HHS Public Access

Author manuscript

Front Appl Math Stat. Author manuscript; available in PMC 2023 October 02.

Published in final edited form as:

Front Appl Math Stat. 2021 ; 7: . doi:10.3389/fams.2021.718607.

An optical flow based left-invariant metric for natural gradient descent in affine image registration

Daniel Tward[1,*]

[1]Brain Mapping Center, University of California Los Angeles, Departments of Computational Medicine and Neurology, Los Angeles, CA, USA

Abstract

Accurate spatial alignment is essential for any population neuroimaging study, and affine (12 parameter linear/translation) or rigid (6 parameter rotation/translation) alignments play a major role. Here we consider intensity based alignment of neuroimages using gradient based optimization, which is a problem that continues to be important in many other areas of medical imaging and computer vision in general. A key challenge is robustness. Optimization often fails when transformations have components with different characteristic scales, such as linear versus translation parameters. Hand tuning or other scaling approaches have been used, but efficient automatic methods are essential for generalizing to new imaging modalities, to specimens of different sizes, and to big datasets where manual approaches are not feasible. To address this we develop a left invariant metric on these two matrix groups, based on the norm squared of optical flow induced on a template image. This metric is used in a natural gradient descent algorithm, where gradients (covectors) are converted to perturbations (vectors) by applying the inverse of the metric to define a search direction in which to update parameters. Using a publicly available magnetic resonance neuroimage database, we show that this approach outperforms several other gradient descent optimization strategies. Due to left invariance, our metric needs to only be computed once during optimization, and can therefore be implemented with negligible computation time.

Keywords

neuroimaging; image registration; optimization; natural gradient descent; Riemannian metric

1 INTRODUCTION

Modern neuroimaging techniques are providing a detailed examination of the nervous system at unprecedented scale. Human studies such as the Alzheimer's Disease Neuroimaging Initiative [1] or Open Access Series of Imaging Studies [2] are providing

*Correspondence: Daniel Tward, dtward@mednet.edu.

AUTHOR CONTRIBUTIONS

DT conceived of the study, developed the methods, carried out the analysis, and wrote the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

hundreds of publicly accessible three dimensional brain images at the millimeter scale to the neuroscience, medical imaging, and computer vision communities. Consortiums such as the BRAIN Initiative Cell Census Network are making petabytes of neuroimaging data available from animal models at the micron and submicron scale [3]. Extracting insight from these massive databases remains a challenge, and establishing common coordinate systems for reporting results is an important step to enable synchronization between different laboratories, between imaging modalities, and across populations[4, 5, 6].

This standardization is enabled by image registration techniques, where optimal spatial transformations are calculated to maximize similarity between new observations and images in standard coordinates. There are many approaches to implementing these techniques. One framework involves building alignments based on point sets including landmarks [7, 8, 9] , curves [10], and surfaces [11, 12]. Another framework involves voxel based imaging data [13, 14, 15, 16, 17]. Additionally hybrid approaches are often used that can combine the strengths of point based and voxel based techniques [18, 19, 20]. There are many more thorough reviews of these techniques, including comparisons in [21].

In this work we focus on voxel intensity based image registration with gradient based optimization. In these approaches robustness to parameter selection is essential. When optimizing over parameters, differences in scales must be accounted for during gradient descent before updating parameters in the negative gradient direction. For example, simple elastix [22] defines an estimate scales routine. Simple elastix is based on the simple ITK wrapper [23] of the powerful ITK library [24, 25]. Simple ITK hard-codes relative scales in manner that tends to give good results for human neuroimages at the millimeter scale. However, efficient automatic methods are essential for generalizing to new imaging modalities, to specimens of different sizes, and to big datasets where manual approaches are not feasible

Here we address this gap by designing a metric for natural gradient descent [26]. In differential geometry, vectors can represent perturbations of parameters, while covectors represent linear maps taking vectors to the real numbers. In gradient based optimization, gradients are covectors, and interpreting them as a perturbation to update parameters is not well formulated. For example, when optimizing over position with parameters with units of “meter”, gradients have units of “per meter”, and should not be added to a quantity with units of “meter”. Using a Riemannian metric (inner product between vectors in a given tangent space), vectors can be converted to covectors and vice versa. Applying this conversion to gradients in optimization before updating is known as natural gradient descent.

In this work we consider affine and rigid transformations, which contain a linear map and a translation. These two components have different units: the linear part is unitless, and the translation part has units of length. This makes choosing stepsizes in gradient based optimization critical. Often orders of magnitude difference in scales means registration algorithms will not converge without laborious and non-reproducible parameter tuning. We introduce a metric for natural gradient descent based on the L^2 norm of optical flow, and we show that it is left invariant, invariant to image padding, and related to Gauss-Newton optimization. We demonstrate the advantages of this approach over more standard methods

on a brain image registration dataset, and discuss our results in the context of related methods.

2 METHODS

In this section we define our coordinate systems and optical flow metric for the 12 parameter affine and 6 parameter rigid case, enumerate several properties, and summarize our validation experiments. We use lowercase letters to denote scalars, boldface lowercase letters to denote vectors, and boldface uppercase letters to denote matrices. Exceptions are I, J which by convention will denote images, and g which by convention will denote a metric tensor.

2.1 The optical flow metric for affine transformations

We consider registration from an atlas image $I: X \subset \mathbb{R}^3 \to \mathbb{R}^m$, to a target image $J: X \to \mathbb{R}^n$. Often in medical imaging $m = n = 1$ (for grayscale images), but it is often 3 (for red, green, blue images), or can be other values.

We work with vectors and affine transforms in homogeneous coordinates, so a point in $\mathbb{R}^3$ is written as $\mathbf{x} = (x_0, x_1, x_2, 1)^T$, and an affine transform is written as

$$\mathbf{A} = \begin{pmatrix} a_{00} & a_{01} & a_{02} & b_0 \\ a_{10} & a_{11} & a_{12} & b_1 \\ a_{20} & a_{21} & a_{22} & b_2 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{1}$$

Images are transformed via the group action $\mathbf{A} \cdot I(\mathbf{x}) = I(\mathbf{A}^{-1}\mathbf{x})$, which is implemented computationally through trilinear interpolation.

We use a coordinate chart based on lexicographic ordering of the above components, φ: affine matrix $\to \mathrm{R}^{12}$ with $\mathbf{A} \mapsto (a_{00}, a_{01}, a_{01}, b_0, a_{10}, \ldots, b_2)^T$. We use the chart induced basis for the tangent space, whereby the standard basis vectors $\mathbf{e}_i$ are pushed forward, $\mathbf{E}_i = D\varphi_A^{-1}\mathbf{E}_i$. In this parameterization the push forward is independent of A. For example, the first two basis vectors are given by

$$\begin{aligned} D\varphi_A^{-1}\mathbf{e}_0 &= D\varphi_A^{-1}\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{E}_0 \\ D\varphi_A^{-1}\mathbf{e}_1 &= D\varphi_A^{-1}\begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{E}_1 . \end{aligned} \tag{2}$$

Tangent vectors can be thought of as small perturbations to the coordinates $\delta\mathbf{a}$, and we write $\delta\mathbf{A} = D\boldsymbol{\varphi}_A^{-1}\delta\mathbf{A}$. A perturbation $\mathbf{A} \mapsto \mathbf{A} + \epsilon\delta\mathbf{A}$ induces optical flow on our template image via

$$\frac{d}{d\epsilon} I\left((\mathbf{A} + \epsilon\delta\mathbf{A})^{-1}\mathbf{x}\right)\Big|_{\epsilon=0} = -DI\left(\mathbf{A}^{-1}\mathbf{x}\right)\mathbf{A}^{-1}\delta\mathbf{A}\mathbf{A}^{-1}\mathbf{x}.$$

We define our inner product g_A between two tangent vectors $\delta\mathbf{A}$, $\delta\mathbf{B}$ at the point $\mathbf{A}$ to be given by the L^2 inner product of the optical flow, scaled by the determinant of $\mathbf{A}$, denoted $|\mathbf{A}|$:

$$g_A(\delta\mathbf{A}, \delta\mathbf{B}) = \frac{1}{|\mathbf{A}|}\int_{AX} \mathbf{x}^T\mathbf{A}^{-T}\delta\mathbf{A}^T\mathbf{A}^{-T}DI^T\left(\mathbf{A}^{-1}\mathbf{x}\right)DI\left(\mathbf{A}^{-1}\mathbf{x}\right)\mathbf{A}^{-1}\delta\mathbf{B}\mathbf{A}^{-1}\mathbf{x}d\mathbf{x}. \tag{3}$$

This metric is left invariant, as can be seen by making a change of variables in the integral, $\mathbf{y} = \mathbf{A}^{-1}\mathbf{x}$, $\mathbf{x} = \mathbf{A}y$, $d\mathbf{x}\,|\mathbf{A}|\,d\mathbf{y}$:

$$g_A(\delta\mathbf{A}, \delta\mathbf{B}) = \int_X \mathbf{y}^T\left(\mathbf{A}^{-1}\delta\mathbf{A}\right)^T DI^T(\mathbf{y})DI(\mathbf{y})\left(\mathbf{A}^{-1}\delta\mathbf{B}\right)\mathbf{y}d\mathbf{y} = g_{\mathrm{id}}\left(\mathbf{A}^{-1}\delta\mathbf{A}, \mathbf{A}^{-1}\delta\mathbf{B}\right). \tag{4}$$

Since $\mathbf{A}^{-1}\delta\mathbf{A}$ is the push forward of the vector $\delta\mathbf{A}$ tangent to the point $\mathbf{A}$ by the map $\mathbf{A}^{-1}$, this implies that this choice of metric is left invariant. An important benefit of this property is that during an optimization procedure we can compute it at identity once, and apply a simple linear transformation to compute it at other locations $\mathbf{A}$.

We compute g_{id} in coordinates as a 12 × 12 matrix via $[g_{\mathrm{id}}]_{ij} = g_{\mathrm{id}}(\mathbf{E}_i, \mathbf{E}_j)$.

2.2 The optical flow metric for rigid transformations

We use *zyx* Euler angles for parameterizing the rotation group, i.e.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & b_x \\ 0 & 1 & 0 & b_y \\ 0 & 0 & 1 & b_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) & 0 \\ 0 & \sin(\theta_x) & \cos(\theta_x) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdots$$
$$\cdots \begin{pmatrix} \cos(\theta_y) & 0 & \sin(\theta_y) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\theta_y) & 0 & \cos(\theta_y) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos(\theta_z) & -\sin(\theta_z) & 0 & 0 \\ \sin(\theta_z) & \cos(\theta_z) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \doteq \mathbf{T}_{x,b_x}\mathbf{T}_{y,b_y}\mathbf{T}_{z,b_z}\mathbf{R}_{x,\theta_x}\mathbf{R}_{y,\theta_y}\mathbf{R}_{z,\theta_z}$$

where $\cdot$ denotes matrix multiplication. We let $\boldsymbol{\varphi}$ be the associated coordinate chart with $\boldsymbol{\varphi}(\mathbf{A}) = (\theta_x, \theta_y, \theta_z, b_0, b_1, b_2)^T$.

At identity (all parameters equal 0), the chart induced basis for the tangent space is:

$$\mathbf{E}_0^{\mathrm{id}} = D\boldsymbol{\varphi}_{\mathrm{id}}^{-1}\mathbf{e}_0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_1^{\mathrm{id}} = D\boldsymbol{\varphi}_{\mathrm{id}}^{-1}\mathbf{e}_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{E}_2^{\mathrm{id}} = D\boldsymbol{\varphi}_{\mathrm{id}}^{-1}\mathbf{e}_2 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{E}_3^{\mathrm{id}} = D\boldsymbol{\varphi}_{\mathrm{id}}^{-1}\mathbf{e}_3 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{E}_4^{\mathrm{id}} = D\boldsymbol{\varphi}_{\mathrm{id}}^{-1}\mathbf{e}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{E}_5^{\mathrm{id}} = D\boldsymbol{\varphi}_{\mathrm{id}}^{-1}\mathbf{e}_5 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

A standard result shows that the chart induced basis at a point A is

$$\begin{aligned}
\mathbf{E}_0^A &= \mathbf{T}_{x,b_x}\mathbf{T}_{y,b_y}\mathbf{T}_{z,b_z}\mathbf{R}_{x,\theta_x}\mathbf{E}_0^{\mathrm{id}}\mathbf{R}_{y,\theta}\mathbf{R}_{z,\theta_z} \\
\mathbf{E}_1^A &= \mathbf{T}_{x,b_x}\mathbf{T}_{y,b_y}\mathbf{T}_{z,b_z}\mathbf{R}_{x,\theta_x}\mathbf{R}_{y,\theta}\mathbf{E}_1^{\mathrm{id}}\mathbf{R}_{z,\theta_z} \\
\mathbf{E}_2^A &= \mathbf{T}_{x,b_x}\mathbf{T}_{y,b_y}\mathbf{T}_{z,b_z}\mathbf{R}_{x,\theta_x}\mathbf{R}_{y,\theta}\mathbf{R}_{z,\theta_z}\mathbf{E}_2^{\mathrm{id}} \\
\mathbf{E}_3^A &= \mathbf{E}_3^{\mathrm{id}}\mathbf{E}_{x,b_x}\mathbf{E}_{y,b_y}\mathbf{E}_{z,b_z}\mathbf{R}_{x,\theta_x}\mathbf{R}_{y,\theta}\mathbf{R}_{z,\theta_z} \\
\mathbf{E}_4^A &= \mathbf{T}_{x,b_x}\mathbf{E}_4^{\mathrm{id}}\mathbf{T}_{y,b_y}\mathbf{T}_{z,b_z}\mathbf{R}_{x,\theta_x}\mathbf{R}_{y,\theta}\mathbf{R}_{z,\theta_z} \\
\mathbf{E}_5^A &= \mathbf{T}_{x,b_x}\mathbf{T}_{y,b_y}\mathbf{E}_5^{\mathrm{id}}\mathbf{T}_{z,b_z}\mathbf{R}_{x,\theta_x}\mathbf{R}_{y,\theta}\mathbf{R}_{z,\theta_z}.
\end{aligned}$$

We can compute g_{id} in coordinates as a 6×6 matrix via $[g_{\mathrm{id}}]_{ij} = g_{\mathrm{id}}(\mathbf{E}_i, \mathbf{E}_j)$ using (3).

2.3 Converting covectors to vectors

For natural gradient descent we convert derivatives of the loss function with respect to the coordinate (which are covectors), to perturbations in the coordinates (which are vectors), by applying the # map, which we define here. If $d\mathbf{A}$ is a covector at the point $\mathbf{A}$ with coordinates $d\mathbf{a}$, and $\delta\mathbf{B}$ is a vector at the point $\mathbf{A}$ with coordinates $\delta\mathbf{a}$, then we can write the action of $d\mathbf{A}$ on $\delta\mathbf{B}$ in coordinates as

$$d\mathbf{A}(\delta\mathbf{B}) = \sum_i d\mathbf{A}_i \delta\mathbf{B}_i.$$

The lift from a covector to vector using the # map is defined implicitly through the relationship

$$d\mathbf{A}(\delta\mathbf{B}) = g_A\left(d\mathbf{A}^\sharp, \delta\mathbf{B}\right).$$

In a given coordinate system this amounts to solving a linear system of equations (i.e. inverting the matrix g_A): $d\mathbf{A}_i^\sharp = \sum_j [g_A]_{ij}^{-1} d\mathbf{A}_j$.

2.4 Computing the metric at *A*

We compute g_{id} by integrating over the image once at the start of optimization, and then use left invariance to define a simple transformation for computing g_A. To do this we need to express the push forward of a vector with $\mathbf{A}^{-1}$ as a matrix multiplication operation. This is computed via

$$\mathbf{M}_A = \underbrace{D\boldsymbol{\varphi}_{\mathrm{id}}}_{N\times 16}\underbrace{\mathbf{A}^{-1}}_{16\times 16}\underbrace{D\boldsymbol{\varphi}_A^{-1}}_{16\times N}$$

where N is 12 for general affine transforms and 6 for rigid transforms. The components of these matrices can be easily computed from their action on basis vectors. We can then write, using multiplication of $N\times N$, matrices:

$$g_A = \mathbf{M}_A^T g_{\mathrm{id}} \mathbf{M}_A$$

2.5 Useful properties of the metric

2.5.1 Invariance to padding—In a typical situation I is an image of a head in air, and air has a signal 0. In this case we can add any amount of zero padding to our image without changing the metric because DI is zero in the padded regions. This may seem trivial, but it is not respected by choices made in other software, such as the "automatic scales estimation" routine in Sec. 2.6 below.

2.5.2 Relationship to Gauss Newton—If the loss function being considered is sum of square error, then this approach is roughly equivalent to Gauss Newton optimization. The minor differences are constant scale factors, numerics of interpolation, and potentially regions over which error is summed. This suggests that choosing a step of order 1 is reasonable for this cost function.

2.5.3 Independence to coordinate origin—Traditionally the positioning of a coordinate origin has been critical to the success of affine registration. If the origin is far from the imaging data (e.g. the corner of the image), then the transformation is very sensitive to changes in the linear parameters.

Because shifting the coordinate origin is equivalent to applying an affine transformation, and our metric is left invariant, our approach is invariant to the coordinate origin. This is illustrated in our experimental results.

2.6 Mapping experiments

We implement a gradient based affine image registration algorithm by minimizing sum of square error, and mutual information loss functions. We use 100 randomly sampled pairs of images from the LPBA40 dataset [5] in native space. This dataset consists brain images

from 40 healthy volunteers, with 20 male and 20 female, between 19.3 and 39.5 years old. Images were 3D Spoiled Gradient Echo MRI volumes from a GE 1.5T system, with 1.5mm mm coronal slice spacing, and 0.86 mm in plane resolution (38 subjects) or 0.78 mm (2 subjects). This dataset has been used for multiple image registration validation studies, most notably [21] which compared 14 different nonrigid registration methods. An example image pair is shown in Fig. 1 (a) and (b), which corresponds to our first randomly selected pair.

We compare the natural gradient descent approach described here to three different gradient descent approaches. In the "vanilla" method we update the linear and translation jointly with no scaling, and in the "alternating" method we update them every other iteration (i.e. we update the linear part on iteration 0, the translation part on iteration one, the linear part on iteration 2, etc.).

We also compare to one other approach, "automatic scales estimation" used in simple elastix [22]. In this approach, the derivative of the transformation's output with respect to each parameter is taken, yielding a 3×12 matrix of partial derivatives at each pixel. Next the sum of squares is computed down each row, yielding a 1×12 vector at each pixel. Last the quantity is averaged over all pixels. This gives a set of scale factors which are used to normalize the components of the gradient before updating parameters. These scale factors are generally only computed once. We note that this is related to our approach with three modifications: (i) we set $I(x) = x$ in (3), (ii) we consider the diagonal elements of g only, and (iii) we define the components of g_A to be equal to those of g_{id}.

In all cases, gradient update steps are performed using a golden section linesearch with a maximum of 10 steps. The minimum in a given search direction is bracketed by a step size of 0, and a step size that would increase the loss. The latter is found by increasing the stepsize found from the previous iteration by a the golden ratio (roughly 1.6) until the loss function increases. Occasionally the step size is found from the golden section search to be 0 within machine precision, in which case we set it to 1×10^{-10} to initialize the procedure at the next gradient descent iteration. Note that in this case optimization will stop, except in the alternating method where it may continue.

For each gradient descent method we consider 12 parameter affine registration (for sum of square error and mutual information loss), and 6 parameter rigid registration (sum of square error only). We also consider placements of the coordinate origin in the image center, half way to the corner, and at the corner. For one case we consider a finer grained sampling of the image origin as discussed in the results section.

In each case we compute the loss function at each step of optimization for 50 steps, and report normalized results that have been shifted and scaled to lie in the range 0 to 1. The minimum for the best method is assigned the value 0, and the initial value of the loss function (equal for all methods) is assigned the value 1. We show these curves directly, and also show the average tied rank at each iteration. The tied rank for method X is defined as 1 plus the number of other methods X does better than, plus half of the number of other methods it performs equal to. This gives a number between 1 (worst) and 4 (best).

Algorithms are implemented in pytorch and gradients with respect to parameters are computed automatically. We use double precision arithmetic to deal with a large dynamic range of step sizes. For metric computation derivatives DI are computed on a voxel grid via centered differences, and integrals are computed by summing over voxels and multiplying by the voxel volume. Algorithms are run on a NVIDIA GeForce RTX 2080 Ti Rev. A GPU device with 11GB of memory.

3 RESULTS

3.1 Methods summary

In the methods section we derived a metric based on the L^2 norm of optical flow of a template image, for the 12 parameter affine and 6 parameter rigid case, and proposed a corresponding natural gradient optimization algorithm. These correspond to 12×12 or 6×6 (respectively) positive definite matrices that depend on an atlas image.

We test our approach using 100 randomly selected pairs of neuroimages from the the LPBA 40 dataset [5]. We consider 4 methods (natural gradient descent, “vanilla” gradient descent, alternating minimization, and automatic scales estimation), 3 placements of coordinate origins (center, half way, and corner), 2 transformation groups (affine and rigid), and 2 loss functions (sum of square error, and mutual information used in the 12 parameter case only).

Our first randomly selected pair of images is shown in Fig. 1 (a) and (b). We also show the initial error (difference between images) in (c), and the error after alignment for the affine (d) and rigid (e) case.

3.2 Metric tensors

For the example images in Fig. 1, the metric tensor at identity for the 12 parameter affine registration is shown in Fig. 2. The three plots show results when the origin is at the center, half way to the corner, and at the corner. Even with the origin in the center, the tensor has significant off diagonal elements, including negative values, suggesting that simply scaling gradient components is suboptimal. With the origin at the corner, there are stronger off diagonal elements, and a much larger difference between the linear and translation parts.

For the example images in Fig. 1, the metric tensor at identity for the 6 parameter rigid registration is shown in Fig. 3. Similar trends are observed, with even stronger negative elements.

3.3 12 parameter affine results with sum of square error

Normalized sum of square error is reported as a function of optimization step in Fig. 4 for the 12 parameter affine registrations. The median for each method is shown as a solid line, the 25th and 75th percentile are shown as a filled area, and all 100 curves are shown in pale colors. One observes that the natural gradient descent approach converges significantly faster than the other methods. Considering different coordinate has a large effect for the alternative methods and no effect for the natural gradient method. The alternative methods

often fail to converge within 50 iterations, with the “vanilla” method failing even with a centered coordinate origin. This illustrates the important lack of robustness in this field.

The bottom right plot shows that the natural gradient method quickly achieves an average rank close to 4 (best), while the vanilla method quickly achieves a rank of 1 (worst). The alternating minimization approach performs better in the long term, while the automatic scales estimation approach performs better in the short term. Relative to the other methods our approach performs worse with a coordinate origin in the center and slightly better for a coordinate origin in the corner, though the effect is minor.

3.4 12 parameter affine results with mutual information

Mutual information is reported as a function of optimization step in Fig. 5 for the 12 parameter affine registrations. The median for each method is shown as a solid line, the 25th and 75th percentile are shown as a filled area, and all 100 curves are shown in pale colors. Trends are very similar to that seen for sum of square error, with the natural gradient descent method converging to the best even more quickly than for sum of square error.

3.5 6 parameter rigid results with sum of square error

Normalized sum of square error is reported as a function of optimization step in Fig. 6 for the 6 parameter rigid registrations. The median for each method is shown as a solid line, the 25th and 75th percentile are shown as a filled area, and all 100 curves are shown in pale colors. One observes that the natural gradient descent approach converges significantly faster than the other methods for every origin placement except the center. Again, considering different coordinate origins has a large effect for the alternative methods and no effect for the natural gradient method.

The bottom right plot shows that the natural gradient method quickly achieves an average rank close to 4 (best) for every case except a centered origin, where it tends to perform roughly equivalent to the alternating minimization scheme. The vanilla method quickly achieves a rank of 1 (worst). The alternating minimization approach performs better in the long term, while the automatic scales estimation approach performs better in the short term, although the order changes much sooner than in the 12 parameter case. Relative to the other methods, our approach performs worse with a coordinate origin in the center, and slightly better for a coordinate origin in the corner.

Because of the similarity between the alternating and natural gradient methods when the origin is centered, we consider comparing these methods as a function of origin position. We consider fraction between center and corner of 0,0.1,0.2,0.3,0.4,0.5 and 1. For each iteration, and for each distance from the center, we use a sign test statistic by averaging across patients. Because this amounts to 350 statistical tests, we control familywise error rate using permutation testing [27]. We randomly flip the sign of each patient 100,000 times, and calculate corrected p values using maximum statistics. These results are shown in Fig. 7. We see that the natural gradient method performs statistically significantly better than the alternating optimization method across all iterations when the origin is shifted by 0.2 or more, and for the first few iterations when the origin is shifted by 0.1 or less. This indicates that even small uncertainties in positioning of the origin, 3.6 cm in this case, can degrade

the performance of traditional approaches. Shifts of this size are reasonable considering that the location of the brain's center is typically unknown until after this type of registration is performed.

4 DISCUSSION

Our results show that the natural gradient method proposed here outperforms other gradient based minimization schemes in almost every case examined, with virtually no computational overhead. This result holds for both 12 parameter affine transformations and 6 parameter rigid transformations, for sum of square error and for mutual information objective functions. The method does not depend on the placement of the coordinate origin, and can be easily implemented within any registration framework to improve convergence speed and reduce the number of hyperparameters that need to be tuned or estimated.

Another interesting finding is that the automatic scales estimation approach, while sophisticated, tends to perform worse than the simpler alternating minimization approach. This is particularly true when the origin is centered, which is the setting recommended by simple elastix [22].

There are few studies of natural gradient descent applied to image registration in the literature. A search on google scholar for *"natural gradient" image registration* returns only two relevant results. The work [28] uses a Fisher information metric for deformable registration and adopts various sampling strategies to define a random process from which it can be computed. The work [29] uses a similar strategy for deformable registration, but considers pairwise registration in a population and therefore computes a Fisher Information metric by taking an expectation across the population. Both rely on some form of stochasticity, and neither of these consider or exploit invariance with respect to a transformation group. In [15], which considers nonlinear registration (Large Deformation Diffeomorphic Image Mapping, LDDMM), a distinction between covectors and vectors in the space of smooth vector fields is made, and gradient descent is performed in the latter by applying a smoothing operation which is equivalent to the inverse of a metric. Other deformable registration approaches with regularization may do this implicitly. This right invariant metric used in LDDMM does not depend on the images being registered, and is based on spatial smoothness which is not relevant in the affine registration case. In related work [30] and [31] developed strategies to learn a metric for deformable image registration, but considered these choices from an accuracy point of view rather than an optimization point of view. In this deformable registration setting, Gauss Newton optimization was used in [32], which is similar to our approach as described in Sec. 2.5.2.

The trend in computer vision and deep learning is to use stochastic gradient descent methods for optimization, and ITK and its derivatives includes several of these approaches for image registration. For example, the work of [33, 34, 35] is used for adaptive stepsize estimation. Other adaptive step size approaches have been developed and applied to deformable image registration as well [36]. These approaches do not consider different scale factors for different parameters. The well known ADAM optimization procedure [37] does update step

sizes on a per parameter basis, but does not allow for mixing of parameters. The importance of this mixing is evident from the significant off diagonal elements seen in Fig. 2 and 3.

A modern trend in image registration is to use deep learning based approaches that replace optimization procedures considered here with prediction using a single forward pass of a deep network. These tools include Voxelmorph [38], Quicksilver [39], and others. While our contribution here does not apply directly to these methods, it could potentially be used to accelerate the training stage. An investigation into this potential will be the subject of future work.

Finally, the simple registration tasks considered here are often only one step in a long pipeline or joint optimization procedure. For example in our work studying serial section images, 3D affine registration is performed jointly with nonlinear alignment and 2D rigid registration on each slice [40, 41]. The approach developed here leads to a reduction by half in the number of parameters that need to be selected manually, allowing our pipeline to generalize to new datasets more effectively. This work demonstrates that straightforward applications of principles from differential geometry can accelerate research in neuroimaging.

FUNDING

This work was supported by the National Institutes of Health (U19MH114821), and the Karen Toffler Charitable Trust through the Toffler Scholar Program.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the Laboratory of Neuroimaging LPBA40 atlas repository https://www.loni.usc.edu/research/atlases.

REFERENCES

[1]. Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 27 (2008) 685–691.

[2]. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience 19 (2007) 1498–1507. [PubMed: 17714011]

[3]. Benninger K, Hood G, Simmel D, Tuite L, Wetzel A, Ropelewski A, et al. Cyberinfrastructure of a multi-petabyte microscopy resource for neuroscience research. Practice and Experience in Advanced Research Computing (2020), 1–7.

[4]. Fonov VS, Evans AC, McKinstry RC, Almli C, Collins D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage (2009) S102.

[5]. Shattuck DW, Mirza M, Adisetiyo V, Hojatkashani C, Salamon G, Narr KL, et al. Construction of a 3d probabilistic atlas of human cortical structures. Neuroimage 39 (2008) 1064–1080. [PubMed: 18037310]

[6]. Wang Q, Ding SL, Li Y, Royall J, Feng D, Lesnar P, et al. The allen mouse brain common coordinate framework: a 3d reference atlas. Cell 181 (2020) 936–953. [PubMed: 32386544]

[7]. Bookstein FL. Morphometric tools for landmark data: geometry and biology (Cambridge University Press) (1997).

[8]. Joshi SC, Miller MI. Landmark matching via large deformation diffeomorphisms. IEEE transactions on image processing 9 (2000) 1357–1370. [PubMed: 18262973]

[9]. Miller MI, Tward DJ, Trouvé A. Coarse-to-fine hamiltonian dynamics of hierarchical flows in computational anatomy. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020), 860–861.

[10]. Glaunes J, Qiu A, Miller MI, Younes L. Large deformation diffeomorphic metric curve mapping. International journal of computer vision 80 (2008) 317–336. [PubMed: 20419045]

[11]. Vaillant M, Glaunes J. Surface matching via currents. Biennial International Conference on Information Processing in Medical Imaging (Springer) (2005), 381–392.

[12]. Charon N, Trouvé A. The varifold representation of nonoriented shapes for diffeomorphic registration. SIAM Journal on Imaging Sciences 6 (2013) 2547–2580.

[13]. Woods RP, Grafton ST, Holmes CJ, Cherry SR, Mazziotta JC. Automated image registration: I. general methods and intrasubject, intramodality validation. Journal of computer assisted tomography 22 (1998) 139–152. [PubMed: 9448779]

[14]. Woods RP, Grafton ST, Watson JD, Sicotte NL, Mazziotta JC. Automated image registration: Ii. intersubject validation of linear and nonlinear models. Journal of computer assisted tomography 22 (1998) 153–165. [PubMed: 9448780]

[15]. Beg MF, Miller MI, Trouvé A, Younes L. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. International journal of computer vision 61 (2005) 139–157.

[16]. Reuter M, Rosas HD, Fischl B. Highly accurate inverse consistent registration: a robust approach. Neuroimage 53 (2010) 1181–1196. [PubMed: 20637289]

[17]. Tward D, Brown T, Kageyama Y, Patel J, Hou Z, Mori S, et al. Diffeomorphic registration with intensity transformation and missing data: Application to 3d digital pathology of alzheimer's disease. Frontiers in neuroscience 14 (2020) 52. [PubMed: 32116503]

[18]. Joshi AA, Shattuck DW, Thompson PM, Leahy RM. Surface-constrained volumetric brain registration using harmonic mappings. IEEE transactions on medical imaging 26 (2007) 1657–1669. [PubMed: 18092736]

[19]. Durrleman S, Prastawa M, Korenberg JR, Joshi S, Trouvé A, Gerig G. Topology preserving atlas construction from shape data without correspondence using sparse parameters. International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer) (2012), 223–230.

[20]. Tward D, Miller M, Trouve A, Younes L. Parametric surface diffeomorphometry for low dimensional embeddings of dense segmentations and imagery. IEEE transactions on pattern analysis and machine intelligence 39 (2016) 1195–1208. [PubMed: 27295651]

[21]. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. Neuroimage 46 (2009) 786–802. [PubMed: 19195496]

[22]. Marstal K, Berendsen F, Staring M, Klein S. Simpleelastix: A user-friendly, multi-lingual library for medical image registration. Proceedings of the IEEE conference on computer vision and pattern recognition workshops (2016), 134–142.

[23]. Yaniv Z, Lowekamp BC, Johnson HJ, Beare R. Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research. Journal of digital imaging 31 (2018) 290–303. [PubMed: 29181613]

[24]. McCormick MM, Liu X, Ibanez L, Jomier J, Marion C. Itk: enabling reproducible research and open science. Frontiers in neuroinformatics 8 (2014) 13. [PubMed: 24600387]

[25]. Yoo TS, Ackerman MJ, Lorensen WE, Schroeder W, Chalana V, Aylward S, et al. Engineering and algorithm design for an image processing api: a technical report on itk-the insight toolkit. Studies in health technology and informatics (2002) 586–592. [PubMed: 15458157]

[26]. Amari SI, Douglas SC. Why natural gradient? Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181) (IEEE) (1998), vol. 2, 1213–1216.

[27]. Nichols T, Hayasaka S. Controlling the familywise error rate in functional neuroimaging: a comparative review. Statistical methods in medical research 12 (2003) 419–446. [PubMed: 14599004]

[28]. Zikic D, Kamen A, Navab N. Natural gradients for deformable registration. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (IEEE) (2010), 2847–2854.

[29]. Wang H, Rusu M, Golden T, Gow A, Madabhushi A. Mouse lung volume reconstruction from efficient groupwise registration of individual histological slices with natural gradient. Medical Imaging 2013: Image Processing (International Society for Optics and Photonics) (2013), vol. 8669, 866914.

[30]. Vialard FX, Risser L. Spatially-varying metric learning for diffeomorphic image registration: A variational framework. International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer) (2014), 227–234.

[31]. Niethammer M, Kwitt R, Vialard FX. Metric learning for image registration. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019), 8463–8472.

[32]. Ashburner J, Friston KJ. Diffeomorphic registration using geodesic shooting and gauss–newton optimisation. NeuroImage 55 (2011) 954–967. [PubMed: 21216294]

[33]. Klein S, Pluim JP, Staring M, Viergever MA. Adaptive stochastic gradient descent optimisation for image registration. International journal of computer vision 81 (2009) 227.

[34]. Qiao Y, Lelieveldt BP, Staring M. An efficient preconditioner for stochastic gradient descent optimization of image registration. IEEE transactions on medical imaging 38 (2019) 2314–2325. [PubMed: 30762536]

[35]. Cruz P Almost sure convergence and asymptotical normality of a generalization of kesten's stochastic approximation algorithm for multidimensional case. arXiv preprint arXiv:1105.5231 (2011).

[36]. Wu J, Tang X. A large deformation diffeomorphic framework for fast brain image registration via parallel computing and optimization. Neuroinformatics (2019) 1–16. [PubMed: 30617757]

[37]. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[38]. Balakrishnan G, Zhao A, Sabuncu MR, Guttag J, Dalca AV. Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging 38 (2019) 1788–1800.

[39]. Yang X, Kwitt R, Styner M, Niethammer M. Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage 158 (2017) 378–396. [PubMed: 28705497]

[40]. Tward DJ, Li X, Huo B, Lee BC, Miller M, Mitra PP. Solving the where problem in neuroanatomy: a generative framework with learned mappings to register multimodal, incomplete data into a reference brain. bioRxiv (2020).

[41]. Tward D, Li X, Huo B, Lee B, Mitra P, Miller M. 3d mapping of serial histology sections with anomalies using a novel robust deformable registration algorithm. Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy (Springer) (2019), 162–173.

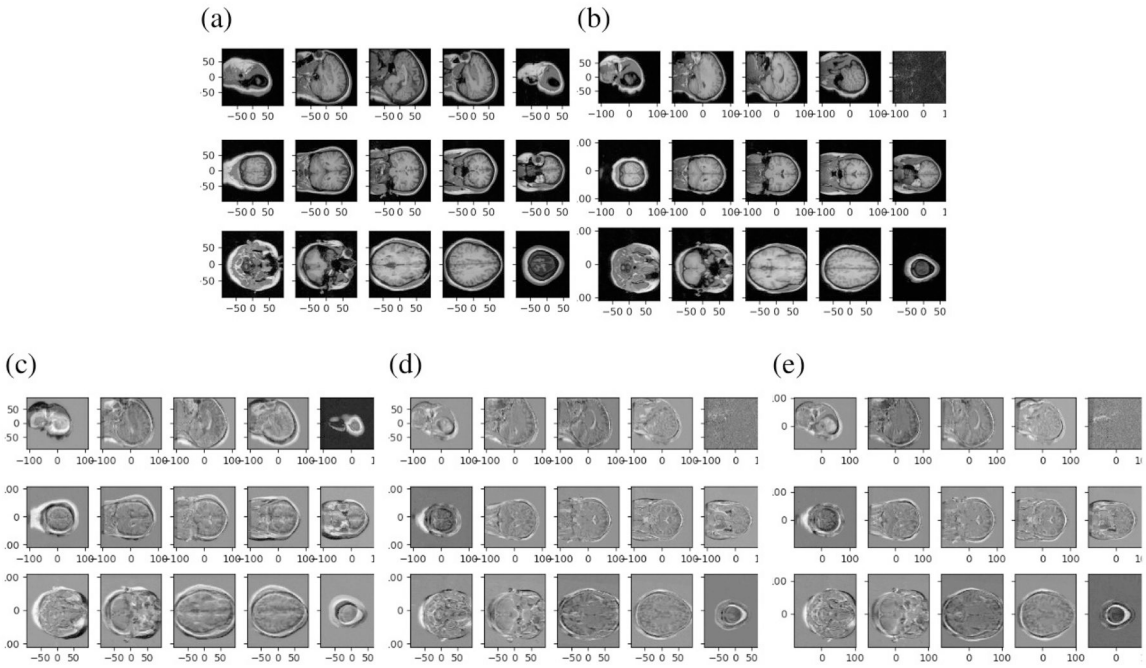

Figure 1.

(a),(b): One example pair of neuroimages used for registration, shown using 5 slices in each of the sagittal (top row) coronal (middle row) and axial (bottom row) planes. Difference between atlas and target images is shown before registration in (c), after affine registration in (d), and after rigid registration in (d).

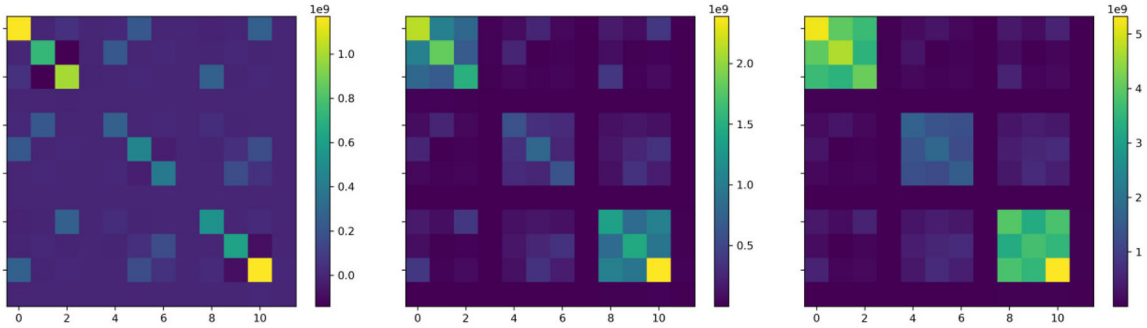

Figure 2.

Metric tensors shown here for 12 parameter affine transformations. Left, origin at the center, right origin at the corner.

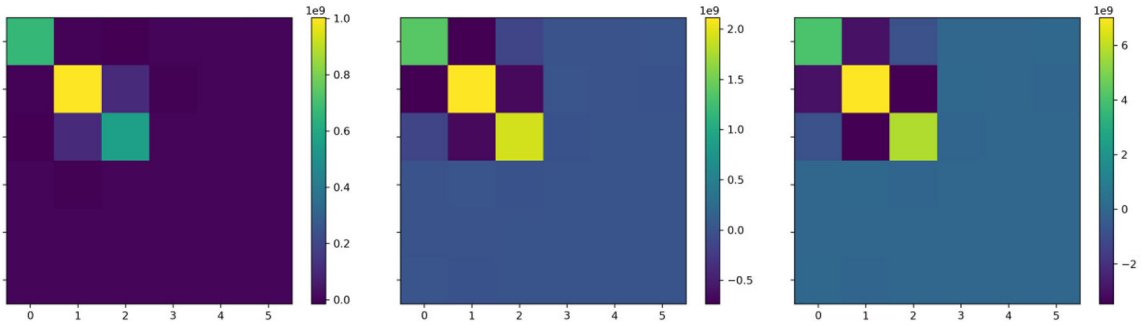

Figure 3.

Metric tensors shown here for 6 parameter rigid transformations. Left, origin at the center, right origin at the corner.

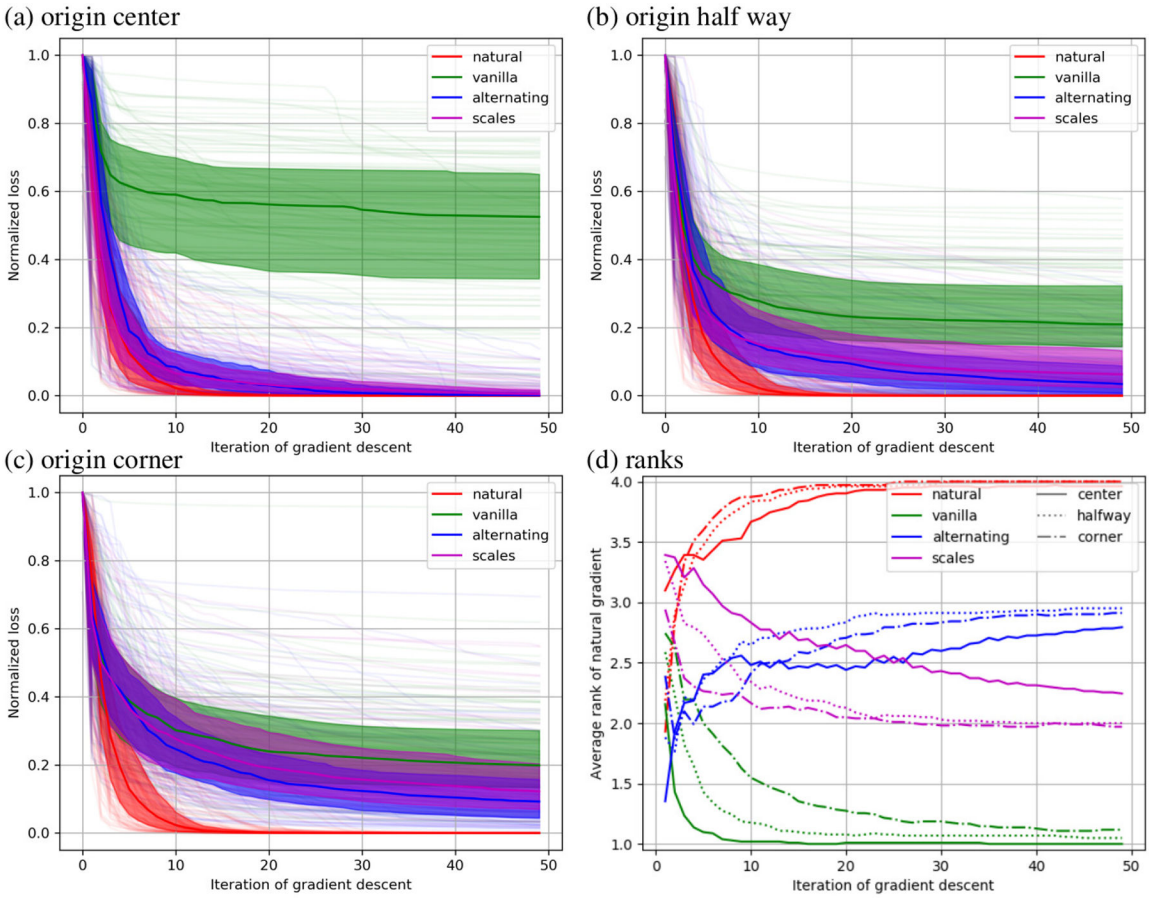

Figure 4.

Results of minimizing sum of square error over 12 parameter affine registrations are shown with the coordinate origin in the center (a), half way to the corner (b), and in the corner (c). In these plots the normalized value of the objective function is shown at each iteration of gradient based optimization. In (d) the rank of each method is shown under each of these conditions.

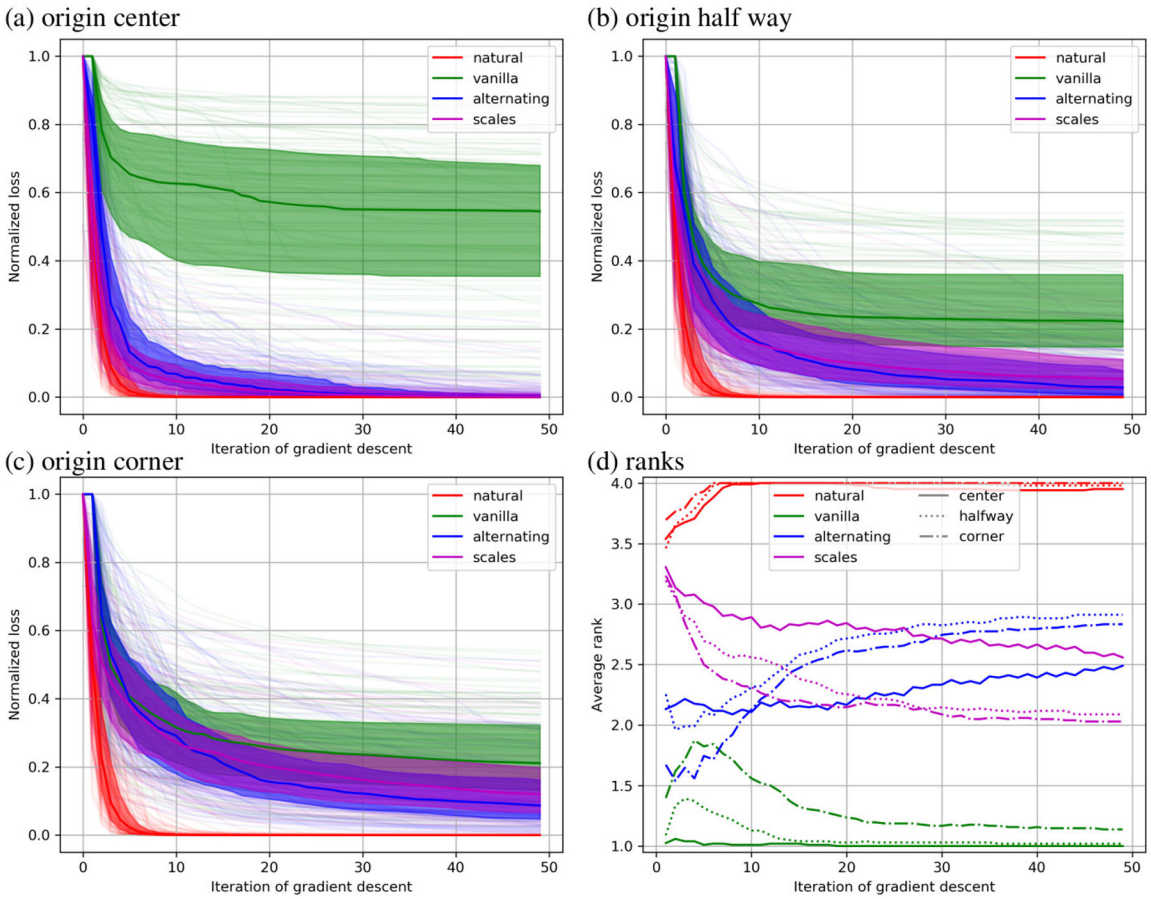

Figure 5.

Results of minimizing negative mutual information 12 parameter affine registrations are shown with the coordinate origin in the center (a), half way to the corner (b), and in the corner (c). In these plots the normalized value of the objective function is shown at each iteration of gradient based optimization. In (d) the rank of each method is shown under each of these conditions.

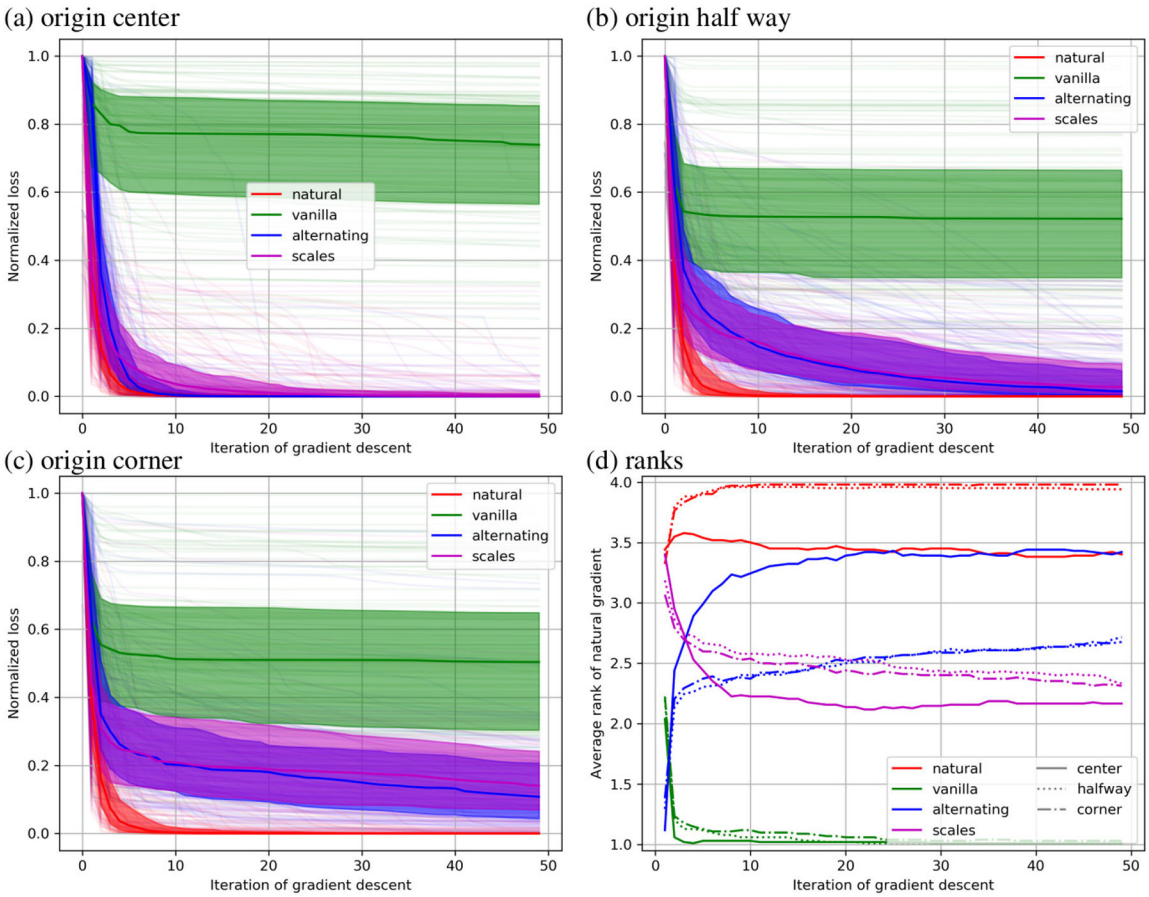

Figure 6.

Results of minimizing sum of square error over 6 parameter rigid registrations are shown with the coordinate origin in the center (a), half way to the corner (b), and in the corner (c). In these plots the normalized value of the objective function is shown at each iteration of gradient based optimization. In (d) the rank of each method is shown under each of these conditions.

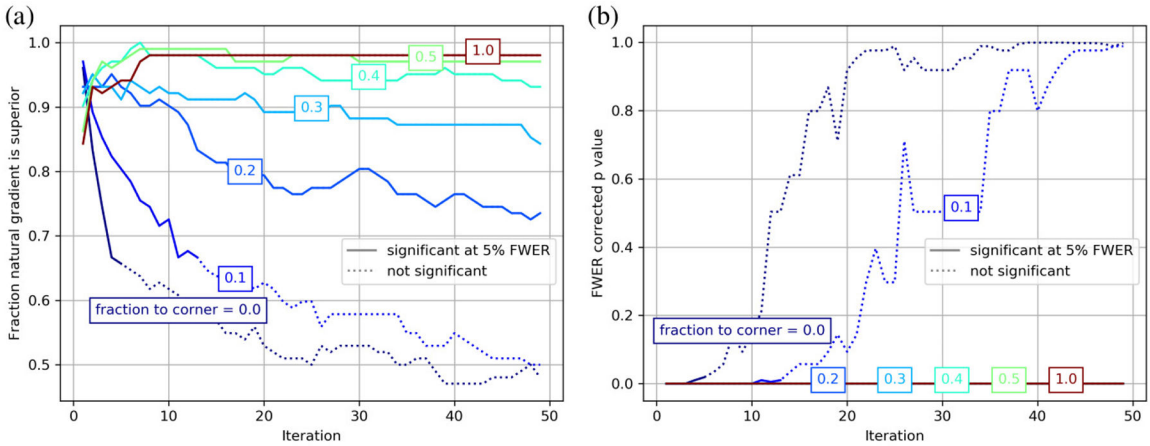

Figure 7.

Comparison between the alternating minimization and natural gradient method are shown at each iteration of optimization for coordinate origin placed at various fractions of the distance between the center and the corner. (a) shows the fraction of the time the natural gradient method performs the best, with statistically significant (familywise error rate corrected p values less than 0.05) samples shown as solid lines. (b) shows the familywise error rate p values under the same conditions.