Similarity Metrics and Efficient Optimization for Simultaneous Registration

Christian Wachinger and Nassir Navab
Computer Aided Medical Procedures (CAMP), Technische Universität München, Germany
wachinge@in.tum.de, navab@in.tum.de

Abstract

We address the alignment of a group of images with simultaneous registration. Therefore, we provide further insights into a recently introduced class of 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 the intrinsically non-squared similarity measures NCC, CR, and MI, in this least-squares optimization framework. Finally, we evaluate the performance of the optimization strategies with respect to the similarity measures, obtaining very promising results for ESM.*

1. Introduction

The analysis of a group or population of images requires their alignment to a canonical pose. Examples are the alignment of 2D face images for their later identification [4], the alignment of 3D tomographic images for the creation of an atlas [14], or the creation of mosaics from ultrasonic volumes [12]. 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 unwanted 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 [5]. Another approach was presented by Wachinger *et al*. [12], which *accumulates pairwise 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, 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 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 report, 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. And finally, we derive efficient gradient-based optimization strategies for simultaneous registration with APE as multivariate similarity framework.

1.1. Related Work

Simultaneous registration has many applications in computer vision and medical imaging when it comes to the alignment of multiple images. Learned-Miller [5] 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. Huang *et al*. [4] applied congealing to align 2D face images, essential for their later identification. Zöllei *et al*. [14] used congealing for the simultaneous alignment of a population of brain images for brain at-

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

las 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*. [3] 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 [14]. Wachinger *et al*. [12] 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.

A good overview of gradient-based optimization methods is provided in Baker and Matthews [1] and Madsen *et al*. [7]. 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 [2]. 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*. [11] achieved good results for the pairwise 2D image alignment with ESM.

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. Due to limited space we only show the major steps, but provide a detailed derivation in the supplementary material [1]. The ML framework for intensity-based registration is:

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

with n images $\mathcal{I} = \{I_1, \ldots, I_n\}$, the transformation parameters $\mathbf{x}$, the joint density function p, and the correct alignment $\hat{\mathbf{x}}$. In the following we will no longer consider $\mathbf{x}$ explicitly in the density function, but it should be clear that it determines the alignment of the images.

2.1. Accumulated Pair-Wise Estimates

APE approximates the joint likelihood function with pair-wise estimates [12]:

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

Assuming a Gaußian distribution of the density p, i.i.d. coordinate samples, and various intensity mappings between the images, popular similarity measures like SSD, NCC, CR, and MI can be derived from the log-likelihood term $\log p(I_j|I_i)$ [9]. APE therefore presents a framework for a class of similarity measures. To deduce it, we take the n-th power of p and then repeatedly apply a combination of product rule and conditional independence of the images:

$$p(I_1, \ldots, I_n)^n = \prod_{i=1}^{n} p(I_i) \cdot \prod_{i=1}^{n} \prod_{j \neq i} p(I_j | I_i). \tag{3}$$

Further, we apply the logarithm to this equation to deduce:

$$\log p(I_1, .., I_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(I_i) + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \log p(I_j | I_i) \tag{4}$$

$$\approx \sum_{i=1}^{n} \sum_{j \neq i} \log p(I_j | I_i) \tag{5}$$

where the approximation is justified because the term $\sum_{i=1}^{n} \log p(I_i)$ remains constant during the optimization and the multiplication with a scalar factor n does not alter the optimization. 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 [5], independent but *not* identical distributions of the coordinate samples $s_k \in \Omega$ in the grid Ω are assumed:

$$p(I_1, \ldots, I_n) = \prod_{s_k \in \Omega} p^k(I_1(s_k), \ldots, I_n(s_k)). \tag{6}$$

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

$$p(I_1, \ldots, I_n) = \prod_{s_k \in \Omega} \prod_{i=1}^{n} p^k(I_i(s_k)). \tag{7}$$

While the consideration of neighboring pixels, surrounding a sample s_k, was already discussed in [5], referred to as pixel cylinder, the consideration of neighboring images has not yet been proposed. So, instead of an independence of images, we assume that each image I_i depends on a certain neighborhood $\mathcal{N}_i$ of images:

$$p(I_1, \ldots, I_n) = \prod_{s_k \in \Omega} \prod_{i=1}^{n} p^k(I_i(s_k) | I_{\mathcal{N}_i}(s_k)). \tag{8}$$

This extension also allows us to derive the voxel-wise extension of SSD proposed in [12], see supplementary material.

[1] http://www.webcitation.org/5fdVLazpg

2.3. Comparison of APE and Congealing

Having APE and congealing derived, the question comes up about their relationship. It is in fact possible to deduce a connection between the two approaches. The detailed proof is stated in the supplementary material. Therein we start with the congealing and derive APE. To make this possible the following assumptions have to be made: 1) complete neighborhood, 2) conditional independence of images, and 3) i.i.d. distribution of coordinate samples. While 3) was explicitly chosen by the design of congealing and 2) by the deduction of APE, the novel part is the neighborhood 1), which relates these two approaches. The extended congealing in equation (8) 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. Extended congealing and APE do not present actual similarity measures, but rather frameworks, where further information about the distribution has to be provided to derive similarity measures. Taking *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 is most versatile.

3. Efficient Optimization Methods

We derive efficient gradient-based optimization methods for 3D rigid transformations, but the parameterization can be easily adapted for different types of alignments.

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 [6, 8]. 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 $\mathbf{x}$ can be seen as 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:

$$\mathbf{x} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \tag{9}$$

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

$$\mathfrak{se}(3) = \left\{ \begin{bmatrix} \mathbf{\Omega} & \mathbf{v} \\ \mathbf{0} & 0 \end{bmatrix} | \mathbf{\Omega} \in \mathbb{R}^{3\times3}, \mathbf{v} \in \mathbb{R}^3, \mathbf{\Omega}^\top = -\mathbf{\Omega} \right\}.$$

The exponential map relates the Lie algebra to the Lie group: $\exp : \mathfrak{se}(3) \rightarrow \mathbb{SE}(3)$. 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.

Let $\mathcal{L} = (\mathbf{l}_1, \ldots, \mathbf{l}_6)$ be the standard basis of $\mathfrak{se}(3)$. Each element $\mathbf{h} \in \mathfrak{se}(3)$ can be expressed as a linear combination of matrices $\mathbf{h} = \sum_{i=1}^{6} h_i \mathbf{l}_i$ with h_i varying over the manifold [13]. 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 $\mathbf{h}$, such that:

$$\mathbf{y} = \mathbf{x} \circ \exp(\mathbf{h}) = \mathbf{x} \circ \exp\left(\sum_{i=1}^{6} h_i \mathbf{l}_i\right). \tag{10}$$

Let us further denote the transformation of a point $\mathbf{p} \in \mathbb{R}^3$ through the mapping $\mathbf{x} \in \mathbb{SE}(3)$ with $w(\Theta(\mathbf{x}), \mathbf{p})$ in the Euclidean embedding space Θ.

3.2. Optimization Methods

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

$$E(\mathbf{x}) = \sum_{i \neq j} F_{i,j}(\mathbf{x}) = \sum_{i \neq j} \frac{1}{2} ||\mathbf{f}_{i,j}(\mathbf{x})||^2 \tag{11}$$

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

Regarding equation (11), we see that we deal with a nonlinear 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 $\mathbf{h}$:

$$E(\mathbf{x} \circ \exp(\mathbf{h})) \approx E(\mathbf{x}) + \mathbf{J}_E(\mathbf{x}) \cdot \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \cdot \mathbf{H}_E(\mathbf{x}) \cdot \mathbf{h} \tag{12}$$

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 general gradient direction $\mathbf{h}$ is a combination of elements from the Lie algebra $\mathbf{h}_i \in \mathfrak{se}(3)$, resulting in $\mathbf{h} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$. The Newton-Raphson (NR) method then has the following compositional update:

$$\mathbf{H}_{F_{i,j}} \mathbf{h}_{i,j}^{\text{NR}} = -\mathbf{J}_{F_{i,j}}^{\top} \qquad \mathbf{x} \leftarrow \mathbf{x} \circ \exp(\mathbf{h}^{\text{NR}}). \tag{13}$$

The global update $\mathbf{h}^{\text{NR}}$ is obtained by summing up the pairwise updates, following the structure of the cost function E in equation (11), leading to

$$\mathbf{h}^{\text{NR}} = \left[\sum_i \mathbf{h}_{i,1}^{\text{NR}}, \ldots, \sum_i \mathbf{h}_{i,n}^{\text{NR}} \right]. \tag{14}$$

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 [1]. In the field of non-linear least squares optimization most of the methods use an approximation of the Hessian [7]. In the following we present different possibilities for approximating the Hessian by a positive definite matrix $\hat{\mathbf{H}}$.

Steepest-Descent The Hessian is approximated by the identity $\hat{\mathbf{H}} = \alpha \cdot \mathbf{I}$, with α the step length, and consequently only considers a first-order Taylor expansion of E. The convergence is linear.

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

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:

$$\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{f}(\mathbf{x}) + \mathbf{J}_{\mathbf{f}}(\mathbf{x}) \cdot \mathbf{h}. \tag{15}$$

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 (11) gives:

$$E(\mathbf{x} \circ \exp(\mathbf{h})) \approx \sum_{i \neq j} \frac{1}{2} ||\mathbf{f}_{i,j}(\mathbf{x} \circ \exp(\mathbf{h}))||^2 \tag{16}$$

$$= \sum_{i \neq j} \frac{1}{2} \mathbf{f}_{i,j}(\mathbf{x} \circ \exp(\mathbf{h}))^{\top} \mathbf{f}_{i,j}(\mathbf{x} \circ \exp(\mathbf{h})) \tag{17}$$

$$= \sum_{i \neq j} \left(F_{i,j}(\mathbf{x}) + \mathbf{h}^{\top} \mathbf{J}_{\mathbf{f}_{i,j}}^{\top} \mathbf{f}_{i,j} + \frac{1}{2} \mathbf{h}^{\top} \mathbf{J}_{\mathbf{f}_{i,j}}^{\top} \mathbf{J}_{\mathbf{f}_{i,j}} \mathbf{h} \right). \tag{18}$$

By comparison with Equation (12), and considering the gradient $\mathbf{J}_F = \mathbf{J}_f^{\top} \mathbf{f}$, we can see that the Hessian is approximated by $\hat{\mathbf{H}} = \mathbf{J}_{\mathbf{f}}^{\top} \mathbf{J}_{\mathbf{f}}$.

We approximate the global Gauß-Newton step $\mathbf{h}^{\text{GN}}$ by the pairwise optimal steps $\mathbf{h}_{i,j}^{\text{GN}}$, analogously to Newton-Raphson, see equation (14). This leads to the update:

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

with $\mathbf{h}^{\text{GN}} = \left[\sum_i \mathbf{h}_{i,1}^{\text{GN}}, \ldots, \sum_i \mathbf{h}_{i,n}^{\text{GN}} \right]$. Gauß-Newton has only in specific cases quadratic convergence [7, 2].

ESM The efficient second-order minimization procedure originally comes from the field of vision-based control [2]. It is an extension of GN and incorporates further knowledge about the specificity of the optimization problem to achieve better results.

More precisely, ESM uses the fact, that when the images are aligned with the optimal transformation $\mathbf{x}^{\text{opt}}$, the images and therefore also their gradients should be very close to each other. This can be used to ameliorate the search direction of the Newton methods. For the standard Newton-Raphson method, the first and second order derivatives around $\mathbf{0}$ are used to build a second-order approximation, see Equation (13). The Gauß-Newton method neglects the second derivative and thus can only build a first-order approximation. In the ESM, the first-order derivatives around $\mathbf{0}$ and $\mathbf{x}^{\text{opt}}$ are used to build a second-order approximation without the necessity of a second-order derivative.

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

$$\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{f}(\mathbf{x}) + \mathbf{J}_{\mathbf{f}}(\mathbf{x}) \cdot \mathbf{h} + \frac{1}{2} \mathbf{h}^{\top} \cdot \mathbf{H}_{\mathbf{f}}(\mathbf{x}) \cdot \mathbf{h}. \tag{19}$$

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

$$\mathbf{J}_{\mathbf{f}}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{J}_{\mathbf{f}}(\mathbf{x}) + \mathbf{H}_{\mathbf{f}}(\mathbf{x}) \cdot \mathbf{h}. \tag{20}$$

Plugging this first-order series in the approximation shown in equation (19) we get the second-order approximation:

$$\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h})) \approx \mathbf{f}(\mathbf{x}) + \frac{1}{2} [\mathbf{J}_{\mathbf{f}}(\mathbf{x}) + \mathbf{J}_{\mathbf{f}}(\mathbf{x} \circ \exp(\mathbf{h}))] \mathbf{h}. \tag{21}$$

Comparing this equation with equation (15) 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 $\mathbf{J}_{\mathbf{f}}^{\text{ESM}} = \frac{1}{2}(\mathbf{J}_{\mathbf{f}}(\mathbf{x}) + \mathbf{J}_{\mathbf{f}}(\mathbf{x} \circ \exp(\mathbf{h})))$ instead of only $\mathbf{J}_{\mathbf{f}}(\mathbf{x})$. Leading to an approximation of the Hessian by $\hat{\mathbf{H}} = \mathbf{J}_{\mathbf{f}}^{\text{ESM}^{\top}} \mathbf{J}_{\mathbf{f}}^{\text{ESM}}$. The compositional update is:

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

with $\mathbf{h}^{\text{ESM}} = \left[\sum_i \mathbf{h}_{i,1}^{\text{ESM}}, \ldots, \sum_i \mathbf{h}_{i,n}^{\text{ESM}} \right]$. ESM has at least quadratic convergence [2].

3.3. Gradient Calculation

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

Steepest-Descent We begin with the gradient for the general cost function E by considering only 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. More precisely we would have to write $I_j(\mathbf{x}_j \circ \exp(\mathbf{h}_j))$ but we continue with the relaxed notation because it should not lead to ambiguities. The gradient has then the form:

$$\begin{aligned}
\mathbf{J}_E(\mathbf{x}) &= \frac{\partial E(\mathbf{x} \circ \exp(\mathbf{h}))}{\partial \mathbf{h}} = \sum_{i \neq j} \frac{\partial F_{i,j}(\mathbf{x} \circ \exp(\mathbf{h}))}{\partial \mathbf{h}} \\
&= \sum_{i \neq j} \frac{\partial \text{SM}(I_i(\mathbf{x}), I_j(\mathbf{x} \circ \exp(\mathbf{h})))}{\partial \mathbf{h}} \qquad (22) \\
&= \sum_{i \neq j} \left.\frac{\partial \text{SM}(I_i, I)}{\partial I}\right|_{I=I_j^\downarrow} \left.\frac{\partial I_j(w(\mathbf{x}); \mathbf{q})}{\partial \mathbf{q}^\top}\right|_{\mathbf{q}=\mathbf{p}} \\
&\qquad \left.\frac{\partial w(\mathbf{y}; \mathbf{p})}{\partial \mathbf{y}^\top}\right|_{\mathbf{y}=\Theta(\text{Id})} \left.\frac{\partial \Theta(\exp(h_k \mathbf{l}_k))}{\partial h_k}\right|_{h_k=0} \\
&= \sum_{i \neq j} \nabla \text{SM}(I_i, I_j^\downarrow) \cdot \nabla I_j^\downarrow \cdot \mathbf{J}_w \cdot \Theta(\mathbf{l}) \qquad (23)
\end{aligned}$$

with setting $I_j^\downarrow = I_j(\mathbf{x} \circ \exp(\mathbf{h}))$ and $I_i = I_i(\mathbf{x})$. ∇SM is the derivative of the used similarity measure, $\nabla I_j^\downarrow$ the gradient of the transformed image $I_j^\downarrow$, $[\mathbf{J}_w]_\mathbf{p}$ the derivative of the transformation, formulated in the Euclidean embedding space, which depends only on the homogeneous coordinates of the considered voxel $\mathbf{p} = [x, y, z, 1]^\top$ (3×12 matrix):

$$[\mathbf{J}_w]_\mathbf{p} = \left.\frac{\partial w(\Theta(\mathbf{x}); \mathbf{p})}{\partial \Theta(\mathbf{x})}\right|_{\mathbf{x}=\Theta(\text{Id})} = \begin{bmatrix} \mathbf{p}^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{p}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{p}^\top \end{bmatrix}$$

and $\Theta(\mathbf{l})$ stacks the basis vectors of $\mathfrak{se}(3)$ expressed in the Euclidean embedding space (12×6 matrix):

$$\Theta(\mathbf{l}) = [\Theta(\mathbf{l}_1), \ldots, \Theta(\mathbf{l}_6)]$$

with the embedding Θ from the Lie group $\mathbb{SE}(3)$ to the Euclidean space $\mathbb{R}^{12}$, $\mathbb{SE}(3) \to \mathbb{R}^{12}$, $\mathbf{x} \mapsto \Theta(\mathbf{x})$.

Gauß-Newton For the derivation of the gradient $\mathbf{J}_\mathbf{f}$, 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 (11). When considering SSD we can simply set $E(\mathbf{x}) = \sum_{i \neq j} \text{SSD}_{i,j}(\mathbf{x})$, since SSD is intrinsically a least-squares problem.

This is not the case for other similarity measures like NCC, CR, and MI. In order to ensure the least-squares nature, we square the similarity measures, leading to $E(\mathbf{x}) = \sum_{i \neq j} \text{SM}_{i,j}^2$, with $\text{SM}_{i,j} = \text{SM}(I_i, I_j^\downarrow)$. Obviously, optimizing the squared similarity measure has far-ranging consequences, which we investigate further in section 3.3.1. The gradient $\mathbf{J}_\mathbf{f}$ at a certain voxel $\mathbf{p}$ in the grid is then calculated as:

$$\begin{aligned}
\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x}) &= \left.\frac{\partial \mathbf{f}_{i,j}(\mathbf{x} \circ \exp(\mathbf{h}))}{\partial \mathbf{h}}\right|_{\mathbf{h}=0} \qquad (24) \\
&= \left.\frac{\partial \text{SM}(I_i(\mathbf{x}), I_j(\mathbf{x} \circ \exp(\mathbf{h})))}{\partial \mathbf{h}}\right|_{\mathbf{h}=0} \qquad (25) \\
&= \nabla \text{SM}_{i,j} \cdot \nabla I_j^\downarrow \cdot \mathbf{J}_w \cdot \Theta(\mathbf{l}). \qquad (26)
\end{aligned}$$

ESM The last gradient that remains is $\mathbf{J}_\mathbf{f}^{\text{ESM}}$ for the ESM. Here we also consider the squared similarity measures like for GN. The calculation of $\mathbf{J}_\mathbf{f}^{\text{ESM}}$ is difficult because part of its calculation is $\mathbf{J}_\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h}))$, which depends on $\mathbf{h}$ that we want to solve for. Taking the optimal transformation $\mathbf{x}^{\text{opt}} = \mathbf{x} \circ \exp(\mathbf{h}^{\text{opt}})$ that is reached after the optimal update step $\mathbf{h}^{\text{opt}}$, the main assumption of ESM is that the gradient of the perfectly aligned image $I_j(\mathbf{x} \circ \exp(\mathbf{h}^{\text{opt}}))$ can be approximated by the gradient of the fixed image $I_i(\mathbf{x})$, so $\nabla I_j(\mathbf{x} \circ \mathbf{h}^{\text{opt}}) \approx \nabla I_i(\mathbf{x})$, leading to

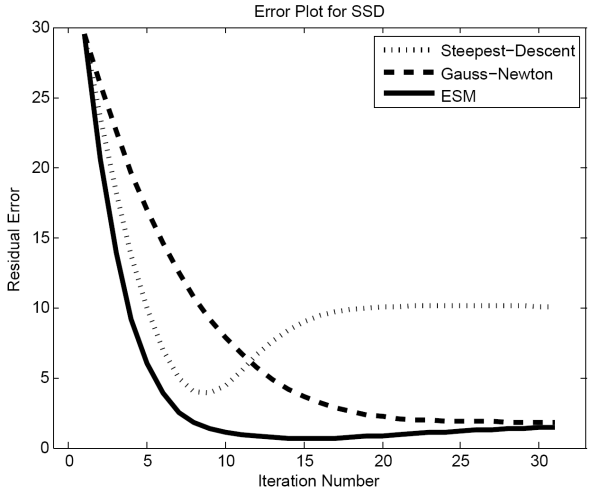
$$\mathbf{J}_{\mathbf{f}_{i,j}}(\mathbf{x} \circ \exp(\mathbf{h}^{\text{opt}})) \approx \nabla \text{SM}_{i,j} \cdot \nabla I_i \cdot \mathbf{J}_w \cdot \Theta(\mathbf{l}). \qquad (27)$$

Obviously, this only makes sense for images of the same modality. Considering the definition of the gradient $\mathbf{J}^{\text{ESM}} = \frac{1}{2}(\mathbf{J}_\mathbf{f}(\mathbf{x}) + \mathbf{J}_\mathbf{f}(\mathbf{x} \circ \exp(\mathbf{h})))$, and equations (26) and (27), we finally get:

$$\mathbf{J}_{\mathbf{f}_{i,j}}^{\text{ESM}}(\mathbf{x}) = \frac{1}{2} \cdot \nabla \text{SM}_{i,j} \cdot (\nabla I_i + \nabla I_j^\downarrow) \cdot \mathbf{J}_w \cdot \Theta(\mathbf{l}). \qquad (28)$$

3.3.1 Gradient of Similarity Measures

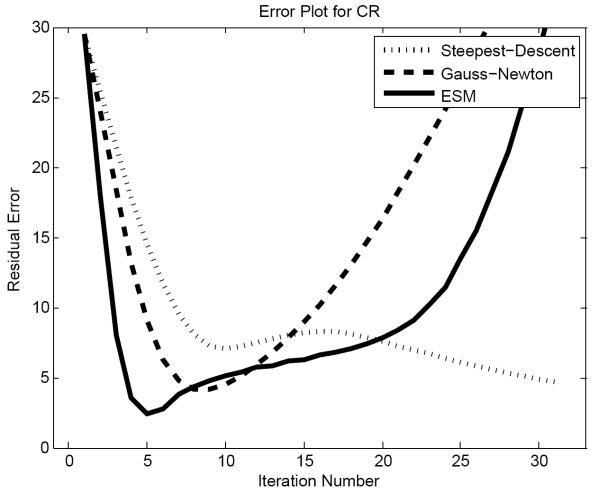
As mentioned in the last section, we optimize the squared similarity measure for normalized cross-correlation (NCC), correlation-ratio (CR), and mutual information (MI) to ensure the least-squares nature of the optimization problem. For sum of squared differences (SSD) this is not necessary. The interesting question is what the consequences are by 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 a lower bound, which is -1 and 0, respectively. To avoid these optimization problems,



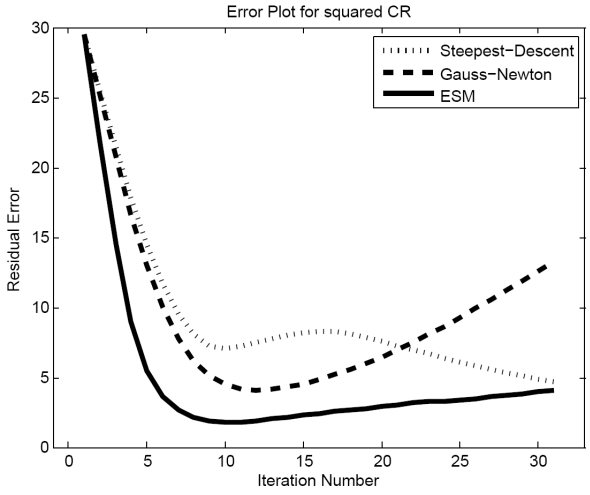
(a) SSD

	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

Figure 1. Plot of the average residual error for each iteration step for SSD, CR, and squared CR. Comparing CR and squared CR shows the much better performance of GN and ESM. ESM converges the fastest and leads to the smallest residual error.

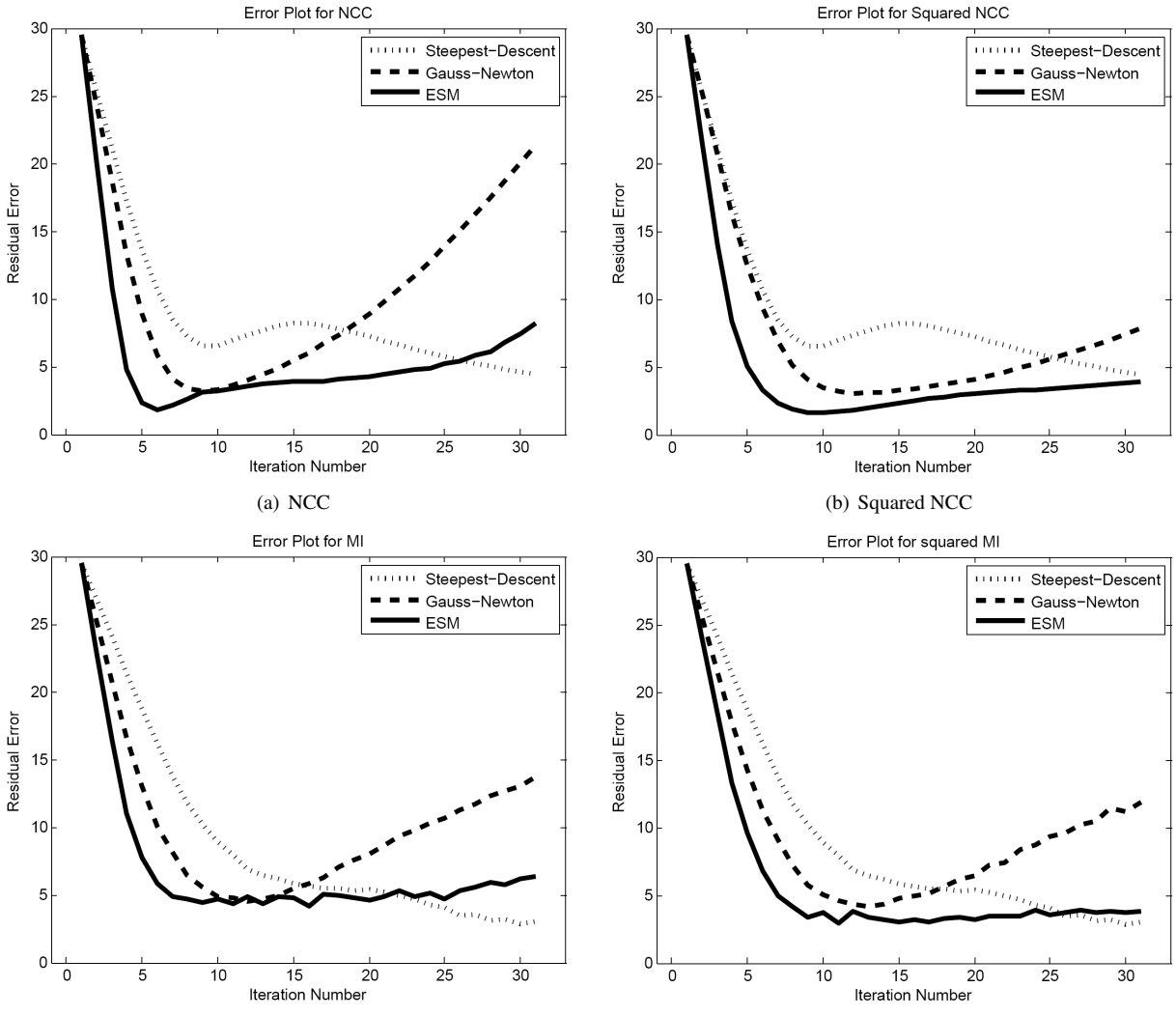
we can simply add a constant ν to the similarity measures $\mathrm{SM}_{i,j} + \nu$, to guarantee that they are in the positive range.

We list the actual derivatives of the similarity measures in the supplementary material. 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 $(\mathbf{J_f}^\top \mathbf{J_f})\mathbf{h} = -\mathbf{J_f}\mathbf{f}$, or a QR-decomposition on $\mathbf{J_f}\mathbf{h} = -\mathbf{f}$. Since the normal equations worsen the numerical condition of the problem, the QR-decomposition presents the stabler choice.

4. Experiments

The experiments were conducted on four 3D ultrasound acquisitions from a baby phantom, having a resolution of $64 \times 64 \times 64$ voxels, see Figure 3. 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. 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 the registration 100 times for each configuration to assess its performance.

In Figure 1 and 2, the averaged residual error is plotted with respect to the iteration number. For SSD, see Figure 1(a), we only have one plot because we do not have to consider the squared variant of it, like already mentioned. The best performance was obtained with ESM, leading to the fastest convergence. But also the Gauß-Newton method lead to a robust convergence. The gradient-descent did not perform well. Although it seems to appraoch the correct



(a) NCC (b) Squared NCC

(c) MI (d) Squared MI

Figure 2. 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 much better for the squared similarity measures. ESM is converging the fastest.

alignment nicely at the beginning, it diverges into another optimum. In the table in Figure 1(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 1(c), the results for GN and ESM are not good. All of the 100 runs diverged. SD, although slower, performed much better. The situation changes a lot, when optimizing the squared function, see Figure 1(d). The ESM quickly approaches the correct alignment and although it diverges a bit afterwards, the error stays below 5. 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 2, 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 were performed on an Intel dual-core 2.4 GHz processor having 2 GB of RAM. The time for one registration, where we allowed for 30 iterations, was below one minute.

5. Discussion

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

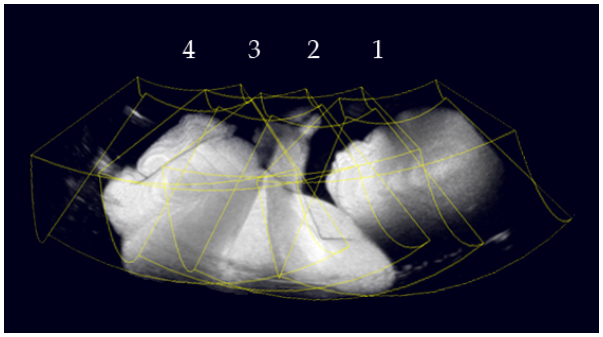


Figure 3. Mosaic of baby phantom from 4 acquisitions.

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 performed 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. Using SD has the additional drawback that the step length α has to be set manually, when no line search is used, which would require further evaluations of the expansive cost function. But surprisingly most of the graphs are not monotonic. Normally, one would expect strictly monotonically decreasing graphs like we obtained it for GN in combination with SSD; approaching the ground truth further with each iteration until the convergence is achieved. In most cases, the graphs are increasing to the end. For ESM the increase is pretty low, though.

We see the reasons for the increase, one the one hand, in the averaging over the 100 registrations, thus diverging algorithms lead to a large residual error that is averaged over. On the other hand, we see the reasons in the complex registration scenario. Even though the structures seem clear, these are still ultrasound volumes we are dealing with, which are inherently contaminated by speckle patterns. This has consequences on the cost function, and more importantly on the gradient calculation, making it a hard registration problem. ESM is more robust in such a noisy scenario because the gradient information of both images are considered.

6. Conclusion

We presented further insights about 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 also presented the optimization of intrinsically non-squared similarity metrics in a least-squares optimization framework. Our experiments showed a superior performance of ESM with respect to speed and accuracy.

7. Acknowledgment

We thank S. Benhimane, D. Zikic, and L. Zöllei for valuable discussions and W. Wein from Siemens Corporate Research for software tools and ultrasound data. This work was partly funded by the European Project "PASSPORT", ref. number 223904.

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[2] S. Benhimane and E. Malis. Real-time image-based tracking of planes using efficient second-order minimization. In *IEEE/RSJ*, pages 943–948, 2004.

[3] T. Cootes, S. Marsland, C. Twining, K. Smith, and C. Taylor. Groupwise Diffeomorphic Non-rigid Registration for Automatic Model Building. In *ECCV*, 2004.

[4] G. Huang, V. Jain, and E. Learned-Miller. Unsupervised Joint Alignment of Complex Images. In *ICCV*, pages 1–8, 2007.

[5] E. G. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 28(2):236–250, 2006.

[6] P. Lee and J. Moore. Gauss-Newton-on-manifold for Pose Estimation. *Journal of Industrial and Management Optimization*, 1(4):565, 2005.

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

[8] R. Mahony and J. Manton. The Geometry of the Newton Method on Non-Compact Lie Groups. *Journal of Global Optimization*, 23(3):309–327, 2002.

[9] A. Roche, G. Malandain, and N. Ayache. Unifying maximum likelihood approaches in medical image registration. *Int J of Imaging Syst and Techn*, 11(1):71–80, 2000.

[10] C. Studholme and V. Cardenas. A template free approach to volumetric spatial normalization of brain anatomy. *Pat Rec Let*, 25(10):1191–1202, 2004.

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

[12] C. Wachinger, W. Wein, and N. Navab. Three-dimensional ultrasound mosaicing. In *MICCAI*, 2007.

[13] M. Zefran, V. Kumar, and C. Croke. On the generation of smooth three-dimensional rigid body motions. *Robotics and Automation, IEEE Transactions on*, 14(4):576–589, 1998.

[14] 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.