IEEE*

Abstract—We address the alignment of a group of images with simultaneous registration. Therefore, we provide further insights into a recently introduced framework for multivariate similarity measures, referred to as *accumulated pair-wise estimates* (APE), and derive efficient optimization methods for it. More specifically, we show a strict mathematical deduction of APE from a maximum-likelihood framework and establish a connection to the congealing framework. This is only possible after an extension of the congealing framework with neighborhood information. Moreover, we address the increased computational complexity of simultaneous registration by deriving efficient gradient-based optimization strategies for APE: Gauss-Newton and the efficient second-order minimization (ESM). We present next to SSD the usage of intrinsically non-squared similarity measures in this least-squares optimization framework. The fundamental assumption of ESM, the approximation of the perfectly aligned moving image through the fixed image, limits its application to mono-modal registration. We therefore incorporate recently proposed structural representations of images, which allow us to perform multi-modal registration with ESM. Finally, we evaluate the performance of the optimization strategies with respect to the similarity measures, leading to very good results for ESM. The extension to multi-modal registration is in this context very interesting because it offers further possibilities for evaluations, due to publicly available data sets with ground-truth alignment.

Index Terms—Registration, Groupwise, Simultaneous, Optimization, Similarity Measures, Multi-modal.

1 Introduction

The analysis of a group or population of images requires their alignment to a canonical pose. Examples are the alignment of handwritten digits or face images for their later identification [1]–[4], the alignment of 3D tomographic images for the creation of an atlas [5], or the creation of mosaics from ultrasonic volumes [6]. First approaches to this groupwise registration problem identified one image as template, and registered all other images to it with a pair-wise approach. While this is a valid strategy for certain applications where such a template exists, in most cases it leads to an undesired introduction of bias with respect to the *a priori* chosen template. *Simultaneous registration* presents a method to circumvent this problem, however, it necessitates *multivariate similarity measures* and an optimization in a higher-dimensional space.

The direct estimation of multivariate measures with high-order joint density functions is prohibitive, because for a reliable estimation of the joint density, the number of samples would have to grow exponentially with the number of images, however, it only grows linearly. Approximations are therefore necessary, like the *congealing* framework presented by Learned-Miller [1]. Another approach was presented by Wachinger *et al.* [6], which *accumulates pair-wise estimates* (APE). The derivation of APE was mainly based on analogies. Moreover, the relationship between congealing and APE has not yet been investigated.

When aligning multiple data sets simultaneously, instead of successively, one has to consider two consequences for the optimization method. First, the registration scenario becomes more complex because the parameter space increases linearly with the number of images. And second, the evaluation of the multivariate similarity measure is more expensive. One is therefore interested in an efficient optimization procedure, which finds the optima robustly and with a minimal amount of evaluations of the objective function. We focus on gradient-based methods because they promise a fast convergence rate due to the guidance of the process by the gradient.

In this article, we address the afore mentioned problems of simultaneous registration. First, we present a strict mathematical deduction of APE from a maximum likelihood framework. Second, we describe an extended version of congealing, enriched with neighborhood information, which allows us to show the connection between APE and congealing. Third, we derive efficient gradient-based optimization strategies for simultaneous registration with APE as multivariate similarity framework.

The direct application of ESM to multi-modal registration is not possible because the addition of gradient images from different modalities is not meaningful. Recently, structural images were proposed [7], [8], which represent the structural information in images, to a certain extent independent of brightness and color. In combination with structural images, we can there-

C. Wachinger is currently with the Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology and the Department of Neurology, Massachusetts General Hospital, Harvard Medical School. This work was completed while he was with the chair for Computer Aided Medical Procedures, Technische Universität München.

N. Navab is with the chair for Computer Aided Medical Procedures, Technische Universität München, Germany.

E-mail: wachinge@in.tum.de, navab@in.tum.de

fore apply ESM for multi-modal registration, because it is transferred to a mono-modal registration problem. The application of multi-modal simultaneous registration gives us additional means for validation because public datasets exist, where particular effort has been taken to measure the ground truth transformation with bone-implanted fiducial markers [9].

1.1 Related Work

Simultaneous registration has many applications in computer vision, pattern recognition, and medical imaging when it comes to the alignment of multiple images. Learned-Miller [1] proposed the congealing framework for the alignment of a large number of binary images from a database of handwritten digits and for the removal of unwanted bias fields in magnetic resonance images. Congealing sums up the entropy of pixel stacks over the image. Huang *et al*. [2] applied congealing to align 2D face images, essential for their later identification. In [1]–[4], a sequential update of the registration parameters is performed. Zöllei *et al*. [5] used congealing for the simultaneous alignment of a population of brain images for brain atlas construction. Studholme and Cardenas [10] construct a joint density function for multivariate similarity estimation, which has the afore mentioned problem for larger image sets. Cootes *et al*. [11] use the minimum description length for the alignment of a group of images in order to create statistical shape models. This criterion demands a great deal of memory so that it only works for a limited number of volumes [5]. Sidorov *et al*. [12] use a stochastic optimization approach for groupwise registration of face images. One image is selected at a time and aligned to the remaining images using a similarity term that is close the voxel-wise variances, see section 2.2.1. It is argued that by randomly selecting the image to update the warp, an approximation to a fully simultaneous registration is achieved. Wachinger *et al*. [6] proposed simultaneous registration for volumetric mosaicing. This poses slightly different requirements on the multivariate similarity measure, because the number of overlapping images varies and can be rather small on specific locations. The therein introduced APE is flexible enough to deal with such situations. APE is a general framework to extend pairwise to multivariate similarity measures. The specific case of APE with sum of squared differences was used by Cox *et al*. [3], [4], referring to it as least squares congealing. Recently, APE was applied for simultaneous deformable registration of time-resolved images [13]. A spatio-temporal registration is performed by embedding the images in 4D space and deforming all of them simultaneously. A similar approach was proposed in [14], working with voxel-wise variances.

A popular class of optimization techniques are gradient-based methods. The utilization of the derivatives of the cost function helps in finding the optimum more efficiently in comparison to techniques that only rely on the function values. Gradient-based techniques are widely applied, where we focus on their application in image alignment [15], [16]. A good overview of gradient-based optimization methods is provided in Baker and Matthews [17] and Madsen *et al*. [18]. Based on their results, we do not consider the Levenberg-Marquardt algorithm because of its very similar behavior to Gauss-Newton. A new method, which is not covered in these articles, comes from the field of vision-based control. It is an efficient-second order optimization method introduced by Benhimane and Malis [19]. They showed that ESM has striking advantages in convergence rate and convergence frequency in comparison to Gauss-Newton (GN) and steepest-descent (SD). Vercauteren *et al*. [20] achieved good results for the pairwise alignment of 2D images with ESM.

Once the update is calculated, either an additive or a compositional scheme for applying the update to the current transformation can be used. In several articles [17], [21]–[23] the advantages of a compositional update are noted, which we consequently also apply in our work.

1.2 Outline

The remainder of the article is structured as follows. In section 2, we present the derivation of multivariate similarity measures from a common probabilistic framework. We deduce APE, as well as, congealing and analyze their relationship by introducing Markov congealing. Subsequently, we focus on APE as similarity framework and derive gradient-based optimization techniques in section 3, most prominently ESM. In section 4, we introduce multi-modal registration with ESM by calculating structural images. The experimental results for ultrasound mosaicing and multi-modal registration are presented in section 5. Parts of this work were previously presented at a conference [24].

2 Multivariate Similarity Metrics

In this section, we present a deduction of APE from a maximum likelihood (ML) framework and show its connection to congealing. Considering n images $\mathcal{I} = \{I_1, \dots, I_n\}$ and the transformation parameters $\mathbf{x}$, the ML framework for intensity-based registration is formulated as:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \log p(I_1, \dots, I_n; \mathbf{x}) \tag{1}$$

with the joint density function p, and the estimated alignment $\hat{\mathbf{x}}$ [25]. We present more details on the actual transformation parameterization in section 3, where for this section it is only important that $\mathbf{x}$ contains the parameters for all images and can represent arbitrary transformations. For notational ease, we will no longer consider $\mathbf{x}$ explicitly in the density function.

2.1 Accumulated Pair-Wise Estimates

APE approximates the joint likelihood function with pair-wise estimates [6]:

$$\log p(I_1, \ldots, I_n) \approx \sum_{i=1}^{n} \sum_{j \neq i} \log p(I_j | I_i). \tag{2}$$

Assuming a Gaussian distribution of the density p, i.i.d. coordinate samples, and various intensity mappings between the images, popular similarity measures such as sum of squared differences (SSD), normalized cross correlation (NCC), correlation ratio (CR), and mutual information (MI) can be derived from the log-likelihood term $\log p(I_j|I_i)$ [26]–[28]. APE therefore presents a framework for similarity measures. We provide more insights on the derivation of similarity measures by deducing SSD in appendix 1.

In order to deduce APE, we first derive a pair-wise approximation with respect to image I_n using the product rule and conditional independence:

$$p(I_1, \ldots, I_n) \overset{\text{Prod.Rule}}{=} p(I_1, \ldots, I_{n-1}|I_n) \cdot p(I_n) \tag{3}$$

$$\overset{\text{Cond.Indep.}}{=} \prod_{i=1}^{n-1} p(I_i|I_n) \cdot p(I_n). \tag{4}$$

Second, we take the n-th power of the joint density function and perform the derivation of equation (4) with respect to each of the images, leading to:

$$p(I_1, \ldots, I_n)^n = \prod_{i=1}^{n} p(I_i) \cdot \prod_{i=1}^{n} \prod_{j \neq i} p(I_j|I_i). \tag{5}$$

Third the logarithm is applied:

$$\log p(I_1, \ldots, I_n)^n = \sum_{i=1}^{n} \log p(I_i) + \sum_{i=1}^{n} \sum_{j \neq i} \log p(I_j|I_i)$$

leading to the desired approximation of the high dimensional density:

$$\log p(I_1, \ldots, I_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(I_i) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \log p(I_j|I_i) \tag{6}$$

$$\approx \sum_{i=1}^{n} \sum_{j \neq i} \log p(I_j|I_i) \tag{7}$$

wherein we no longer consider the multiplicative constant $\frac{1}{n}$ and the prior term $\sum_{i=1}^{n} \log p(I_i)$. The prior may, however, be used in future applications to incorporate further knowledge about the registration problem. The presented deduction is not limited to similarity measures and presents a general approximation of higher order densities by pairwise ones.

2.2 Congealing

In the congealing framework [1], independent but *not* identical distributions of the coordinate samples $s_k \in \Omega$ in the grid Ω are assumed:

$$p(I_1, \ldots, I_n) = \prod_{s_k \in \Omega} p^k(I_1(s_k), \ldots, I_n(s_k)). \tag{8}$$

Assuming further i.i.d. input images I_i leads to:

$$p(I_1, \ldots, I_n) = \prod_{s_k \in \Omega} \prod_{i=1}^{n} p^k(I_i(s_k)). \tag{9}$$

In the following, we derive a more general form of congealing that applies, instead of the assumption of independent images, the Markov property. This means that images are independent, if we know a certain local neighborhood of images around the current one. While the consideration of neighboring pixels, surrounding a sample s_k, was already discussed in [1], referred to as pixel cylinder, the consideration of neighboring images has not yet been proposed. So, instead of independent images, we assume that each image I_i depends on a certain neighborhood $\mathcal{N}_i$ of images:

$$p(I_1, \ldots, I_n) = \prod_{s_k \in \Omega} \prod_{i=1}^{n} p^k(I_i(s_k)|I_{\mathcal{N}_i}(s_k)). \tag{10}$$

We refer to this approximation as Markov-congealing. The size of the neighborhood depends on the structure in the image stack. If there is no further information about the images, considering a maximal neighborhood $I_{\mathcal{N}_i} = (I_1, \ldots, I_{i-1}, I_{i+1}, \ldots, I_n)$ seems reasonable. If there is, however, a certain order or evolution in the stack (camera parameters, motion, etc.), the neighborhood can be chosen appropriately to reflect this structure.

2.2.1 *Voxel-wise Variances*

The Markov-congealing allows us to derive the voxel-wise variances as proposed in [6] and applied in [12], [14]. The term *voxel-wise* estimation [5] is used, since the approach taken in the congealing framework focuses on certain pixel or voxel locations at a time. Voxel-wise variances combine the approach of a voxel-wise similarity estimation with the assumptions underlying SSD, which are Gaussian distributed intensity values and the identity as intensity mapping.

We incorporate the neighborhood information by estimating the mean μ_k for each voxel location s_k with:

$$\mu_k = \frac{1}{n} \sum_{l=1}^{n} I_l(s_k). \tag{11}$$

Following the formal definition of a local neighborhood $\mathcal{N}_i$ in the Markov sense, the calculation of the mean should not include the image I_i itself [29]. This leads, however, to higher computational costs because for each image and for each voxel location a different mean has to be calculated. We therefore go ahead with the more practical approximation, leading to:

$$p(I_1, \ldots, I_n) = \prod_{s_k \in \Omega} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(I_i(s_k) - \mu_k)^2}{2\sigma^2}\right) \tag{12}$$

with variance σ^2. This leads to the formula for voxel-wise SSD:

$$\log p(I_1, \ldots, I_n) \approx - \sum_{s_k \in \Omega} \sum_{i=1}^{n} (I_i(s_k) - \mu_k)^2. \tag{13}$$

Looking at the formula, we can see that voxel-wise SSD leads to the calculation of the variance at each location and subsequently accumulates the values [30]. The variance is one of the measures to express the *statistical dispersion* of a random variable [31]. In contrast to entropy, which measures the structuredness of a variable, it can only deal with mono-modal matchings. An interesting equality exists between voxel-wise SSD and APE SSD:

$$\sum_{s_k \in \Omega} \sum_{i=1}^{n} (I_i(s_k) - \mu_k)^2 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s_k \in \Omega} (I_i(s_k) - I_j(s_k))^2, \tag{14}$$

with the proof shown in appendix 2.

2.3 Comparison of APE and Congealing

In the last sections, we discussed APE and congealing as separate approximations to the high dimensional density with a connection for SSD. In this section, we investigate if there is a direct theoretical relationship between the two approaches. It is in fact possible to deduce a connection between APE and Markov-congealing. The detailed proof is stated in appendix 3. Therein we start with the Markov-congealing and derive APE. In order to make the derivation possible, the following assumptions have to be incorporated: (a) maximal neighborhood, (b) conditional independence of images, and (c) i.i.d. distribution of coordinate samples. While (c) was explicitly chosen by the design of congealing and (b) by the deduction of APE, the novel part is the neighborhood (a), which relates these two approaches. The Markov-congealing in equation (10) presents therefore an intermediate between APE and congealing.

To conclude, for congealing no specific distribution has to be selected, because the similarity can directly be calculated with the sample entropy. Markov-congealing and APE do not present actual similarity measures, but frameworks, where further information about the distribution has to be provided to derive similarity measures. Incorporating *e.g.* a Gaussian distribution and an identity intensity mapping leads to an SSD like extension. APE, in contrast to congealing, assumes an identical distribution of coordinate samples, which makes a reliable estimation for a small number of overlapping images possible. For congealing, a larger number is necessary, because the estimation is done with the information at one location at a time. Consequently, the choice, which multivariate similarity approximation to choose, is application dependent. We will focus on APE in the remaining article because it seems most versatile.

3 Efficient Optimization Methods

In this section, we present efficient gradient-based optimization methods for simultaneous registration. More precisely, we focus on APE as similarity measure and 3D rigid transformations as transformation model, where the parameterization can be easily adapted to different types of alignments. In contrast to [1]–[4], we do not update one parameter at a time, but update all parameters at once. Problems with the sequential update are illustrated in [4].

3.1 Transformation Parameterization

We parameterize the spatial transformations with Lie groups because 3D rigid transformations do not form a vector space. We perform a geometric optimization using local canonical coordinates. It has the advantage that the geometric structure of the group is taken care of intrinsically [32], [33]. This enables us to use an unconstrained optimization. Alternatively, one could embed them into the Euclidean space and perform a constrained optimization with Lagrange multipliers.

Each rigid 3D transformation $\mathbf{x}$ is an element of $\mathbb{SE}(3)$, the special Euclidean group. It is possible to describe them with a 4×4 matrix having the following structure:

$$\mathbf{x} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \tag{15}$$

with the rotational part $\mathbf{R}$, element of the special orthogonal group $\mathbb{SO}(3)$, and the translational part $\mathbf{t} \in \mathbb{R}^3$.

$\mathbb{SE}(3)$ forms a manifold and is a group under standard matrix multiplication, therefore it is a Lie group. On Lie groups, the tangent space at the group identity defines a Lie algebra. The Lie algebra captures the local structure of the Lie group. The Lie algebra of $\mathbb{SE}(3)$ is denoted by $\mathfrak{se}(3)$, and is defined by:

$$\mathfrak{se}(3) = \left\{ \begin{bmatrix} \mathbf{\Omega} & \mathbf{v} \\ \mathbf{0} & 0 \end{bmatrix} | \mathbf{\Omega} \in \mathbb{R}^{3\times 3}, \mathbf{v} \in \mathbb{R}^3, \mathbf{\Omega}^\top = -\mathbf{\Omega} \right\}.$$

The standard basis of $\mathfrak{se}(3)$ is $\mathcal{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_6\}$ with:

$$\mathbf{l}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{l}_4 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{16}$$

$$\mathbf{l}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{l}_5 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{17}$$

$$\mathbf{l}_3 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{l}_6 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{18}$$

Each element $\mathcal{H} \in \mathfrak{se}(3)$ can be expressed as a linear combination of matrices $\mathcal{H} = \sum_{i=1}^{6} h_i \mathbf{l}_i$ with h_i varying over the manifold [34] and $\mathbf{h} = [h_1, \dots, h_6]^\top$. The exponential map relates the Lie algebra to the Lie group:

$$\exp : \mathfrak{se}(3) \to \mathbb{SE}(3) \tag{19}$$

$$\mathcal{H} \mapsto \exp\left(\sum_{i=1}^{6} h_i \mathbf{l}_i\right) = \sum_{j=0}^{\infty} \frac{1}{j!}\left(\sum_{i=1}^{6} h_i \mathbf{l}_i\right)^j .$$

It exists an open cube V around $\mathbf{0}$ in $\mathfrak{se}(3)$ and an open neighborhood U of the identity matrix $\mathbf{I} \in \mathbb{SE}(3)$ such that the group exponential is smooth and one-to-one onto, with a smooth inverse, therefore a diffeomorphism. An explicit expression for the calculation of the exponential for elements in $\mathbb{SE}(3)$ exists, as shown in [35, pp.413]. For the restriction to $\mathbb{SO}(3)$, the explicit formula is known as Rodrigues' formula.

Using the local coordinate charts, there exists for any $\mathbf{y} \in \mathbb{SE}(3)$ in some neighborhood of $\mathbf{x}$ a vector in the tangent space $\mathcal{H} \in \mathfrak{se}(3)$, such that:

$$\mathbf{y} = \mathbf{x} \circ \exp(\mathcal{H}) = \mathbf{x} \circ \exp\left(\sum_{i=1}^{6} h_i \mathbf{l}_i\right). \tag{20}$$

Let us further denote the transformation of a point $\mathbf{p} = [x, y, z, 1]^\top \in \mathbb{R}^4$ in homogeneous coordinates through the mapping $\mathbf{y} \in \mathbb{SE}(3)$ with $w(\mathbf{y}, \mathbf{p})$:

$$w : \mathbb{SE}(3) \times \mathbb{R}^4 \to \mathbb{R}^4 \tag{21}$$

$$(\mathbf{y}, \mathbf{p}) \mapsto w(\mathbf{y}, \mathbf{p}) = \mathbf{p}'. \tag{22}$$

Finally, for ease of notation we define an extension of the exponential, enabling the direct application of the parameter vector $\exp(\mathbf{h}) := \exp(\mathcal{H})$.

3.2 Optimization Methods

The global transformation $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, with $\mathbf{x}_i \in \mathbb{SE}(3)$, maps the points from each of the image spaces to the joint image space, $\mathbb{R}^4 \to \mathbb{R}^4, \mathbf{p} \mapsto w(\mathbf{x}_i, \mathbf{p})$. The cost function E that we want to optimize is a sum of squared smooth functions:

$$E(\mathbf{x}) = \sum_{i \neq j} F_{i,j}(\mathbf{x}) = \sum_{i \neq j} \frac{1}{2} \|\mathbf{f}_{i,j}(\mathbf{x})\|^2 \tag{23}$$

with $F_{i,j}$ representing the pair-wise similarity measure.

Regarding equation (23), we see that we deal with a non-linear least-squares problem. Therefore, efficient optimization methods were proposed that achieve in many cases linear, or even quadratic, convergence without the explicit calculation of the second derivatives.

The starting point from all the following optimization methods is a Taylor expansion of the cost function around the current transformation $\mathbf{x}$ along the gradient direction $\mathbf{h}$:

$$E(\mathbf{x} \circ \exp(\mathbf{h})) \approx E(\mathbf{x}) + \mathbf{J}_E(\mathbf{x}) \cdot \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \cdot \mathbf{H}_E(\mathbf{x}) \cdot \mathbf{h} \tag{24}$$

with $\mathbf{J}_E(\mathbf{x}) = \frac{\partial E(\mathbf{x} \circ \exp(\mathbf{h}))}{\partial \mathbf{h}}\Big|_{\mathbf{h}=\mathbf{0}}$ and $\mathbf{H}_E(\mathbf{x}) = \frac{\partial^2 E(\mathbf{x} \circ \exp(\mathbf{h}))}{\partial \mathbf{h}^2}\Big|_{\mathbf{h}=\mathbf{0}}$ the Jacobian and Hessian, respectively, of E at the point $\mathbf{x}$. The global gradient direction $\mathbf{h}$ is a combination of local elements $\mathbf{h}_i$, resulting in $\mathbf{h} = [\mathbf{h}_1, \dots, \mathbf{h}_n]$. The Newton (NT) method then has the following compositional update:

$$\mathbf{H}_{F_{i,j}} \mathbf{h}_{i,j}^{\mathrm{NT}} = -\mathbf{J}_{F_{i,j}}^\top \qquad \mathbf{x} \leftarrow \mathbf{x} \circ \exp(\mathbf{h}^{\mathrm{NT}}). \tag{25}$$

The global update $\mathbf{h}^{\mathrm{NT}}$ is obtained by summing up the pairwise updates, following the structure of the cost function E in equation (23), leading to

$$\mathbf{h}^{\mathrm{NT}} = \left[\sum_i \mathbf{h}_{i,1}^{\mathrm{NT}}, \dots, \sum_i \mathbf{h}_{i,n}^{\mathrm{NT}}\right]. \tag{26}$$

Unfortunately, the explicit calculation of the Hessian causes problems because it is numerically not well-behaved and computationally expensive, so that its usage is not recommended [17]. In the field of non-linear least squares optimization most of the methods use an approximation of the Hessian [18]. In the following we present different possibilities for approximating the Hessian by a positive definite matrix $\hat{\mathbf{H}}$.

3.2.1 Steepest-Descent

For SD, the Hessian is approximated by the identity $\hat{\mathbf{H}} = \alpha \cdot \mathbf{I}$, leading to the update:

$$\alpha \cdot \mathbf{h}^{\mathrm{SD}} = -\mathbf{J}_E^\top(\mathbf{x}) \qquad \mathbf{x} \leftarrow \mathbf{x} \circ \exp(\mathbf{h}^{\mathrm{SD}})$$

with α the step length. Consequently, SD only considers a first-order Taylor expansion of E and has linear convergence.

3.2.2 Gauss-Newton

The approximation of the Hessian for Gauss-Newton is based on a linear approximation of the components of $\mathbf{f}$ in a neighborhood of $\mathbf{x}$. For small $\|\mathbf{h}\|$ we obtain from the Taylor expansion:

$$\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{f}(\mathbf{x}) + \mathbf{J}_{\mathbf{f}}(\mathbf{x}) \cdot \mathbf{h}. \tag{27}$$

For notational ease, we often write $\mathbf{f}$ instead of $\mathbf{f}_{i,j}$ when no reference to the images i and j is necessary. Setting this linear approximation in our cost function E, as defined in equation (23), gives:

$$E(\mathbf{x} \circ \exp(\mathbf{h})) \approx \sum_{i \neq j} \frac{1}{2} \|\mathbf{f}_{i,j}(\mathbf{x} \circ \exp(\mathbf{h}))\|^2 \tag{28}$$

$$= \sum_{i \neq j} \frac{1}{2} \mathbf{f}_{i,j}(\mathbf{x} \circ \exp(\mathbf{h}))^\top \mathbf{f}_{i,j}(\mathbf{x} \circ \exp(\mathbf{h})) \tag{29}$$

$$= \sum_{i \neq j} \left(F_{i,j}(\mathbf{x}) + \mathbf{h}^\top \mathbf{J}_{\mathbf{f}_{i,j}}^\top \mathbf{f}_{i,j} + \frac{1}{2} \mathbf{h}^\top \mathbf{J}_{\mathbf{f}_{i,j}}^\top \mathbf{J}_{\mathbf{f}_{i,j}} \mathbf{h}\right). \tag{30}$$

By comparison with equation (24), and considering the gradient $\mathbf{J}_F = \mathbf{J}_{\mathbf{f}}^\top \mathbf{f}$, we can see that the Hessian is approximated by $\hat{\mathbf{H}} = \mathbf{J}_{\mathbf{f}}^\top \mathbf{J}_{\mathbf{f}}$.

We get the global Gauss-Newton step $\mathbf{h}^{\mathrm{GN}}$ by the pairwise optimal steps $\mathbf{h}_{i,j}^{\mathrm{GN}}$, analogously to the Newton method, see equation (26). This leads to the update:

$$(\mathbf{J}_{\mathbf{f}_{i,j}}^\top \mathbf{J}_{\mathbf{f}_{i,j}}) \mathbf{h}_{i,j}^{\mathrm{GN}} = -\mathbf{J}_{\mathbf{f}_{i,j}}^\top \mathbf{f}_{i,j} \qquad \mathbf{x} \leftarrow \mathbf{x} \circ \exp(\mathbf{h}^{\mathrm{GN}})$$

with $\mathbf{h}^{\text{GN}} = \left[\sum_i \mathbf{h}^{\text{GN}}_{i,1}, \ldots, \sum_i \mathbf{h}^{\text{GN}}_{i,n}\right]$. Gauss-Newton has only in specific cases quadratic convergence [18], [19].

3.2.3 ESM

The efficient second-order minimization procedure originally comes from the field of vision-based control [19]. It is very related to GN because it also considers least-squares problems. ESM achieves, however, significantly better results by incorporating further knowledge about the specificity of the optimization problem. It was shown that ESM has a cubic convergence rate [36].

More precisely, ESM uses the fact, that if the images are aligned with the optimal transformation, the images and therefore also their gradients should be very similar to each other. This can be used to ameliorate the search direction of the Newton methods. For the standard Newton method, the first and second order derivatives around $\mathbf{x}$ are used to build a second-order approximation, see equation (25). The Gauss-Newton method considers only the first derivative of $\mathbf{f}$ around $\mathbf{x}$ and can therefore only build a first-order approximation. For ESM, the first-order derivatives of $\mathbf{f}$ around $\mathbf{x}$ and $\mathbf{x} \circ \exp(\mathbf{h})$ are used to build a second-order approximation without the necessity of an explicit calculation of a second-order derivative.

To deduce ESM, we start with a second-order Taylor approximation of the function $\mathbf{f}$:

$$\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{f}(\mathbf{x}) + \mathbf{J_f}(\mathbf{x}) \cdot \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \cdot \mathbf{H_f}(\mathbf{x}) \cdot \mathbf{h}. \quad (31)$$

Subsequently, we do a second Taylor expansion around $\mathbf{x}$, but this time of the Jacobian of $\mathbf{f}$:

$$\mathbf{J_f}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{J_f}(\mathbf{x}) + \mathbf{H_f}(\mathbf{x}) \cdot \mathbf{h}. \quad (32)$$

Plugging this first-order series in the approximation shown in equation (31), we get a second-order approximation without second-order derivatives:

$$\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{f}(\mathbf{x}) + \frac{1}{2}[\mathbf{J_f}(\mathbf{x}) + \mathbf{J_f}(\mathbf{x} \circ \exp(\mathbf{h}))]\mathbf{h}. \quad (33)$$

The problem about this approximation is the calculation of the Jacobian $\mathbf{J_f}(\mathbf{x} \circ \exp(\mathbf{h}))$, which is dependent on the update $\mathbf{h}$ that we want to solve for, and therefore do not know yet. We illustrate a solution for this problem in section 3.3.3.

Comparing equations (27) and (33) shows the similarity between the Gauss-Newton and ESM procedure. For the development of the update rule we proceed therefore analogously to Gauss-Newton. The only difference is the usage of $\mathbf{J}^{\text{ESM}}_{\mathbf{f}} = \frac{1}{2}[\mathbf{J_f}(\mathbf{x}) + \mathbf{J_f}(\mathbf{x} \circ \exp(\mathbf{h}))]$ instead of only $\mathbf{J_f}(\mathbf{x})$. This leads to an approximation of the Hessian by $\hat{\mathbf{H}} = \mathbf{J}^{\text{ESM}^\top}_{\mathbf{f}} \mathbf{J}^{\text{ESM}}_{\mathbf{f}}$. The compositional update is:

$$\left(\mathbf{J}^{\text{ESM}^\top}_{\mathbf{f}_{i,j}} \mathbf{J}^{\text{ESM}}_{\mathbf{f}_{i,j}}\right) \mathbf{h}^{\text{ESM}}_{\mathbf{f}_{i,j}} = -\mathbf{J}^{\text{ESM}^\top}_{\mathbf{f}_{i,j}} \mathbf{f}_{i,j} \qquad \mathbf{x} \leftarrow \mathbf{x} \circ \exp(\mathbf{h}^{\text{ESM}})$$

with $\mathbf{h}^{\text{ESM}} = \left[\sum_i \mathbf{h}^{\text{ESM}}_{i,1}, \ldots, \sum_i \mathbf{h}^{\text{ESM}}_{i,n}\right]$.

3.3 Gradient Calculation

In the last section, we introduced the gradients $\mathbf{J}_E$, $\mathbf{J_f}$, and $\mathbf{J}^{\text{ESM}}_{\mathbf{f}}$ without further explaining their calculation. This will be the subject of this part, together with an analysis on how the gradient calculation changes for different similarity measures.

3.3.1 Steepest-Descent

We begin with the gradient for the general cost function E by considering one moving image at a time. W.l.o.g., we assume I_i as fixed and I_j as moving image leading to $F_{i,j}(\mathbf{x} \circ \exp(\mathbf{h})) = \text{SM}(I_i(\mathbf{x}), I_j(\mathbf{x} \circ \exp(\mathbf{h}))$, with SM a pair-wise similarity measure. The point-wise gradient has the form:

$$\begin{aligned}
[\mathbf{J}_E(\mathbf{x})]_\mathbf{p} &= \left[\left.\frac{\partial E(\mathbf{x} \circ \exp(\mathbf{h}))}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{0}} \right]_\mathbf{p} \quad (34)\\
&= \sum_{i \neq j} \left.\frac{\partial \text{SM}(I_i(\mathbf{x}), I_j(\mathbf{x} \circ \exp(\mathbf{h})); \mathbf{p})}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{0}}\\
&= \sum_{i \neq j} [\mathbf{J}_{\text{SM}_{i,j}}(\mathbf{x})]_\mathbf{p} \cdot [\mathbf{J}_{I_j}(\mathbf{x})]_\mathbf{p} \cdot [\mathbf{J}_w(\mathbf{x})]_\mathbf{p} \cdot \mathbf{J}_e(\mathbf{x}).
\end{aligned}$$

The Jacobian $[\mathbf{J}_{\text{SM}}(\mathbf{x})]_\mathbf{p}$ is a scalar value, corresponding to the derivative of the similarity measure:

$$\begin{aligned}
[\mathbf{J}_{\text{SM}_{i,j}}(\mathbf{x})]_\mathbf{p} &= \left.\frac{\partial \text{SM}(I_i(\mathbf{x}), I_j(\mathbf{x} \circ \exp(\mathbf{h})); \mathbf{p})}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{0}}\\
&= \left.\frac{\partial \text{SM}(I_i(\mathbf{x}), I; \mathbf{p})}{\partial I}\right|_{I=I_j(\mathbf{x} \circ \exp(\mathbf{0}))=I_j(\mathbf{x})}\\
&= \nabla \text{SM}(I_i(\mathbf{x}), I_j(\mathbf{x}); \mathbf{p}). \quad (35)
\end{aligned}$$

The Jacobian $[\mathbf{J}_{I_j}(\mathbf{x})]_\mathbf{p}$ is a matrix of dimension (1×3), corresponding to the spatial derivative of the moving image under the current transformation $\mathbf{x}$:

$$\begin{aligned}
[\mathbf{J}_{I_j}(\mathbf{x})]_\mathbf{p} &= \left.\frac{\partial I_j(w(\mathbf{x} \circ \exp(\mathbf{h}), \mathbf{p}))}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{0}} \quad (36)\\
&= \left.\frac{\partial I_j(w(\mathbf{x}, w(\exp(\mathbf{h}), \mathbf{p})))}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{0}} \quad (37)\\
&= \left.\frac{\partial I_j(w(\mathbf{x}, \mathbf{z}))}{\partial \mathbf{z}}\right|_{\mathbf{z}=w(\exp(\mathbf{0}),\mathbf{p})=\mathbf{p}} \quad (38)\\
&= \nabla I_j(w(\mathbf{x}, \mathbf{p})). \quad (39)
\end{aligned}$$

The Jacobian $[\mathbf{J}_w(\mathbf{x})]_\mathbf{p}$ is of dimension (3×16), corresponding to the derivative of the vector $w(\mathbf{Z}, \mathbf{p})$ with respect to the elements of the matrix $\mathbf{Z}$:

$$\begin{aligned}
[\mathbf{J}_w(\mathbf{x})]_\mathbf{p} &= \left.\frac{\partial w(\mathbf{x} \circ \exp(\mathbf{h}), \mathbf{p})}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{0}} \quad (40)\\
&= \left.\frac{\partial w(\mathbf{Z}, \mathbf{p})}{\partial \mathbf{Z}}\right|_{\mathbf{Z}=\mathbf{x} \circ \exp(\mathbf{0})=\mathbf{x}} \quad (41)\\
&= \begin{bmatrix} \mathbf{p}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{p}^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{p}^\top & \mathbf{0} \end{bmatrix}. \quad (42)
\end{aligned}$$

The Jacobian $\mathbf{J}_e(\mathbf{x})$ is of dimension (16×6), corresponding to the derivative of the exponential mapping with respect to each of the transformation parameters h_i:

$$\mathbf{J}_e(\mathbf{x})_i = \left.\frac{\partial \exp(\mathbf{h})}{\partial h_i}\right|_{\mathbf{h}=\mathbf{0}} = \left.\frac{\partial \exp(\sum_{i=1}^{6} h_i \mathbf{l}_i)}{\partial h_i}\right|_{\mathbf{h}=\mathbf{0}} \tag{43}$$

$$= \exp\left(\sum_{i=1}^{6} h_i \mathbf{l}_i\right)\Bigg|_{\mathbf{h}=\mathbf{0}} \cdot \mathbf{l}_i = \mathbf{l}_i \tag{44}$$

Stacking the vectorized basis vectors $[\mathbf{l}_i]_v$ of $\mathfrak{se}(3)$ leads to:

$$\mathbf{J}_e(\mathbf{x}) = [[\mathbf{l}_1]_v, \ldots, [\mathbf{l}_6]_v] \tag{45}$$

In appendix 4, we provide further insights into the form and dimensionality of the Jacobian matrices, when not focusing on a point-wise derivative. Based on this analysis we calculate the Jacobian of the cost function as an accumulation of all point-wise Jacobians:

$$\mathbf{J}_E(\mathbf{x}) = \sum_{\mathbf{p}\in\Omega} [\mathbf{J}_E(\mathbf{x})]_{\mathbf{p}}. \tag{46}$$

3.3.2 Gauss-Newton

For the derivation of the gradient $\mathbf{J_f}$, which is part of the Gauss-Newton optimization, we have to guarantee that the cost function fulfills further presumptions. The Gauss-Newton procedure was deduced by starting at a least-squares problem $E(\mathbf{x}) = \sum_{i\neq j} \frac{1}{2}\|\mathbf{f}_{i,j}(\mathbf{x})\|^2$, see equation (23). When considering SSD we can simply set $E(\mathbf{x}) = \sum_{i\neq j} \mathrm{SSD}_{i,j}(\mathbf{x})$, since SSD is intrinsically a least-squares problem.

This is not the case for other similarity measures like correlation ratio or mutual information. In order to ensure the least-squares nature, we square the similarity measures, leading to $E(\mathbf{x}) = \sum_{i\neq j} \|\mathrm{SM}_{i,j}(\mathbf{x})\|^2$. Obviously, optimizing the squared similarity measure has far-ranging consequences, which we investigate further in section 3.3.4. Moreover, issues concerning the conditioning of the Hessian may arise, which we discuss in section 3.3.5. The gradient $\mathbf{J_f}$ is calculated as:

$$[\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x})]_{\mathbf{p}} = \left.\frac{\partial \mathbf{f}_{i,j}(\mathbf{x}\circ\exp(\mathbf{h}))}{\partial \mathbf{h}}\right|_{\mathbf{h}=0} \tag{47}$$

$$= \left.\frac{\partial \mathrm{SM}(I_i(\mathbf{x}), I_j(\mathbf{x}\circ\exp(\mathbf{h})))}{\partial \mathbf{h}}\right|_{\mathbf{h}=0} \tag{48}$$

$$= [\mathbf{J}_{\mathrm{SM}_{i,j}}(\mathbf{x})]_{\mathbf{p}} \cdot [\mathbf{J}_{I_j}(\mathbf{x})]_{\mathbf{p}} \cdot [\mathbf{J}_w(\mathbf{x})]_{\mathbf{p}} \cdot \mathbf{J}_e(\mathbf{x}) \tag{49}$$

Stacking all the point-wise derivatives leads to the Jacobian of $\mathbf{f}$:

$$\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x}) = \begin{pmatrix} [\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x})]_{\mathbf{p}_1} \\ \vdots \\ [\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x})]_{\mathbf{p}_{|\Omega|}} \end{pmatrix} \tag{50}$$

3.3.3 ESM

The last gradient that remains is $\mathbf{J}_{\mathbf{f}}^{\mathrm{ESM}}$ for the ESM. Here we also consider the squared similarity measures like for GN. The calculation of $\mathbf{J}_{\mathbf{f}}^{\mathrm{ESM}}$ is difficult because part of its calculation is $\mathbf{J_f}(\mathbf{x}\circ\exp(\mathbf{h}))$, which depends on $\mathbf{h}$ that we want to solve for. In order to address this issue, Benhimane and Malis [19] consider the optimal update step $\mathbf{h}^{\mathrm{opt}}$ for the current location $\mathbf{x}$, leading to the perfect alignment $\mathbf{x}^{\mathrm{opt}} = \mathbf{x}\circ\exp(\mathbf{h}^{\mathrm{opt}})$. To consider the influence of this optimal update step for the product:

$$\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x}\circ\exp(\mathbf{h}^{\mathrm{opt}}))\cdot\mathbf{h}^{\mathrm{opt}} = \left.\frac{\partial \mathbf{f}_{i,j}(\mathbf{x}\circ\exp(\mathbf{h}))}{\partial \mathbf{h}}\right|_{\mathbf{h}=\mathbf{h}^{\mathrm{opt}}} \cdot \mathbf{h}^{\mathrm{opt}},$$

we have to analyze each of the four factors resulting from the derivation, see equation (49).

We proceed from right to left, starting with the $\mathbf{J}_e$. In [37, pp. 157], a proof is presented that:

$$\mathbf{J}_e(\mathbf{x})\cdot\mathbf{h}^{\mathrm{opt}} = \mathbf{J}_e(\mathbf{x}^{\mathrm{opt}})\cdot\mathbf{h}^{\mathrm{opt}}$$

utilizing the properties of the Lie algebra and the exponential map. Next, the derivative of the transformation $\mathbf{J}_w$ is the same for $\mathbf{x}$ and $\mathbf{x}\circ\exp(\mathbf{h}^{\mathrm{opt}})$ [19]. In order to have an approximation of the third term, the main assumption of ESM is incorporated. The gradient of the perfectly aligned image $\nabla I_j(\mathbf{x}\circ\exp(\mathbf{h}^{\mathrm{opt}}))$ can be approximated by the gradient of the fixed image $\nabla I_i(\mathbf{x})$, leading to:

$$\mathbf{J}_{I_j}(\mathbf{x}\circ\exp(\mathbf{h}^{\mathrm{opt}})) \approx \mathbf{J}_{I_i}(\mathbf{x}). \tag{51}$$

This takes the specificity of our optimization problem into account, because for image registration the possibility exists to approximate this gradient. Naturally, this is only feasible for images of the same modality. The last term is the derivative of the similarity measure, which we approximate by $\mathbf{J}_{\mathrm{SM}}(\mathbf{x})$. This finally leads to the overall approximation:

$$\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x}\circ\exp(\mathbf{h}^{\mathrm{opt}})) \approx \mathbf{J}_{\mathrm{SM}_{i,j}}(\mathbf{x})\cdot\mathbf{J}_{I_i}(\mathbf{x})\cdot\mathbf{J}_w(\mathbf{x})\cdot\mathbf{J}_e(\mathbf{x}). \tag{52}$$

Considering the definition of the gradient $\mathbf{J}_{\mathbf{f}}^{\mathrm{ESM}} = \frac{1}{2}(\mathbf{J_f}(\mathbf{x}) + \mathbf{J_f}(\mathbf{x}\circ\exp(\mathbf{h})))$, and equations (49) and (52), we finally get:

$$\mathbf{J}_{\mathbf{f}_{i,j}}^{\mathrm{ESM}} = \frac{1}{2}\cdot\mathbf{J}_{\mathrm{SM}_{i,j}}\cdot[\mathbf{J}_{I_i} + \mathbf{J}_{I_j}]\cdot\mathbf{J}_w\cdot\mathbf{J}_e. \tag{53}$$

evaluated at the current alignment $\mathbf{x}$.

3.3.4 Gradient of Similarity Measures

As mentioned in the last section, we optimize the squared similarity measure for the statistic metrics to ensure the least-squares nature of the optimization problem. For sum of squared differences this is not necessary. The interesting question is about the consequences of optimizing the squared function instead. Assuming a function ϕ and its squared version $\Phi = \phi^2$. The first and second derivatives of Φ are $\Phi' = 2\cdot\phi\cdot\phi'$ and $\Phi'' = 2\cdot(\phi')^2 + 2\cdot\phi\cdot\phi''$. Problematic is the introduction of new extrema for $\phi = 0$ and the change of their type for $\phi < 0$. NCC, CR, and MI have lower bounds, which are -1, -1, and 0, respectively. We avoid these optimization problems, by adding a constant ν to the similarity measures $\mathrm{SM}_{i,j} + \nu$, guaranteeing that they are in the positive range.

We list the actual derivatives of the similarity measures in the supplementary material for completeness, but they can also be found in *e.g.* [38]. Especially MI has frequently been applied for registration with a steepest-descent approach [38]–[41]. The densities are estimated following a kernel-based Parzen window approach [42]. For the estimation of the kernel window size, several methods were proposed. One common technique considers the maximization of a pseudo likelihood [43], which has the drawback of a trivial maximum at zero [40]. Instead, a leave-one-out strategy was proposed [38], [43]. The reported Parzen window kernel size that led to best results was five [40], which we adopted in our experiments. Note that for the calculation of the update $\mathbf{h}$ of the least-squares problems, either an LU- or Cholesky-decomposition could be used on the normal equations $(\mathbf{J}_{\mathbf{f}}^{\top}\mathbf{J}_{\mathbf{f}})\mathbf{h} = -\mathbf{J}_{\mathbf{f}}\mathbf{f}$, or a QR-decomposition on $\mathbf{J}_{\mathbf{f}}\mathbf{h} = -\mathbf{f}$. Since the normal equations worsen the numerical condition of the problem, the QR-decomposition presents the stabler choice.

3.3.5 Singularity of Hessian Approximation

For Gauss-Newton, equation (30) shows that the gradient is $\mathbf{J}_F = \mathbf{J}_{\mathbf{f}}^{\top}\mathbf{f}$ and the Hessian is approximated by $\hat{\mathbf{H}} = \mathbf{J}_{\mathbf{f}}^{\top}\mathbf{J}_{\mathbf{f}}$. For the perfect alignment $\mathbf{x}^{\mathrm{opt}}$, the gradient $\mathbf{J}_F = \mathbf{J}_{\mathbf{f}}^{\top}\mathbf{f}$ is zero. In contrast to SSD, the similarity function $\mathbf{f}$ is nonzero at $\mathbf{x}^{\mathrm{opt}}$ for NCC, CR, and MI. Consequently, $\mathbf{J}_{\mathbf{f}}$ has to be zero at $\mathbf{x}^{\mathrm{opt}}$, leading to an approximation of the Hessian $\hat{\mathbf{H}}$ that is singular at $\mathbf{x}^{\mathrm{opt}}$. The same analysis holds for ESM, because the factor driving the gradient to zero is $\mathbf{J}_{\mathrm{SM}}$, equally appearing in the gradient of ESM. In practice, it is unlikely that the matrix is singular due to image noise, image interpolation, and approximation of the perfect alignment. We analyzed the condition number close to the ground truth alignment, with results shown in appendix 5. In our experiments, we noticed only a slight increase of the condition number, while approaching the optimum. The condition number represents the sensitivity of the solution to errors.

Newton optimization for pairwise registration with mutual information was reported in [44], [45]. A first-order approximation of the Hessian was used, which is identical to the approximation of the Hessian in Gauss-Newton. It was noted in [46] that it is more appropriate to work with the actual Hessian of mutual information, instead of the approximation, because problems with the definiteness can occur. We studied the definiteness of the Hessian in our experiments, but have not observed problems. Helpful when running into such problems may be the usage of a line-search procedure.

3.3.6 Relationship to Forward/Inverse Compositional Update

In [17], the differences between forward and inverse compositional updates are discussed. The inverse update scheme has computational advantages, because the image gradient can be pre-computed. In our case of simultaneous registration, where all images move, we are no longer able to pre-compute the gradient image, and consequently the difference vanishes. It is, however, interesting to look at ESM from this perspective, because instead of either considering the gradient of the fixed or moving image, both gradients are combined. Hence, ESM presents a combination of forward and inverse compositional update.

4 Multi-Modal Registration with ESM

The fundamental assumption of ESM, the approximation $\mathbf{J}_{I_j}(\mathbf{x}\circ\exp(\mathbf{h}^{\mathrm{opt}})) \approx \mathbf{J}_{I_i}(\mathbf{x})$, prevents its direct application to multi-modal registration. The reason is that for multi-modal images, the gradient directions and orientations are not comparable. A solution to address this issue has, however, recently been proposed with the creation of structural images [7], [8]. The idea is to create representations of images that are focusing on the structures in the images, and not on the intensities or colors in which they are depicted. A standard intensity-based registration is subsequently performed on the structural representations, however, not with sophisticated similarity measures like MI, but with mono-modal metrics like SSD. The structural representations are therefore converting a multi-modal registration problem into a mono-modal one. The positive effect on ESM is that they make the approximation in equation (51) meaningful and ESM therefore applicable. In [7], the multi-modal registration of T1, T2, PD-weighted MR and CT images was investigated, which we also focus on in this study. The application in other scenarios of multi-modal registration depends on the availability of suitable structural representations.

A positive aspect of structural representations is that simple similarity measures can be applied. This obviates the problems of working with more complex similarity measures in a least-squares framework, as discussed in sections 3.3.4 and 3.3.5. Moreover, it enables a faster similarity evaluation, which is even more important for simultaneous registration, where the multivariate similarity metrics are more demanding to calculate. For the case of APE, the influence of the faster similarity evaluation is quadratic, since all combinations of pairwise estimates are calculated in equation (23).

Finally, structural representations in combination with ESM yield advantages for the validation of registration results. The validation of rigid registration is generally easier than deformable registration, because under the assumption of a rigid object, it is possible to measure the camera pose to obtain ground truth data. The drawback is, however, that there are rarely volumetric acquisitions of group of images from a static object. One example are volumetric ultrasound acquisitions. Another interesting application is the alignment of multi-modal volumes for neurosurgery. Effort has been taken to exactly measure the location with bone-implanted fiducial markers in

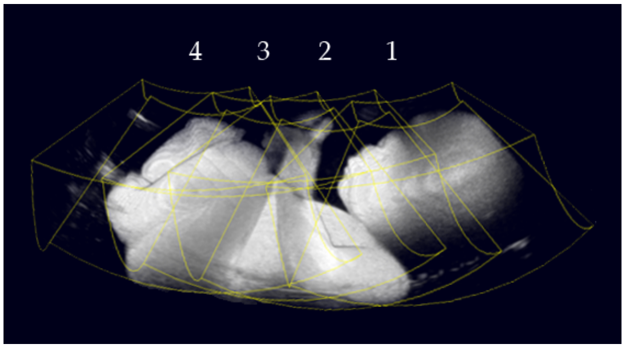


Fig. 1. Mosaic of baby phantom from 4 acquisitions.

order to provide ground truth data [9]. This data is well suited for rigid registration experiments because the acquisitions are acquired with a short time difference, and further, the skull provides a rigid frame, avoiding deformations. This validation is, however, only possible in combination with structural representations, as mentioned previously. The structural representation that we employ in combination with ESM are entropy images. We briefly explain their calculation in the following section.

4.1 Structural Representations with Entropy Images

For entropy images the entropy is calculated on local sub-volumes and subsequently stored in form of a dense descriptor. Be I an image and D_I the corresponding entropy image. For each spatial location s_k in the grid of D_I we set the intensity value to be

$$D_I(s_k) = \mathrm{H}(I|_{\mathcal{N}(s_k)}), \tag{54}$$

with $\mathcal{N}(s_k)$ a local neighborhood around s_k. The entropy H is then calculated on the sub-volume $I|_{\mathcal{N}(s_k)}$.

We use $9 \times 9 \times 9$ sub-volumes for the case of isotropic voxel spacing. For anisotropic spacing we adapt the neighborhood accordingly. Further, we perform a spatially weighted density estimation using a Gaussian weighting scheme together with a kernel-based Parzen window method. We select 64 bins and a global normalization of the intensity values. The Shannon entropy was chosen to measure the entropy.

5 Experiments

We perform experiments for two different applications to test APE in combination with the described optimization procedures. One application is the registration of multiple ultrasound volumes for volumetric ultrasound mosaicing, and the second application is the alignment of a group of multi-modal volumes.

5.1 Ultrasound Mosaicing

For ultrasound mosaicing, the experiments were conducted on four 3D ultrasound acquisitions from a baby phantom, having a resolution of $64 \times 64 \times 64$ voxels and a bit depth of one byte, see figure 1 and the video in the supplementary material. The registration of ultrasound images is challenging because of the degradation with speckle noise and the viewing angle dependent nature of the volumes. We displaced the volumes randomly from the correct position, guaranteeing an accumulated residual error of 30 over all the volumes. The correct position is obtained from manual alignment. We weight 1mm equal to 1° to make translational and angular displacement from the ground truth comparable. Starting from the random initial position we run 100 simultaneous registrations for each configuration to assess its performance.

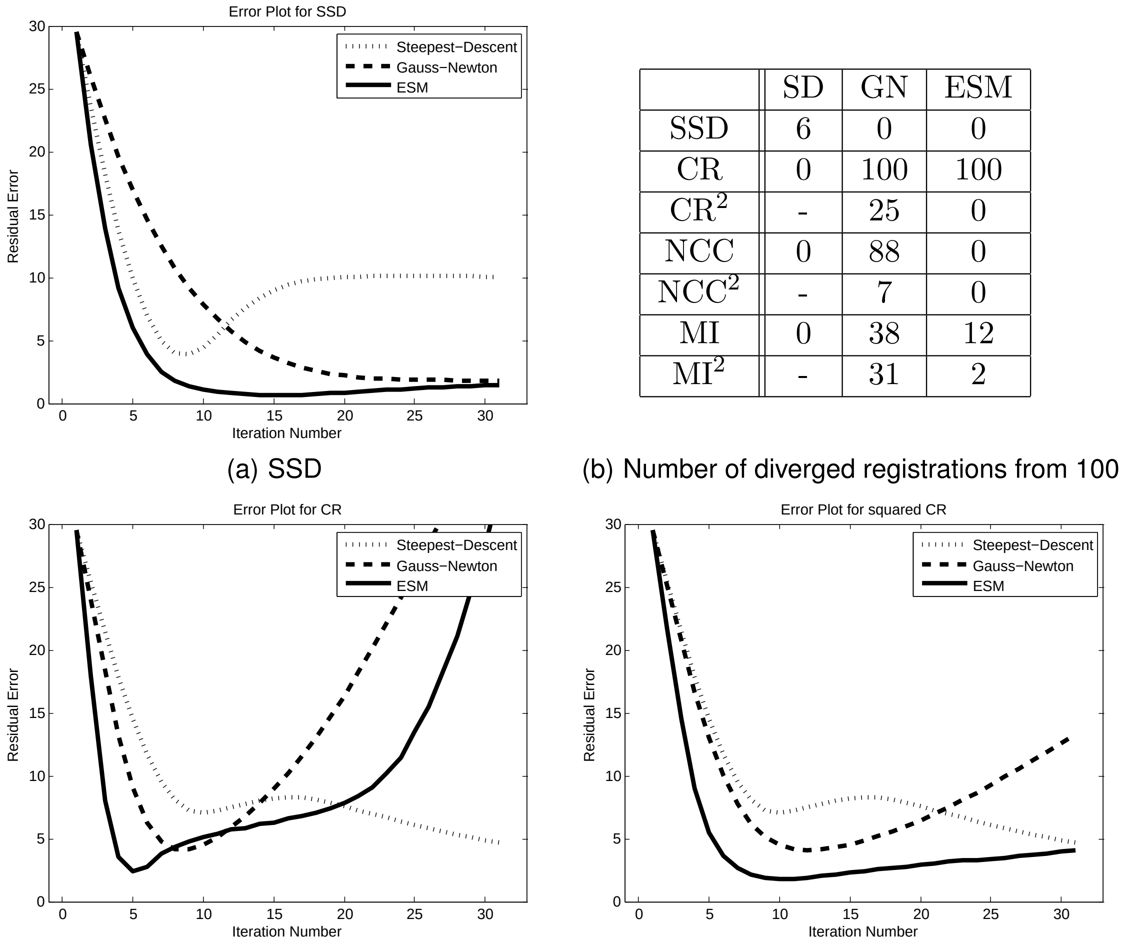
In figures 2 and 3, the average residual error is plotted with respect to the iteration number. For the calculation of the residual error, we first compensate for a global shift between the ground truth and registration result. Subsequently, we evaluate the root mean squared error (RMSE) between the ground truth and the registration result. Finally, the mean RMSE over all trials gives the average residual error. Note that diverging trials lead to a large residual error that is averaged over, causing an increasing instead of decreasing error curve. We discuss this further in section 6.

For SSD, see figure 2(a), we only have one plot because we do not have to consider the squared variant of it, like already mentioned. The best performance is obtained with ESM, leading to the fastest convergence. But also the Gauss-Newton method leads to a robust convergence. The gradient-descent does not perform well. Although it seems to approach the correct alignment nicely at the beginning, it diverges into another optimum. In the table in figure 2(b), the number of registrations that diverged are listed. We consider a registration diverged, when the residual error after 30 steps is larger than half the initial error.

For CR, see figure 2(c), the results for GN and ESM are not good. All of the 100 runs diverged. Steepest-descent, although slower, performs much better. The situation changes a lot, when optimizing the squared function, see figure 2(d). The ESM quickly approaches the correct alignment. Also GN improves, but the result is still not good. We also plot the curve for SD as reference, although it is the one of CR, because we do not use the squared variant for SD.

For NCC and MI, see figure 3, the situation is pretty similar to CR. The performance of GN and ESM when using the non-squared similarity measures is insufficient, leading to a high divergence rate. The situation improves enormously when optimizing the squared function instead. ESM always performs better than GN, both, with respect to speed and robustness. Furthermore, the performance of SD is interesting. Although the convergence is slower, compared to the others, it is in most cases robust.

All the registrations are performed on an Intel dual-core 2.4 GHz processor having 2 GB of RAM. The time



(a) SSD

	SD	GN	ESM
SSD	6	0	0
CR	0	100	100
CR^2	-	25	0
NCC	0	88	0
NCC^2	-	7	0
MI	0	38	12
MI^2	-	31	2

(b) Number of diverged registrations from 100

(c) CR

(d) Squared CR

Fig. 2. Plot of the average residual error for each iteration step for SSD, CR, and squared CR. Comparing CR and squared CR shows the improved performance of GN and ESM. ESM converges the fastest and leads to the smallest residual error. We plot the curve of SD also for squared CR to facilitate the comparison.

for one registration, where we allowed for 30 iterations, is below one minute.

5.2 Multi-Modal Registration

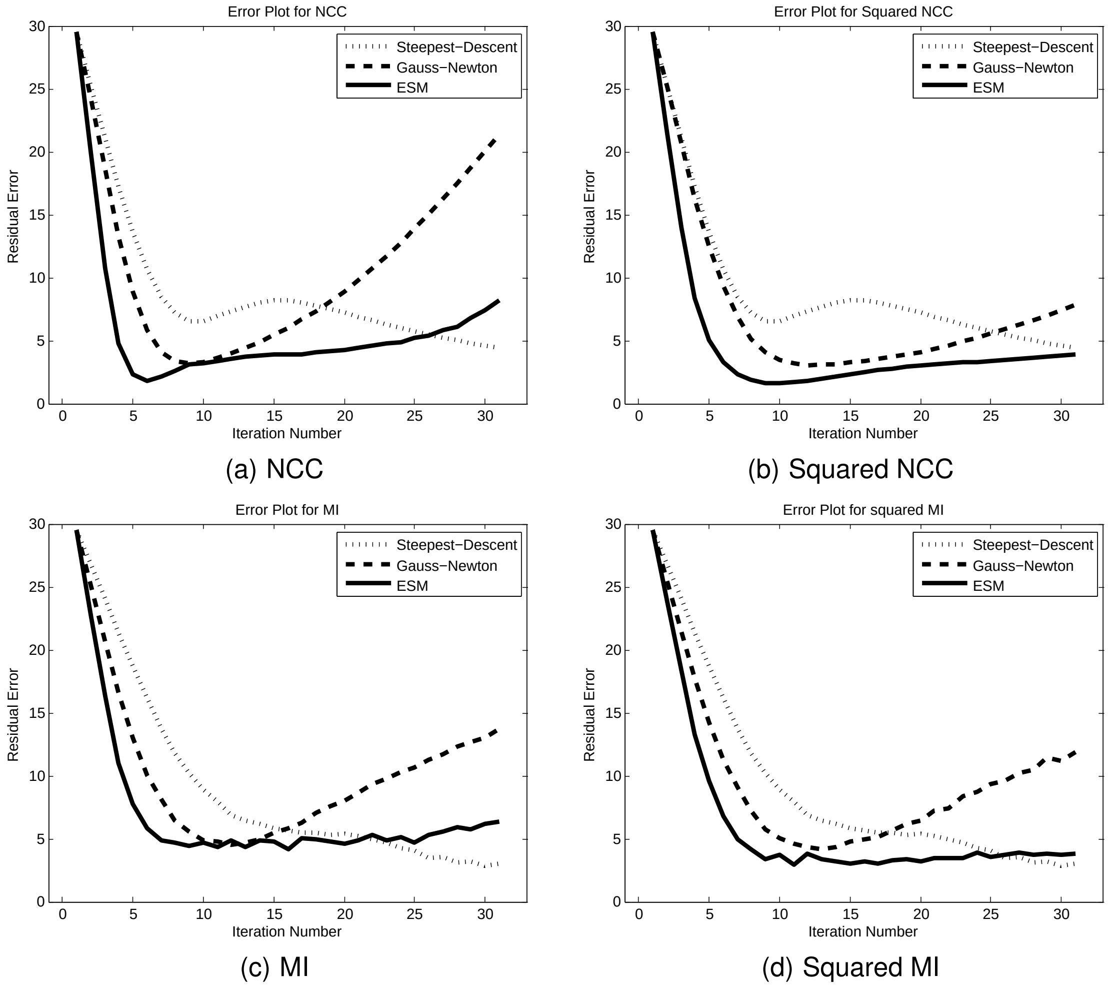
For multi-modal registration, we conduct experiments on T1, T2, and PD-weighted MR images from the BrainWeb database[1] and CT, T1, T2, and PD images from the Retrospective Image Registration Evaluation (RIRE) database[2]. We work with BrainWeb images containing 3% noise and 20% intensity non-uniformity, in order to achieve realistic results. For both databases the ground truth alignment is provided. Cross-sectional slices of the original volumes and entropy volumes from Brainweb are shown in figure 4 and from RIRE in figure 5.

In the simultaneous registration study, we compare ESM on entropy volumes to Gauss-Newton on entropy and original volumes. For Gauss-Newton on the original volumes, we select mutual information as similarity metric, which is corresponding to the sate-of-the-art configuration. Further, we use SSD as similarity measure for the registration with entropy volumes. We run 50 registrations, each starting from a random initial position. Each initial position has an accumulated RMS error of 45 over all volumes from the correct alignment, again weighting 1mm equal to 1°. The average residual error for each step is shown in figure 6. We observe that for the BrainWeb dataset the convergence of GN on the original volumes with MI and GN on the entropy images with SSD is identical, as to be expected. ESM converges, however, significantly faster than GN. On the RIRE data, most registrations do not converge for GN on the original volumes. GN with entropy images leads to good registration results. The convergence of ESM is, however, once again significantly faster than the one of GN.

6 Discussion

The experiments show the good performance of simultaneous registration using the APE framework and

[1] http://www.bic.mni.mcgill.ca/brainweb/
[2] http://www.insight-journal.org/rire/



(a) NCC

(b) Squared NCC

(c) MI

(d) Squared MI

Fig. 3. Plot of the average residual error for each iteration step for NCC, squared NCC, MI, and squared MI. The convergence of GN and ESM is significantly improved for the squared similarity measures. ESM is converging the fastest. We plot the curve of SD also for squared NCC and MI to facilitate the comparison.

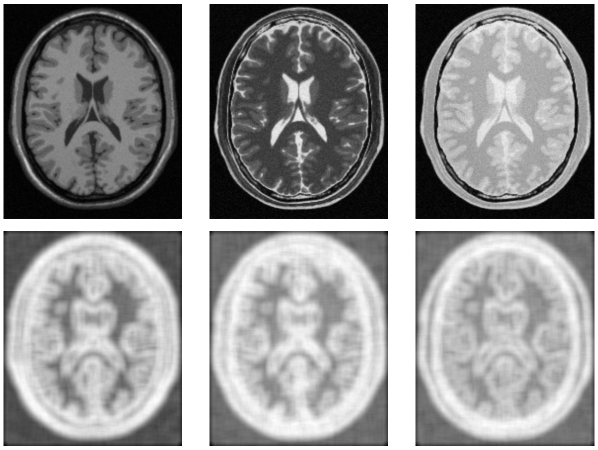


Fig. 4. Multi-modal images (T1, T2, PD) from BrainWeb dataset together with entropy images.

gradient-based optimization. The performance of the optimization methods, however, depends on the chosen similarity measure. In our experiments, the squared versions of NCC, CR, and MI performs better for GN and ESM. For all measures, the fastest approximation to the correct results are obtained with ESM. In most cases GN was faster than SD.

The convergence graphs are not all monotonic, as one would expect; approaching the ground truth further with each iteration until the convergence is achieved. The reasons for the increase lie, on the one hand, in the averaging over the 100 registrations, thus diverging trials lead to a large residual error that is averaged over. And on the other hand, we see the reasons in the conditioning of the Hessian approximation and the complex registration scenario. For ultrasound mosaicing, the volumes are inherently contaminated by speckle patterns, making it a difficult registration problem. Analogously, the registration of multi-modal volumes is challenging. The performance of ESM indicates that it is more robust in such noisy scenarios because the gradient information of both images is considered. Finally, our results show that structural representations and ESM nicely complement

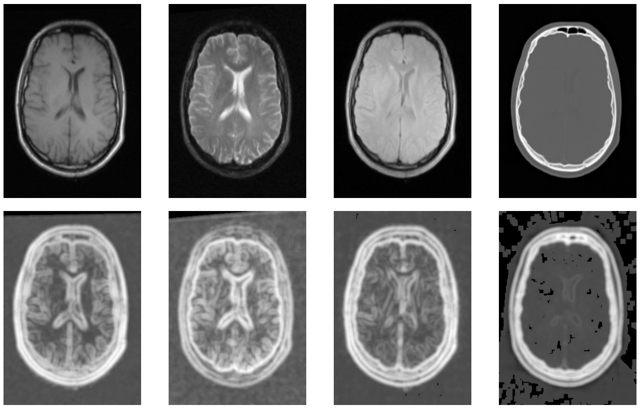


Fig. 5. Images from RIRE dataset (T1, T2, PD, CT) together with entropy images.

each other.

7 Conclusion

We presented further insights into multivariate similarity measures and optimization methods for simultaneous registration of multiple images. First, we deduced APE from a ML framework and showed its relation to the congealing framework. This required an extension of the congealing framework with neighborhood information. Second, we focused on efficient optimization methods for APE. We started the deduction of the optimization methods from the same Taylor expansion, to provide the reader a good overview of the methods and further insights into the relatively unknown ESM. We further presented the optimization of intrinsically non-squared similarity metrics in a least-squares optimization framework. Finally, we illustrated the usage of ESM for multimodal registration with structural representations. Our experiments showed a superior performance of ESM with respect to speed and accuracy for the case of ultrasound mosaicing and multi-modal registration.

For further illustration, we attach a video encoded with the Xvid codec (www.xvid.org) showing a comparison of optimization procedures for simultaneous registration. Also in the supplementary material is a list of derivatives of similarity measures.

Acknowledgments

We thank S. Benhimane and D. Zikic for valuable discussions as well as Siemens Corporate Research for ultrasound data. We are very grateful to the unknown reviewers for their insightful comments. This work was partly funded by the European Project ”PASSPORT” (ref. number 223904) and the Humboldt foundation.

References

[1] E. G. Learned-Miller, “Data driven image models through continuous joint alignment,” *IEEE Trans on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 236–250, 2006.

[2] G. Huang, V. Jain, and E. Learned-Miller, “Unsupervised Joint Alignment of Complex Images,” in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[3] M. Cox, S. Lucey, S. Sridharan, and J. Cohn, “Least squares congealing for unsupervised alignment of images,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.

[4] M. Cox, S. Sridharan, S. Lucey, and J. Cohn, “Least-squares congealing for large numbers of images,” in *IEEE International Conference on Computer Vision*, 2009, pp. 1949–1956.

[5] L. Zöllei, E. Learned-Miller, E. Grimson, and W. Wells, “Efficient Population Registration of 3D Data,” in *Computer Vision for Biomedical Image Applications, ICCV*, 2005.

[6] C. Wachinger, W. Wein, and N. Navab, “Three-dimensional ultrasound mosaicing,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Brisbane, Australia, October 2007.

[7] C. Wachinger and N. Navab, “Structural image representation for image registration,” in *CVPR Workshops, IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA)*, San Francisco, California, Jun. 2010.

[8] ——, “Manifold learning for multi-modal image registration,” in *11st British Machine Vision Conference (BMVC)*, 2010.

[9] J. West, J. Fitzpatrick, M. Wang, B. Dawant, C. Maurer Jr, R. Kessler, R. Maciunas, C. Barillot, D. Lemoine, A. Collignon *et al.*, “Comparison and evaluation of retrospective intermodality brain image registration techniques,” *Journal of Computer Assisted Tomography*, vol. 21, no. 4, p. 554, 1997.

[10] C. Studholme and V. Cardenas, “A template free approach to volumetric spatial normalization of brain anatomy,” *Pattern Recogn. Lett.*, vol. 25, no. 10, pp. 1191–1202, 2004.

[11] T. Cootes, S. Marsland, C. Twining, K. Smith, and C. Taylor, “Groupwise Diffeomorphic Non-rigid Registration for Automatic Model Building,” in *European Conference on Computer Vision*, 2004.

[12] K. Sidorov, S. Richmond, and D. Marshall, “An Efficient Stochastic Approach to Groupwise Non-rigid Image Registration,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[13] M. Yigitsoy, C. Wachinger, and N. Navab, “Temporal Groupwise Registration for Motion Modeling,” in *Information Processing in Medical Imaging (IPMI)*, 2011.

[14] C. Metz, S. Klein, M. Schaap, T. Van Walsum, and W. Niessen, “Nonrigid registration of dynamic medical imaging data using nd+t b-splines and a groupwise optimization approach,” *Medical Image Analysis*, vol. 15, no. 2, pp. 238–249, 2011.

[15] B. Lukas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” in *Image Understanding Workshop*, 1981.

[16] B. Horn and B. Schunck, “Determining optical flow,” *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[17] S. Baker and I. Matthews, “Lucas-Kanade 20 Years On: A Unifying Framework,” *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[18] K. Madsen, H. Nielsen, and O. Tingleff, *Methods for Non-Linear Least Squares Problems*, 2nd ed. Technical University of Denmark, 2004.

[19] S. Benhimane and E. Malis, “Real-time image-based tracking of planes using efficient second-order minimization,” in *IEEE/RSJ*, 2004, pp. 943–948.

[20] T. Vercauteren, X. Pennec, E. Malis, A. Perchant, and N. Ayache, “Insight into efficient image registration techniques and the demons algorithm,” *IPMI*, 2007.

[21] R. Stefanescu, X. Pennec, and N. Ayache, “Grid powered nonlinear image registration with locally adaptive regularization,” *Medical Image Analysis*, vol. 8, no. 3, pp. 325–342, 2004.

[22] C. Chefd’hotel, G. Hermosillo, and O. Faugeras, “Flows of diffeomorphisms for multimodal image registration,” in *IEEE International Symposium on Biomedical Imaging*, 2002, pp. 753–756.

[23] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Diffeomorphic demons: Efficient non-parametric image registration,” *NeuroImage*, vol. 45, no. 1, pp. 61–72, 2009.

[24] C. Wachinger and N. Navab, “Similarity Metrics and Efficient Optimization for Simultaneous Registration,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

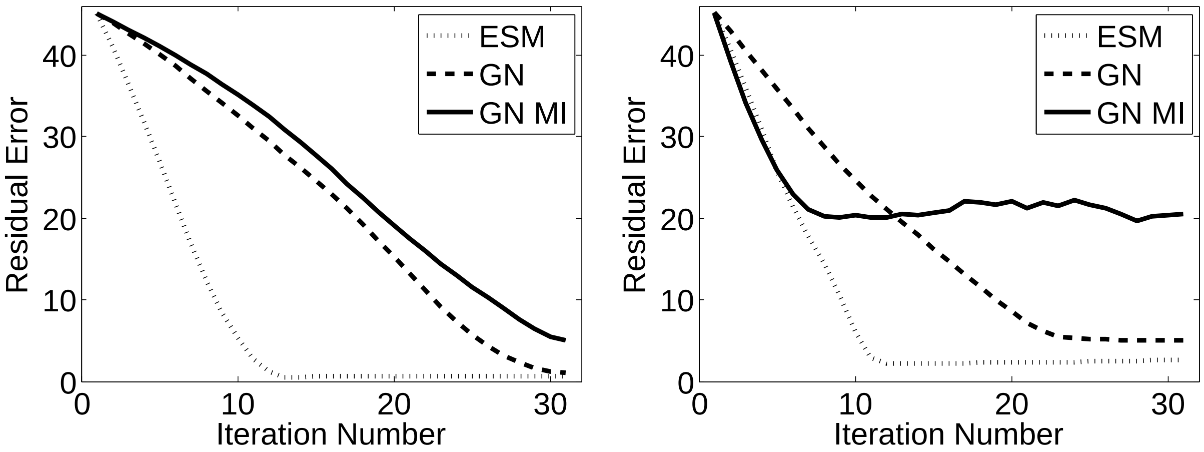


Fig. 6. Multi-modal simultaneous registration for BrainWeb (left) and RIRE (right) volumes. For 'GN' and 'ESM' SSD on the entropy volumes is used, for 'GN MI' mutual information on the original volumes.

[25] L. Zöllei, "A unified information theoretic framework for pair- and group-wise registration of medical images," Ph.D. thesis, MIT; MIT-CSAIL, 2006.

[26] P. A. Viola, "Alignment by maximization of mutual information," Ph.D. thesis, Massachusetts Institute of Technology, 1995.

[27] A. Roche, G. Malandain, and N. Ayache, "Unifying maximum likelihood approaches in medical image registration," *International Journal of Imaging Systems and Technology: Special Issue on 3D Imaging*, vol. 11, no. 1, pp. 71–80, 2000.

[28] C. Wachinger and N. Navab, "A Contextual Maximum Likelihood Framework for Modeling Image Registration," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[29] S. Li, *Markov random field modeling in image analysis*. Springer-Verlag New York Inc, 2009.

[30] C. Wachinger, "Three-dimensional ultrasound mosaicing," Master's thesis, Technische Universität München, March 2007.

[31] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[32] P. Lee and J. Moore, "Gauss-Newton-on-manifold for Pose Estimation," *Journal of Industrial and Management Optimization*, vol. 1, no. 4, p. 565, 2005.

[33] R. Mahony and J. Manton, "The Geometry of the Newton Method on Non-Compact Lie Groups," *Journal of Global Optimization*, vol. 23, no. 3, pp. 309–327, 2002.

[34] M. Zefran, V. Kumar, and C. Croke, "On the generation of smooth three-dimensional rigid body motions," *Robotics and Automation, IEEE Transactions on*, vol. 14, no. 4, pp. 576–589, 1998.

[35] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*. CRC Press, 1994.

[36] E. Malis, "Méthodologies d'estimation et de commande à partir d'un système de vision," Habilitation, Nice-Sophia Antipolis, 2008.

[37] S. Benhimane, "Vers une approche unifiee pour le suivi temps-reel et l'asservissement visuel," Docteur en Sciences - Specialite: Informatique Temps-reel, Automatique et Robotique, Ecole Nationale Superieure des Mines de Paris, 2006.

[38] G. Hermosillo, C. Chefd'Hotel, and O. Faugeras, "Variational Methods for Multimodal Image Matching," *International Journal of Computer Vision*, vol. 50, no. 3, pp. 329–343, 2002.

[39] W. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Medical Image Analysis*, 1996. [Online]. Available: http://citeseer.ist.psu.edu/354937.html

[40] E. D'Agostino, F. Maes, D. Vandermeulen, and P. Suetens, "A viscous fluid model for multimodal non-rigid image registration using mutual information," *Medical image analysis*, vol. 7, no. 4, pp. 565–575, 2003.

[41] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual information based registration of medical images: A survey." *IEEE Trans. Med. Imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[42] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[43] B. Turlach, "Bandwidth selection in kernel density estimation: A review," *CORE and Institut de Statistique*, vol. 19, no. 4, pp. 1–33, 1993.

[44] P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Transactions on Image Processing*, vol. 9, pp. 2083–2099, 2000.

[45] G. Panin and A. Knoll, "Mutual information-based 3d object tracking," *International Journal of Computer Vision*, vol. 78, no. 1, pp. 107–118, 2008.

[46] A. Dame and E. Marchand, "Accurate real-time tracking using mutual information," in *IEEE Int. Symp. on Mixed and Augmented Reality, ISMAR'10*, Seoul, Korea, October 2010, pp. 47–56.

Christian Wachinger received his MS and PhD degrees from Technische Universität München (TUM), Germany in 2007 and 2011, respectively. During this time, he spent a year at Telecom ParisTech and Ecole Centrale Paris as well as six months at Siemens corporate research, Princeton. He is currently with the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT) and the Department of Neurology, Harvard Medical School. His research interests include image registration, probabilistic modeling, spectral analysis, segmentation, and its medical applications.

Nassir Navab is a full professor and director of the institute for Computer Aided Medical Procedures (CAMP) at Technical University of Munich (TUM) with a secondary faculty appointment at its Medical School. He is also acting as Chief Scientific Officer for SurgicEye. In November 2006, he was elected as a member of board of directors of MICCAI society. He has served on the Steering Committee of the IEEE Symposium on Mixed and Augmented Reality between 2001 and 2008. He is the author of hundreds of peer reviewed scientific papers and over 40 US and international patents. He has received Siemens Inventor of the Year award in 2001, and SMIT technology award in 2010 and co-authored many awarded papers in prestigious conferences.