Optimization of Mutual Information for Multiresolution Image Registration

Philippe Thévenaz, *Member, IEEE,* and Michael Unser, *Fellow, IEEE*

***Abstract*—We propose a new method for the intermodal registration of images using a criterion known as mutual information. Our main contribution is an optimizer that we specifically designed for this criterion. We show that this new optimizer is well adapted to a multiresolution approach because it typically converges in fewer criterion evaluations than other optimizers. We have built a multiresolution image pyramid, along with an interpolation process, an optimizer, and the criterion itself, around the unifying concept of spline-processing. This ensures coherence in the way we model data and yields good performance. We have tested our approach in a variety of experimental conditions and report excellent results. We claim an accuracy of about a hundredth of a pixel under ideal conditions. We are also robust since the accuracy is still about a tenth of a pixel under very noisy conditions. In addition, a blind evaluation of our results compares very favorably to the work of several other researchers.**

***Index Terms*—B-spline, intermodal volume alignment, Marquardt–Levenberg, Parzen window, pyramid.**

I. Introduction

IMAGE registration addresses the following problem: given two images (or volumes), find a geometric transformation that maps the first image into the second one [1]. This problem often occurs in biomedical applications [2], [3]. When the difference between the two images is only the condition of the subject (e.g., resting versus performing a task), we speak of intramodal registration [4], [5]. Alternatively, when the subject is imaged in essentially two different ways (e.g., local glucose uptake versus proton density), we speak of intermodal registration [6]. This second task is more difficult than the first one because of the lack of a direct relation between the intensities of the two images. Another area where image registration plays an important role is remote sensing [7]. There, intramodal registration is often applied to mosaicking applications [8], the registration of images with different ground resolutions [9], and the detection of changes in the landscape [10], while intermodal registration is necessary to correct for band-to-band misregistration [11].

Recently, an elegant solution has been proposed independently by Viola *et al.* [12] and Collignon *et al.* [13], which is based on the maximization of the statistical dependence—or mutual information—of corresponding voxel intensities in the images to register. This information-theoretic criterion does not depend on any assumption on the data (other than stationarity), does not assume specific relations between intensities in different modalities and can be applied without modification to any pair of modalities.

The purpose of this paper is to present a new, highly-efficient optimizer for the maximization of mutual information. It is designed to converge in very few criterion evaluations to the desired optimum when initialized with good starting conditions. We formulate the mutual-information criterion as a continuous and differentiable function of the registration parameters using Parzen windows. The optimizer takes advantage of the differentiability of the criterion to get a global understanding of the behavior of the criterion near the optimum. This is exploited in a Marquardt–Levenberg-type of iterative procedure and exhibits superlinear convergence when close enough to the optimum [14].

Another goal of this paper is to investigate the issue of accuracy. While the criterion alone more or less determines the accuracy of registration on a one-pixel scale, the interpolation model plays an essential role when sub-pixel accuracy is desired. In the intramodal case, correlation methods allowed for high accuracy when combined with high-order interpolation models [5]. Thus, the question arises whether it is possible to reach a similar accuracy in a reasonable time for the intermodal case. Since the main drawback of high-order interpolation models is their computational cost, it is necessary to develop an optimizer that is fast without compromising accuracy. Multiresolution is a natural solution to this problem, but not all optimizers are equally suitable: the best candidates are those optimizers that converge in very few criterion evaluations when initialized with good starting conditions. This requirement rules out many optimizers, for example those that first need to explore around the initial condition before eventually becoming superlinear, or those that consider only one parameter at a time.

Our optimizer works in conjunction with a high-quality multiresolution representation of the image based on cubic splines. Optimization is first performed on a coarse scale with few data and then refined at finer scales, gradually taking more data into account. Our image pyramid is such that there are as few differences as possible between levels. Together, this least-squares pyramid and the immediate superlinear convergence of our optimizer tend to diminish the number of criterion evaluations; the simultaneous processing of all parameters further reduces this number.

We investigate the performance of our optimizer on intramodal and intermodal images with simulated misregistrations.

Manuscript received May 7, 1999; revised June 12, 2000. The images and the gold-standard transformations of Section V-D were provided as part of the project, "Evaluation of Retrospective Image Registration," National Institutes of Health, 1 R01 NS33926-02, Principal Investigator, J. M. Fitzpatrick, Vanderbilt University, Nashville, TN. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Pierre Moulin.

The authors are with the Biomedical Imaging Group, Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland (e-mail: philippe.thevenaz@epfl.ch).

Publisher Item Identifier S 1057-7149(00)10073-9.

We demonstrate the accuracy of the method on real-world data from the Vanderbilt database, for which the correct registration solution is approximately known [2]. We also compare our optimizer to the method of Maes *et al.* [15] using the standard Powell optimization algorithm and show an increase in performance with a factor of about 6 without loss of precision.

This paper is organized as follows: in Section II, we introduce mutual information and we design a way to achieve its computation. In Section III, we present a continuous model for the images and we discuss the benefits of a multiresolution approach. In Section IV, we describe a new algorithm that optimizes the mutual information between the two images to register. In Section V, we perform several experiments and compare our results to those of other researchers. In Section VI, we relate this paper to the work of other researchers. We conclude in Section VII.

II. Mutual Information

A. Definitions

1) Parzen Window: Let w be a function with unit integral $(\int_{-\infty}^{\infty} w(\xi)\,d\xi = 1)$. Further, let $\{x_i\}$ be a set of N samples of a random variable X with probability density function $p(x)$. Then, the so-called Parzen estimate of p is

$$\tilde{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{w((x-x_i)/\varepsilon(N))}{\varepsilon(N)} \tag{1}$$

where ε is a strictly positive scaling factor that controls the width of the Parzen window w. From (1), it is easy to see that $\tilde{p}(t_1)$ takes a large value at some position t_1 where many samples x_i happen to cluster so tightly that their associated Parzen windows $w((t_1 - x_i)/\varepsilon)/\varepsilon$ overlap often. In the contrary, at some other position t_2 where the samples happen to be not dense, few overlap takes place and $\tilde{p}(t_2)$ has a lower value. This process is particularly easy to understand if we ask that the Parzen window be positive $(w(\xi) \geq 0,\ \forall\,\xi \in \mathbb{R})$. We shall satisfy this positivity constraint throughout this paper, even though this is not required to ensure that $\tilde{p}_N$ converges to p when enough samples are available. Note that other technical conditions on w and on the dependence of $\varepsilon(N)$ on N are required for this convergence [16]. The underlying principle is as follows: when N is large, many samples are available and ε is made small, which leads to a scaled Parzen window $w(\cdot/\varepsilon)/\varepsilon$ that is Dirac-like. In turn, p can be captured in great details because the contribution of the samples are very local. In the contrary, when only few samples are available, ε is made large. This corresponds to a widening of the Parzen window w such that the influence of any sample x_i has a larger support, which tends to obliterate the details of p.

2) Histogram Estimation: Let $f_T(\mathbf{x})$ be a test image we want to align to a reference image $f_R(\mathbf{x})$. These images are defined on a continuous domain $\mathbf{x} \in V^c$ that may have any number of dimensions (e.g., surface, volume). The coordinates $\mathbf{x}_i$ are samples of V^c; the discrete set of these samples is called V. Let $\mathbf{g}(\mathbf{x};\, \mu_1,\, \mu_2,\, \cdots)$ be some geometric transformation with associated parameters $\boldsymbol{\mu} = (\mu_1,\, \mu_2,\, \cdots)$. Let L_T and L_R be discrete sets of intensities associated to the test and the reference image, respectively. Let w be a separable Parzen window defined as above. Then, we define the joint discrete Parzen histogram as

$$h(\iota,\, \kappa;\boldsymbol{\mu}) = \frac{1}{\varepsilon_T\, \varepsilon_R} \sum_{\mathbf{x}_i\in V} w(\iota/\varepsilon_T - f_T(\mathbf{g}(\mathbf{x}_i;\, \boldsymbol{\mu}))/\varepsilon_T) \cdot w(\kappa/\varepsilon_R - f_R(\mathbf{x}_i)/\varepsilon_R) \tag{2}$$

where $\iota \in L_T$ and $\kappa \in L_R$, and where ε_T is related to $\mathrm{card}(L_T)$ and ε_R to $\mathrm{card}(L_R)$. Hence, the contribution to the joint histogram of a single pair of pixels with intensities $(f_T,\, f_R)$ which can take values in a continuum, is distributed over several discrete bins $(\iota,\, \kappa)$ at once by the window function w. This joint histogram is proportional to the discrete Parzen probability (or frequency) p given by

$$p(\iota,\, \kappa;\, \boldsymbol{\mu}) = \alpha(\boldsymbol{\mu})\, h(\iota,\, \kappa;\, \boldsymbol{\mu}) \tag{3}$$

where we have introduced the normalization factor

$$\alpha(\boldsymbol{\mu}) = \frac{1}{\displaystyle\sum_{\iota\in L_T}\sum_{\kappa\in L_R} h(\iota,\, \kappa;\boldsymbol{\mu})}. \tag{4}$$

This normalization factor takes up the role of the factor $1/N$ in (1). It is required because it may happen that $\sum_{\xi} w(\xi + f) \neq 1$ for some f, even though every admissible Parzen window w satisfies $\int w(\xi)\, d\xi = 1$. The marginal discrete probabilities and histograms are given by

$$p_T(\iota;\, \boldsymbol{\mu}) = \alpha(\boldsymbol{\mu})\, h_T(\iota;\, \boldsymbol{\mu}) = \sum_{\kappa\in L_R} p(\iota,\, \kappa;\, \boldsymbol{\mu}), \tag{5}$$

$$p_R(\kappa;\, \boldsymbol{\mu}) = \alpha(\boldsymbol{\mu})\, h_R(\kappa;\, \boldsymbol{\mu}) = \sum_{\iota\in L_T} p(\iota,\, \kappa;\, \boldsymbol{\mu}). \tag{6}$$

3) Mutual Information: The negative S of the mutual information between the transformed test image and the reference image is

$$S(\boldsymbol{\mu}) = -\sum_{\iota\in L_T}\sum_{\kappa\in L_R} p(\iota,\, \kappa;\, \boldsymbol{\mu}) \cdot \log_2\left(\frac{p(\iota,\, \kappa;\, \boldsymbol{\mu})}{p_T(\iota;\, \boldsymbol{\mu}) p_R(\kappa;\, \boldsymbol{\mu})}\right). \tag{7}$$

The mutual-information registration criterion states that the transformed test image $f_T(\mathbf{g}(\mathbf{x}))$ is correctly aligned with the reference image by the parameter $\hat{\boldsymbol{\mu}}$ for which S is minimal.

B. Illustration

To give a concrete example of the objects defined above, we propose to simplify the situation as much as possible—in the context of the present illustrative section. First of all, we consider that the intensities of the test image f_T and that those of the reference image f_R consist of two levels only, given by $\{\iota_0,\, \iota_1\}$, and $\{\kappa_0,\, \kappa_1\}$, respectively. To further simplify, we also assume that $L_T = \{\iota_0,\, \iota_1\}$ and that $L_R = \{\kappa_0,\, \kappa_1\}$. Then, we assume that these images are defined on an infinite domain, while the discrete domain V onto which we shall perform computations is limited to some finite aperture. Two arbitrary images satisfying these conditions are given on top of Fig. 1, where it should be obvious that they are misregistered by a translation of

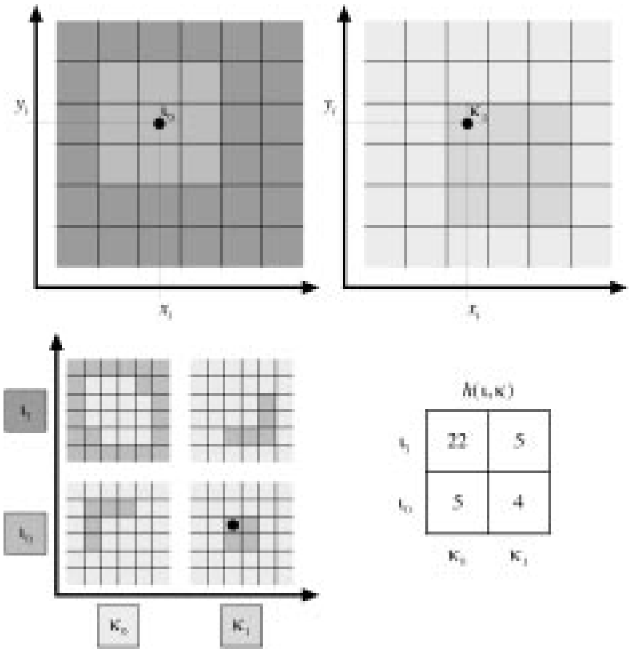


Fig. 1. Determination of the joint histogram needed to compute the mutual-information criterion.

1 pixel, both horizontally and vertically, and where we have that $\text{card}(V) = 36$. Then, we select as Parzen window a centered square pulse such that $w(\iota, \kappa) = 1$ for $|\iota|, |\kappa| < 1/2$, and such that $w(\iota, \kappa) = 0$ elsewhere. In addition, we set $\varepsilon_T = \varepsilon_R = 1$. For this trivial choice of Parzen window, the joint histogram defined in (2) is no different from a traditional one, where intensities would be first quantized and then would increment some discrete counter (this process is sometimes called binning).

The bottom-left part of Fig. 1 presents a pictorial description of the paired elements $\{(\iota_0, \kappa_0), (\iota_0, \kappa_1), (\iota_1, \kappa_0), (\iota_1, \kappa_1)\}$ that contribute to the joint histogram computed according to (2). To help focus the attention, some specific entry with spatial location (x_i, y_i) and with intensities (ι_0, κ_1) is shown with a black dot. The bottom-right part of the same figure gives the resulting joint histogram itself. After application of (3)–(7), the configuration shown in Fig. 1 results in the negated mutual information value

$$S = -\frac{5}{36}\log_2\frac{5\cdot 36}{27\cdot 9} - \frac{4}{36}\log_2\frac{4\cdot 36}{9\cdot 9} - \frac{22}{36} \\ \cdot \log_2\frac{22\cdot 36}{27\cdot 27} - \frac{5}{36}\log_2\frac{5\cdot 36}{27\cdot 9} \cong -0.045.$$

We leave to the reader the verification that, should the two images be aligned without misregistration, the mutual-information criterion (which is the negative of the mutual information itself) would reach its minimal value; in the case of this pair of images, it is given by $S \cong -0.865$. To undertake this task, it is necessary to remember the relation $\lim_{p\to 0_+} p \log p = 0$.

C. *Partition of Unity*

An unfortunate consequence of computing p_R as a marginal probability is that it makes it depend explicitly on the transformation parameters $(\mu_1, \mu_2, \cdots)$, even if the discrete images were to be of infinite spatial extent V. Although the reference image doesn't change with a variation in these parameters, p_R is sensitive to them because of the coupling introduced by the separable Parzen window w. One way to avoid this effect is to introduce the partition of unity constraint

$$\sum_{\xi\in\mathbb{Z}} w(\xi + f) = 1, \qquad \forall f \in \mathbb{R}. \tag{8}$$

Note that this constraint should not be confounded with the unit-integral constraint imposed upon every admissible Parzen window. When the partition of unity is satisfied for any sample value f, the marginal probability p_R becomes independent of the transformation parameters $(\mu_1, \mu_2, \cdots)$. From (6), (3), and (2), we determine that

$$\begin{aligned} p_R(\kappa; \boldsymbol{\mu}) &= \frac{\alpha(\boldsymbol{\mu})}{\varepsilon_T\,\varepsilon_R} \sum_{\iota\in L_T} \sum_{\mathbf{x}_i\in V} w(\iota/\varepsilon_T - f_T(\mathbf{g}(\mathbf{x}_i; \boldsymbol{\mu}))/\varepsilon_T) \\ &\quad \cdot w(\kappa/\varepsilon_R - f_R(\mathbf{x}_i)/\varepsilon_R) \\ &= \frac{\alpha(\boldsymbol{\mu})}{\varepsilon_T\,\varepsilon_R} \sum_{\mathbf{x}_i\in V} w(\kappa/\varepsilon_R - f_R(\mathbf{x}_i)/\varepsilon_R) \\ &\quad \cdot \underbrace{\sum_{\iota\in L_T} w(\iota/\varepsilon_T - f_T(\mathbf{g}(\mathbf{x}_i; \boldsymbol{\mu}))/\varepsilon_T)}_{1} \end{aligned} \tag{9}$$

where ε_T and ε_R have been chosen such that $\iota/\varepsilon_T \in \mathbb{Z}$ and $\kappa/\varepsilon_R \in \mathbb{Z}$ for $\iota \in L_T$ and $\kappa \in L_R$, respectively. Hence, we finally have that

$$\begin{aligned} p_R(\kappa) = \frac{\alpha(\boldsymbol{\mu})}{\varepsilon_T\,\varepsilon_R} \sum_{\mathbf{x}_i\in V} w(\kappa/\varepsilon_R - f_R(\mathbf{x}_i)/\varepsilon_R), \\ \forall(\mu_1, \mu_2, \cdots). \end{aligned} \tag{10}$$

This happens irrespectively of the extent of V, be it infinite or finite. Another advantage is that the normalization factor $\alpha(\boldsymbol{\mu})$ now takes a constant value α that doesn't depend on $(\mu_1, \mu_2, \cdots)$ as shown in (11) at the bottom of the page, where $\text{card}(V)$ denotes the number of samples $\mathbf{x}_i$ in V.

$$\alpha = \frac{1}{\displaystyle\sum_{\mathbf{x}_i\in V} \frac{1}{\varepsilon_T} \underbrace{\sum_{\iota\in L_T} w(\iota/\varepsilon_T - f_T(\mathbf{g}(\mathbf{x}_i; \boldsymbol{\mu}))/\varepsilon_T)}_{1} \frac{1}{\varepsilon_R} \underbrace{\sum_{\kappa\in L_R} w(\kappa/\varepsilon_R - f_R(\mathbf{x}_i)/\varepsilon_R)}_{1}} = \frac{\varepsilon_T\,\varepsilon_R}{\text{card}(V)} \tag{11}$$

D. Partial Overlap

In practice, the images f_T and f_R are defined over the continuous but finite domains V_T^c and V_R^c, respectively. It follows naturally that the domain on which it is possible to determine S is limited to their intersection

$$V^c = V_R^c \cap \mathbf{g}^{-1}(V_T^c;\, \boldsymbol{\mu}) \tag{12}$$

the extent of which will vary during the course of optimization, and is at most as large as V_R^c. To implement (2), these continuous domains are then sampled to yield the discrete set V, with size $\mathrm{card}(V)$. Although (12) seems to imply that there exist a dependence between V and some component μ of $\boldsymbol{\mu}$, it is necessary to understand that, due to the discrete nature of V, this dependence is not continuous: an infinitesimal variation $d\mu$ does not result in an infinitesimal variation dV. Rather, $\mathrm{card}(V)$ varies in an incremental way. Being discrete and finite, the set V is necessarily countable; thus, the ratio $dV/d\mu$ is zero almost everywhere (a.e.). Therefore, we shall ignore this last contribution in the calculus of the gradient of S with respect to $\boldsymbol{\mu}$.

Nevertheless, even if the infinitesimal variation is zero a.e., the incremental variation cannot be ignored. We acknowledge that fact by taking into account the actual value of the geometric parameter $\boldsymbol{\mu}$ while recomputing the volume of overlap V each time $\boldsymbol{\mu}$ is modified. Then, we recompute $h(\iota,\, \kappa;\, \boldsymbol{\mu})$, $p(\iota,\, \kappa;\, \boldsymbol{\mu})$, $p_T(\iota;\, \boldsymbol{\mu})$, $p_R(\kappa;\, \boldsymbol{\mu})$, and $S(\boldsymbol{\mu})$ accordingly.

E. B-Splines

B-spline functions $\beta^n(x)$ have many interesting properties [17], [18]. Of particular relevance for this paper is the fact that they satisfy the constraint for the partition of unity (8), while remaining positive, thus being admissible Parzen windows. In addition, they have the advantage of being smooth functions with explicit derivatives and a finite support. They are piecewise polynomials of degree $n \geq 0$ and can be recursively defined as the convolution of the B-spline of degree $(n-1)$ with β^0

$$\begin{aligned}\beta^n(x) &= (\beta^{n-1} * \beta^0)(x)\\ &= \int_{-\infty}^{\infty} \beta^{n-1}(x)\beta^0(x-t)\,dt, \qquad n > 0\end{aligned} \tag{13}$$

where β^0 is a unit square pulse

$$\beta^0(x) = \tfrac{1}{2}\left(\mathrm{sign}\left(x+\tfrac{1}{2}\right)\mathrm{sign}\left(x-\tfrac{1}{2}\right)\right) \tag{14}$$

and where the sign function is defined by

$$\mathrm{sign}(x) = \begin{cases} -1, & x < 0\\ 0, & x = 0\\ 1, & x > 0.\end{cases} \tag{15}$$

Not only will these B-splines be used as Parzen window, but they will also provide the basis functions for representing continuous images given by a set of samples.

III. Multiresolution

A. Image Model

Let us assume that an image $f(\mathbf{x})$ is known from a set of samples $f_i = f(\mathbf{x}_i)$ that are regularly spaced on a Cartesian grid. To be useful, an image model must satisfy several constraints. First, it must allow one to interpolate an image, which links the samples f_i and their location $\mathbf{x}_i$ to the continuous function $f(\mathbf{x})$. This property is typically needed when performing the geometric transformation $f \to f(\mathbf{g}(\mathbf{x}_i))$. Second, given some continuous function $y(\mathbf{x})$, there must exist a procedure to recover a set of samples y_i at locations $\mathbf{x}_i$ such that the model based on this set would reconstruct a close approximation to $y(\mathbf{x})$. A typical application of this requirement arises when one computes a resolution pyramid, for in this case the procedure can be sketched by $(f_i,\, \mathbf{x}_i) \to f(\mathbf{x}) \to f(2\mathbf{x}) = y(\mathbf{x}) \to (y_i,\, \mathbf{x}_i)$.

We base our image model on the B-spline functions of degree n introduced in Section II-E. Specifically, we have that

$$f(\mathbf{x}) = \sum_{\mathbf{x}_i \in V} c(\mathbf{x}_i)\, \beta^n(\mathbf{x} - \mathbf{x}_i) \tag{16}$$

where $\beta(\mathbf{x})$ is a separable convolution kernel given by the product $\beta^n(x_1) \cdot \beta^n(x_2) \cdot \cdots$, and where the expansion B-spline coefficients $c_i = c(\mathbf{x}_i)$ are computed from the sample values f_i by recursive digital filtering [18]. This model is continuous, differentiable a.e. for $n \geq 0$, and differentiable for $n > 1$. It serves three purposes. First, its rescaled versions yield the image pyramid that we use for our multiresolution approach [19]. Second, it allows us to resample the transformed image $f(\mathbf{g}(\mathbf{x}_i))$. Finally, it is used in computing the image gradient needed during optimization.

B. Model Degree

The model degree determines the quality of the approach. The lowest-possible degree $n = 0$ is called nearest-neighbor. Used to compute the resolution pyramid, it results in aliasing. Used to compute $f(\mathbf{g}(\mathbf{x}))$, it results in blocking artifacts. Used to compute S, it results in a discontinuous criterion, which is hard to optimize. Also, the optimum is generally not uniquely defined. The next degree $n = 1$ corresponds to linear interpolation. It results in less aliasing, and oversmoothing substitutes for blocking. Meanwhile, the criterion is better-behaved. In these two cases, the computation of the B-spline coefficients c is trivial. For higher degrees, this computation is slightly more involved, but aliasing is reduced substantially. Blocking and smoothing are gradually replaced by ringing. At the extreme, when $n \to \infty$, aliasing disappears altogether but ringing is strongly present (sinc, or Shannon interpolation [20]). A good compromise between all these issues is to select a cubic B-spline β^3 as model kernel.

There are three major reasons why the choice of a high-quality model is essential to the proper behavior of a multiresolution registration method. First, consider performing optimization at a coarse level of the pyramid. The steps made by the optimizer at this level correspond to big strides at the finest level. It follows that precision is of utmost importance at this coarse level, and subpixel interpolation must be faithful. This calls for a degree n that is higher than what is traditionally selected. Second, consider having found the optimal parameter $\hat{\boldsymbol{\mu}}$ at some level l. The optimal parameter at the next finer level $l+1$ is not identical because data are more detailed, and the added details call for some corrective action. It is however desired that the corrections be as small as possible, which is achieved by minimizing the amount of detail distinguishing level l from level $l+1$. Thus, it is best to limit the aliasing inherent in the size-reduction operation, which again calls for a high model degree n. We shall see in Section IV that our optimizer requires a differentiable kernel. We prefer to avoid having to sample a derivative where it is discontinuous, which could sometimes arise with linear interpolation. This is one more reason to select a high model degree n.

C. Grey Cone

One of the benefits of a multiresolution strategy is the reduction of the amount of data when the resolution is coarse. However, this data simplification is detrimental to the robustness of the estimation of the joint probability $p(\iota, \kappa)$ from which mutual information is computed. A compromise must be found between, on one hand, too few data and too many grey-levels in (L_T, L_R), and on the other hand, an abundance of data but a coarse grey-level quantization. For these reasons, we feel it is appropriate to extend the concept of geometric multiresolution (image pyramid) to the concept of grey-level multiresolution (grey cone). A reasonable approach is to keep, at any resolution level, a constant ratio between the amount of available data at that particular resolution, and the number of entries in the discrete joint probability p. One consequence to keep in mind is that not only do we change data when we switch from one level to the next, but we also change the criterion itself, since we now let L_T and L_R depend on the actual resolution level. However, we still expect that the true optimal alignment parameters $(\hat{\mu}_1, \hat{\mu}_2, \cdots)$ will vary only slightly between levels. The sets (L_T, L_R) are constructed by regular sampling of the grey-level range of (f_T, f_R).

D. Underlying Assumptions

To use multiresolution with some success in performing the registration of two datasets, it is necessary that their coarse representation be discriminant enough. In an imaging context, it is well-known that most of the signal energy is concentrated toward low spatial frequencies, which are essentially preserved while switching from a fine resolution to a coarser one. Therefore, it is legitimate to first register the large-scale features at coarse resolution, and then only to refine registration by taking fine-scale features into account.

In addition to multiresolution, the pseudo-quantization introduced by the grey cone results in a registration criterion that tends to be dominated by high-contrast features at coarse scales. However, this trend can be partially compensated so as to restore some sensitivity to low-contrast features. Suppose for a while that the discrete set of intensities L has been so badly chosen that all intensities would fall in the same bin, should a traditional histogram be constructed: in this case apparently, no information can be used for registration. By contrast, the histogram construct proposed in (2), together with a cubic spline in the role of the Parzen window, spreads data contributions over several bins, which allows the recovery of more information than with traditional binning. In fact, the cubic spline β^3 satisfies not only the partition of unity $\sum_{\xi \in \mathbb{Z}} \beta^3(\xi - x) = 1$, but also additional relations related to the Strang–Fix theory of approximation [21]. Those are as follows:

$$\sum_{\xi \in \mathbb{Z}} \xi\, \beta^3(\xi - x) = x \tag{17}$$

$$\sum_{\xi \in \mathbb{Z}} \xi^2\, \beta^3(\xi - x) = \tfrac{1}{3} + x^2 \tag{18}$$

$$\sum_{\xi \in \mathbb{Z}} \xi^3\, \beta^3(\xi - x) = x^3 + x. \tag{19}$$

It is trivial to derive from these relations the fact that the empirical average computed over the bins $\mu = \varepsilon \cdot \sum_{\xi \in L} \xi\, p(\xi)$ is identical to the empirical average computed over the data $\mu = (1/\mathrm{card}(V)) \cdot \sum_{\mathbf{x}_i \in V} f(\mathbf{x}_i)$, no matter how bad the choice of the set L is; in particular, this shows that the presence of the grey cone has no influence on the data average. Moreover, it can also be shown that the L-based variance is simply a biased version of the empirical one.[1] More precisely, if the former is $\sigma_L^2 = \varepsilon \cdot \sum_{\xi \in L} (\xi - \mu)^2\, p(\xi)$ and if the latter is $\sigma_V^2 = (1/\mathrm{card}(V)) \cdot \sum_{\mathbf{x}_i \in V} (f(\mathbf{x}_i) - \mu)^2$, then we have that $\sigma_V^2 = \sigma_L^2 - \varepsilon^2/3$. Furthermore, the third moments are exactly equal, whether computed over the bins or over the data

$$\varepsilon \cdot \sum_{\xi \in L} (\xi - \mu)^3\, p(\xi) = (1/\mathrm{card}(V)) \cdot \sum_{\mathbf{x}_i \in V} (f(\mathbf{x}_i) - \mu)^3.$$

It should be clear from these relations of statistical equivalence up to third order—disregarding the constant bias in the case of the variance—that reducing the number of levels while working within the coarse region of the grey cone is not as detrimental as it may seem at first.

IV. Optimization

In addition to optimizing at the coarse levels, the multiresolution strategy does not preclude optimization at the finest one. For this strategy to be efficient in terms of computation time, it is required that, at the finest level, the number of criterion evaluations necessary to reach some registration precision be less than the number needed to solve the same problem without a multiresolution strategy. From this consideration, it follows that it is important to select an optimizer that benefits strongly from good starting conditions. As examples of bad candidates, one can think of many direction-set methods (e.g., conjugate-gradient with or without explicit derivatives), where the optimizer often needs to sequentially explore several directions in the $(\mu_1, \mu_2, \cdots)$ space, before even starting to really optimize. With such algorithms, especially when the conditions are nearly optimal, many criterion evaluations are wasted simply to assess that these conditions are, well, nearly optimal.

The main contribution of this paper is an optimization algorithm based on the same strategy as that of the Marquardt–Levenberg optimizer which is characterized by a global "understanding" of its immediate surroundings [14]. It benefits from superlinear convergence, a regime in which the optimizer converges quadratically (or better) when the optimum is close enough. An important difference between the present optimizer

[1] We consider here that $\mathrm{card}(V)$ is sufficiently large so as to have $1/(\mathrm{card}(V) - 1) \cong 1/\mathrm{card}(V)$.

and Marquardt–Levenberg's is that our specific registration problem is non least-squares.

Our optimizer is iterative; it proceeds by trying potentially better solutions around a given initial condition. Apart from the propagation of the final solution from a coarser level to the next finer level, where it will be used as initial condition, the existence of an underlying image pyramid and of a grey cone is ignored while optimizing within any given level. Hence, we present this algorithm out of the multiresolution context.

A. Criterion Model

As a first step, let us express the mutual information (7) by a Taylor expansion

$$\begin{aligned} S(\boldsymbol{\mu}) = {} & S(\boldsymbol{\nu}) + \sum_i \frac{\partial S(\boldsymbol{\nu})}{\partial \mu_i}\,(\mu_i - \nu_i) \\ & + \frac{1}{2} \sum_{i,\,j} \frac{\partial^2 S(\boldsymbol{\nu})}{\partial \mu_i}\,\partial \mu_j\,(\mu_i - \nu_i)\,(\mu_j - \nu_j) + \cdots. \end{aligned} \tag{20}$$

We then simplify (20) by ignoring all terms above second-order. Thus, the residual error will decay like $\|\boldsymbol{\mu} - \boldsymbol{\nu}\|^3$, provided $\partial^3 S(\boldsymbol{\nu})/(\partial \mu_i\,\partial \mu_j\,\partial \mu_k)$ is bounded. This happens in particular when the Parzen window is a B-spline of degree $m \geq 3$. If both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are not too far from the optimum, this simplified quadratic model is known to be quite appropriate.

B. Gradient

Let us define the gradient ∇S as

$$\nabla S = \left[\frac{\partial S}{\partial \mu_1}, \frac{\partial S}{\partial \mu_2}, \cdots\right]. \tag{21}$$

In general, a component of ∇S is given by

$$\begin{aligned} \frac{\partial S}{\partial \mu} = {} & -\alpha^2(\boldsymbol{\mu}) \sum_{\iota \in L_T} \sum_{\kappa \in L_R} \\ & \cdot \left[\frac{1}{\alpha(\boldsymbol{\mu})}\,\frac{\partial h(\iota,\,\kappa;\,\boldsymbol{\mu})}{\partial \mu}\,\log_2\left(\frac{e\,h(\iota,\,\kappa;\,\boldsymbol{\mu})}{\alpha(\boldsymbol{\mu})\,h_T(\iota;\,\boldsymbol{\mu})\,h_R(\kappa;\,\boldsymbol{\mu})}\right)\right. \\ & + h(\iota,\,\kappa;\,\boldsymbol{\mu})\,\log_2\left(\frac{e\,\alpha(\boldsymbol{\mu})\,h_T(\iota;\,\boldsymbol{\mu})\,h_R(\kappa;\,\boldsymbol{\mu})}{h(\iota,\,\kappa;\,\boldsymbol{\mu})}\right) \\ & \cdot \sum_{\zeta \in L_T} \sum_{\eta \in L_R} \frac{\partial h(\zeta,\,\eta;\,\boldsymbol{\mu})}{\partial \mu} - \frac{1}{\log_e(2)} \\ & \cdot \left(\frac{\partial h_T(\iota;\,\boldsymbol{\mu})}{\partial \mu}\,h_R(\kappa;\,\boldsymbol{\mu}) + h_T(\iota;\,\boldsymbol{\mu})\,\frac{\partial h_R(\kappa;\,\boldsymbol{\mu})}{\partial \mu}\right) \\ & \left. \cdot \frac{h(\iota,\,\kappa;\,\boldsymbol{\mu})}{\alpha(\boldsymbol{\mu})\,h_T(\iota;\,\boldsymbol{\mu})\,h_R(\kappa;\,\boldsymbol{\mu})}\right] \end{aligned} \tag{22}$$

where e is the exponential constant. This expression can be simplified because, in our formulation of the mutual-information criterion, $h_R = \sum h(\iota,\,\kappa;\,\boldsymbol{\mu})$ and $\alpha = \sum \sum h(\iota,\,\kappa;\,\boldsymbol{\mu})$ do not dependent on $\boldsymbol{\mu}$. By selecting a B-spline of degree m as a Parzen window satisfying the partition of unity condition, we get that

$$\frac{\partial S}{\partial \mu} = -\sum_{\iota \in L_T} \sum_{\kappa \in L_R} \frac{\partial p(\iota,\,\kappa;\,\boldsymbol{\mu})}{\partial \mu}\,\log_2\left(\frac{p(\iota,\,\kappa;\,\boldsymbol{\mu})}{p_T(\iota;\,\boldsymbol{\mu})}\right). \tag{23}$$

The derivation from (22) to (23) is detailed in the Appendix. Then, we can expand the gradient of the joint probability distribution

$$\begin{aligned} & \frac{\partial p(\iota,\,\kappa;\,\boldsymbol{\mu})}{\partial \mu} \\ & \quad = \frac{1}{\operatorname{card}(V)} \sum_{\mathbf{x}_i \in V} \\ & \quad \cdot \left[\beta^m(\kappa/\varepsilon_R - f_R(\mathbf{x}_i)/\varepsilon_R)\,\left.\frac{\partial \beta^m(\xi)}{\partial \xi}\right|_{\xi = \iota/\varepsilon_T - f_T(\mathbf{g}(\mathbf{x}_i;\,\boldsymbol{\mu}))/\varepsilon_T}\right. \\ & \qquad \left. \cdot \frac{1}{\varepsilon_T}\left(\left.\frac{-df_T(\mathbf{t})}{d\mathbf{t}}\right|_{\mathbf{t} = \mathbf{g}(\mathbf{x}_i;\,\boldsymbol{\mu})}\right)^T \frac{\partial \mathbf{g}(\mathbf{x}_i;\,\boldsymbol{\mu})}{\partial \mu}\right] \end{aligned} \tag{24}$$

where it is possible to introduce the explicit expression for the derivative of a B-spline derived from (13)

$$\frac{\partial \beta^m(\xi)}{\partial \xi} = \beta^{m-1}(\xi + 1/2) - \beta^{m-1}(\xi - 1/2) \tag{25}$$

and where the spatial gradient of an image $df(\mathbf{t})/d\mathbf{t}$ is given by the B-spline model of degree n

$$\begin{aligned} & \frac{df_T(\mathbf{t})}{d\mathbf{t}} \\ & \quad = \sum_{\mathbf{x}_i \in V} c(\mathbf{x}_i)\,\left.\frac{d\beta^n(\mathbf{u})}{d\mathbf{u}}\right|_{\mathbf{u} = \mathbf{t} - \mathbf{x}_i} \\ & \quad = \sum_{\mathbf{x}_i \in V} c(\mathbf{x}_i) \begin{bmatrix} \left.\dfrac{\partial \beta^n(u)}{\partial u}\right|_{u = (\mathbf{t} - \mathbf{x}_i)_1} \beta^n((\mathbf{t} - \mathbf{x}_i)_2) \cdots \\ \beta^n((\mathbf{t} - \mathbf{x}_i)_1)\,\left.\dfrac{\partial \beta^n(u)}{\partial u}\right|_{u = (\mathbf{t} - \mathbf{x}_i)_2} \cdots \\ \vdots \end{bmatrix}. \end{aligned} \tag{26}$$

The last unexplained term in (24) is $\partial \mathbf{g}(\mathbf{x}_i;\,\boldsymbol{\mu})/\partial \mu$, which describes the variation in position due to a variation in parameter. This term depends on geometry alone. Finally, the gradient of the marginal joint density can be expressed by

$$\begin{aligned} \frac{\partial p_T(\iota)}{\partial \mu} & = \sum_{\kappa \in L_R} \frac{\partial p(\iota,\,\kappa)}{\partial \mu} \\ & = \frac{1}{\operatorname{card}(V)} \sum_{\mathbf{x}_i \in V} \left.\frac{\partial \beta^m(\xi)}{\partial \xi}\right|_{\xi = \iota/\varepsilon_T - f_T(\mathbf{g}(\mathbf{x}_i;\,\boldsymbol{\mu}))/\varepsilon_T} \\ & \quad \cdot \frac{1}{\varepsilon_T}\left(\left.\frac{-df_T(\mathbf{t})}{d\mathbf{t}}\right|_{\mathbf{t} = \mathbf{g}(\mathbf{x}_i;\boldsymbol{\mu})}\right)^T \frac{\partial \mathbf{g}(\mathbf{x}_i;\,\boldsymbol{\mu})}{\partial \mu}. \end{aligned} \tag{27}$$

C. Hessian

Let us define the matrix of the second derivative of S as its Hessian $\nabla^2 S$

$$\nabla^2 S = \begin{bmatrix} \dfrac{\partial^2 S}{\partial \mu_1 \, \partial \mu_1} & \dfrac{\partial^2 S}{\partial \mu_1 \, \partial \mu_2} & \cdots \\ \dfrac{\partial^2 S}{\partial \mu_2 \, \partial \mu_1} & \dfrac{\partial^2 S}{\partial \mu_2 \, \partial \mu_2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}. \tag{28}$$

With the same assumptions as before, including a Parzen window satisfying the partition of unity condition, we determine a component of the Hessian by

$$\begin{aligned} &\frac{\partial^2 S}{\partial \mu_1 \, \partial \mu_2} \\ &\quad = -\left(\sum_{\iota \in L_T} \sum_{\kappa \in L_R} \frac{\partial^2 p(\iota, \kappa)}{\partial \mu_1 \, \partial \mu_2} \log_2 \left(\frac{p(\iota, \kappa)}{p_T(\iota)} \right) \right) \\ &\qquad + \frac{1}{\log_e(2)} \left(\sum_{\iota \in L_T} \frac{\partial p_T(\iota)}{\partial \mu_1} \frac{\partial p_T(\iota)}{\partial \mu_2} \frac{1}{p_T(\iota)} \right) \\ &\qquad - \frac{1}{\log_e(2)} \left(\sum_{\iota \in L_T} \sum_{\kappa \in L_R} \frac{\partial p(\iota, \kappa)}{\partial \mu_1} \frac{\partial p(\iota, \kappa)}{\partial \mu_2} \frac{1}{p(\iota, \kappa)} \right). \end{aligned} \tag{29}$$

The first term of (29) depends on the second-order variation of the joint probability when a pair of registration parameters varies jointly. We will ignore this term, which amounts to linearizing the variation of p with respect to $\boldsymbol{\mu}$. Another motivation for dropping the first term in (29) arises when one considers the situation at ideal registration of two dependent images. In this case, we have that $p(\iota, \kappa) = p_T(\iota) \, p_R(\kappa)$. Then, the partition of unity condition implies

$$\begin{aligned} &\sum_{\iota \in L_T} \sum_{\kappa \in L_R} \frac{\partial^2 p(\iota, \kappa)}{\partial \mu_1 \, \partial \mu_2} \log_2 \left(\frac{p(\iota, \kappa)}{p_T(\iota)} \right) \\ &\quad = \sum_{\kappa \in L_R} \frac{\partial^2 p_R(\kappa)}{\partial \mu_1 \, \partial \mu_2} \log_2 (p_R(\kappa)) = 0 \end{aligned} \tag{30}$$

and the first term in (29) vanishes. We note that the remaining terms do still contribute and that the Hessian does not globally vanish at ideal registration. This is important to keep superlinear convergence near the optimum. Finally, in this paper we use the following simplified form

$$\begin{aligned} &\frac{\partial^2 S}{\partial \mu_1 \, \partial \mu_2} \\ &\quad \approx \frac{1}{\log_e(2)} \left(\sum_{\iota \in L_T} \frac{\partial p_T(\iota)}{\partial \mu_1} \frac{\partial p_T(\iota)}{\partial \mu_2} \frac{1}{p_T(\iota)} \right) \\ &\qquad - \frac{1}{\log_e(2)} \left(\sum_{\iota \in L_T} \sum_{\kappa \in L_R} \frac{\partial p(\iota, \kappa)}{\partial \mu_1} \frac{\partial p(\iota, \kappa)}{\partial \mu_2} \frac{1}{p(\iota, \kappa)} \right). \end{aligned} \tag{31}$$

Comparing this last expression with (24) and (28), one sees that every term needed by our simplified Hessian has been already precomputed while determining the value of the gradient. Thus, another fortunate consequence of ignoring the second-order term in (29) is that the Hessian $\nabla^2 S$ comes at essentially no additional computational cost with respect to that of the gradient $\boldsymbol{\nabla} S$.

D. Standard Optimizers

The steepest-gradient descent is a minimization algorithm that can be succinctly described by

$$\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} - \Gamma \, \boldsymbol{\nabla} S \left(\boldsymbol{\mu}^{(k)} \right). \tag{32}$$

Its local convergence is guaranteed, although it may be very slow. A key problem is the determination of the appropriate scaling diagonal matrix Γ.

The Newton method can be described by

$$\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} - \left(\nabla^2 S \left(\boldsymbol{\mu}^{(k)} \right) \right)^{-1} \boldsymbol{\nabla} S \left(\boldsymbol{\mu}^{(k)} \right). \tag{33}$$

Its convergence to an optimum is not guaranteed: it may converge to a saddle point (at the same time a maximum for some parameter μ_1 and a minimum for another parameter μ_2). Even worse, it diverges from the desired solution when the problem is not convex. In return, it is extremely efficient when the criterion is locally quadratic convex, for in this case it finds the optimum after a single criterion evaluation.

E. Marquardt–Levenberg Strategy

The Marquardt–Levenberg strategy is a convenient way to combine the advantages of the gradient method with those of the Newton method, preserving the efficiency of the latter when the conditions are nearly optimal, and the robustness of the former when they are not.

Let us introduce a modified Hessian $\mathcal{H}S$ in which we retain the off-diagonal entries of $\nabla^2 S$ and multiply its diagonal entries by some factor

$$[\mathcal{H}S(\boldsymbol{\mu})]_{i,\,j} = [\nabla^2 S(\boldsymbol{\mu})]_{i,\,j} \, (1 + \delta_{i,\,j} \, \lambda) \tag{34}$$

where $\delta_{i,\,j}$ is the Kronecker symbol, and where λ is a tuning factor that represents the compromise between the gradient method and the Newton method. Suppose we now determine the new update $\boldsymbol{\mu}^{(k+1)}$ as in

$$\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} - \left(\mathcal{H}S \left(\boldsymbol{\mu}^{(k)} \right) \right)^{-1} \boldsymbol{\nabla} S \left(\boldsymbol{\mu}^{(k)} \right). \tag{35}$$

Depending on the value of λ, one can distinguish two extreme cases. When $\lambda \to 0$, one sees that (35) and (33) are identical. When $\lambda \to +\infty$, the diagonal terms of the modified Hessian $\mathcal{H}S$ dominate, and we are in the situation of (32). Note that, although the magnitude of the update is adapted to each component by the virtue of the normalizing term $[\nabla^2 S]^{-1}_{i,\,i}$, the steps are vanishingly small in this second case. This is not a problem because it is easy to adapt λ between these two extremes in order to achieve a good compromise between the efficiency (but lack of robustness) of the Newton approach, and the size of the steps of the robust (but generally inefficient) gradient approach. In this paper, the mechanism to adapt λ is identical to the original Marquardt–Levenberg proposition.

V. Experiments

We want first to illustrate the performance of our approach with experiments in which the true alignment is known *a priori*, and in which the grey-level correspondence between the modalities is controlled. This will allow for an objective measurement of the quality of our algorithm. Then, we propose a case where the true alignment is approximately known, but where the grey-level correspondence between the modalities is not. Finally, we present results in which the true alignment has been estimated by another registration technique that is believed to be very accurate, but that is generally unavailable because it requires planning before data acquisition. This last validation approach has been used as a benchmark by several other researchers [2], and is representative of a typical application for this algorithm.

A. Warping Index

The general procedure for validating our algorithm will be to start with two images that are supposed to be in perfect registration. We then destroy this correspondence by applying a known geometric transformation. The goal is to recover its inverse by our registration method.

Consider a test image f_T and a reference image f_R that are already in perfect correspondence. Rather than transforming a single image in this pair, we prefer to transform both because this tends to lessen any bias that would otherwise result from the introduction of interpolation artifacts into one image only. Therefore, we compute

$$g_R(\mathbf{x}) = f_R(\mathbf{g}_0(\mathbf{x})), \quad g_T(\mathbf{x}) = f_T(\mathbf{g}_0^{-1}(\mathbf{x})) \qquad (36)$$

where $\mathbf{g}_0$ is a rigid-body transformation consisting of a random translation and a random rotation around the center of the image. It follows that the correct registration of g_T to g_R involves the transformation $\mathbf{g} = \mathbf{g}_0 \circ \mathbf{g}_0$ such that $g_R(\mathbf{x}) = g_T(\mathbf{g}(\mathbf{x}))$. In our case, since $\mathbf{g}_0$ is rigid-body, so is $\mathbf{g}$.

Next, we estimate a rigid-body transformation $\tilde{\mathbf{g}}$ out of the data (g_T, g_R). Our aim is now to determine the precision of each estimation. We achieve this goal by introducing a warping index ϖ that measures an average geometric error ϖ

$$\varpi = \frac{1}{\operatorname{card}(V)} \sum_{\mathbf{x}_i \in V} \left\| \mathbf{g}^{-1}(\mathbf{x}_i) - \tilde{\mathbf{g}}^{-1}(\mathbf{x}_i) \right\| \qquad (37)$$

where $\|\cdot\|$ stands for the Euclidean distance. After having performed several registrations with different realizations of the random transformation $\mathbf{g}_0$, we average together the values ϖ and report a pooled warping index. For this paper, there are 100 warping indexes to pool for each experiment. Meanwhile, $\mathbf{g}_0$ has a translation that is uniformly distributed in $[-2.5, 2.5]$, and a rotation around the center of the image that is uniformly distributed in $[-\pi/36, \pi/36]$. Hence, the maximal excursion of $\mathbf{g} = \mathbf{g}_0 \circ \mathbf{g}_0$ is about seven pixels of translation and 10° of rotation.

B. Objective Validation

We start our series of experiments in a controlled environment where we know *a priori* both the geometry and the grey-level correspondence between the test and the reference image. First, we attempt to register identical images and show that our registration algorithm performs well in this intramodal case. Then, we simulate the intermodal case by performing a nonlinear, nonmonotonic transformation on the grey-levels of one of the image, and show that our registration algorithm performs as well as in the intramodal case. Finally, we add white Gaussian noise to the simulated intermodal images and show that our registration algorithm is robust with respect to this type of noise. We perform these experiments in 2-D with widely available images. Although unrealistic, this controlled environment offers a clear framework for the interpretation of the results.

1) Intramodal Case: Selecting the 512×512 Lena image in the role of both f_R and f_T, we expect a multiresolution approach to influence at least two aspects of the registration method. First, it should improve the robustness of nonstochastic optimization procedures such as ours. Second, it should improve its speed. To observe these effects, we compare the success of our registration method when we vary the number of levels in the image pyramid. Since the goal is to investigate the interaction between robustness and precision, we prescribe a fixed overall computation time and adapt the number of criterion evaluations at each level such that this resource is shared between levels in an adequate fashion.

Table I presents the results of these experiments where the first block of lines corresponds to a strategy using a one-level "pyramid," and the last block of lines to a six-level pyramid (the coarsest level is a 16 × 16 image). The geometric unit of the warping index for the intramodal case ϖ_{intra} is 1.0 pixel at the finest resolution, and the time unit is 1.0 CPU second on a Sun ULTRA 30 workstation. The processing time reported in this table includes the overhead time necessary for computing the pyramid (starting each time from the finest level). For example, performing 64 criterion evaluations on an image down-sized from $\operatorname{card}(V) = 512 \times 512$ to $\operatorname{card}(V) = 32 \times 32$ requires 3.9 s, while performing twice as many criterion evaluations with the same overhead requires only 5.9 s. We also give $\operatorname{card}(L) = \operatorname{card}(L_T) = \operatorname{card}(L_R)$ the number of quantized intensities at each level of the grey cone. To determine this number of grey-levels, we observe the rule $\operatorname{card}(L_R)\operatorname{card}(L_T)/\operatorname{card}(V) = R$ that we proposed at Section III-C, and we select a constant ratio $R = 1/8$.

The purpose of this experiment is to show the degree to which multiresolution is able to ameliorate the performance. We observe that our algorithm is essentially unable to converge within the allotted computation time when the pyramid consists of its finest level only. This was to be expected, since we spent most of our attention to building an optimizer that works well when it is close to the solution, without concerning ourselves with its behavior when the solution is remote. With two levels, the accuracy improves but is still insufficient. In the three-level case, the order of magnitude of the accuracy reached by our algorithm is about a pixel. With four levels, this accuracy still improves to about a couple hundredth of a pixel; it reaches a hundredth of a pixel for five levels. An additional sixth level does not bring additional gains. It is also interesting to note that, although half of the processing time is spent at the finest level, the biggest improvements result from computations performed at coarser

TABLE I
INFLUENCE OF MULTIRESOLUTION ON THE ROBUSTNESS OF REGISTRATION IN AN IDEAL CASE. ϖ_{intra}: ORIGINAL LENA VERSUS ORIGINAL LENA. ϖ_{inter}: MODIFIED LENA VERSUS ORIGINAL LENA

	Initial	Coarse	$\cdots$	$\cdots$	$\cdots$	$\cdots$	Fine
ϖ_{intra}	12.7 ± 6.5						12.1 ± 6.6
ϖ_{inter}	12.7 ± 6.5						12.5 ± 6.6
Time/Iter.							65.3/8
Total Time							65.3
$\mathrm{card}(L)/\sqrt{\mathrm{card}(V)}$							64/512
ϖ_{intra}	12.7 ± 6.5					9.4 ± 6.6	9.1 ± 6.7
ϖ_{inter}	12.7 ± 6.5					11.3 ± 6.9	11.2 ± 6.9
Time/Iter.						33.1/16	31.7/4
Total Time						33.1	64.8
$\mathrm{card}(L)/\sqrt{\mathrm{card}(V)}$						32/256	64/512
ϖ_{intra}	12.7 ± 6.5				2.5 ± 2.9	1.5 ± 2.4	1.4 ± 2.3
ϖ_{inter}	12.7 ± 6.5				6.4 ± 5.9	5.8 ± 5.9	5.7 ± 5.9
Time/Iter.					17.5/32	16.4/8	31.7/4
Total Time					17.5	33.9	65.6
$\mathrm{card}(L)/\sqrt{\mathrm{card}(V)}$					16/128	32/256	64/512
ϖ_{intra}	12.7 ± 6.5			0.58 ± 0.73	0.079 ± 0.10	0.022 ± 0.028	0.018 ± 0.019
ϖ_{inter}	12.7 ± 6.5			0.72 ± 1.2	0.096 ± 0.23	0.024 ± 0.055	0.013 ± 0.022
Time/Iter.				9.7/64	9.5/16	16.4/8	31.7/4
Total Time				9.7	19.2	35.6	67.3
$\mathrm{card}(L)/\sqrt{\mathrm{card}(V)}$				8/64	16/128	32/256	64/512
ϖ_{intra}	12.7 ± 6.5		1.2 ± 1.5	0.16 ± 0.36	0.028 ± 0.039	0.0095 ± 0.014	0.0094 ± 0.013
ϖ_{inter}	12.7 ± 6.5		1.7 ± 0.86	0.19 ± 0.083	0.033 ± 0.018	0.013 ± 0.0061	0.0090 ± 0.0045
Time/Iter.			5.9/128	5.8/32	9.5/16	16.4/8	31.7/4
Total Time			5.9	11.7	21.2	37.6	69.3
$\mathrm{card}(L)/\sqrt{\mathrm{card}(V)}$			4/32	8/64	16/128	32/256	64/512
ϖ_{intra}	12.7 ± 6.5	12.2 ± 6.5	2.6 ± 2.5	0.36 ± 0.40	0.051 ± 0.088	0.017 ± 0.027	0.014 ± 0.019
ϖ_{inter}	12.7 ± 6.5	12.6 ± 6.5	3.5 ± 3.0	0.33 ± 0.48	0.041 ± 0.033	0.013 ± 0.0062	0.0090 ± 0.0045
Time/Iter.		4.0/256	3.9/64	5.8/32	9.5/16	16.4/8	31.7/4
Total Time		4.0	7.9	13.7	23.2	39.6	71.3
$\mathrm{card}(L)/\sqrt{\mathrm{card}(V)}$		2/16	4/32	8/64	16/128	32/256	64/512

levels; the improvements at fine levels are however the hardest to obtain, and our optimizer excels at getting them. As planned, the overall computation time is about the same in all cases.

2) Intermodal Ideal Case: Now, we keep the same image f_R as before, and we synthesize f_T by applying a grey-level transformation that is nonlinear and nonmonotonic, thus possessing no inverse. By this operation we try to confuse the algorithm. The synthesized test image is

$$f_T = \frac{1 - \cos(2\,\pi\, f_R)}{2} \tag{38}$$

where the range of intensities found in f_R is assumed to be $[0,\, 1]$. We performed the same experiments as in the intramodal case. Table I shows the results where the warping index for the intermodal case is given by ϖ_{inter}. We observe that the optimization is trapped in local minima far from the true optimum for pyramids consisting of one, two and three levels. Using at least four levels solves this problem. We see that the performance of the algorithm is essentially the same for the intermodal ideal case as it was for the intramodal one.

3) Intermodal Noisy Case: Trying to confuse the algorithm even more, we now add independent realizations of white Gaussian noise both to the reference image f_R, and to the test image f_T synthesized according to (38). We measure the amount of added noise as a signal-to-noise ratio (SNR) r expressed in decibels, according to

$$r = 10\,\log_{10}\,\frac{\sigma^2(f)}{\sigma^2(n)} \tag{39}$$

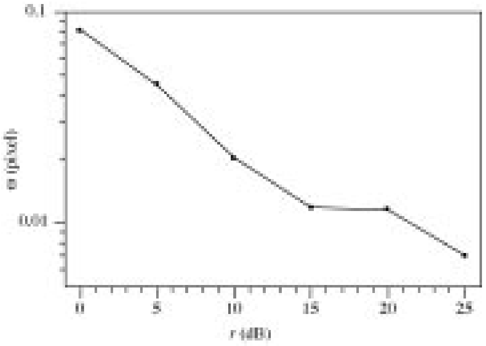


Fig. 2. Warping index ϖ versus amount of added noise (noisy modified Lena versus noisy original Lena).

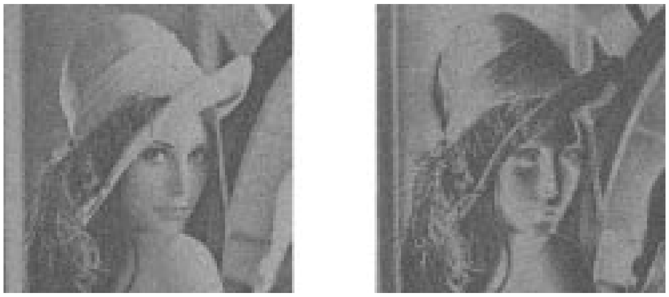


Fig. 3. Lena image corrupted with 0 dB noise. Left: reference. Right: modified histogram.

where $\sigma^2(f)$ is the variance of an image and $\sigma^2(n)$ is the variance of the added noise. We corrupted the test and reference image in such a way that they exhibit the same SNR. Fig. 2 shows the resulting dependence of the warping index ϖ on r, using the same methodology as in the previous experiments, and with a five-level pyramid. We observe that the algorithm is left undisturbed by moderate amounts of noise (SNR better than 15 dB). Moreover, the degradation is graceful when more noise is added. For example, even when the variance of the noise is as big as the variance of the signal itself (0 dB case), the warping index is still no more than a tenth of a pixel. Fig. 3 displays f_R and f_T in this particular 0 dB case. Note that the actual realization of the noise is different in all 100 experiments performed for each data point of Fig. 2.

We can now compare the performance of the present intermodal image registration algorithm with those of the intramodal one that we presented in [5]. We conclude that their precision is essentially the same in presence of strong noise (both reach a tenth of a pixel in the 0 dB case), while the intermodal algorithm applied to the intramodal noiseless case of Section V-B1 performs worse than the intramodal algorithm (a hundredth of a pixel instead of a thousandth of a pixel for a comparable allotted computation time). This relative loss of precision is largely compensated for by the fact that it is impossible to register images like those presented at three with the intramodal algorithm. Also, for many practical purposes, the precision of a hundredth of a pixel reached by the present intermodal algorithm is often sufficient.

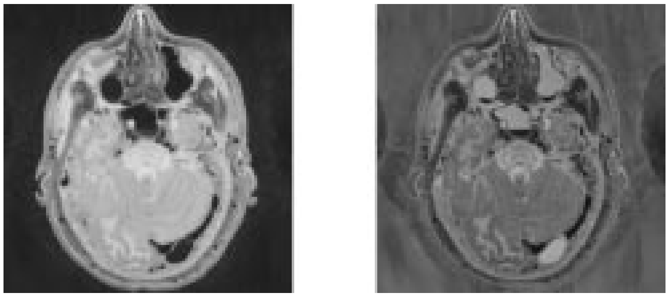


Fig. 4. Cryosection of a human brain in an RGB representation. Left: red channel. Right: blue channel.

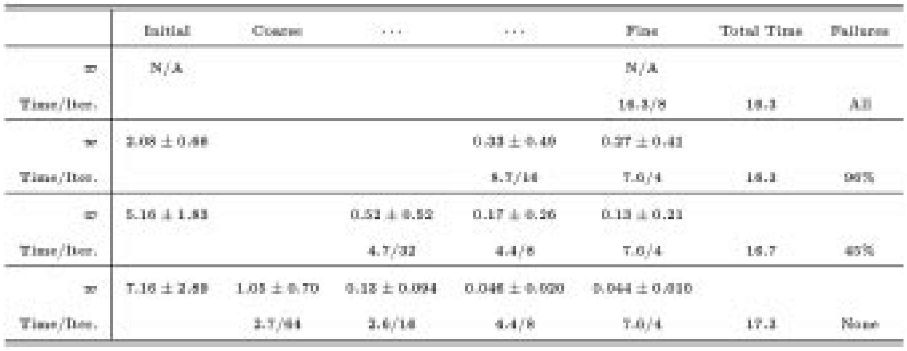
C. Known Geometry with Unknown Grey Correspondence

We now use a pair of biomedical 2-D images coming from different modalities. Fig. 4 shows such a pair, where the left image is the red channel of the cryosection of a human brain (Slice 4125 of the Visible Human Project), and the right image is its blue channel. Since they come from the same 24 bit color photograph, we can have some *a priori* confidence in their overall correct prior alignment. However, we may also have to mitigate this statement since inaccuracies in the scanner interfere with the level of geometric precision we are interested in. To alleviate this problem somewhat, we have reduced the image size threefold to 256×256, which tends to reduce any original mismatch in the color channels by as much. Contrary to the experiments of Section V-B, the correspondence between the intensities recorded into the red and the blue channel is unknown.

1) Multiresolution: Table II presents the results of the same experiments as before performed with the images of Fig. 4. Out of 100 trials, we retain in this table only those for which the quality of the registration is subpixel; we consider the rejected cases to be failures. We name capture range the largest pre-registration warping index that leads to a subpixel post-registration warping index. Trying again to assess the gain in robustness brought by multiresolution, we observe that our algorithm is unable to converge within the allotted computation time when the pyramid consists of its finest level only. With two levels, some cases are within the capture range but the number of failures is still very significant. Both accuracy and capture range improve with the introduction of a third level, where we still experience about as many failures as successes. With a fourth level however, the capture range is maximal and every of the 100 random transformations leads to a successful subpixel registration. The residual error, which might also be due to inaccuracies in the scanning device itself, reaches a half tenth of a pixel.

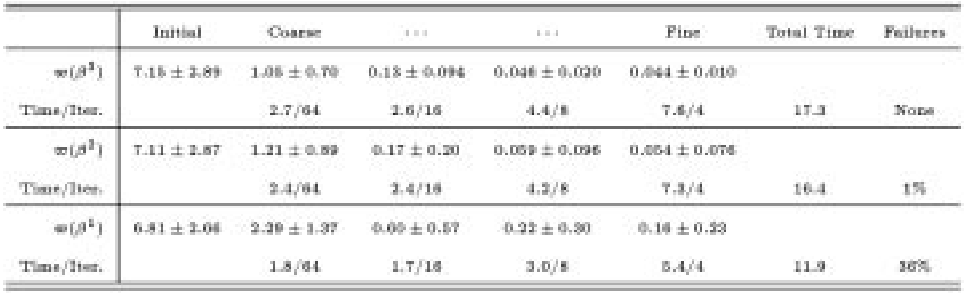
Since we retain in Table II only those experiments with initial conditions that lead to subpixel registration at full resolution, it is easy to read in the first data column of this table the maximal amount of initial misregistration our method can cope with. This maximal amount clearly depends on the number of levels of the pyramid; its general trend is to double for each additional level, which is consistent with the use of a dyadic multiresolution scheme. As a rule of thumb, we observe that our method works well as soon as the initial misregistration is subpixel at any given level; this subpixel score has then to be scaled from the actual spatial resolution to the final spatial resolution to indicate the true range of misregistration where our method is ef-

TABLE II
INFLUENCE OF MULTIRESOLUTION ON THE ROBUSTNESS OF REGISTRATION (CRYOSECTION BLUE CHANNEL VERSUS CRYOSECTION RED CHANNEL)



	Initial	Coarse	⋯	⋯	Fine	Total Time	Failures
ϖ	N/A				N/A		
Time/Iter.					16.3/9	16.3	All
ϖ	2.08 ± 0.66			0.33 ± 0.49	0.27 ± 0.41		
Time/Iter.				8.7/16	7.6/4	16.3	96%
ϖ	5.16 ± 1.83		0.52 ± 0.52	0.17 ± 0.26	0.13 ± 0.21		
Time/Iter.			4.7/32	4.4/8	7.6/4	16.7	45%
ϖ	7.16 ± 2.89	1.05 ± 0.70	0.13 ± 0.094	0.046 ± 0.020	0.044 ± 0.010		
Time/Iter.		2.7/64	2.6/16	4.4/8	7.6/4	17.3	None

TABLE III
INFLUENCE OF THE MODEL DEGREE ON THE ROBUSTNESS OF REGISTRATION (CRYOSECTION BLUE CHANNEL VERSUS CRYOSECTION RED CHANNEL)



	Initial	Coarse	⋯	⋯	Fine	Total Time	Failures
$\varpi(\beta^3)$	7.15 ± 2.89	1.05 ± 0.70	0.13 ± 0.094	0.046 ± 0.020	0.044 ± 0.010		
Time/Iter.		2.7/64	2.6/16	4.4/8	7.6/4	17.3	None
$\varpi(\beta^2)$	7.11 ± 2.87	1.21 ± 0.89	0.17 ± 0.20	0.059 ± 0.096	0.054 ± 0.076		
Time/Iter.		2.4/64	2.4/16	4.2/8	7.3/4	16.4	1%
$\varpi(\beta^1)$	6.81 ± 2.06	2.29 ± 1.37	0.60 ± 0.57	0.22 ± 0.30	0.16 ± 0.23		
Time/Iter.		1.8/64	1.7/16	3.0/8	5.4/4	11.9	36%

ficient. For example, we see in Table II that the capture range of a four-level multiresolution pyramid is about $2^{4-1} = 8$ pixels, as computed according to Definition (37).

2) Quality of the Model: We expect the quality of the image model to reflect itself in the quality of registration, particularly at the coarse levels of the pyramid. To investigate this hypothesis, we construct Table III, where we show the results of registration using cubic, quadratic and linear models, respectively. The number of levels and the number of criterion evaluations are identical in these three cases.

The quality of the model affects both interpolation and pyramid computation. One can see that the difference between a cubic and a quadratic model is not striking when dealing with the finer levels of the pyramid. For the coarser levels however, the difference is more marked. This tends to show that the main advantage of using a cubic model (with respect to the quadratic one) is not so much due to interpolation, but rather to reduced aliasing in the pyramid. Note that quadratic and cubic models have essentially the same computational cost, while a linear model is somewhat cheaper. The gain in speed is not dramatic however, and has to be weighed against a sharp reduction in accuracy. Moreover, since the algorithm sometimes failed to converge with a linear model, robustness is also decreased. For all these reasons we advocate the use of a cubic model.

3) Powell Optimizer: We want now to compare the accuracy and efficiency of our proposed optimizer to the Powell algorithm that has also been used in the context of image registration based on mutual information [15]. The goal of this presentation is to show the reason why a Powell algorithm fails to take full advantage of a multiresolution approach, while the optimizer proposed in this paper succeeds. The unfortunate corollary to this proposition is that our optimizer is inefficient out of a multiresolution context; in particular, at the first (coarsest) pyramid level, it is less robust and slower than many other optimizers. To benefit from both of best worlds, we suggest a compromise where a robust, but eventually evaluation-hungry optimizer, is used at the coarsest resolution, followed by the efficient use of our accurate, evaluation-savvy optimizer at finer levels. This suggestion is not further pursued here; rather, we concentrate on the performance of our optimizer alone.

The structure of this presentation is as follows: we first show a case where the traditional Powell optimizer yields good results (no multiresolution). Since this case does not correspond to the context in which our optimizer has been developed, we experience much worse performance. We then introduce multiresolution in a way that tends to be very favorable to Powell. We observe that this optimizer performs better than without multiresolution; at the same time, we observe that our algorithm yields good performances too. The important point comes last: while it is not possible to further enhance Powell, it is still possible to have a large increase in performance with our algorithm. In conclusion, we outperform the best (multiresolution) Powell result by a factor three with respect to time, without any compromise in accuracy.

The Powell algorithm computes no criterion derivatives while attempting to recover the gradient $\boldsymbol{\nabla} S$ and the Hessian $\nabla^2 S$, which makes it an attractive candidate when closed forms of the derivatives are not available or when their computation cost is prohibitive. It is known as a direction-set method, where the parameter space span($\boldsymbol{\mu}$) is explored along straight lines exclusively [linear combinations of $(\mu_1, \mu_2, \cdots)$].

Line minimizations require a bracketing of the minimum along the considered line before being able to start the opti-

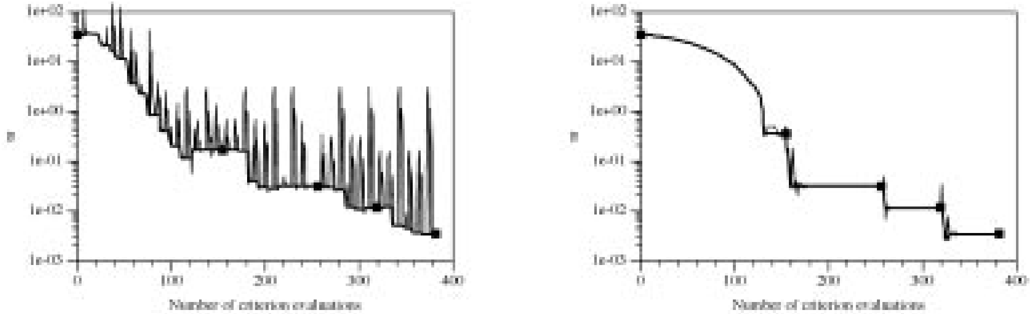


Fig. 5. Left: accuracy of the multiresolution Powell algorithm during optimization versus number of criterion evaluations. Right: accuracy of the proposed algorithm during optimization versus number of criterion evaluations. Thick line: best solution. Thin line: current attempt. Square dots: change of resolution level.

mization itself. This bracketing alone is worth several criterion estimations, which is inefficient when using a multiresolution strategy. Because these initial criterion estimations are necessary for the proper behavior of the Powell optimizer, regardless of whether the starting conditions are good or not, it is not possible to arbitrarily reduce their number. In addition, no convergence can be detected before at least as many line minimizations as free parameters have been performed. Obviously, the fact that the Powell algorithm uses estimates for the derivatives rather than their true values tends to further reduce its efficiency.

We conduct an experiment where the transformation $\mathbf{g}$ corresponds to an initial displacement of 10 pixels along each axis and to a rotation of $\pi/18 = 10°$. The image model is cubic for both algorithms (including Powell), and the joint histogram and the mutual information are computed according to (2) and (7), respectively. Thus, we expect to reach the same accuracy with either optimizer since the absolute optimum is defined only by the criterion and by the interpolation technique. The images are the same as in Section V-B2, with the same grey-cone strategy.

First, we attempt registration without multiresolution. We observe that the Powell algorithm needs 196 criterion evaluations to converge when its working conditions are identical to those found in [15], that is, at most 10 criterion evaluations for each line minimization, and convergence thresholds set to 10^{-5} for Powell and 10^{-3} for Brent minimization routine, respectively [22]. Allowing for the same number of criterion evaluations, the algorithm proposed in this paper is unable to converge at all, which demonstrates its lack of efficiency and robustness when it is far from the solution.

We then attempt registration in a multiresolution context. Using a four-level pyramid, we observe the evolution of the warping index ϖ during the course of registration, both for our algorithm and for Powell. Fig. 5 shows the result of this experiment, where Powell has been allowed to freely decide for convergence at each level, and where the number of criterion evaluations performed by our optimizer has been set equal to those observed while letting Powell converge. We can clearly see that both algorithms reach a satisfying solution ($\varpi = 0.003\,338$ in both cases), which demonstrates the gain in robustness brought by multiresolution. Thick square dots indicate the last result reached before a change of level. From the coarsest to the finest level, Powell claimed convergence after 155, 102, 62, and 62 criterion evaluations.

The bracketing episodes of the Powell algorithm can be easily identified as big excursions of ϖ. Those are necessary because this algorithm has no indication of the correct scale of the optimization problem and has to start with wild guesses each time a new direction is tried. The reward is a reduction in complexity, since no explicit derivative computations are performed. This translates in a reduction of the time per criterion evaluation. With a number of evaluations set to match those of the Powell algorithm, we need 716 s to perform the computations, while Powell is done in half the time (363 s). This last value has to be compared to the time needed by Powell to reach convergence without multiresolution (821 s).

We observe that both algorithms can be characterized by bursts of an efficient optimization mode alternating with more static periods. It is important to point out that, as soon as the initial conditions are good (about one pixel), our algorithm converges almost instantly when compared to Powell. In fact, the 62 criterion evaluations performed by the latter on the two finest resolution levels represent the smallest possible amount of computation, because Powell needs a first sweep through three parameters (with ten criterion evaluations each) to optimize for the added image details that distinguish a resolution level from the next, and one additional sweep to decide for convergence. By contrast, our algorithm is not constrained by line minimizations; it can stop at any time during optimization, and starts to simultaneously optimize for all parameters from the very first criterion evaluation on. This suggests that it can converge with a much reduced number of evaluations at each level, but for the first one.

We then propose an optimization strategy where the number of evaluations performed at finer levels of the multiresolution pyramid has been sharply reduced. Fig. 6 shows a case where 128, 32, 16 and 8 evaluations have been performed (so few evaluations make no sense in the context of a Powell optimizer, so we provide no direct comparison). Our accuracy is as good, or better than Powell (we reach $\varpi = 0.002\,591$). Moreover, since we remove many of those criterion evaluations that make for the longest computation time, we are able to reach convergence much earlier than Powell, both in terms of time and criterion evaluations. We need no more than 132 s to perform the whole optimization procedure, which is about the third of Powell in a multiresolution context, and about six times faster than the traditional Powell optimizer. Fig. 7 substantiates these results and show that the time spent at coarse resolution is essentially irrel-

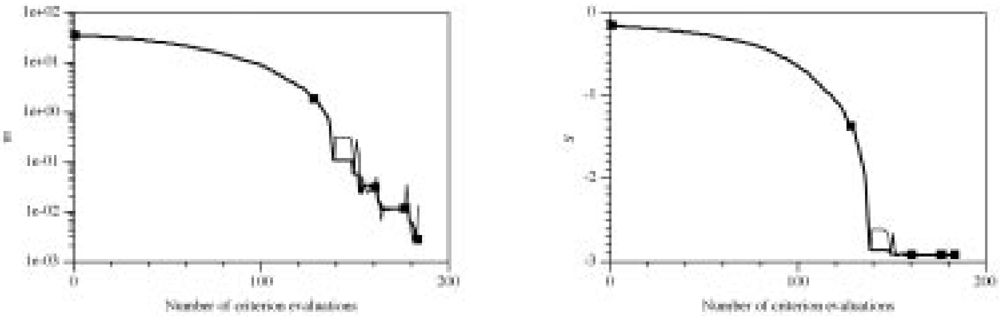


Fig. 6. Left: accuracy of the proposed algorithm during optimization versus number of criterion evaluations. Right: observed value of the criterion. The number of allowed criterion evaluations at each level is less than those demanded by the Powell optimizer. Thick line: best solution. Thin line: current attempt. Square dots: change of resolution level.

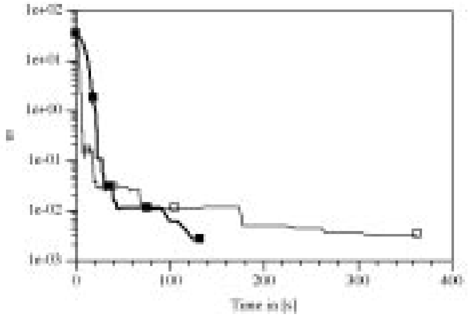


Fig. 7. Comparison of accuracies during optimization versus computation time. Thick line: proposed optimizer. Thin line: Powell optimizer. Square dots: change of resolution level.

TABLE IV
RESIDUAL DIFFERENCE IN mm BETWEEN A FIDUCIAL-MARKER PROSPECTIVE REGISTRATION TECHNIQUE AND THE PROPOSED RETROSPECTIVE ALGORITHM (VOLUMETRIC BRAIN DATA)

	CT-T1	CT-PD	CT-T2	CT-T1r	CT-PDr	CT-T2r
Median	1.5	1.7	1.1	0.9	1.0	0.9
Max	2.9	4.2	4.2	3.1	1.6	2.8
	PET-T1	PET-PD	PET-T2	PET-T1r	PET-PDr	PET-T2r
Median	3.0	2.7	2.6	1.9	2.0	1.9
Max	14.3	9.7	6.0	4.9	9.6	8.4

evant. Obviously, there are many—perhaps more robust—optimizers other than ours that could be used for this first level, including Powell. Thereafter, an optimizer that takes strong benefit of starting conditions, such as ours, is an absolute necessity for a successful multiresolution strategy.

We also take advantage of this experiment to show the relationship between the measure of geometric accuracy ϖ and the value taken by the criterion S during the course of optimization by our algorithm. It can be seen in Fig. 6 that ϖ closely follows S. It is also very likely that the most efficient behavior of our algorithm has been while optimizing on the range $\varpi \in [0.1, 1]$, because there it needed few evaluations to head its way toward the optimum.

D. Prospective and Retrospective Registration

We also applied our algorithm to the registration of volumes acquired by computed tomography (CT) or positron emission tomography (PET) with respect to three different magnetic resonance imaging (MRI) modalities: proton density (PD), T1 relaxation time (T1), and T2 relaxation time (T2) . The goal was to align the CT or the PET volumes with the MRI ones, which represents very different measurements since the former use X-rays, respectively, the decay of injected radioactive isotopes, while the latter deals with the interaction between spin and magnetic field. The MRI volumes were available in two versions: raw (PD, T1, T2), and corrected (rectified) for scanner-dependent geometric distortion (PDr, T1r, T2r). There were seven patients in each case.

We compare the results of our intermodal brain image registration algorithm to those of several other approaches published in the literature. The comparison is based on a methodology proposed by West *et al.* [2], who let selected researchers access a standard set of volumes to be registered. They also act as a repository for the ideal registration transformations (gold-standard) acquired by a prospective method using physical markers. These markers are erased before the volumes are disclosed to the investigators, who then face a retrospective blind registration task. After registration, they report back a set of transformation parameters that are compared to the gold-standard. This results in a geometric error measured in mm, and allows for a simple ranking of the competing algorithms—from an accuracy point of view. Although we carried out our registration some time after the researchers listed in the paper by West *et al.*, we were blinded in exactly the same way.

Table IV shows the results obtained by our algorithm and give the median and the maximum error over about ten cases for each pair of modalities. The registration of the images was crudely initialized by an exhaustive search procedure. With a proper heuristic to decide for convergence of the optimization, the typical execution time for a $256 \times 256 \times 28$ CT-MR registration is about 4 min on a Macintosh 9600 clocked at 350 MHz, including about 40 s of data preprocessing (e.g., determination of a mask over which to carry the optimization, nonlinear intensity modification of the CT data to spread their distribution more evenly). For a $128 \times 128 \times 15$ PET-MR registration, the typical execution time is about 40 s, of which 10 s are spent in preprocessing and 30 s in performing the realignment itself. These results compare very favorably to those of other investigators published in the literature [2].

We show the performance of the other investigators in Figs. 8 and 9, where the labels are the same as in [2], and where we have represented ourselves by the label TH. The accuracy of the gold-

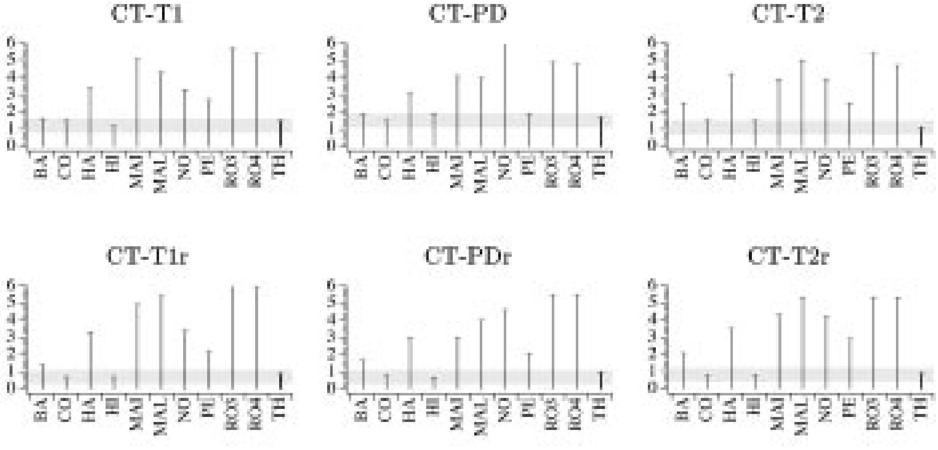


Fig. 8. Residual difference in mm between the prospective gold-standard and several retrospective registration algorithms (CT versus other modalities). The algorithm proposed in this paper is labeled TH.

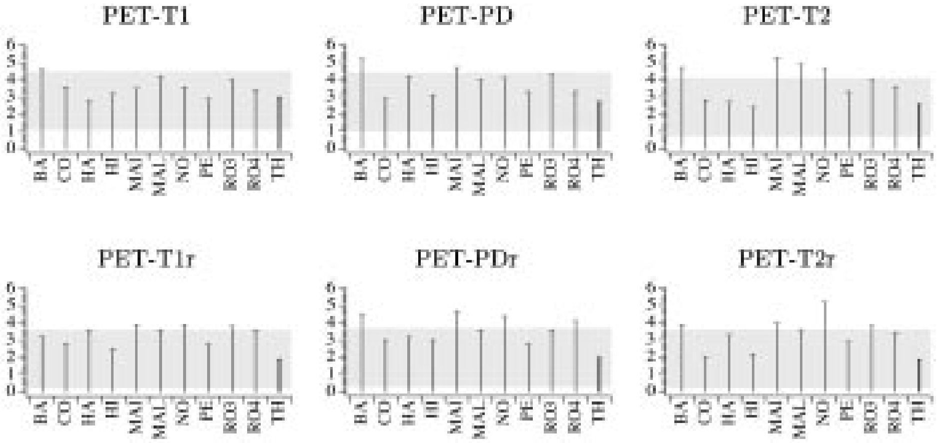


Fig. 9. Residual difference in mm between the prospective gold-standard and several retrospective registration algorithms (PET versus other modalities). The algorithm proposed in this paper is labeled TH.

standard has been estimated in the paper previously cited. Thus, for a given pair of modalities, we can surround the best result of all 11 investigators by a tolerance band of the corresponding size. This band is shown in grey in Figs. 8 and 9.

We can draw several comments with respect to ranking. First, no single algorithm is the best in all cases. Moreover, there are only three algorithms that stay within the tolerance band defined above (CO [13], HI [23] and TH); those three algorithms are all implementations of mutual information. Among all algorithms, there are 2 that come first the most often (HI and TH); however, TH is the only algorithm that is beaten by the least number of better results. It is also the best algorithm with respect to pooled median errors (for precise numeric results and experimental conditions, see [2]). Nevertheless, no algorithm taking part in this study clearly outperformed the others, and a better gold-standard or a larger number of datasets would be necessary to get more confidence in these comparative results.

VI. Discussion

A. Choice of the Criterion

While the mutual-information measure is based on the joint histogram, some measures from other researchers [6], [23] are based on the joint histogram, too. Maes *et al.* [15] show relations between them. Other criterions are also suggested in the same paper, such as the I_α-divergence, the I_α-information, the χ^2-divergence, the χ^2-information, the difference between the joint entropy and the mutual information, and the ratio between the mutual information and the sum of the marginal entropies. The authors could not establish a clear preference for either of these.

Studholme *et al.* [23]–[25] present a comparative study of several voxel-based registration criterions (e.g., various correlation measures, corresponding variance, moments of the joint histogram, joint entropy H, mutual information S). In term of robustness, they conclude that mutual information performs extremely well when compared to the other measures. In another paper [26], they propose to use $Y = 1 + S/H$, the ratio between mutual information and the joint entropy, which hints at even better performances. Due to its apparent robustness to the partial overlap problem, this last criterion could be a good candidate to initiate registration at the coarsest level of a pyramid approach.

B. Computation of the Criterion

Viola *et al.* [12], [27] propose to estimate the joint histogram $P^*(z)$ on the basis of Parzen windows made of Gaussian density functions which do not satisfy the partition of unity. Thus, their scheme does not benefit from the simplified expression of derivatives that we presented at Section IV, which is particularly relevant when computing the second-order derivatives needed for the Hessian.

While Viola and Wells produce an estimate of the joint histogram that is entirely continuous, thanks to Parzen windows and thanks to the direct use of unquantized grey values, Collignon *et al.* [13], [15], [28] represent the joint histogram in an essentially discrete fashion: they use binning with regularly spaced bins. Our work is a compromise between these two extreme views, because our representation of the histogram is continuous, like in Viola's approach, and at the same time it is described by a set of discrete and regularly-spaced intensity values, like in Collignon's approach.

C. Optimizer

Viola *et al.* [12], [27] propose a stochastic estimate of the mutual information between two datasets. They proceed by drawing two population data samples (or sets) A and B, which hold, respectively, N_A and N_B elements. Each element consists of a pair (u, v) of pixels located at identical coordinates in the two images to register. The purpose of the first set A is to provide an estimate of the joint histogram $P^*(z)$, while the purpose of the second set B is to estimate the mutual information $h(z) \propto \sum \log P^*(z_i)$. This leads to a computational load that is quadratic in the number of elements. Thus, in their approach it is impractical to sample the data in an exhaustive way, which constrains the optimization to use noisy estimates of both the criterion itself and of its derivatives. Other approaches (ours, and those presented below) do not suffer this limitation.

Studholme *et al.* [23]–[25] perform experiments based on a hill-climbing optimization algorithm that requires no derivative estimates. In the terminology of Hooke and Jeeves [29], their algorithm is best described as direct search without pattern search. This simple optimizer, where only the exploratory phase is retained, is embedded in a multiresolution framework. Their pyramid is computed by an averaging-downsampling scheme akin to Haar's wavelet. By contrast, we take in this paper full advantage of a pyramid that is optimal in a least-squares sense and that can be computed at a very modest computational cost (typically, less than a single criterion evaluation), while outperforming a Gaussian pyramid and its associated oversmoothing drawbacks, even considering an idealized, nontruncated Gaussian kernel.

Collignon *et al.* [13], [15], [28] use a Powell optimization algorithm to search for the best alignment of data. This optimizer is based on a series of line minimizations in the parameter space and suffers from sensitivity to the initial order in which the parameters are optimized. This order must be tuned to the data, which detracts from the general applicability of the mutual-information criterion. The optimizer developed in this paper is insensitive to that aspect because all parameters are considered simultaneously. In addition, it offers savings in the number of needed criterion evaluations when compared to a Powell optimizer because none is wasted in bracketing and because our algorithm is free to stop at any time. Nevertheless, a clear advantage of the Powell algorithm is its robustness. This suggests a global optimization strategy where Powell is used at the coarsest level of a multiresolution pyramid to bring robustness, and where our algorithm is used at all finer levels for faster convergence.

VII. Conclusions

We have developed a new optimizer for solving the problem of intermodal image registration. This optimizer takes benefit of the Marquardt–Levenberg strategy, while extending its capabilities to a specific problem that does not involve a least-squares criterion. The optimized criterion is the mutual information between the two images to register. We propose to compute its value by using separable Parzen windows. We show that the selection of a Parzen window that satisfies the partition of unity simplifies several aspects of the problem. It allows us to find a tractable closed-form expression for the gradient of the criterion with respect to the transformation parameters, and to justify a simplified form for its Hessian as well. Moreover, the partition of unity guarantees that the marginal histogram of the fixed reference image does not depend on the geometric transformation applied on the test image. We have introduced a coherent framework based on a continuous image model for applying the transformations and for computing the derivatives of the criterion. The same model is used for performing the registration in a multiresolution context. Both model and Parzen windows are based on B-splines. We have shown experimentally that our new optimizer is well adapted to multiresolution processing, which brings robustness and speed to the whole approach. We reach a better accuracy in less time than previously published methods.

Appendix

We provide here the steps that link (22) to (23). First, we concentrate on the middle term of (22) and determine that

$$\begin{aligned}\sum_{\zeta\in L_T}\sum_{\eta\in L_R}\frac{\partial h(\zeta,\eta;\boldsymbol{\mu})}{\partial\mu} &= \sum_{\eta\in L_R}\left(\sum_{\zeta\in L_T}\frac{\partial h(\zeta,\eta;\boldsymbol{\mu})}{\partial\mu}\right)\\ &= \sum_{\eta\in L_R}\frac{\partial h_R(\eta;\boldsymbol{\mu})}{\partial\mu}=0\end{aligned}$$

where we have taken into account the definition (6) and the conditions that lead to (10). Introducing (3) into (22), and taking advantage of the independence of p_r on μ, we get that

$$\begin{aligned}\frac{\partial S}{\partial\mu} = \sum_{\iota\in L_T}\sum_{\kappa\in L_R}\Bigg[&-\frac{\partial p(\iota,\kappa;\boldsymbol{\mu})}{\partial\mu}\\ &\cdot\left(\log_2(e)+\log_2 p(\iota,\kappa;\boldsymbol{\mu})\right.\\ &\left.-\log_2 p_T(\iota;\boldsymbol{\mu})-\log_2 p_R(\kappa;\boldsymbol{\mu})\right)\\ &+\frac{1}{\log_e(2)}\,\frac{\partial p_T(\iota;\boldsymbol{\mu})}{\partial\mu}\,\frac{p(\iota,\kappa;\boldsymbol{\mu})}{p_T(\iota;\boldsymbol{\mu})}\Bigg].\end{aligned}$$

By reorganization of the terms, we write that

$$\begin{aligned}\frac{\partial S}{\partial \mu} = &\left[\sum_{\kappa\in L_R} \left(-\log_2(e) + \log_2 p_R(\kappa;\boldsymbol{\mu})\right) \sum_{\iota\in L_T} \frac{\partial p(\iota,\kappa;\boldsymbol{\mu})}{\partial\mu}\right]\\ &- \left[\sum_{\iota\in L_T}\sum_{\kappa\in L_R} \frac{\partial p(\iota,\kappa;\boldsymbol{\mu})}{\partial\mu}\log_2\frac{p(\iota,\kappa;\boldsymbol{\mu})}{p_T(\iota;\boldsymbol{\mu})}\right]\\ &+ \left[\frac{1}{\log_e(2)}\sum_{\iota\in L_T}\frac{\partial p_T(\iota;\boldsymbol{\mu})}{\partial\mu}\frac{1}{p_T(\iota;\boldsymbol{\mu})}\sum_{\kappa\in L_R}p(\iota,\kappa;\boldsymbol{\mu})\right].\end{aligned}$$

The first of these three terms disappears because $\sum_\iota \partial p(\iota,\kappa)/\partial\mu = \partial p_R(\kappa)/\partial\mu = 0$. For the last term, we get from the definition (5) that

$$\begin{aligned}&\frac{1}{\log_e(2)}\sum_{\iota\in L_T}\frac{\partial p_T(\iota;\boldsymbol{\mu})}{\partial\mu}\frac{1}{p_T(\iota;\boldsymbol{\mu})}\sum_{\kappa\in L_R}p(\iota,\kappa;\boldsymbol{\mu})\\ &\quad= \frac{1}{\log_e(2)}\sum_{\iota\in L_T}\frac{\partial p_T(\iota;\boldsymbol{\mu})}{\partial\mu}.\end{aligned}$$

We also observe that

$$\begin{aligned}\sum_{\iota\in L_T}\frac{\partial p_T(\iota;\boldsymbol{\mu})}{\partial\mu} &= \sum_{\iota\in L_T}\sum_{\kappa\in L_R}\frac{\partial p(\iota,\kappa;\boldsymbol{\mu})}{\partial\mu}\\ &= \sum_{\kappa\in L_R}\left(\sum_{\iota\in L_T}\frac{\partial p(\iota,\kappa;\boldsymbol{\mu})}{\partial\mu}\right)\\ &= \sum_{\kappa\in L_R}\frac{\partial p_R(\kappa;\boldsymbol{\mu})}{\partial\mu} = 0\end{aligned}$$

which concludes the equivalence between (22) and (23).

References

[1] L. Gottesfeld Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, pp. 325–376, Dec. 1992.

[2] J. West, J. M. Fitzpatrick, M. Y. Wang, B. M. Dawant, C. R. Maurer, R. M. Kessler, R. J. Maciunas, C. Barillot, D. Lemoine, A. Collignon, F. Maes, P. Suetens, D. Vandermeulen, P. A. van den Elsen, S. Napel, T. S. Sumanaweera, B. Harkness, P. F. Hemler, D. L. G. Hill, D. J. Hawkes, C. Studholme, J. B. A. Maintz, M. A. Viergever, G. Malandin, X. Pennec, M. E. Noz, G. Q. Maguire Jr., M. Pollack, C. A. Pelizzari, R. A. Robb, D. Hanson, and R. P. Woods, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J. Comput. Assist. Tomogr.*, vol. 21, pp. 554–566, July/Aug. 1997.

[3] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Imag. Anal.*, vol. 2, pp. 1–36, Apr. 1998.

[4] J. V. Hajnal, S. Nadeem, E. J. Soar, A. Oatridge, I. R. Young, and G. M. Bydder, "A registration and interpolation procedure for subvoxel matching of serially acquired MR images," *J. Comput. Assist. Tomogr.*, vol. 19, pp. 289–296, Mar.–Apr. 1995.

[5] P. Thévenaz, U. E. Ruttimann, and M. Unser, "A pyramid approach to sub-pixel registration based on intensity," *IEEE Trans. Image Processing*, vol. 7, pp. 27–41, Jan. 1998.

[6] R. P. Woods, J. C. Mazziotta, and S. R. Cherry, "MRI-PET registration with automated algorithm," *J. Comput. Assist. Tomogr.*, vol. 17, pp. 536–546, July–Aug. 1993.

[7] J. Le Moigne, W. Xia, S. Chettri, T. El-Ghazawi, E. Kaymaz, B.-T. Lerner, M. Mareboyana, N. Netanyahu, J. Pierce, S. Raghavan, J. C. Tilton, W. J. Campbell, and R. F. Cromp, "Towards an intercomparison of automated registration algorithms for multiple source remote sensing data," in *Proc. Image Registration Workshop*, J. Le Moigne, Ed. Greenbelt, MD: NASA Goddard Space Flight Center, Nov. 20–21, 1997, vol. NASA/CP-1998-206 853, pp. 307–316.

[8] K. Watson, "Processing remote sensing images using the 2-D FFT-noise reduction and other applications," *Geophysics*, vol. 58, pp. 835–852, June 1993.

[9] J.-P. Djamdji, A. Bijaoui, and R. Manière, "Geometrical registration of images: The multiresolution approach," *Photogram. Eng. Remote Sensing*, vol. 59, pp. 645–653, May 1993.

[10] J. R. G. Townshend, C. O. Justice, C. Gurney, and J. McManus, "The impact of misregistration on change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 30, pp. 1054–1060, Sept. 1992.

[11] M. Berman, L. Bischof, and S. J. Davies, "Estimating band-to-band misregistrations in aliased imagery," *Graph. Models Image Process.*, vol. 56, pp. 479–493, Nov. 1994.

[12] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," in *Proc. 5th Int. Conf. Computer Vision*, Boston, MA, June 20–23, 1995, pp. 16–23.

[13] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal, "Automated multi-modality image registration based on information theory," in *Information Processing in Medical Imaging*. Norwell, MA: Kluwer, 1995, pp. 263–274.

[14] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, pp. 431–441, 1963.

[15] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, pp. 187–198, Apr. 1997.

[16] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, Sept. 1962.

[17] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Part I—Theory," *IEEE Trans. Signal Processing*, vol. 41, pp. 821–832, Feb. 1993.

[18] ——, "B-spline signal processing: Part II—Efficient design and applications," *IEEE Trans. Signal Processing*, vol. 41, pp. 834–848, Feb. 1993.

[19] ——, "The L_2 polynomial spline pyramid," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 364–379, Apr. 1993.

[20] A. Aldroubi, M. Unser, and M. Eden, "Cardinal spline filters: Stability and convergence to the ideal sinc interpolator," *Signal Process.*, vol. 28, pp. 127–138, Aug. 1992.

[21] G. Strang and G. Fix, "A Fourier analysis of the finite element variational method," in *Constructive Aspects of Functional Analysis*. Erice/Rome, Italy: Centro Internazionale Matematico Estivo, Edizioni Cremonese, June 27–July 7, 1971, pp. 796–830.

[22] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes, The Art of Scientific Computing*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[23] D. L. G. Hill, C. Studholme, and D. J. Hawkes, "Voxel similarity measures for automated image registration," in *Proc. SPIE Visualization Biomedical Computing*, vol. 2359, R. A. Robb, Ed., Rochester, MN, Oct. 4–7, 1994, pp. 205–216.

[24] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Automated 3-D registration of MR and CT images of the head," *Med. Imag. Anal.*, vol. 1, pp. 163–175, 1996.

[25] ——, "Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures," *Med. Phys.*, vol. 24, pp. 25–35, Jan. 1997.

[26] C. Studholme, D. J. Hawkes, and D. L. G. Hill, "A normalized entropy measure for multi-modal image alignment," in *Proc. SPIE Conf. Image Processing*, vol. 3338, San Diego, CA, Feb. 23–26, 1998, pp. 132–143.

[27] W. M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Med. Imag.Anal.*, vol. 1, pp. 35–51, 1996.

[28] A. Collignon, D. Vandermeulen, P. Suetens, and G. Marchal, "3D multimodality medical image registration using feature space clustering," in *Proc. Computer Vision, Virtual Reality, Robotics Medicine*, N. Ayache, Ed. Nice, France, Apr. 1995, pp. 195–204.

[29] R. Hooke and T. A. Jeeves, "A direct search solution of numerical and statistical problems," *J. Assoc. Comput. Mach.*, vol. 8, pp. 212–229, 1961.

Philippe Thévenaz (M'95) was born in Lausanne, Switzerland, in 1962. He received the Diploma degree in microengineering in January 1986, from the Lausanne Swiss Federal Institute of Technology (EPFL), and the Ph.D. degree in June 1993, with a thesis on the use of the linear prediction residue for text-independent speaker recognition from the Institute of Microtechnology (IMT), University of Neuchâtel, Switzerland.

He was with IMT where he worked in the domains of image processing (optical flow) and speech processing (speech coding and speaker recognition). He was then a Visiting Fellow with the Biomedical Engineering and Instrumentation Program, National Institutes of Health (NIH), Bethesda, MD, where he developed research interests that include splines and multiresolution signal representations, geometric image transformations, and biomedical image registration. Since 1998, he has been First Assistant with the Lausanne Swiss Federal Institute of Technology.

Michael Unser (M'88–SM'94–F'99) received the M.S. (summa cum laude) and Ph.D. degrees in electrical engineering in 1981 and 1984, respectively, from the Swiss Federal Institute of Technology, Lausanne, Switzerland.

From 1985 to 1997, he was with the Biomedical Engineering and Instrumentation Program, National Institutes of Health, Bethesda, MD, where he was Head of the Image Processing Group. He is now Professor and Head of the Biomedical Imaging Group, Swiss Federal Institute of Technology, Lausanne. His main research area is biomedical image processing. He has a strong interest in sampling theories, multiresolution algorithms, wavelets, and the use of splines for image processing. He is the author of over 80 published journal papers in these areas. He is on the editorial boards of *Signal Processing*, the *Journal of Visual Communication and Image Representation*, and *Pattern Recognition*. He serves as regular Chair for the SPIE Conference on Wavelet Applications in Signal and Image Processing, which has been held annually since 1993.

Dr. Unser is an Associate Editor for the IEEE TRANSACTIONS ON MEDICAL IMAGING. He is a former Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (1992–1995), the IEEE SIGNAL PROCESSING LETTERS (1994–1998), and was a member of the IMDSP Committee of the IEEE Signal Processing Society (1993–1999). He received the Dommer Prize for Excellence from the Swiss Federal Institute of Technology in 1981, the Research Prize of the Brown–Boveri Corporation, Switzerland, for his thesis in 1984, and the IEEE Signal Processing Society's 1995 Best Paper Award. In January 1999, he was elected Fellow of the IEEE with the citation "For contributions to the theory and practice of splines in signal processing."