Abstract—We present an algorithm for fast elastic multidimensional intensity-based image registration with a parametric model of the deformation. It is fully automatic in its default mode of operation. In the case of hard real-world problems, it is capable of accepting expert hints in the form of soft landmark constraints. Much fewer landmarks are needed and the results are far superior compared to pure landmark registration. Particular attention has been paid to the factors influencing the speed of this algorithm. The B-spline deformation model is shown to be computationally more efficient than other alternatives.

The algorithm has been successfully used for several 2D and 3D registration tasks in the medical domain, involving MRI, SPECT, CT, and ultrasound image modalities. We also present experiments in a controlled environment, permitting an exact evaluation of the registration accuracy. Test deformations are generated automatically using a random hierarchical fractional wavelet-based generator.

Index Terms—Image registration, elastic registration, splines, landmarks

I. INTRODUCTION

Image registration is the task of finding a correspondence function mapping coordinates from a reference image to coordinates of homologous points in a test image [1]. We call the registration elastic [2] if the family of correspondence functions is sufficiently general, capable of expressing essentially arbitrary nonlinear relations. Image registration is applied in the areas of motion analysis [4–6], video compression and coding [7], object tracking [8], or image stabilization. It leads to algorithms for segmentation [9], depth reconstruction from stereo images [10, 11], and for general 3D reconstruction. In the biomedical domain, there is a frequent need for comparing images for analysis and diagnostic purposes. This is accomplished by registering the images and aligning them by warping using the correspondence function identified.

Applications include intra-subject [12], inter-subject [13, 14], and inter-modality analysis [15–17], registration with annotated atlases [18, 19], quantification and qualification of feature shapes and sizes [20], distortion compensation [21, 22] and motion detection [23, 24] and compensation [25].

Various nonlinear registration algorithms for brain warping applications are presented by Warfield et al. [9]. Bayesian interpretation of elastic matching was reviewed by Gee [19], also in the context of human neuroanatomy. Articles by Van den Elsen et al. [26] and Maintz and Viergever [27] contain a very comprehensive and detailed classification of available methods for medical imaging applications. Lester and Arridge [28] treat the hierarchical aspects of the algorithms.

The deformation models of elastic registration algorithms fall into two basic categories. The first type are non-parametric, local methods — the deformation function is basically unconstrained and belongs to a very large and unrestrictive functional space. These methods can be formulated as variational, defining a scalar criterion that completely determines the final solution [2]. More generally, they can be also expressed using partial differential equations (PDE) [29–32].

The presented algorithm belongs to a second group of methods that use parametric models, representing the deformation by a moderate number of parameters, often in the multi-scale setting. Specific examples include hierarchical basis functions by Moulin et al. [7], quadtree-splines [5], multiresolution subspaces [33, 34], and wavelets [35, 36]. Splines are well suited for this kind of problems; they have appeared in various incarnations. In this paper we use a multiresolution B-spline representation, as was initially suggested in the pioneering work of Szeliski et al. [5, 10].

A. Proposed algorithm

The algorithm described in this article is a synthesis of several ideas. First, it is a generalization to multiple dimensions of the unidirectional registration algorithm we described in [22]. Its main features are the use of B-splines to describe both the image and the deformation, a double multiresolution strategy (for both the image and the deformation), a scalar pixel-based difference measure, and an iterative multidimensional optimization algorithm [37, 38]. The deformation model has been generalized and the whole algorithm re-engineered for faster execution.

Second, we present the idea of semi-automatic registration, targeted to more difficult registration problems. We ask an expert to identify a small number of corresponding points in both images. The points are also called landmarks [3, 12, 39, 40]. We add a term to the data part of the criterion, to steer the algorithm towards the correct solution indicated by the landmarks.

B. Organization of this article

In Section II, we describe the concept of registration by minimization, the difference measure, the B-spline image model, and the structure of the deformation model. In Section III
we justify our choice of B-splines as basis functions for the deformation model. We present the optimization method in Section IV, where we also describe the multiresolution strategy. Section V is devoted to the semi-automatic mode incorporating landmark information into the global criterion. We deal with implementation issues in Section VI and present experiments and applications in Section VII. For more details, we refer the reader to the first author’s thesis report [38] and its associated web page.

II. PROBLEM FORMULATION

The input images are given as two $N$-dimensional discrete signals $f_r(i)$ and $f_l(i)$, where $i \in I \subset \mathbb{Z}^N$, and $I$ is an $N$-dimensional discrete interval representing the set of all pixel coordinates in the image. We call $f_r$ and $f_l$ reference and test images, respectively. We suppose that the test image is a geometrically deformed version of the reference image, and vice versa.\(^2\) This is to say that the points with the same coordinate $x$ in the reference image $f_r(x)$ and in the warped test image $f_w(x) = f_r(g(x))$ should correspond. Here, $f_r$ is a continuous version of the test image and $g(x)$ is a deformation (correspondence) function to be identified.

A. Cost function

The two images $f_r, f_w$ will not be identical because of noise and also because the assumption that there is a geometrical mapping between the two images is not necessarily correct. Therefore, we define the solution to our registration problem as the result of the minimization $g = \arg \min_{g \in G} E(g)$, where $G$ is the space of all admissible deformation functions $g$. We have chosen the SSD (sum of squared differences) criterion

$$E = \frac{1}{|I|} \sum_{i \in I} e_i^2 = \frac{1}{|I|} \sum_{i \in I} (f_w(i) - f_r(i))^2 \quad (1)$$

because it is fast to evaluate and yields a smooth criterion surface which lends itself well to optimization. Minimization of (1) yields the optimal solution $g$ in the ML (maximum likelihood) sense under the assumption that $f_r$ is a deformed (warped) version of $f_l$ with i.i.d. (independent and identically distributed) Gaussian noise added to each pixel. The SSD criterion proved to be robust enough, especially if preprocessing was used to equalize the image values — mostly applied high-pass filtering and histogram normalization [22]. In principle, there is no difficulty in extending our method for more sophisticated pixel-based similarity measures, such as information-based measures [41], especially mutual information [17], or weighted $\ell_p$ norms. Only the evaluation of the criterion and its derivatives (gradient) needs to be changed.

B. Image interpolation

In accordance with [22], we choose to interpolate the image using uniform B-splines:\(^3\)

$$f_r^* (x) = \sum_{i \in I \subseteq \mathbb{Z}^N} b_i \beta_n (x - i) \quad (2)$$

where $\beta_n$ is a tensor product of B-splines of degree $n$, that is $\beta_n(x) = \prod_{k=1}^N \beta_n(x_k)$, with $x = (x_1, \ldots, x_N)$.

C. Deformation model structure

So far, we have considered the deformation function $g$ to be an arbitrary admissible function $\mathbb{R}^N \rightarrow \mathbb{R}^N$. We will restrict it now to a family of functions described by a finite number of parameters $c_j$:

$$g(x) = x + \sum_{j \in J} c_j \varphi_j(x) \quad (3)$$

where $J$ is a set of parameter indexes and $\varphi_j$ are the corresponding basis functions. This transforms a variational problem into a much easier finite-dimensional minimization problem, for which numerous algorithms exist [43]. Moreover, the restriction of the family $G$ of all possible functions $g$ can already guarantee some useful properties, such as the regularity (smoothness) of the solution. Note that the addition of $x$ in the above equation makes the set of zero parameters correspond to identity.

D. Existence, unicity, and regularization

Note that the criterion $E$ is non-negative and continuous and $f_r^*$ is periodic due to boundary conditions. Consequently, $E$ has a minimum; i.e., the proposed problem has a solution. However, depending on the images at hand, the solution does not have to be unique and there can be local minima. Fortunately, this does not pose problems in practice thanks to a multiresolution approach (Section IV-B) which smoothes out images at coarse levels and brings us sufficiently close to the solution at fine resolution levels. The algorithm will find a solution if started within the attraction basin of that solution. The virtual springs (Section V) play a role of an a priori information and a regularization term; extra regularization can be applied [44] if desired.

III. DEFORMATION BASIS

The purpose of this section is to motivate our choice of (cubic) B-splines [42] as the most adequate basis functions $\varphi_j$ to represent the deformation in model (3). The alternative possibilities that come to mind are polynomials [45], harmonic functions [18,46], radial basis functions [3,47], and wavelets [35,48,49].

\(^2\)In the multimodal case, which we are not considering here, there can be also an intensity mapping between the two images.

\(^3\)Uniform symmetric B-splines [42] of degree $n$ are piecewise polynomials of degree $n$. The polynomial pieces are delimited by uniformly placed knots. B-splines of degree $n$ have continuous derivatives up to order $n-1$ everywhere. Their integer shifts form a basis. The first (degree zero) symmetric B-spline is defined as $\beta_0(x) = 1$ for $x \in (-\frac{1}{2}, \frac{1}{2})$ and 0 otherwise. Higher order B-splines are defined recursively as $\beta_{n+1}(x) = \beta_n(x) \beta_n(0)$; and their support is $(-\frac{n+1}{2}, \frac{n+1}{2})$. 

It is highly desirable to have as few basis functions as possible to contribute to each particular point, while keeping the approximation quality. First, short basis functions have small overlap. This reduces the interdependency between the coefficients (parameters) and consequently makes the minimization problem easier to solve. Small overlap also makes the approximation quality. First, short basis functions have coefficients (parameters) and consequently makes the minimization problem easier to solve. Small overlap also makes the Hessian (the matrix of second partial derivatives, needed for some optimizers) more sparse and therefore potentially faster to invert.

Second, the size of the support of the basis functions directly influences the speed of the calculation. The evaluation of the deformation function (3) at \( N_{\text{pix}} \) points costs \( O(N_{\text{pix}}N_{\text{cnt}}) \) operations, where \( N_{\text{cnt}} \) is the number of functions \( \varphi_j \) contributing to a single point.\(^4\) The cost of evaluating the gradient \( \nabla E \) of the criterion \( E \) with respect to the coefficients is higher but asymptotically equivalent, because each of the \( N_{\text{pix}} \) pixels contributes to exactly \( N_{\text{cnt}} \) components of the gradient. Note that this cost is independent of the total number \( N_{\text{tot}} \) of the basis functions (unless \( N_{\text{tot}} = N_{\text{cnt}} \)). The cost of evaluating the Hessian is \( O(N_{\text{pix}}N_{\text{cnt}}^2) \) operations. (See also Section VI-A.)

Figure 1 shows the generating functions needed to calculate a value at one point (denoted by the vertical bar) for various bases; only functions that are non-zero at that point are considered. Except for the Fourier basis, we choose basis functions of the same degree (cubic), generating the same space. We see clearly that the least number of contributing functions (four) is in the B-spline case. This effect turns out to be even more dramatic in higher dimensions.

The reasoning above rules out the polynomials because no fast algorithm is known for their evaluation and the brute-force evaluation is slow due to their long support. As for the radial basis functions, although there are algorithms with reduced asymptotical complexity for evaluation of radial basis functions [50–53], their overhead is still non-negligible. We decided against the harmonic (Fourier) basis functions because of their lack of localization (the fact that any two of them overlap). Another argument against the Fourier basis is that it cannot express linear functions (affine deformations). The only two remaining candidate basis are therefore B-splines and B-spline wavelets.

### A. Splines versus wavelets

To make a fair comparison between B-spline and wavelet bases, we consider compactly supported cubic B-spline wavelets [54] spanning the same cubic spline space. First, let us analyze the task of evaluating the deformation at a single point. For simplicity, we will work in 1D. There are only four participating B-splines altogether while there are four participating B-spline wavelets at each level, plus four scaling functions (cubic B-splines) at the coarsest level. Second, to evaluate the deformation at a set of equally spaced points (this corresponds to a regular grid in multiple dimensions), the direct B-spline representation is also the most efficient, the interpolation requiring only four multiplications per pixel.

\(^4\)We assume that the cost of evaluating the basis function itself is constant or that their values can be precalculated.

This is better than all alternatives available when using the B-spline wavelets, including iterative filterbank and FFT-based algorithms.

Note that the complexity of evaluation of the gradient of the criterion corresponds to the complexity of the evaluation of the deformation because the same type of formula is involved (see Section VI-A).

### B. B-spline deformation model

The B-spline deformation model is obtained by substituting a scaled version of the B-spline (or tensor product thereof) in (3)

\[
g(x) = x + \sum_{j \in I_c \subset \mathbb{Z}^N} c_j \beta_{n_m}(x/h - j)
\]

where \( n_m \) is the degree of splines used, \( h \) is the knot spacing, and the division is taken elementwise. This corresponds to placing the knots on a regular grid over the image. We require the node spacing \( h \) to be integer, which together with the separability of \( \beta_{n_m}(x) \) implies that the values of the B-spline \( \beta_{n_m}(x) \) are only needed at a very small number of points \( (n_m + 1)h \) and can be precalculated. We can evaluate \( g \) on the whole grid with the cost of only \( N(n_m + 1) \) multiplications per pixel.

![Fig. 1. Basis functions involved in evaluating the value of a 1D function at one point (denoted by a vertical line): radial basis functions |x|\(^3\) (a), harmonic functions (b), cubic B-splines (c), cubic B-spline wavelets (d).](image-url)
The B-spline model has good approximation properties and is fast to evaluate. It is physically plausible, for example cubic splines minimize the ‘strain energy’ \( \|g''\|_2 \) \([55, 56]\). It can encode all affine transformations, including rigid body motion. Moreover, B-splines are scalable in the sense that any coarse level deformation can be represented at a finer scale without any loss of information given an integer ratio between scales. The expansion operator (Section VI-C) is therefore exact.

IV. OPTIMIZATION STRATEGY

A. Optimization algorithm

Recall from (1) and (4) that we need to minimize a criterion \( E \) with respect to a finite number of parameters \( \mathbf{c} \). To determine which of the many available algorithms performs best in our context, we tested four local iterative algorithms which can be cast into a common framework: At each step \( i \) we take the actual estimate \( \mathbf{c}(i) \) and calculate a proposed update \( \Delta \mathbf{c}(i) \). If the step is successful, then the proposed point is accepted, \( \mathbf{c}(i+1) = \mathbf{c}(i) + \Delta \mathbf{c}(i) \). Otherwise, a more conservative update \( \Delta \mathbf{c}(i) \) is calculated, and the test is repeated.

1) Gradient descent with feedback step size adjustment with update rule: \( \Delta \mathbf{c}(i) = -\mu \nabla E(\mathbf{c}(i)) \). After a successful step, \( \mu \) is multiplied by \( \mu_f \), otherwise it is divided by \( \mu_f \).

2) Gradient descent with quadratic step size estimation. We choose a step size \( \mu^r \) minimizing the following approximation of the criterion around \( \mathbf{c}(i) \): \( E(\mathbf{c}(i) + \mathbf{x}) = E(\mathbf{c}(i)) + \mathbf{x}^T \nabla E(\mathbf{c}(i)) + \alpha \| \mathbf{x} \|^2 \), where \( \alpha \) is identified from the two last calculated criterion values \( E \). As a fallback strategy, the previous step size is divided by \( \mu_f \), as above.

3) Conjugated gradient. This algorithm \([43]\) chooses its descent directions to be mutually conjugate so that moving along one does not spoil the result of previous optimizations. To work well, the step size \( \mu \) has to be chosen optimally. Therefore, at each step, we need to run another internal one-dimensional minimization routine which finds the optimal \( \mu \); this makes it the slowest algorithm in our setting.

4) Marquardt-Levenberg. The most effective algorithm in the sense of the number of iterations was a regularized Newton method inspired by the Marquardt-Levenberg algorithm (ML), as in \([22]\). Various approximations of the Hessian matrix \( \nabla^2 E \) were examined (see also Section VI-A).

As the behavior of all optimizers is comparable at the beginning of the optimization process (see Figure 2), the main factor determining the speed is the cost of a single iteration. The evaluation costs are presented in Table I; for the ML algorithm, the cost of the Hessian matrix inversion (which grows with the cube of the number of parameters) must be added. It follows that the gradient descent (GD) iterations are the least costly, the difference between the two variants being minimal. We therefore recommend to use the GD algorithm with the quadratic step size estimation (which works better than the feedback adjustment) and we use it for experiments in the remainder of the paper. One additional pleasant property of the GD algorithm is its tendency to leave uninfluential coefficients intact, unlike the ML algorithm. Consequently, less regularization is needed for the GD algorithm.

Under different constraints, when a small number of parameters is sought, the criterion is smooth, and high precision is needed, the ML algorithm performs the best. This is because its higher cost per iteration is compensated for by a smaller number of iterations due to the quadratic convergence. An example of such a situation is shown in Figure 3. (See also \([57]\)). Among Marquardt-Levenberg (ML) algorithms, we found the performance to be superior when using the full Hessian.

B. Multiresolution

As in \([22]\), we use the multiresolution approach for both the image and deformation models. We start with the coarsest resolution versions of both, and alternatively refine the image and the deformation model every time convergence is reached, until the finest level. The coarse versions of images are generated using a reduction operator (see Section VI-C). Conversely, coarse level solutions are extrapolated to finer levels using an expansion operator (cubic spline interpolation).
corresponding to their stiffnesses, and landmark positions in the reference, resp. test images. The spring factors $E_s$ should aim for a compromise between landmark pairs. We propose to start with all the criterion experimentally to get the most satisfactory results. We augment the data part of the model the additional overhead is negligible. Another difference is that thanks to our parametric deformation information at landmark points where it is really known.

The landmarks are added when the automatic algorithm

does not succeed in making the algorithm to converge to the right solution, and $\alpha_i$ too high that forces the solution to a landmark position that is perhaps not sufficiently precise.

As an example, we tried to register an MRI slice from an atlas\(^7\) with a sample MRI test image\(^8\). The atlas is a labeled and annotated collection of images. To identify the same structures in the test image, we register it with the unlabeled version of the atlas. Once the geometric correspondence is established, the structures and their labels from the atlas can be projected onto the test image. Prior to registration, the histogram of the test image was matched to that of the reference. The unsupervised registration correctly registers some of the structures but misses others; in particular the skull boundary (see Figure 4). We then identified several landmarks in both images (Figure 5). Using this minute hint, the semi-automatic algorithm could recover a plausible deformation, even though the landmark information alone (using e.g., thin-plate splines) would not have been enough [38]. We gave the weight 1.0 to all landmarks except the landmark at the bottom left part of the skull which had a weight of 0.2. This made the final positions of the landmarks coincide with the target ones to within about 2 pixel for the least weighted landmark and about 1 pixel for all the others.

Adding the spring term privileges likely solutions based on our a priori knowledge and makes the problem better-posed. The points need not to be image-dependent landmarks. For example anchoring the four corners of the image prevents the solution from degenerating. In this way, the springs play in part the role of a regularization factor.

The landmarks are added when the automatic algorithm cannot solve the problem by itself and an input from a human expert is needed. For this reason, we decided to accept the landmark data as trustworthy and definitive. This is unlike in [58, 59], where the landmarks come from an automatic process, such as iterative closest-point algorithm (ICRP), and therefore cannot be regarded as definitive. However, it is possible to give a certain feedback to the expert, for example the value of the criterion in landmark neighborhoods. This could be also used to reject misplaced landmarks.

VI. IMPLEMENTATION ISSUES

The purpose of this section is to describe some specific aspects of our implementation. These are mostly independent of the main philosophy of the algorithm but can have a major impact on its performance.

A. Explicit derivatives

For the optimization algorithm, we need to calculate the partial derivatives of $E$, as they form the gradient vector $\nabla_s E(c^{(i)})$ and the Hessian matrix $\nabla^2_s E(c^{(i)})$. Starting from equation (1), we obtain the first partial derivatives

$$\frac{\partial E}{\partial e_{i,m}} = \frac{1}{|I|} \sum_{i \in I} \frac{\partial f_w(i)}{\partial x_m} \left. \frac{\partial f^c(x)}{\partial x_m} \right|_{x=g(i)} \frac{\partial g_m(i)}{\partial e_{i,m}}$$

\(^7\)Courtesy of Harvard Medical School, http://www.med.harvard.edu/AANLIB/home.html

\(^8\)We use a proton density MR image from the Visible Human project http://www.meddean.luc.edu/lumen/meded/grossanatomy/cross_section/index.html
Fig. 4. The reference MRI proton density brain slice from the atlas with (a) and without labels (b). The sample test slice of a corresponding region (c). The superposition (in red and green) of the two images before (d) and after the registration (e). The deformation field (f). Cubic splines were used with knot spacing of $h = 32$. The image size was 512 x 512 pixels. The difference between images is only partially corrected by the unsupervised registration. Misalignment of several structures is clearly visible.
The superimposed images after registration using the semi-automatic algorithm structures are well identified; the alignment is clearly superior to that in Figure 4.

Fig. 5. The reference (a) and test (b) images with superimposed landmarks (in red). The superimposed images after registration using the semi-automatic algorithm (c) and the deformation field found (d). Corresponding anatomical structures are well identified; the alignment is clearly superior to that in Figure 4.

as well as the second partial derivatives

$$\frac{\partial^2 E}{\partial c_{j,m} \partial c_{k,n}} = \frac{1}{\|I\|} \sum_{i \in I} \left( \frac{\partial^2 f_1^e}{\partial f_w(i)^2} \frac{\partial f_1^c}{\partial x_m} \frac{\partial f_1^c}{\partial x_n} \right. + \frac{\partial c_i}{\partial f_w(i)} \frac{\partial^2 f_1^c}{\partial x_m \partial x_n} \left) \frac{\partial g_m}{\partial c_{j,m}} \frac{\partial g_n}{\partial c_{k,n}} \right) \left(7\right)$$

From (1) defining the SSD criterion, we get $$\frac{\partial E}{\partial f_w(i)} = 2(f_w(i) - f_t(i))$$ and $$\frac{\partial^2 E}{\partial f_w(i)^2} = 2$$. The derivative of the deformation function (4) is simply $$\frac{\partial g_m}{\partial c_{j,m}} = \beta_m(x/h - j)$$. The deformation model is linear and all its second derivatives are therefore zero; that is the reason for the simplicity of (7). The partial derivatives of $$f_1^c$$ in (6) and (7) can be calculated from (2) as a tensor product $$\frac{\partial f_1^c}{\partial x_m}(x) = \sum_{k \in I} b_k \beta_n(x_m - k_m) \prod_{l \notin m} \beta_n(x_l - k_l)$$. Second-order partial derivatives of $$f_1^c$$ are obtained in a similar fashion.

The Marquardt-Levenberg approximation of the Hessian assumes that the term $$\frac{\partial^2 E}{\partial f_w(i)^2}$$ is negligibly small or that it sums to zero on average, which justifies omitting this term from (7), see [43]. Another simplification is to consider only diagonal terms $$\frac{\partial^2 E}{\partial c_{j,m}^2}$$ of which the diagonal Hessian is negligible small or that it sums to zero on average, which justifies omitting this term from (7). Both $$\frac{\partial E}{\partial c_{j,m}}$$ can be transformed into a discrete separable convolution $$\left\{ \frac{\partial E}{\partial c_{j,m}} \right\}_j = \sum_i w(i)b(j \cdot h - i) = (w \ast b) \lfloor j \cdot h \right\}$$, where we have substituted $$w$$ for the first two factors in (6), $$b(q) = \beta_n(-q/h)$$, and $$\lfloor \cdot \right\}$$ indicates downsampling as defined by the formula, with elementwise multiplication $$\cdot \cdot \cdot$$. The convolution kernel $$b$$ is separable and the convolution can be calculated as a sequence of $$N$$ unidimensional convolutions $$\left((w \ast b_1) \ast \cdots \ast b_n\right)_{\lfloor j \cdot h \right\}}$$.

Because of the downsampling, calculating one output value at step $$k$$ consists of a scalar product with a filter $$b_k$$ of length $$(n_m + 1)h_k$$ and shifting this filter by $$h_k$$.

C. Multiresolution spline representation

To deploy the multiresolution strategy (see Section IV-B), we need to specify expansion and reduction operators. We will use the same approach for both the deformation model and the image model. The expansion can be performed exactly; we choose to do optimal reduction in the $$L_2$$ sense [60]. Both expansions and reductions can be performed efficiently using FIR and recursive IIR filters. To cope with the finite extent of our signals, we put extra B-splines outside the interval of interest. This allows for complete control of the signal within the interval of interest, see [38] for details.

D. Fast spline calculations

It is essential to take full advantage of the properties of splines. First, specialized routines are used to calculate the
values of a B-spline of a specific order using a minimum number of operations. Second, as we are using tensor products of B-splines as our basis functions, many operations can be performed in a separable fashion, reducing the complexity of operations from $O(k^N)$, where $N$ is the number of dimensions and $k$ the size of the data, to $O(kN)$. This is the case for the prefiltering step required to find the B-spline coefficients, and also for the interpolation of values of a function given by its B-spline coefficients. Third, the compact support of B-splines simplifies many of the infinite sums in the expressions given earlier, reducing them to sums over just a small number of elements.

E. Stopping criterion

To get a fast optimization algorithm, particular attention has to be paid to the stopping criterion. This holds for both GD and ML algorithms. Classically, the relative and absolute improvement of the criterion value is compared with a fixed threshold [43]. For our class of problems, we found it to be advantageous to base the stopping criterion on the changes $\Delta \mathbf{c}$ of parameter values. We stop when the step size falls below an $a$ priori given threshold $\varepsilon$. The size of a step that fails gives an indication of the accuracy of the result and is therefore easy to set. Typically, we would use the threshold of $\varepsilon = 10^{-1} \sim 10^{-3}$ pixels for the finest level an slightly more for coarser levels, as there is usually not enough details and coherence between levels.

F. Masking

A substantial gain in speed comes from considering only important pixels when calculating the data criterion (1) and its derivatives. It is possible to determine an $a$ priori mask of significant pixels, for example $10 \sim 50$ $\%$ of the total number of pixels, and to consider only those pixels in subsequent calculations. The contributions of individual pixels to the change of the criterion is directly proportional to the amplitude of the directional derivatives at the respective points, see (6). Therefore, a reasonable strategy is to construct the mask by thresholding the gradient of the image at each multiresolution level.

VII. Experiments

This section presents a series of experiments in a controlled environment to assess the accuracy, speed, and robustness of our algorithm. We show the SSD criterion (1) we minimize, and also a warping index $\varpi = \sqrt{\frac{1}{|R|} \sum_{i \in R} \| \mathbf{g}(\mathbf{r}) - \mathbf{g}_c(\mathbf{r}) \|^2}$; that is, the mean geometric error between the true and the recovered deformation. The mean is only calculated over a region $R$, the part of the image containing useful data (object); an example of a region can be seen in Figure 6, bottom left.

A. Registration of MRI brain slices

To illustrate the behavior of the algorithm, we show its performance when recovering a known deformation of a 2D slice of an anatomical spin-echo MRI volume of the brain.\(^9\) We use here artificially deformed images because the knowledge of the ground truth permits us to better judge the performance of the algorithm.

The original image of size $256 \times 256$ pixels is shown in Figure 6, top left. We use a cubic spline control grid with one knot for every 32 pixels. We warp the image with a deformation belonging to the warp space and consisting of displacements up to 15 pixels (1 pixel corresponds to approximately 0.9 mm). The warped image is superimposed on the original in Figure 6, top right. Then the automatic registration algorithm is run. The stopping threshold is set to 0.5 pixels for all levels except the last, where we set it to 0.1 pixels. The recovered deformation was used to warp again the original image. Its warped version is shown superimposed on the image warped with the true deformation in Figure 6, bottom right. We note that the deformation was well recovered with no perceptible difference.

The spatial distribution of the resulting geometrical error is shown in Figure 8. The maximum error is about 1.5 pixels, while the mean geometric error (warping index $\varpi$) over the total of the brain is about 0.4 pixels. We generally observe that the error is concentrated in areas with little detail in the image. Other, high-contrast regions such as edges are resolved much more precisely than indicated by the value of $\varpi$, often with subpixel accuracy. On the other hand the agreement in the zones with low-contrast will be worse and often only coincidental, since there is little or no information to guide the algorithm.

The evolution of the optimization can be studied from the graphs in Figure 7. We observe the steady and correlated descent of the observable criterion being optimized ($E$) and of the warping index ($\varpi$), the quantity measuring the quality of the registration. The abrupt changes in the curves are caused by the transitions between levels of the multiresolution progression; they are small thanks to the accuracy of the spline model.

Note that the final values of both $E$ and $\varpi$ depend strongly on the preset stopping threshold, which in turn influences the optimization time. The threshold value is a subjective compromise between the accuracy and computation time. It is perfectly possible to stop optimizing only after 7 s and skip the finest resolution level altogether, if the precision of $\varpi = 0.7$ pixels is acceptable. On the other hand, after about 4 more minutes of iteration, the error $\varpi$ descends to less than $10^{-4}$ pixels. However, in the authors’ opinion, such super subpixel accuracy is almost never achievable on real images, because of the noise and the unknown characteristics of the acquisition process.

B. Deformation generator

We have implemented a fractional wavelet based random deformation generator. It yields deformations with a prescribed smoothness (regularity), characterized by a Sobolev exponent $r$ — the maximum number of (fractional) derivatives in the

\(^9\)First author’s brain. Images courtesy of Arto Nirkko from Inselspital Hospital, Bern, Switzerland.
$L_2$ sense. This is guaranteed if the Fourier transform decreases asymptotically at least as $1/\omega^{r+0.5}$. We express the random displacement $g(x) - x$ in an orthogonal wavelet basis. We use orthonormal symmetric fractional B-spline wavelets [61, 62] of degree $\alpha = r - 0.5$, which have precisely the desired regularity and Fourier decay at infinity. We let the wavelet coefficients $\theta_{j,k}$ be random (zero mean, independent, and normally distributed) with standard deviation decreasing as $2^{-rj}$, where $j$ denotes the scale. This makes the Fourier spectrum of the displacement decrease as required over the whole frequency range and ensures that the (mean) displacement belong to the Sobolev space $W^{r}_{2}$ [63].

To obtain corresponding 2D deformation fields, we use separable 2D wavelet transforms with the same basis functions and the same decrease of amplitude of the coefficients in each component as in the 1D case. We can observe in Figure 9 how the deformation gets progressively more smooth and regular with increasing $r$.

### C. Out-of-space deformation

The true deformation is not guaranteed to lie in the space where we are looking for it and can therefore never be recovered exactly. The associated error is called an approximation error. We performed various experiments to compare the approximation error with the overall registration error.

We generated a random hierarchical deformation using the wavelet methodology from the previous section (with $r = 2$) and projected it into the space with knot spacing $h = 8$. We deformed the MRI image (Figure 6) with this deformation and tried to recover it in spaces with knot spacings $h = 8 \sim 256$. Figure 10 shows the recovered deformations and the residual differences between the reference image and the warped test images for different values of the knot spacing $h$. We observe that the deformation can be recovered almost completely when we search in the correct space ($h = 8$); important errors arise when we search in different, coarser spaces. Ultimately, for $h = 256$, we can express only deformations close to affine, which is obviously not enough to capture all the details of the true deformation.

We now compare the error that our algorithm yields with the smallest error it could possibly achieve, given the search space. To find the best achievable approximation of some deformation, given the knot spacing and spline degree, we will use the fact that the warping index is in fact the $\ell_2$ (Euclidean) distance. Therefore, the best approximation is an orthogonal projection of the deformation onto the search space and can be calculated easily.

The warping index resulting from the registration process is compared with the best achievable one in a given space in Figure 11. We see that although the ideal values are not

---

Fig. 7. The evolution of the optimization process. The left column displays the evolution with respect to the number of iterations, while the right column represents the same quantity respect to time. The first row shows the SSD criterion $E$, the second row the warping index $\omega$. The step changes correspond to the changes in the model and image resolutions. We observe good correlation between all four graphs.
Fig. 6. From top to bottom: The original slice of anatomical MRI brain image, original superimposed over the true deformation, the recovered deformation versus the true deformation, and the mask used to calculate the warping index (bottom left).

Fig. 8. The geometrical error after registration (green) with superposed contours of the original MRI image (red). The maximum (green) intensity corresponds to an error of 1.5 pixels.

$r = 1.4$
$r = 1.6$
$r = 1.8$
$r = 2.0$
$r = 2.2$
$r = 2.4$

Fig. 9. Examples of randomly generated fractal-like deformations for various Sobolev exponents $r$. Observe how the deformation gets smoother with increasing $r$. 
Fig. 10. The deformation recovered using progressively smaller (coarser) deformation spaces (left column), and the corresponding residual error (right column). The knot spacing $h$ and warping index $\varepsilon$ are shown. Cubic splines were used.

attained, the difference is within the range of half a pixel. In real situations, the true deformation space is not known. However, thanks to the good approximation properties of splines, we can reasonably expect that by using a sufficiently small value of $h$, we can reduce the approximation error to an acceptable value.

D. Choosing the spline degree

The choice of the spline degree for the image and deformation models is a trade-off between the accuracy and speed. Here too we generated hierarchical random deformations (see Section VII-C) with varying smoothness and applied them on the MRI image. We recover the deformations in spline spaces with grid spacing $h = 32$ pixels for linear, quadratic, and cubic spline deformation models, with $\varepsilon = 0.01$ pixels. We observe (see Figure 12) that higher order splines perform better, while the difference between linear and quadratic is much more important than between quadratic and cubic splines. The sample registration times were 20.1 s, 26.7 s, and 48.9 s, for linear, quadratic, and cubic splines, respectively. This indicates that to use quadratic splines for the deformation model might be a good compromise between approximation properties and speed. Note that the task of recovering fastly changing deformations is doubly difficult, as they cannot be represented well by the deformation model and they do not have a pronounced effect on the image because of its lack of details at small scale in many regions. Note also that as the deformation gets smooth, the geometric error of the recovered deformation gets almost as small as the minimum achievable error.

E. Choosing the grid spacing

Thanks to the properties of our deformation model and the optimization algorithm, the grid spacing and thus the number of parameters influences the execution time only mildly. Therefore, the main criterion for choosing the grid spacing $h$ should be the estimated intrinsic resolution (smoothness) of the deformation to be recovered. A control grid that is too coarse is not able to express the deformation in all details. On the other hand, too fine a control grid is overcompensating for
true image differences and noise. The effect of the grid spacing is less pronounced for smoother deformations, see Figure 13.

F. Noise dependence

We added various levels of noise to the test images (i.e., after the warping has been performed) to demonstrate the influence of the SNR (signal to noise ratio) on the registration results. We used 60 random deformation with \( r = 2.0 \), cubic splines with knot spacing \( h = 32 \), and stopping criterion \( \varepsilon = 0.01 \). We observed that for SNR better than 10 dB, the influence of the noise is very small (Figure 14).

G. Starting point

The following experiment evaluates the robustness of the algorithm with respect to the starting point. Here, we tried to recover the deformation from Section VII-A (MRI images) optimizing only at the finest level. We linearly varied the starting point of the optimizer between identity and the true deformation and observed the attained warping index \( \infty \) for a stopping threshold of \( \varepsilon = 0.01 \) pixels. Figure 15 shows the warping index of the deformation used as a starting point and the warping index of the recovered deformation. We observe that although the final result does depend on the starting point, it is most likely only the influence of the stopping criterion.

The algorithm therefore proves to be very robust, even without the help of a multiresolution: it converged in all cases to the desired solution. On the other hand, the elapsed time and the number of iterations differed significantly, from 2 iterations when starting from the true solution, to several hundreds when starting from identity.

H. Statistical distribution of errors

To evaluate the behavior of the algorithm on a larger set of test cases, we generated a series of random hierarchical deformations (see Section VII-B), warped the MRI slice with them, and applied our registration algorithm to recover the deformation. We used the stopping threshold \( \varepsilon = 0.01 \) pixels and a warping space which contained the deformation. We then compared the warping index corresponding to the recovered deformation with the initial warping index, that is, the distance between the true deformation and identity. In Figure 16 we present the scatter plot describing the relation between the initial and final warping indexes. We observe that the algorithm gives results with accuracy consistently better than 0.1 pixels.

I. Experiments with real data

We applied our algorithm to various problems involving medical images of several modalities. We developed a registration procedure for ECD\(^{10}\) and Xenon inhalation SPECT images [64] in the view of atlas creation [37]. Figure 17 shows the resulting alignment obtained after registering two

\(^{10}\)ECD (Technetium Ethylene Cysteine Diethylester) is a radioactively marked intravenously injected agent.
slices of anatomical (spin-echo) MRI images of two different subjects.\textsuperscript{11}

To further illustrate the use of our algorithm, we present registered MRI images from a heart beat sequence\textsuperscript{12}, see Figure 18. The extracted deformation field can be used to extract trajectories of various points in the heart which is important for diagnostic purposes. Analyzing this field also permits the determination of the velocity and derived parameters, such as the accumulated displacement, and strain. We also analyzed standard 2D ultrasound sequences of the heart [65]. The algorithm proved to be robust to the occasional change of structure (topology) due to the underlying 3D nature of the true movement.\textsuperscript{13}

Another technique for assessing cardiac performance is myocardial perfusion by MRI [66, 67]. A sequence of MRI images\textsuperscript{14} is acquired with at high speed to assess the diffusion of the agent. A role of the registration is to compensate for the (heart) motion to provide the time profiles of the intensities at each tissue point. The profiles are subsequently analyzed to yield the physical (absorption) parameters of the tissue. Figure 19 shows a few selected images of the sequence. It also shows differences between images; we observe a significant amount of motion artifacts. Most of these artifacts are compensated for in the corrected sequence, where each of the images was registered with (and warped towards) its already corrected predecessor. Ideally, the corrected sequence should appear static, except for the movement of the agent. In this application, a number of virtual springs with carefully chosen weights was used, to make the deformation compensate for the movement of the tissues, but not for the movement of the contrast agent.

Let us end with a 3D example: the registration of two computer tomography (CT) head volumes.\textsuperscript{15} Due to the large size of the original volumes (512×512×45 voxels), it was impractical to perform the registration directly. We chose instead to perform the registration on reduced volumes (128×128×45) which took about 10 minutes to complete\textsuperscript{16} with the control knots placed every 8×8×8 voxels and stopping threshold of ε = 0.01 pixels. We then interpolated this deformation to the original volume size.\textsuperscript{17}

We observe that it is difficult to do any meaningful comparison of the volumes prior to registration, see Figure 20. However, once the registration is performed, even small differences are clearly apparent (Figure 21). Moreover, the deformation field itself can provide valuable quantitative information about the relative sizes and shapes of various parts of the anatomy from the two volumes. Note that the control grid spacing must be adapted to the task at hand because it influences the amount of the agent.

\textsuperscript{11}Images courtesy of Arto Nirkko, Inselspital Hospital, Bern, Switzerland.
\textsuperscript{12}LECB, NIH, http://www-lecb.ncifcrf.gov/flickr/
\textsuperscript{13}Analyzing directly 3D ultrasound heart sequences would avoid this problem. However, 3D heart sequence acquisitions are much more rare in the clinical use.
\textsuperscript{14}Courtesy of J.-P. Vallée, Unité d’imagerie numérique, University Hospital, Geneva, Switzerland.
\textsuperscript{15}Images courtesy of Philippe Thévenaz, EPFL, Lausanne, Switzerland. The images were acquired using the same machine and the same protocol, but not preregistered.
\textsuperscript{16}On a 700 MHz Pentium based computer.
\textsuperscript{17}Registering directly the undecimated volumes on the same computer takes about 3 hours with very minor increase in quality as relatively smooth deformations are sought. We are currently working on an optimized reimplementation of the algorithm that should reduce these times considerably.
Fig. 19. The first line presents original images number 6, 9, 11, and 14 from a sequence of originally 60 images of myocardial perfusion MRI. The second line presents the difference images between the original images and their immediate predecessors; movement artifacts can be clearly seen. On the third line you can see the difference images from the motion corrected sequence using our algorithm; the movement artifacts are significantly reduced. The same effect is also visible comparing the differences of the sequence images with the first image of the sequence on the original (fourth line) and corrected (fifth line) sequences.
Fig. 18. The reference MRI image from a heart sequence with superimposed contours (a). The same contours over another image (the test image) from the same sequence before the registration (b) and after (c). The deformation field (d). Quadratic splines were used with knot spacing of $h = 64$, image size was $256 \times 256$ pixels.

Fig. 20. The axial, sagittal, and coronal views of the two CT brain volumes (one in red, second one in green) prior to registration.

VIII. CONCLUSIONS

We developed a fully automatic elastic registration algorithm. We extended the idea from [22] to multidimensional data, and streamlined the algorithm to accelerate it. We designed a new step-prediction formula for the gradient descent algorithm and showed its efficiency for our application. A double multiresolution strategy brings speed and robustness and additionally eliminates the need for an initial rigid registration as the coarse grid deformation itself plays this role.

We introduced the concept of virtual springs, yielding a semi-automatic registration method, capable of using expert hints in the form of landmarks to solve particularly difficult problems where the fully automatic algorithm may be mislead. This is a powerful combination of the ideas of manual landmark registration and the pixel-based registration using splines.

We applied the algorithm to a wide range of artificially generated problems involving deformations with varying smoothness applied to anatomical MRI images to demonstrate the algorithm’s speed, robustness, and accuracy. Furthermore, we presented several medical applications using various image modalities.

We believe that by producing a specialized program taking advantage of a specific configuration, the run time can be decreased by an additional factor of 2 to 10. This will enable truly interactive operation of automatic and semi-automatic elastic image registration with numerous applications in medicine, biology, and any other field where deformed images need to be compared.
ACKNOWLEDGMENTS

We are grateful to Dr. Philippe Thévenaz for helpful discussions.

REFERENCES


