Multimodal Image Registration by Maximization of the Correlation Ratio

Alexis Roche, Grégoire Malandain, Nicholas Ayache, Xavier Pennec

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Multimodal Image Registration by Maximization of the Correlation Ratio

Alexis Roche — Grégoire Malandain — Xavier Pennec — Nicholas Ayache

N° 3378

Août 98

THÈME 3

ISSN 0249-6399

INRIA
SOPHIA ANTIPOLIS

Multimodal Image Registration by Maximization of the Correlation Ratio

Alexis Roche* , Grégoire Malandain , Xavier Pennec , Nicholas Ayache

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Epidaure

Rapport de recherche n° 3378 — Août 98† — 42 pages

Abstract: Over the last five years, new "voxel-based" approaches have allowed important leaps in multimodal image registration, notably due to the increasing use of information-theoretic similarity measures. Their wide success has led to the progressive abandon of measures using standard image statistics (mean and variance). Until now, such measures have essentially been based on heuristics. In this paper, we address the determination of a new measure based on standard statistics from a theoretical point of view. We show that it naturally leads to a known concept of probability theory, the *correlation ratio*. In our derivation, we take as the hypothesis the functional dependence between the image intensities. This means that one image is considered as a model of the other. Although such a hypothesis is not validate in every circumstance, it enables us to incorporate implicitely an a priori smoothness model. We also demonstrate preliminary results of multimodal rigid registration involving Magnetic Resonance (MR), Computed Tomography (CT), and Positron Emission Tomography (PET) images. These results suggest that the correlation ratio provides a good trade-off between accuracy and robustness.

Key-words: registration, medical images, multimodality, similarity measures, random variables geometry, correlation ratio.

* Email: Alexis.Roche@sophia.inria.fr

† Bien qu'ayant été enregistré au mois de mars 1998, ce rapport de recherche est resté confidentiel jusqu'à la date de sa publication, à savoir courant août 1998.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : 04 92 38 77 77 - International : +33 4 92 38 77 77 —Fax : 04 92 38 77 65 - International : +33 4 92 38 77 65

Recalage d'images multimodales par maximisation du rapport de corrélation

Résumé : Au cours des cinq dernières années, de nouvelles approches "orientées voxel" ont permis d'importantes avancées en recalage d'images multimodales, notamment grâce à l'émergence de mesures héritées de la théorie de l'information. Leur large succès a conduit à l'abandon progressif des mesures faisant appel à des statistiques standard (moyenne et variance). Jusqu'à présent, ces dernières ont été essentiellement basées sur des heuristiques. En nous plaçant dans un cadre théorique, nous proposons ici de déterminer une nouvelle mesure de similarité "stantard". Nous sommes naturellement amenés à un concept classique de la théorie des probabilités, le *rapport de corrélation*. L'obtention du critère fait appel à une hypothèse de dépendance fonctionnelle entre les intensités des images. Cela signifie que l'une des images est considérée comme un modèle de l'autre. Bien qu'une telle hypothèse ne soit pas vérifiée en toute circonstance, elle nous permet d'incorporer implicitement un modèle de continuité *a priori*. Pour finir, nous présentons nos premiers résultats de recalage avec des images par résonance magnétique, scannographie et tomographie par émission de positon. Ces résultats suggèrent que le rapport de corrélation est un bon compromis entre la précision et la robustesse.

Mots-clés : recalage, images médicales, images multimodales, mesures de similarité, géométrie des variables aléatoires, rapport de corrélation.

Contents

1 Introduction

1.1 Overview

Registration is a basic task in image processing. Given two images representing the *same* or even *analogous* objects, it consists in geometrically transforming one image (the *floating* image) to the other (the *reference* image) so that voxels representing the same physical structure may be superimposed. In the field of medical imaging, registration is useful for the study of diseases, diagnosis, evaluation of treatment efficacy, and even computer-assisted surgery.

Most of the existing registration techniques address the ideal case where the images are similar in intensity up to an unknown geometrical transformation. Some of them take into account elastic deformations of the tissues [19, 18]. However, in many cases, we are faced with highly dissimilar images that makes the above methods inefficient in practice. We can identify four different sources of dissimilarity in medical images:

- *The representation of information.* The intensity set that represents a tissue depends on the imaging modality. For example, bones may appear with low intensities in MR[1] T1 weighted images while they correspond to high intensity values in CT[2] images.
- *The non-redundancy of information.* Images of different modalities provide complementary information. For example, anatomical information is available in MR images, while functional information is available in PET[3] or SPECT[4] images.
- *The measurement noise:* the non-reproducible part of the imaging process. It implies distortions of the signal and is not necessarily additive, Gaussian, or stationary. For example, bias is frequent in MR images due to the magnetic field inhomogeneities. Ultrasound (US) images are known to be corrupted with speckle, a highly tissue-dependent, noise.
- *Occlusion.* This is a classical situation which can occur due to the growth of a tumor, the diffusion of a contrast agent, or simply when parts of the tissues are masked.

For the last five years, some evidence has been shown that similarity measure-based approaches should provide satisfactory results when the images present considerable dissimilarities. Their general principle consists of quantifying the quality of a registration with respect to the statistical "similarity" of the images' overlapping voxels. It involves the design of a *similarity measure* that is assumed to be maximal when the images are correctly aligned. These approaches are thus often implemented using an optimization scheme. Other implementations use a dynamical process [10].

Many similarity measures have been proposed in the literature (see [9, 4, 21, 3] for reviews). Considering the elementary problem of matching two identical images, the first idea was to use a least squares criterion. Simple correlation measures were proposed in order to extend this method to situations where the image intensities are

[1] Magnetic Resonance.
[2] Computed Tomography.
[3] Positron Emission Tomography.
[4] Single Photon-Emission Computed Tomography.

no longer identical, but linearly correlated. These similarity measures have been used extensively, notably in medical imaging. However, they make a strong assumption regarding the relationship that exists between the images. They generally do a good job only in monomodal problems.

More recently, other measures have been devised to achieve registration when the images present important dissimilarities, which is typically the case of multimodal images. The Woods criterion [26, 25] provided a powerful algorithm for matching PET images with MR images, although it needs some manual segmentation to work efficiently. A recently proposed adaptation using robust estimators [12] proved that the approach could lead to fully automatic registration of SPECT and MR images.

1.2 Mutual information

Mutual information [24, 23, 8, 17, 11] is today probably the most popular similarity measure for multimodal registration. It has been successfully applied to several modality combinations including MR, CT, PET, and SPECT. Given two images X and Y, one may define their joint probability density function (joint pdf), $P(i,j)$, by simple normalization of their 2D-histogram (other approaches are possible, see appendix A). Let $P_x(i)$ and $P_y(j)$ denote the corresponding marginal probability density functions (pdf's). Mutual information between X and Y is given by [2]:

$$I(X,Y) = \sum_{i,j} P(i,j) \ \log_2 \frac{P(i,j)}{P_x(i)P_y(j)}.$$

The mutual information measure can be considered very general since it makes very few assumptions regarding the relationship that exists between the image intensities (see [22] for an excellent discussion). It does not assume linear correlation, nor even functional correlation, but only statistical dependence.

Nevertheless, one of its pitfalls is to treat intensity values in a purely *qualitative* way, without considering any notion of proximity in the intensity space. In a real image, one tissue is never represented by a single intensity value but rather by a certain interval. Thus, nearby intensities convey a lot of *spatial* information, which it might be risky to be forgiving about. Figure 1 presents a synthetic situation in which mutual information is not well-adapted.

In some cases, one can reasonably make additional hypotheses about the relationship that exists between the images; then mutual information is under-constrained. Practically, one often observes its tendency to handle many local maxima. In this paper, we investigate the case where a *functional* correlation can be assumed. However, minimal assumptions are made regarding the nature of the function. The similarity measure we propose is inherited from probability theory and is known as the *correlation ratio*. We will show that, in addition to a relative generality, the correlation ratio takes into account proximity in the intensity space, resulting in robustness at relatively low resolution and, thus, attractive computing time.

After introducing our formalism in section 2, the theoretical concept of correlation ratio is then presented in section 3. We describe the registration algorithm in section 4. Section 5 draws a theoretical comparison with the Woods criterion, mutual information and the weighted neighbor likelihood, while section 6 proposes a discussion in the form of two synthetic registration examples. Preliminary results involving MR, CT, and PET images are shown and discussed in section 7.

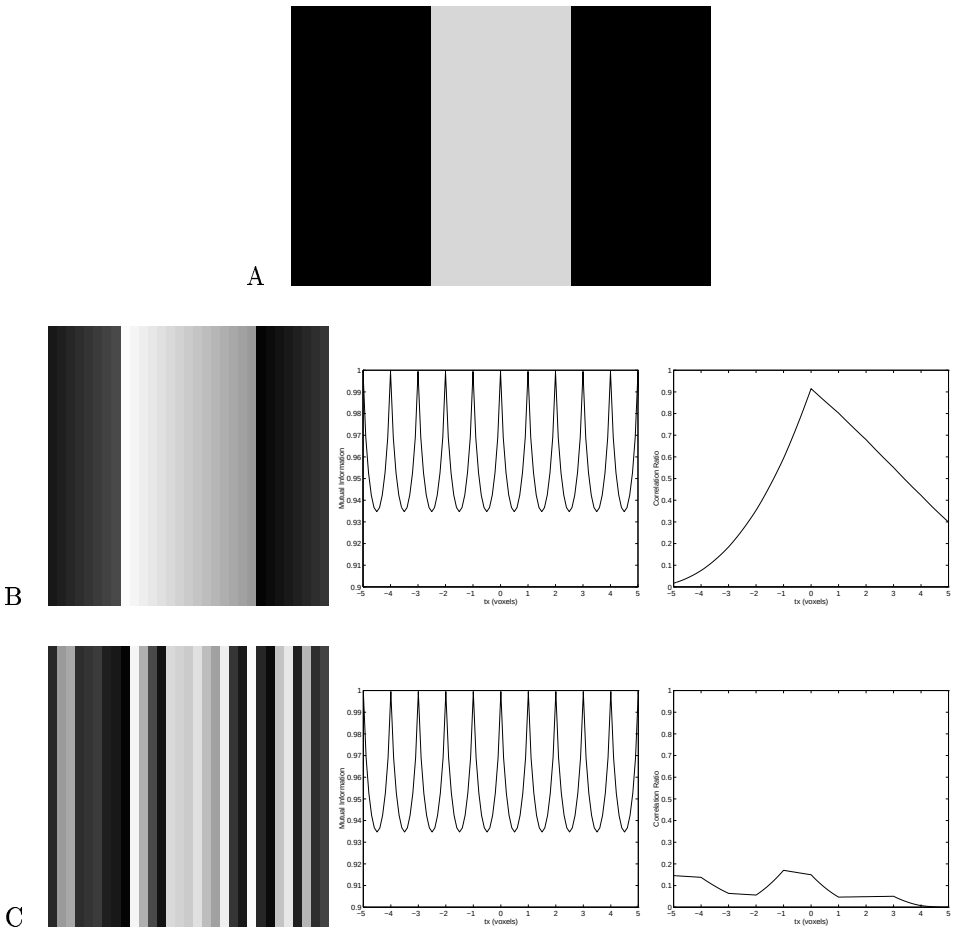


Figure 1: *"Grey stripe" registration experiment. (A), binary image (40 by 30 pixels). (B), gradation image of the stripe (30 by 30 pixels): the intensity is uniform in a given column but each column has a different intensity. Mutual information $I(A, B_T)$ and the correlation ratio (later explained) have been computed for various horizontal translations of B using bilinear interpolation for non-integer translations. Plots: Left, mutual information vs. horizontal translation. Right, the correlation ratio. By convention, the null translation corresponds to the case where the stripes completely overlap. Notice that for any integer translation, $I(A, B_T)$ is equal to 1, its upper bound (provided that B_T totally falls into the region delimited by A). For non-integer translations, smaller values are observed due to interpolation. Mutual information does not explain how to align the stripe in B with the stripe in A. (C), image obtained by randomly permuting the columns of B. Mutual information behaves in the same way as for (B) while the correlation ratio now takes non-meaningful values.*

2 Images as random variables

As statistical concepts provide a convenient framework for the computation of similarity, we "artificially" consider the images as random variables (see appendix A). This corresponds to interpreting an image histogram as a probability density function (pdf). Moreover, we also consider the 2D-histogram of an image pair as a joint pdf. In the context of similarity measure-based methods, most authors generally interpret images as the results of a random experience. We insist that our point of view is slightly different. This means that when randomly selecting one voxel in an image X (each of them having the same probability), the probability of getting an intensity i is simply proportional to the number of voxels, N_i, in X having the intensity i:

$$P(i) = \frac{N_i}{N}. \tag{1}$$

This formalism is not compulsory. It seems paradoxical to consider data as random variables. But it is a convenient interpretation given that probabilistic concepts turn out to be powerful for designing similarity measures. One may also notice that this "artificial" randomness must not be confused with the stochastic nature of the noise that corrupts the images.

For convenience, we will suppose in this section that we deal with 3D images. The presented method can easily be adapted to 2D images. In order to define the joint pdf of an image pair, let us consider two images (X, Y) and a spatial transformation T that maps the grid of Y, Ω_y, to the grid of X, Ω_x. Both Ω_x and Ω_y are finite tridimensional sets of voxels. X and Y take their intensity values in a finite set $\mathcal{A}$ which can be assumed to be the same for the two images. Typically, $\mathcal{A} = \{0, 1, 2, \ldots, 255\}$.

$$X : \Omega_x \to \mathcal{A}, \qquad Y : \Omega_y \to \mathcal{A}.$$

Applying the transformation T to image Y, we define a new mapping from the transformed positions of Y to $\mathcal{A}$:

$$\begin{aligned} Y_T : T(\Omega_y) &\to \mathcal{A}, \\ \omega &\mapsto Y\left[T^{-1}(\omega)\right]. \end{aligned}$$

We should be able to find the intensities that a given point of $T(\Omega_y)$ *simultaneously* takes in X and Y_T. As we want to deal with continuous spatial transformations, points of Ω_y generally do not transform to points of Ω_x. Thus, for a given transformation T, interpolation is needed in order to *define* the joint pdf of the images.

In this work, we estimate the intensity in X using trilinear interpolation and rounding towards the nearest integer in order to get values in $\mathcal{A}$[5]. Points of $T(\Omega_y)$ that don't have eight neighbors in Ω_x are rejected. Let $T(\Omega_y)^*$ denote the subset of accepted points and $\tilde{X}$ denote the interpolation of X. We define the image pair as the following random couple:

$$\begin{aligned} Z_T : T(\Omega_y)^* &\to \mathcal{A}^2, \\ \omega &\mapsto \left(\tilde{X}(\omega),\ Y\left[T^{-1}(\omega)\right]\right). \end{aligned}$$

[5] Other, most accurate methods like spline or sinc interpolation are too time consuming; still it is probably possible to improve trilinear interpolation while keeping an acceptable complexity.

Then we define the joint pdf of X and Y_T like we did for a single image eq (1):

$$P_T(i,j) = \frac{\text{Card } \{x | Z_T(x) = (i,j)\}}{\text{Card } T(\Omega_y)^*}. \tag{2}$$

The marginal pdf's of the images X and Y_T are entirely determined by the joint pdf $P_T(i,j)$. However, they are not equal *a priori* to those we would have gotten by considering single images. Due to interpolation, they depend on the transformation T :

$$P_{x,T}(i) = \sum_j P_T(i,j), \qquad P_{y,T}(j) = \sum_i P_T(i,j).$$

3 Random variables geometry

In section 2, we have defined the joint pdf of two images. We now study a method for evaluating some kind of dependence between two random variables when their joint pdf is known. For this purpose, the geometry of L^2 random variables presents two advantages. First, the norm on L^2 naturally imposes a constraint of proximity in the sample space unlike information-theoretic measures. This is consistent with the assumption that the images are smooth. Second, there exists a simple method for quantifying the *functional* dependence between two random variables. We are aware that the notion of functional dependence is more restrictive than the general notion of statistical dependence.

This section summarizes the fundamental concepts of random variables geometry. More details will be found in [16].

3.1 L^2 space

L^2 is defined as the space of square integrable real variables[6], that is the variables which verify:

$$E(X^2) = \int_\Omega X^2 \, dPr < +\infty,$$

where E denotes the expectation operator. One shows that L^2 is a Hilbert space with respect to the dot product $\langle X, Y \rangle = E(XY)$. Thus, the corresponding norm is the second-order moment of a variable:

$$\|X\|_2 = \sqrt{E(X^2)}. \tag{3}$$

The L^2 norm is closely related to the classical notions of expectation, variance and standard deviation. Equation (3) can be rewritten:

$$\|X - E(X)\|_2 = \sqrt{Var(X)} = StdDev(X).$$

Due to its Hilbertian structure, L^2 has interesting geometric properties. One can define a notion of orthogonality between two variables:

$$X \perp Y \iff E(XY) = 0.$$

[6]In fact, L^2 is the quotient space of square integrable variables with respect to the "almost everywhere" equivalence relationship. This distinction has no importance here.

The meaning of orthogonality in L^2 relates with the notion of independence, but in a less restrictive way. Recall that two variables X and Y are said to be independent if their joint pdf is equal to the product of their marginal pdf's, that is $P(x,y) = P(x)P(y)$. One easily shows that $E(XY) = E(X)E(Y)$ for two such variables. Thus:

$$X \text{ and } Y \text{ independent} \implies X - E(X) \perp Y - E(Y).$$

However, the converse is false. Orthogonality in L^2 is a weaker constraint than independence. It may be seen as a notion of independence *on the average*.

In a general way, the angle α between two variables X and Y is defined thanks to a basic property of dot products:

$$\langle X, Y \rangle = \|X\|_2 \|Y\|_2 \cos\alpha. \tag{4}$$

3.2 Expectation

L^2 contains the one-dimensional space Δ of deterministic variables, i.e. variables which are constant on the state space Ω. Given a variable X, its expectation is:

$$E(X) = \int_{\mathbb{R}} x\, p(x)\, dx = \langle X, 1 \rangle.$$

Therefore, $E(X)$ is nothing but the orthogonal projection of X onto Δ. In the sense of the L^2 norm, it is the *constant* variable which best approximates X (classical notion of mean).

$$E(X) = \arg\min_{C \in \Delta} \|X - C\|_2^2.$$

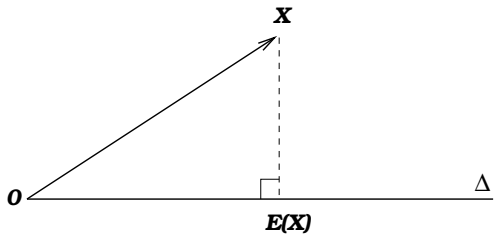


Figure 2: *Geometric interpretation of expectation. $E(X)$ is the orthogonal projection of X onto the constant direction Δ.*

3.3 Correlation coefficient

A quick method to evaluate (approximately) the degree of dependence between two variables is to compute their correlation coefficient. Given two variables X and Y, it is defined as:

$$\rho(X,Y) = \frac{Cov(X,Y)^2}{Var(X)Var(Y)} = \frac{[E(XY) - E(X)E(Y)]^2}{[E(X^2) - E(X)^2]\,[E(Y^2) - E(Y)^2]}.$$

From a geometric point of view, we write it:

$$\rho(X,Y) = \frac{\langle\, X - E(X), Y - E(Y)\,\rangle^2}{\|X - E(X)\|_2^2 \|Y - E(Y)\|_2^2}.$$

Using eq (4), the correlation coefficient between X and Y can be interpreted in a geometric way. Let α denote the angle between $X - E(X)$ and $Y - E(Y)$. We have:

$$\rho(X,Y) = \cos^2 \alpha.$$

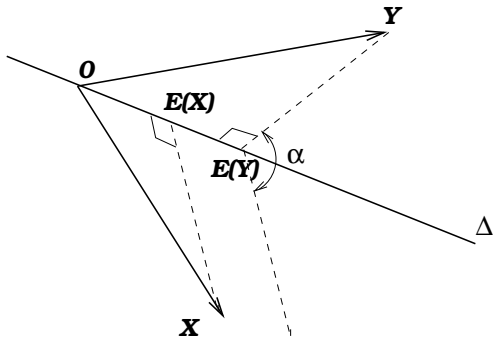


Figure 3: *Geometric interpretation of the correlation coefficient. We have:* $\rho(X,Y) = \cos^2 \alpha$. *The constant (or deterministic) direction is denoted by* Δ.

We see that $\rho(X,Y)$ is larger as the angle α is small. It reaches 1 if $X - E(X)$ and $Y - E(Y)$ are colinear. This is to say that the correlation coefficient measures the *linear* dependence between two variables. As we want to take into account general functions between X and Y, possibly non-linear and non-invertible, this is not a good measure of *functional* dependence.

3.4 Conditional expectation

Evaluating the functional dependence between two variables comes down to an interpolation problem with no constraints. Suppose we want to estimate a variable Y with another variable X. A natural approach would be: (1) find the function $\phi^*(X)$ that best fits Y among *all* possible functions of X; (2) quantify the quality of the estimate $\phi^*(X)$ with respect to Y. The notion of conditional expectation provides a straightforward method for such an evaluation, without having to test every possible function of X.

If X and Y are not independent, knowing an event $X = x$ should provide some new information about Y. Any event $X = x$ induces a *conditional* pdf for Y, that is

$$p(y|x) \stackrel{def}{=} \frac{p(x,y)}{p(x)}.$$

Then, the corresponding *a posteriori* expectation of Y is:

$$\phi^*(x) = E(Y|X = x) = \int y\, p(y|x)\, dy.$$

To any possible realization of X corresponds an *a posteriori* expectation of Y. Thus, one defines a function of X, which is the conditional expectation of Y in terms of X:

$$E(Y|X) = \phi^*(X).$$

Notice that $E(Y|X)$ is *also* a random variable. It is easy to verify that it is an unbiased estimate (see appendix C), i.e.,

$$E\left[E(Y|X)\right] = E(Y).$$

The conditional expectation's major interest is that it is the optimal approximator in the sense of the L^2 norm. We show in appendix B that $E(Y|X)$ is the measurable function of X that has the smallest distance to Y :

$$\phi^* = \arg\min_{\phi} \|Y - \phi(X)\|_2.$$

Minimal constraints are imposed on the functions ϕ. The only requirement is that they be measurable, which is a much more general property than continuity.

3.5 Total variance theorem

Let us now give a geometric interpretation of the conditional expectation. We consider the sub-space L_x of every possible function ϕ of X (provided that it remains in L^2):

$$L_x = L^2 \cap \{\phi(X) \mid \phi : \mathbb{R} \to \mathbb{R}\}.$$

Notice that every constant variable is a (constant) function of X, so that:

$$\Delta \subset L_x.$$

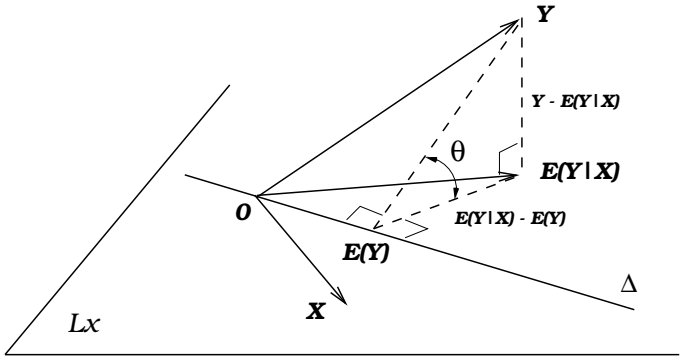


Figure 4: *Geometric interpretation of the conditional expectation. It is the orthogonal projection onto L_x.*

As the conditional expectation $E(Y|X)$ minimizes the distance between Y and L_x, $E(Y|X)$ is the *orthogonal projection* of Y onto L_x (see figure 4). This is due to the Hilbertian structure of L^2. This simple geometrical property allows us to compute easily the distance between Y and L_x. Indeed, $Y - E(Y|X)$ is orthogonal to any vector of L_x by definition of the orthogonal projection. Notably,

$$Y - E(Y|X) \;\perp\; E(Y|X) - E(Y).$$

Therefore, the triangle whose vertices are Y, $E(Y)$, and $E(Y|X)$ is right-angled in $E(Y|X)$. Applying the Pythagorean theorem, we retrieve a result known as the total variance theorem:

$$Var(Y) \quad = \quad \|Y - E(Y)\|_2^2,$$

$$
\begin{aligned}
&= \|E(Y|X) - E\left[E(Y|X)\right]\|_2^2 + \|Y - E(Y|X)\|_2^2, \\
&= \|E(Y|X) - E(Y)\|_2^2 + \int_{\mathbb{R}}\int_{\mathbb{R}} \left[y - \phi^*(x)\right]^2 \, p(x,y)\, dxdy, \qquad (5) \\
&= Var\left[E(Y|X)\right] + \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \left[y - \phi^*(x)\right]^2 \, p(y|x)\, dy \right) p(x)\, dx. \qquad (6)
\end{aligned}
$$

Step (5) relies on the fact that $E\left[E(Y|X)\right] = E(Y)$. Let us denote:

$$
Var(Y|X=x) \stackrel{def}{=} \int_{\mathbb{R}} \left[y - \phi^*(x)\right]^2 \, p(y|x)\, dy,
$$

$$
E_X, \text{ the operator defined by } \forall \psi : \mathbb{R} \to \mathbb{R}, \quad E_X(\psi) = \int_{\mathbb{R}} \psi(x)\, p(x)\, dx.
$$

Then, eq (6) can be rewritten in a more compact way:

$$
Var(Y) = Var\left[E(Y|X)\right] + E_X\left[Var(Y|X=x)\right].
$$

This may be seen as an energy conservation equation. The variance of Y is decomposed as a sum of two "energy" terms:

- $Var\left[E(Y|X)\right]$ is the variance of the conditional expectation $E(Y|X)$. It measures the part of Y which is predicted by X.
- Conversely, the term $E_X\left[Var(Y|X=x)\right]$, which is called the conditional variance, represents the square distance of Y to the space L_x. It measures the part of Y which is functionally independent of X.

3.6 Correlation ratio

We now can design a measure of functional dependence between X and Y. Accounting for the interpretation of the total variance theorem in terms of energy, it seems natural to compare the "explained" energy of Y with its total energy. This leads to the definition of the *correlation ratio* between X and Y:

$$
\eta(Y|X) = \frac{Var\left[E(Y|X)\right]}{Var(Y)}.
$$

The correlation ratio also has a simple geometric interpretation. Let θ denote the angle between $Y - E(Y)$ and the space L_x. By definition, θ is also the angle between $Y - E(Y)$ and $E(Y|X) - E(Y)$ (see figure 4). We have

$$
\cos^2\theta = \frac{Var\left[E(Y|X)\right]}{Var(Y)}, \qquad \sin^2\theta = \frac{E_X\left[Var(Y|X=x)\right]}{Var(Y)}.
$$

Then:

$$
\eta(Y|X) = \cos^2\theta.
$$

Unlike the correlation coefficient which measures the *linear* dependence between two variables, the correlation ratio measures the *functional* dependence. It takes on values between 0 and 1. A value near 1 indicates a high functional dependence, while a value near 0 indicates a low functional dependence. The two extreme cases are:

$$
\begin{aligned}
&\eta(Y|X) = 1 \iff \exists \phi \;\; Y = \phi(X), \\
&\eta(Y|X) = 0 \iff E(Y|X) = constant = E(Y).
\end{aligned}
$$

By nature, the correlation ratio is asymmetric since the two variables fundamentally do not play the same role in the functional relationship. Thus, in general :

$$\eta(Y|X) \neq \eta(X|Y).$$

Some additional properties of the correlation ratio are summarized below:

- $\eta(Y|X) = \rho\left[E(Y|X), Y\right]$
- $\eta(Y|X) \geq \rho(X,Y)$
- $\eta(Y|X) = \rho(X,Y)$ if and only if $E(Y|X)$ is linear[7] with respect to X, i.e. $\exists(\alpha,\beta) \quad E(Y|X) = \alpha + \beta X$

4 Algorithm

4.1 Initialization

Given two images X and Y to be registered, one has to make two choices before starting the registration process:

1. Which one image should be the *reference* image or, conversely, which should be the *floating* image?
2. As correlation ratio is not symmetrical, should we compute $\eta(Y|X)$ or $\eta(X|Y)$? This requires the choice of a *template* image that is used to *estimate* the other image in the sense of the conditional expectation (see section 3).

Depending on the choices for image X, we have four different possibilities for computing the correlation ratio:

Image X	*reference*	*floating*
template	$\eta(Y_T\|X)$	$\eta(Y\|X_T)$
estimated	$\eta(X\|Y_T)$	$\eta(X_T\|Y)$

As the reference image is interpolated (see section 2), it should be the one with the lowest frequency band. However, this choice turned out to have little importance in the experiments we made. In contrast, the choice of the template image is crucial. It involves an ad-hoc hypothesis that one image is a good model of the other. For simplicity, we assume in the following that X both denotes the reference image and the template image.

4.2 Computation of the correlation ratio

In order to compute $\eta(Y_T|X)$ for a given transformation T, in practice we use the equation:

$$1 - \eta(Y_T|X) = \frac{E_X\left[Var(Y_T|X=x)\right]}{Var(Y_T)}.$$

Thus, having defined the joint pdf of X and Y_T (see section 2), we compute $\eta(Y_T|X)$ using the following formula:

$$1 - \eta(Y_T|X) = \frac{1}{\sigma^2} \sum_i \sigma_i^2 \; P_{x,T}(i),$$

[7] To be accurate, one should rather say *affine* instead of *linear*.

with:

$$\sigma^2 = \sum_j j^2 P_y(j) - m^2, \qquad m = \sum_j j P_y(j),$$

$$\sigma_i^2 = \frac{1}{P_x(i)} \sum_j j^2 P(i,j) - m_i^2, \qquad m_i = \frac{1}{P_x(i)} \sum_j j P(i,j).$$

4.3 Maximization scheme

We want to find the transformation that maximizes $\eta(Y_T|X)$:

$$T^* = \arg\max_{T \in \mathcal{T}} \eta(Y_T|X).$$

In the present work, we restricted the search space $\mathcal{T}$ to the set of 3D rigid transformations. Since the correlation ratio is assumed to be maximal when the images are correctly aligned, the registration process is performed via an optimization scheme. Our current implementation is similar to the MIRIT algorithm proposed in [8], employing Powell's multidimensional direction set method coupled with Brent's line optimization [15]. Although it has the advantage of requiring no gradient computation, Powell's method is intended for convex criteria. Therefore, the maximized criterion should hold no attraction basin except from the one corresponding to the correct transformation. There is no evidence that correlation ratio verifies this property. Therefore, a robust maximization scheme would probably work better. However, in the cases we treated, the correlation ratio demonstrated low sensitivity to being trapped in local maxima.

4.4 Parameterization

As we deal with rigid transformations, the correlation ratio between X and Y_T depends on a six-dimensional parameter μ representing the transformation T_μ :

$$J(\mu) = \eta(Y_{T_\mu}|X), \qquad \mu = (\mathbf{t}, \mathbf{r}),$$

where $\mathbf{t}$ is a translation vector and $\mathbf{r}$ a rotation vector [1]. Numerical instabilities corresponding to small rotations are handled with a Taylor expansion [14].

In order to ignore image intensities which are not supposed to be pertinent for registration, intensity regions of interest can be specified. Intensities which stand apart are not taken into account in the computation of the joint pdf. It also allows a speed up in the calculation with a minimal loss in terms of information.

5 Theoretical comparison with other measures

5.1 Woods criterion

The criterion devised by Woods et al. [26] is a heuristic that was originally designed for PET-MR registration. Isolating a given iso-intensity set in the MR image, their approach consists of measuring the corresponding set intensity dispersion in the PET image. According to the formalism introduced in section 3, we can write the Woods criterion as follows:

$$W(Y|X) = E_X \left(\frac{\sqrt{Var(Y_x)}}{E(Y_x)} \right), \tag{7}$$

where Y_x is the conditional variable induced on Y by the event $X = x$. We use the notation $W(Y|X)$ in order to emphasize that the criterion is asymmetrical, such as the correlation ratio.

The dispersion measure $\sqrt{Var(Y_x)}/E(Y_x)$ is a normalized standard deviation. It has the property of being invariant to multiplicative changes in the PET image. The criterion is obtained by averaging the various measured dispersions and is assumed to be *minimal* at the registration position. Considering eq (8), this turns out to be a strategy analogous to ours. In fact, it turns out to be a strategy analogous to ours. Equation (7) should be compared to the following characterization of correlation ratio:

$$1 - \eta(Y|X) = \frac{1}{Var(Y)} \, E_X \left[Var(Y_x) \right] . \tag{8}$$

Though different, eq (7) and eq (8) express the same basic idea. Even so, we can identify two differences. First, the correlation ratio sums variances while the Woods criterion sums normalized standard deviations. Second, normalization is achieved in the correlation ratio via a global division by $Var(Y)$; in the Woods criterion, every term of the sum is divided by a mean.

One can wonder whether the Woods criterion is equivalent to the correlation ratio. The following theorem shows that this is not true.

Theorem 1 *The Woods criterion and the Correlation Ratio are not equivalent in the sense that there exists no invertible mapping f, such that,*

$$\forall (X, Y) \in L^2 \quad \eta(Y|X) = f \left[W(Y|X) \right] .$$

Proof We give a counter-example. Let X and Y be independent L^2 variables with expectation 1 and variance 1. Such variables exist. As X and Y are independent, $\eta(Y|X) = 0$. Moreover, it is easy to check out that $W(Y|X) = 1$. Let now $Z = Y + 1$. Z is independent from X and we have: $E(Z) = 2$, $Var(Z) = 1$. Thus, $\eta(Z|X) = 0$ and $W(Z|X) = 1/2$, which completes the proof. □

A geometric interpretation of the Woods criterion can be given by noticing that the ratio in eq (7) corresponds to a tangent (see figure 5):

$$\begin{aligned} \frac{\sqrt{Var(Y_x)}}{E(Y_x)} &= \tan \theta_x , \\ W(Y|X) &= E_X(\tan \theta_x). \end{aligned}$$

5.2 Mutual information

We have seen that the correlation ratio is a measure which relates to the variance. Mutual information relates with another measure of randomness: entropy. Figures 6 and 7 qualitatively demonstrate the difference between entropy and variance. While entropy measures the average "surprise" provided by some random events, variance evaluates their average dispersion around a mean value. Entropy implicitly introduces a notion of predictability; variance implicitly introduces a notion of proximity.

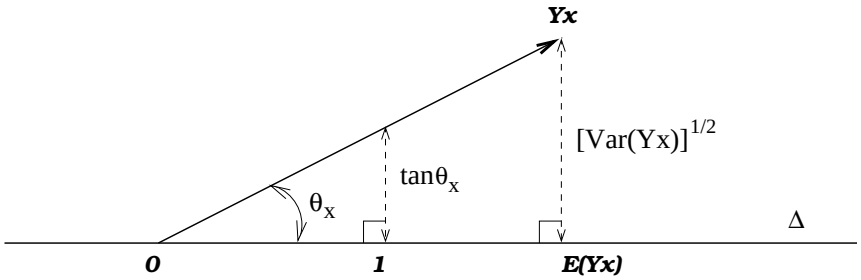


Figure 5: *Geometric interpretation of the Woods criterion. It is the average tangent between the conditional variables* Y_x *and the constant direction* Δ.

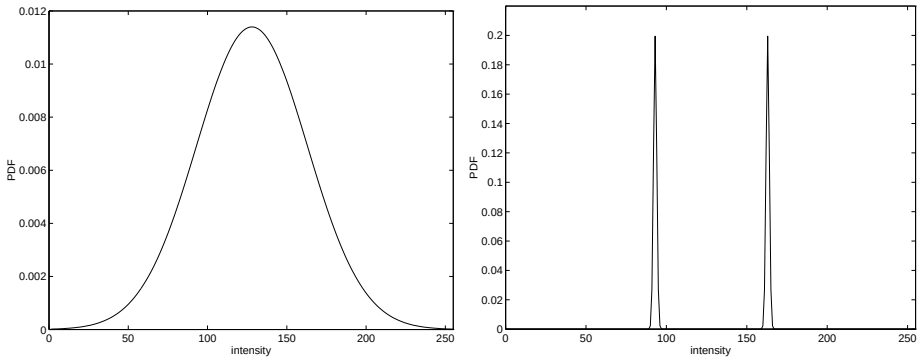


Figure 6: *These two distributions have the same variance, but not the same entropy.* $Var = 35^2$. *Left,* $Entropy = 7.17$. *Right,* $Entropy = 3.05$.

Given two variables X and Y, their mutual information can be expressed as the difference between an entropy term and a conditional entropy term [2, 6]:

$$I(X,Y) = H(Y) - E_X\left[H(Y|X=x)\right]. \tag{9}$$

This relation is often interpreted by stating that $I(X,Y)$ measures the uncertainty reduction on Y due to the knowledge of X (and the other way round). Equation (9) is structurally analogous to the total variance theorem:

$$Var\left[E(Y|X)\right] = Var(Y) - E_X\left[Var(Y|X=x)\right].$$

In other words, the variance of the conditional expectation, $Var\left[E(Y|X)\right]$, also measures some kind of uncertainty reduction on Y due to the knowledge of X. But this measure is made in the sense of variance and not entropy. Recall that the correlation ratio is a normalized version of $Var\left[E(Y|X)\right]$ (see section 3). However, an important difference between $I(X,Y)$ and $Var\left[E(Y|X)\right]$ is the symmetry property:

$$I(X,Y) = I(Y,X), \quad \text{while, in general,} \quad Var\left[E(Y|X)\right] \neq Var\left[E(X|Y)\right].$$

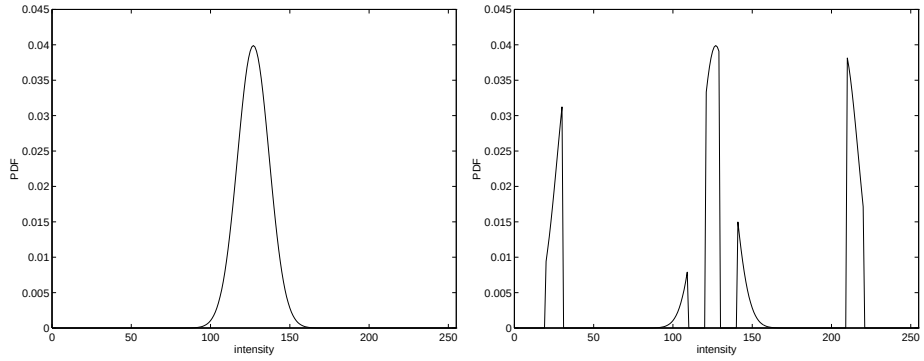


Figure 7: *These two distributions have the same entropy, but not the same variance.* $H = 5.37$. *Left,* $Var = 10^2$. *Right,* $Var = 68.01^2$.

The following theorem shows that this analogy is more than structural in the case of Gaussian couples.

Theorem 2 *Let* (X,Y) *be a pair of Gaussian continuous random variables. Then:*

(i) The correlation ratio is symmetrical with respect to X *and* Y *and is equal to the correlation coefficient:*

$$\eta(X|Y) = \eta(Y|X) = \rho(X,Y).$$

(ii) Mutual information[8] *and the correlation coefficient are related by a strictly increasing mapping:*

$$I(X,Y) = -\frac{1}{2}\ln\left[1-\rho(X,Y)\right].$$

[8] Here, we consider the continuous definition of mutual information which is equivalent to the discrete definition up to an infinite "constant" (see [2, 6]).

Proof Property (i) relies on a classical result: if (X, Y) is Gaussian, then $E(Y|X)$ is linear with respect to X [13, 16]. Let us now prove property (ii). The entropy of an n-variate Gaussian distribution, G, is given by [2]:

$$H(G) = \frac{1}{2} \ln \left[(2\pi e)^n \det \Sigma \right],$$

where Σ is the covariance matrix of G. Both X and Y are Gaussian since (X, Y) is a Gaussian pair. Using the identity

$$I(X, Y) = H(X) + H(Y) - H(X, Y),$$

we can express the mutual information between X and Y:

$$\begin{aligned} & I(X, Y) \\ &= \frac{1}{2} \ln \left[(2\pi e) Var(X) \right] + \frac{1}{2} \ln \left[(2\pi e) Var(Y) \right] - \frac{1}{2} \ln \left\{ (2\pi e)^2 [Var(X) Var(Y) - Cov(X, Y)^2] \right\}, \\ &= -\frac{1}{2} \ln \left[\frac{Var(X) Var(Y) - Cov(X, Y)^2}{Var(X) Var(Y)} \right], \\ &= -\frac{1}{2} \ln \left[1 - \rho(X, Y) \right], \end{aligned}$$

which completes the proof. □

This equivalence between mutual information and the correlation ratio in the Gaussian case has a purely theoretical flavor for us since medical images are never Gaussian (see for example [5]). Practically, maximizing the correlation ratio is not equivalent to maximizing mutual information.

5.3 Weighted neighbor likelihood

In [22], Viola already proposed performing registration by evaluating the degree of functional dependence between two images. This approach is very analogous to that we have proposed in section 3. First, a *weighted neighbor approximator* is used to estimate the Y image in terms of the X image. Second, a similarity measure is obtained by considering the estimation likelihood (under hypotheses we won't discuss here). We successively analyze these two steps.

5.3.1 Weighted neighbor approximator

We demonstrate the analogy of the weighted neighbor approximator with an estimation using the conditional expectation. Given two images X and Y, the weighted neighbor approximator F^* (to be compared with $\phi^* = E(Y|X)$) depends on a parameter ν according to:

$$\begin{aligned} F^*_\nu(x) &= \sum_\omega W(x, X(\omega))\, Y(\omega), \qquad (10) \\ \text{with} \quad W(x, x') &= \frac{g_\nu(x - x')}{\sum_\omega g_\nu(x - X(\omega))}, \end{aligned}$$

where g_ν denotes a Gaussian pdf[9] with mean 0 and standard deviation ν.

[9] In Viola's work, other types of density can be used. However, the Gaussian case can be generalized very easily.

Theorem 3 *Considering images X and Y as discrete random variables (see appendix A), F_ν^* is the conditional expectation of Y in terms of $X+N$, N being independent from (X,Y), and Gaussian with mean 0 and standard deviation ν.*

Proof Let X, Y, N be three L^2 variables so that N is Gaussian and is independent of (X,Y). Let $p(x,y)$ denote the joint pdf of (X,Y) and $g_\nu(x)$ denote the pdf of N. As N is independent of (X,Y), the pdf of $X+N$, $\tilde{p}(x)$, is the convolution of the marginal pdf of X, $p(x)$, with $g_\nu(x)$:

$$\tilde{p}(x) = p * g_\nu\,(x).$$

Equivalently, the joint pdf $\tilde{p}(x,y)$ of $(X+N,Y)$ is defined by:

$$\tilde{p}(x,y) = p(.,y) * g_\nu\,(x,y),$$

where the convolution only concerns the x coordinate.

Let us rewrite $F_\nu^*(x)$. Applying our definition of images as random variables, we can formulate eq (10) in terms of probabilities. Let C_Ω denote the cardinal of the state space Ω. We have:

$$\begin{aligned} F_\nu^*(x) &= C_\Omega \int W(x, X(\omega))\, Y(\omega)\, dPr, \\ &= C_\Omega \int\int W(x,x')\, y\, p(x',y)\, dx' dy, \\ &= C_\Omega \int \left(\int W(x,x')\, p(x',y)\, dx' \right) y\, dy. \end{aligned}$$

Now,

$$W(x,x') = \frac{g_\nu(x-x')}{C_\Omega \int g_\nu(x-x')\, p(x')\, dx'} = \frac{g_\nu(x-x')}{C_\Omega\, \tilde{p}(x)}. \tag{11}$$

Thus,

$$\int W(x,x')\, p(x',y)\, dx' = \frac{\tilde{p}(x,y)}{\tilde{p}(x)} = \tilde{p}(y|x).$$

And finally,

$$F_\nu^*(x) = \int \tilde{p}(y|x)\, y\, dy = E(Y|\,(X+N=x)).$$

Therefore, we have $F_\nu^* = E(Y|\,X+N)$, which completes the proof. □

We see that there exists a close relationship between the weighted neighbor approximator F_ν^* and the conditional expectation ϕ^*. In the case $\nu = 0$, $F_0^* = \phi^*$. Thus, F_ν^* is a generalization of the conditional expectation. The larger the parameter ν is, the more it imposes a constraint of smoothness to F_ν^*. If $\nu = 0$, there is no constraint.

Although such a constraint of smoothness seems intuitively reasonable, this is actually achieved by corrupting X with an additive Gaussian noise. Here, that sounds absurd because we have supposed that X is a discrete random variable. From our discrete point of view, it is clear that $F_\nu^*(X)$ is the best estimate of Y (in the sense of L^2) only if $\nu = 0$.

Nevertheless, considering non-null ν parameters might be relevant too. In fact, the addition of noise (ν parameter) is a way to define X as a continuous variable. This technique is known as the Parzen Window method [7]. Instead of directly computing the pdf of X by simple normalization of the histogram, one may smooth it in order to get a continuous pdf.

5.3.2 Likelihood

The similarity measure $L_\nu(X,Y)$ derived by Viola is the log likelihood of the weighted neighbor approximator (under some assumptions we won't discuss here [22]). It is given by

$$L_\nu(X,Y) = -k \sum_{\omega'} \left[\sum_{\omega} W(X(\omega), X(\omega')) \, (Y(\omega) - Y(\omega')) \right]^2 .$$

where k is a constant that plays no role in the optimization.

Is there a connection between $L_\nu(X,Y)$ and the correlation ratio? We can rewrite $L_\nu(X,Y)$ according to our formalism in terms of random variables (see appendix A):

$$\begin{aligned}
L_\nu(X,Y) &= -kC_\Omega^2 \int\int \left[\int\int W(x,x') \, (y-y') \, p(x',y') \, dx'dy' \right]^2 p(x,y) \, dxdy, \\
&= -k \int\int \left[\int\int \frac{g_\nu(x-x')}{\tilde{p}(x)} \, (y-y') \, p(x',y') \, dx'dy' \right]^2 p(x,y) \, dxdy, \\
&= -k \int\int \left[\int (y-y') \, \frac{\tilde{p}(x,y')}{\tilde{p}(x)} \, dy' \right]^2 p(x,y) \, dxdy, \\
&= -k \int\int \left[y - F_\nu^*(x) \right]^2 p(x,y) \, dxdy.
\end{aligned}$$

Let us consider the Gaussian variable N we have introduced in theorem 3. We have:

$$\begin{aligned}
L_\nu(X,Y) &= -k \int\int \left[y - E(Y \,|\, X+N=x) \right]^2 p(x,y) \, dxdy, \\
&= -k \, E_{X+N} \left[Var(Y \,|\, X+N=x) \right].
\end{aligned}$$

The case $\nu = 0$ corresponds to

$$L_0(X,Y) = -k \, E_X \left[Var(Y \,|\, X=x) \right].$$

It turns out that, if $\nu = 0$, maximizing the weighted neighbor likelihood is equivalent to minimizing the conditional variance $E_X \left[Var(Y|X=x) \right]$, that is the numerator in eq (8). However, the correlation ratio involves a division by $Var(Y)$, which plays a critical role in registration problems since it depends on the size of the images' overlapping region. If $Var(Y)$ is not taken into account, every transformation that completely disconnects images corresponds to a global optimum of the criterion. Notice that for exactly the same reasons, mutual information is preferred to conditional entropy.

6 Synthetic experiments

In this section, we propose to discuss the difference between the measures we talked about using two synthetic experiments. This is not intended to draw comparisons in real medical applications, but only to give an intuitive understanding. We use the following abbreviations: MI (Mutual Information), CR (Correlation Ratio), CC (the correlation coefficient), and OW (Opposite of Woods criterion[10]).

[10] Recall that the Woods criterion is assumed to be *minimal* at the registration position. This is why we consider its opposite which has to be maximized like MI, CR, and CC.

6.1 Experiment 1

We extracted a coronal slice from a 3-D brain MR stored with one byte per voxel (image X). We then artificially constructed a second image (image Y_1) using the following scheme:

For each voxel of coordinates (x, y),
  If $|X(x-6, y) - X(x, y)| < 18$ and $(x-6) \geq 0$, $\quad Y_1(x, y) := X(x-6, y)$
  Else, $\quad Y_1(x, y) := X(x, y)$

Applying this distortion, 62% of the voxels do not have the same intensity in image X and image Y_1. However, X and Y_1 intensities do not differ by more than 18 so that it is visually difficult to detect any difference between them (see figure 8). We computed several similarity measures between X and Y_1 for various horizontal translations of Y_1 ranging from -20 voxels to $+20$ voxels with a step of 0.1 (see figure 9). As intuitively expected, CR, OW, and CC reach their maximum for the null translation. But MI has an absolute maximum corresponding to a translation of $+6$ voxels and only a local maximum for the null translation.

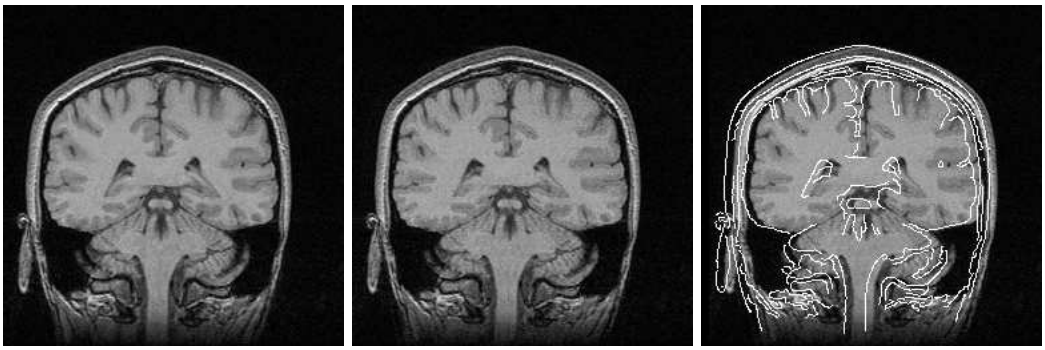


Figure 8: *Left, the reference image is a 2-D brain MR scan (image X). Middle, the floating image is the same scan that we have imperceptibly distorted (image Y_1). Right, the transformed image Y_1 corresponding to the absolute maximum of MI is superimposed with contours extracted from the image X: a registration error in the horizontal direction is visible.*

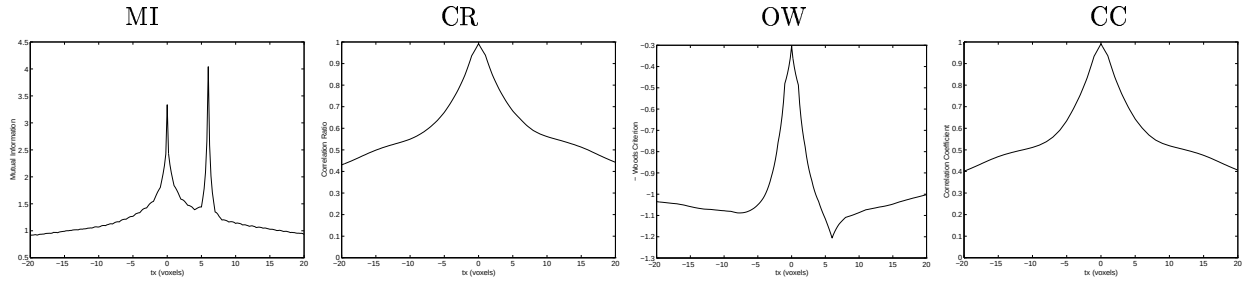


Figure 9: *Plot of four similarity measures versus horizontal translation in experiment 1. From left to right, MI, CR, OW, and CC.*

Looking at the joint pdf's represented in figure 10, one can explain this result. The left joint pdf corresponds to a null translation of Y_1. It is easy to guess a linear correlation between intensities. There are many outliers

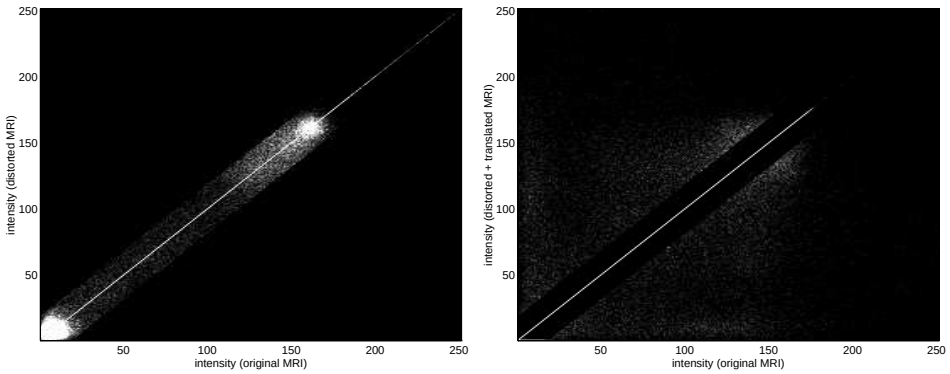


Figure 10: *Joint pdf of the original MR scan (image X) and its artificially distorted version (image Y_1). Left, no translation. Right, a 6 voxel horizontal translation corresponding to the absolute maximum of mutual information.*

but they are all gathered in a small region of the intensity space. In the right pdf, which corresponds to a translation of 6 voxels, there are less outliers but they cover a much larger region of the intensity space. MI "prefers" this second situation because it does not favour intensity clusters. In contrast, the 6 voxels translation is particularly discouraged by OW : it corresponds to a minimum. Indeed, in this situation, the conditional expectations m_i corresponding to low intensities in X are especially small with respect to their corresponding standard deviations σ_i. As a consequence, the ratios σ_i/m_i are large. This phenomenon does not occur with the correlation ratio since no division by the m_i's is involved. Finally, we note that CC behaves like CR. Indeed, in this case, the intensity correlation is approximately linear. However, these measures are generally not equivalent as is demonstrated in the following experiment.

6.2 Experiment 2

Taking the same reference slice as in experiment 1, we applied a quadratic intensity transformation to create a new image Y_2 (see figure 11):

For each voxel of coordinates (x, y),

$$Y_2(x,y) := \left[\bar{X}_1(x,y)^2 + 3.2\,\bar{X}_1(x,y) - 3538\right]/5$$

with $\bar{X}_1 = X_1 - 64$.

In fact, this corresponds to constructing an orthogonal vector to X in the space L_X (see section 3). Y_2 is a non-linear, non-monotonic function of X. As in experiment 1, we studied the behavior of MI, CR, OW, and NCC with respect to translations of Y_2. We computed several similarity measures for translations of Y_2 ranging from -5 voxels to $+5$ voxels with a step of 0.1 (see figure 12). MI, CR, and OW reach their maximum for the correct null translation. But CC is now *minimal* which indicates that it cannot align the images. Moreover, the attraction basin is very narrow for OW, which makes it difficult to maximize globally, and it is larger for CR than for MI. One must also notice that MI handles local maxima corresponding to integer translations. This is due to interpolation but this does not occur for CR, OW and CC.

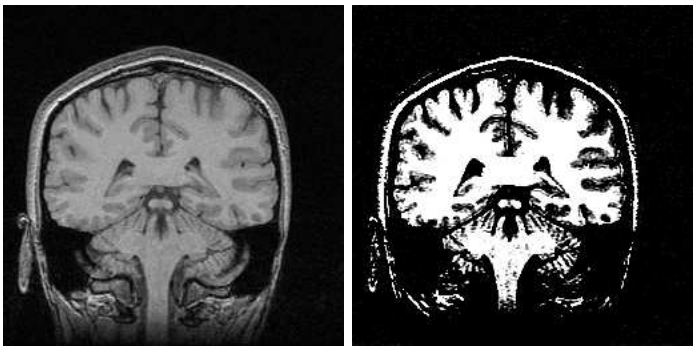


Figure 11: *Left, image X. Right, image Y_2 obtained by applying a quadratic intensity transformation to X.*

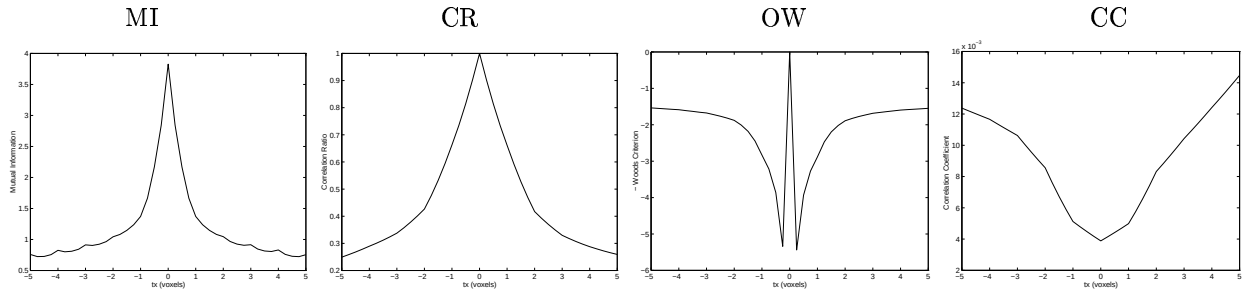


Figure 12: *Plot of four similarity measures versus horizontal translation in experiment 2. From left to right, MI, CR, OW and CC.*

7 Preliminary results

We now present three real registration experiments involving multimodal brain images of the same patient. Images are:

- MR, T1 weighted ($256 \times 256 \times 24$ voxels of $1.25\,mm \times 1.25\,mm \times 4\,mm$)
- MR, T2 weighted ($256 \times 256 \times 24$ voxels of $1.25\,mm \times 1.25\,mm \times 4\,mm$)
- CT ($512 \times 512 \times 29$ voxels of $0.65\,mm \times 0.65\,mm \times 4\,mm$)
- PET ($128 \times 128 \times 15$ voxels of $2.59\,mm \times 2.59\,mm \times 8\,mm$)

All images were stored with one byte per voxel. Geometrical distortions in MR images were *not* rectified. The gold standard transformations between each modality were known thanks to a prospective, marker-based registration method. However, the images we used all had traces of the markers removed.

We used the algorithm described in section 4 to perform registration between images. Experiments are summarized in the following table.

protocol	reference and template	floating and estimated	accuracy tables	figures
MR-T1 to MR-T2	MR-T2	MR-T1	1	13, 14, 17, 18
CT to MR-T1	MR-T1	CT	2	15, 19, 20
PET to MR-T1	MR-T1	PET	3	16, 21, 22

Notice that the distinction between the template and the estimated image only concerns CR and OW since they are non-symmetrical measures. No preprocessing of the images was done. Notably, we used OW without removing non-brain structures in the MR image as proposed in [26] for PET-MR registration.

7.1 Accuracy study

Subsampling of the floating image was tested at various resolutions. This was achieved by taking voxels at regular intervals with lengths f_x, f_y and f_z in the x, y and z direction, respectively. No filtering was applied. As the images we used had few slices, we did not subsample with respect to the z direction. Notice that subsampling implies a loss of information but, in counterparts, the CPU time required for registration is divided by $f_x.f_y.f_z$ with respect to the maximal resolution. In all experiments, the transformation was initialized as the identity.

The registration errors were computed according to the marker-based transformations. Thus, we supposed that these transformations were much more accurate than the ones found by our automatic registration. We selected eight "typical" points in the floating image, that is points with coordinates

$$\frac{1+\alpha_i}{3}\,d_x,\ \frac{1+\beta_i}{3}\,d_y,\ \frac{1+\gamma_i}{3}\,d_z,$$

where $\alpha_i = \{0,1\}$, $\beta_i = \{0,1\}$, $\alpha_i = \{0,1\}$, $\gamma_i = 0,1$, and d_x, d_y, d_z are the dimensions of the image. For each of the so defined points P_i, the error was computed as follows:

$$\epsilon_i = \|(T_g - T_r)(P_i)\|_{euclidean},$$

where T_g and T_r denote respectively the genuine marker-based transformation and the one found by automatic registration.

Results for each protocol are shown in tables 1, 2, and 3. In each case, we only retained the similarity measures which gave sensible results, that is:

- MR-T1 to MR-T2 registration: MI, CR, OW, and CC
- CT to MR-T1 registration: MI and CR
- PET to MR-T1 registration: MI, CR, and OW

Some aberrant solutions are given. They typically correspond to the cases where the algorithm converged towards a local but non-global maximum.

Table 1: *Maximum and mean errors for MR-T1 to MR-T2 registration based on positions of stereotaxic markers.*

Sub. factors	Measure	Max error (mm)	Mean error (mm)
$(1,1,1)$	MI	2.89	1.58
	CR	4.35	2.36
	OW	13.63	9.45
	CC	6.49	3.79
$(4,4,1)$	MI	19.00	10.99
	CR	3.29	1.77
	OW	5.46	2.92
	CC	5.33	3.17
$(8,8,1)$	MI	21.32	17.98
	CR	5.46	4.39
	OW	12.70	7.81
	CC	15.16	8.95

Table 2: *Maximum and mean errors for CT to MR-T1 registration based on positions of stereotaxic markers.*

Sub. factors	Measure	Max error (mm)	Mean error (mm)
$(1,1,1)$	MI	1.74	1.15
	CR	2.04	1.21
$(2,2,1)$	MI	7.21	4.25
	CR	5.48	3.03
$(4,4,1)$	MI	33.53	23.91
	CR	5.53	3.09
$(8,8,1)$	MI	*	*
	CR	9.65	5.21

7.2 Robustness study

For each protocol, we studied the behavior of MI, CR, OW, and CC in the neighborhood of the genuine transformation. One parameter was expanded while the five others were fixed. In each case, several resolutions were tested with subsampling of the floating image. Results are shown in figures 13, 14 (MR-T1-to-MR-T2 registration), 15 (CT-to-MR-T1 registration) and 16 (PET-to-MR-T1 registration).

They qualitatively demonstrate that CR was less sensitive to subsampling than MI and OW. Subsampling typically introduces local maxima so that the global maximum becomes difficult to track. Another effect is the displacement of the global maximum. This was very perceptible for OW in every protocol, for CC in CT-to-T1

Table 3: *Maximum and mean errors for PET to MR-T1 registration based on positions of stereotaxic markers.*

Sub. factors	Measure	Max error (mm)	Mean error (mm)
$(1,1,1)$	MI	4.40	3.38
	CR	3.86	2.95
	OW	9.09	5.22
$(2,2,1)$	MI	136.31	97.88
	CR	4.96	3.33
	OW	10.25	6.84

and PET-to-T1 registration, and for MI in PET-to-T1 registration. CR demonstrated good stability in all experiments.

7.3 Synthesis

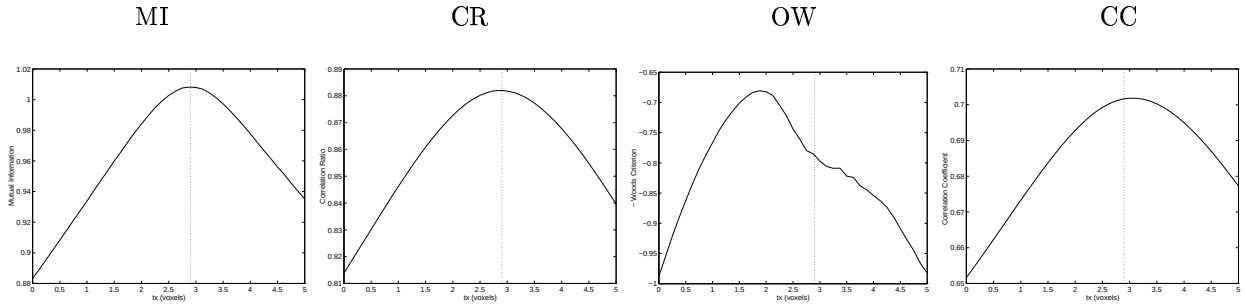
In these experiments, MI and CR demonstrated comparable accuracy levels at maximal resolution, for each protocol. MI gave the best results for T1-to-T2 and CT-to-T1 registration but CR was better for PET-to-T1 registration. OW demonstrated much weaker performances while CC was irrelevant, except for T1-to-T2 registration. Moreover, CR allowed good registration at relatively low resolutions in each protocol. In contrast, subsampling drastically affected the performances of MI and OW. However, we are aware that the behavior of every criterion may strongly rely on the interpolation method which is used. There is no doubt that this issue can be improved with no (or little) additional computing time [8].

The case of PET images is particular because they are much more distorted than MR or CT. This might explain why mutual information lacks robustness in PET-T1 registration. It is generally admitted that today the Woods criterion is the best similarity measure for this specific problem. In some way, our results corroborate this observation, suggesting that the taking into account of nearby intensities might be crucial for PET images. Mutual information seems to be better adapted for high-resolution images. For example, a good strategy for CT-MR registration could be to use the correlation ratio for a quick guessing of the correct transformation (using subsampling), and then mutual information for probably more accurate alignment.

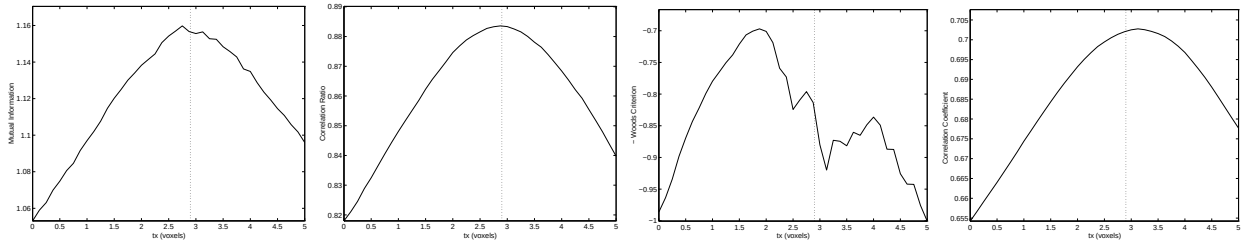
The discrepancies found experimentally between the correlation ratio and the Woods criterion are surprising since these two measures are formally based on similar considerations (see section 5). It seems that the correlation ratio gives not only a theoretical justification to the Woods criterion but also perceptible practical improvements. We also verified that CC was not equivalent to CR in practice. Therefore, generalizing the notion of linear correlation is interesting.

Finally, our experiments tend to show that assuming a functional correlation between certain multimodal images is not critical. Even if this implies an approximation (see e.g. [24] for a discussion in the CT-MR case), the point is to know whether this is acceptable or not. A preprocessing step might enforce the validity of such a hypothesis. In the case of CT-MR registration, van den Elsen et al. [20] have proposed a simple intensity

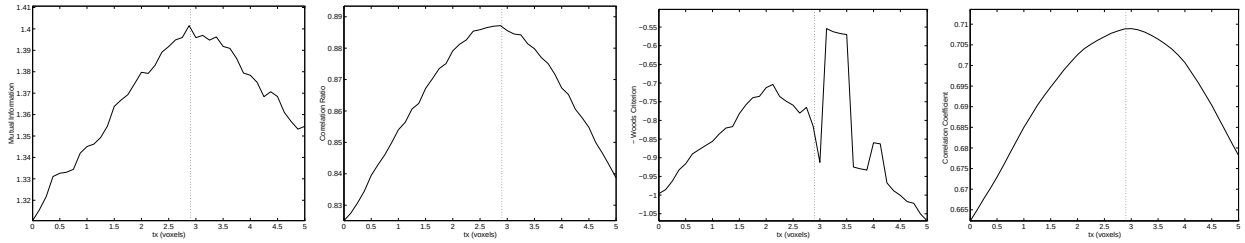
mapping to the original CT image so that bone and air appear in the same intensity range as is the case in MR images. Then, low intensities in MR (air and bone) may project to clustered intensities in CT. In PET-MR registration, Woods et al. [26] remove beforehand non-brain structures from the MR image for analogous reasons. Such treatments might improve the correlation ratio accuracy.



(a) Maximal resolution



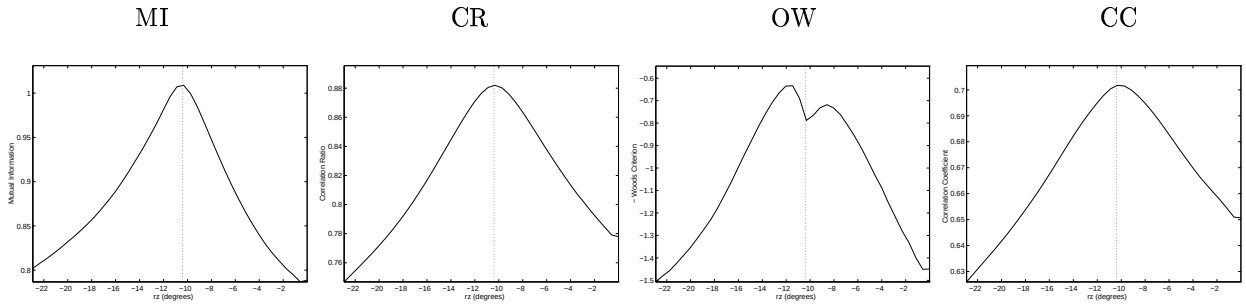
(b) Subsampling of MR-T1 image by factors $4 \times 4 \times 1$



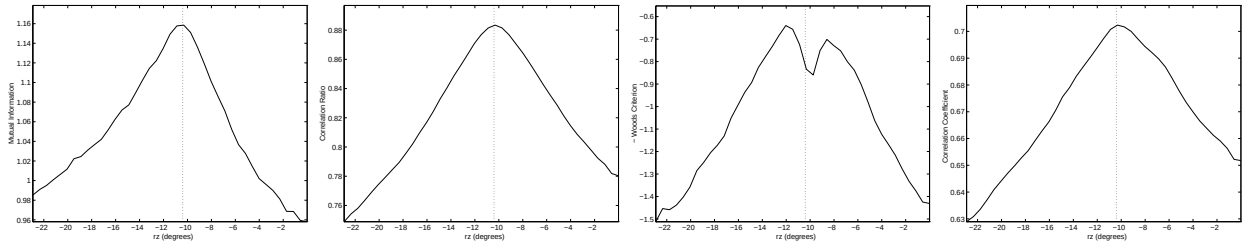
(c) Subsampling of MR-T1 image by factors $8 \times 8 \times 1$

Figure 13:

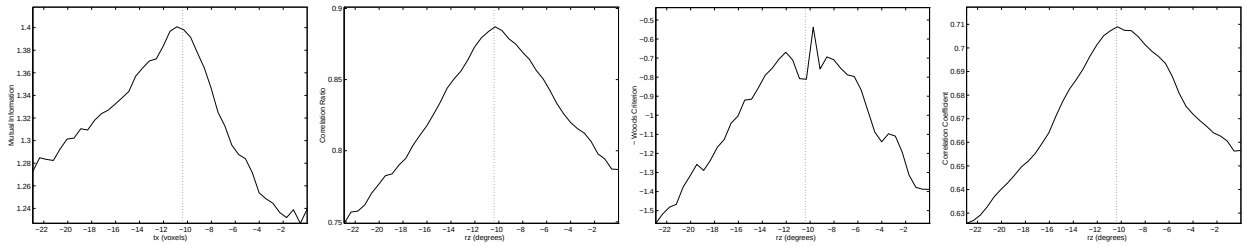
MR-T1 to MR-T2 registration experiment. Plots of MI, CR, OW, and CC vs. horizontal translation in the neighborhood of the gold standard transformation (gold standard parameter $t_x \approx 2.9mm$*).*



(a) Maximal resolution



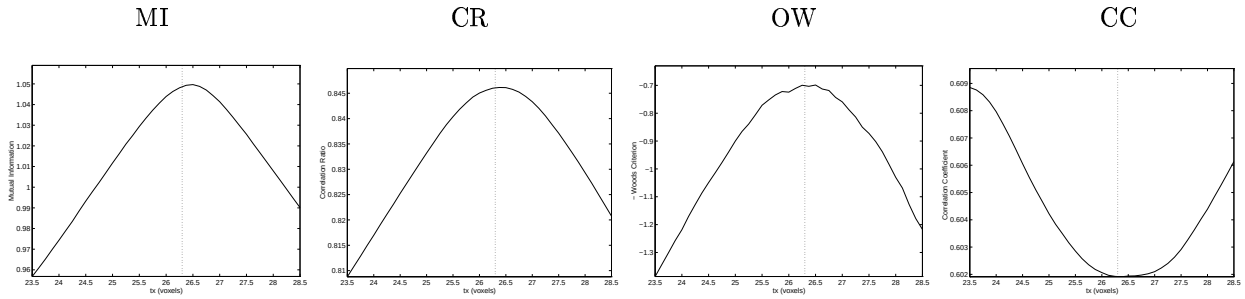
(b) Subsampling of MR-T1 image by factors $4 \times 4 \times 1$



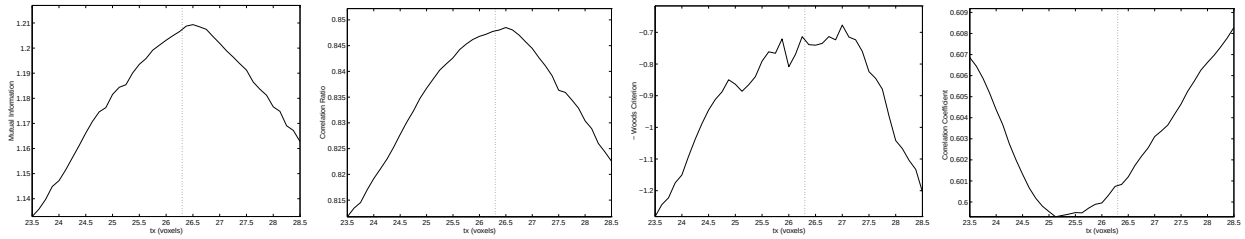
(c) Subsampling of MR-T1 image by factors $8 \times 8 \times 1$

Figure 14:

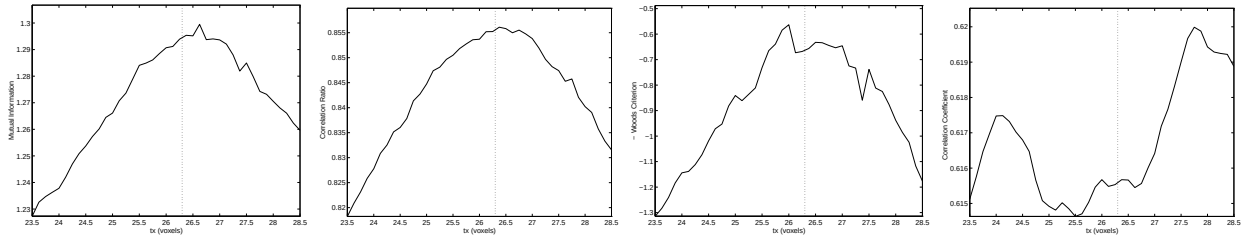
MR-T1 to MR-T2 registration experiment. Plots of MI, CR, OW, and CC vs. axial rotation in the neighborhood of the gold standard transformation (gold standard parameter $r_z \approx -10.4\,°$*).*



(a) Subsampling of CT image by factors $2 \times 2 \times 1$



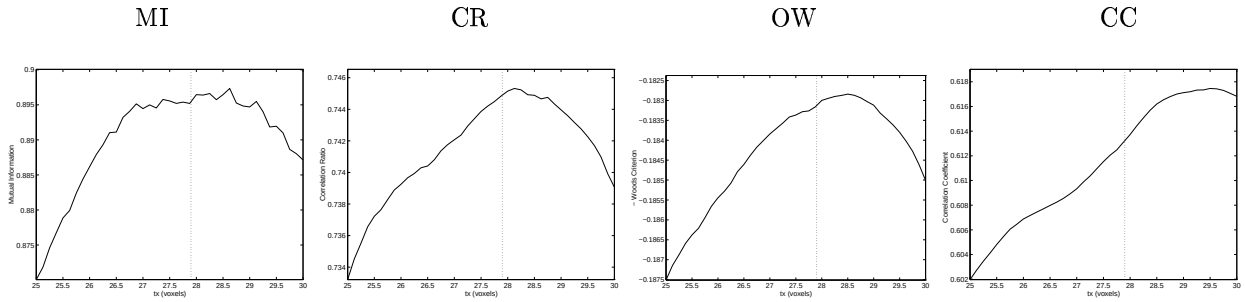
(b) Subsampling of CT image by factors $8 \times 8 \times 1$



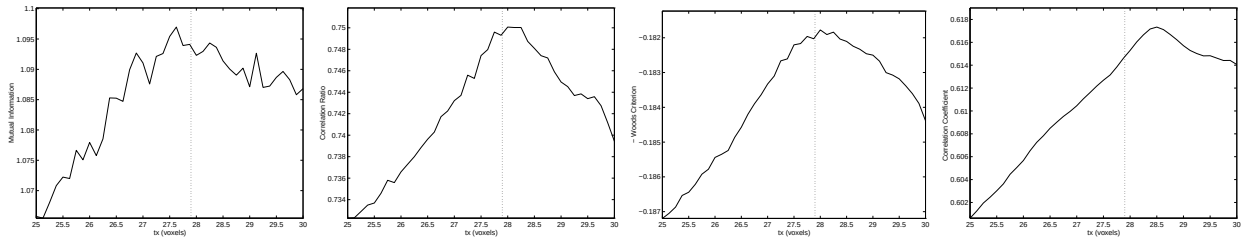
(c) Subsampling of CT image by factors $12 \times 12 \times 1$

Figure 15:

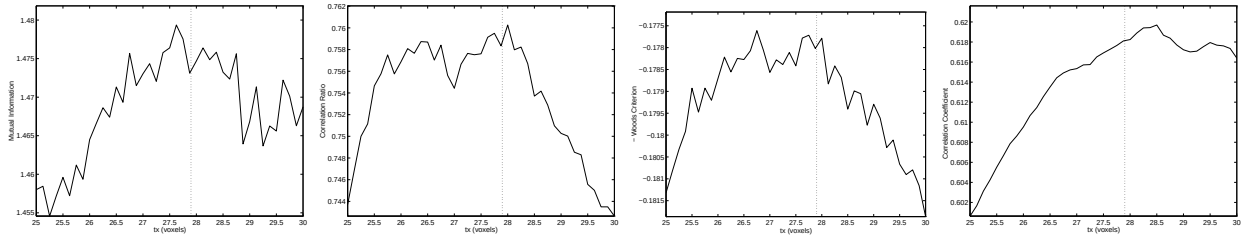
CT to MR-T1 registration experiment. Plots of MI, CR, OW, and CC vs. horizontal translation in the neighborhood of the gold standard transformation (gold standard parameter $t_x \approx 26.3mm$).



(a) Maximal resolution



(b) Subsampling of PET image by factors $2 \times 2 \times 1$



(c) Subsampling of PET image by factors $4 \times 4 \times 1$

Figure 16:

PET to MR-T1 registration experiment. Plots of MI, CR, OW, and CC vs. horizontal translation in the neighborhood of the gold standard transformation (gold standard parameter $t_x \approx 27.9mm$*).*

8 Conclusion

We have proposed to use the *correlation ratio* as a similarity measure for multimodal image registration. Its derivation has been shown to be consistent if a functional dependence is assumed between the image intensities. In other words, our method applies if one image can be considered as a good model of the other. As a consequence, the images play different roles in the registration process. An ideal model image would be an accurate, multi-channel segmentation of the tissues. However, the method demonstrated good performances when choosing a real image as a model. This enables us to use it without any pre-processing step.

We have pointed out the difficulty of designing a similarity measure that copes with large dissimilarities of images while taking into account the spatial information provided by similar-in-value intensities (which is a smoothness prior). Thus, the correlation ratio may be seen as a compromise between these two requirements. Furthermore, we have shown that the framework of L^2 random variables provides geometric interpretations of some existing similarity measures.

The registration experiments we made with MR, CT, and PET images yielded encouraging results. The correlation ratio gave performances that were comparable to mutual information in each case, and much better than the Woods criterion and the correlation coefficient. Our results suggest that a measure based on standard image statistics (means and variances) is not necessarily less accurate than an information-theoretic one. As we dealt with different types of images, we believe that the correlation ratio might be useful in a wide class of registration problems. One of its main interests is a great robustness at low resolutions, enabling considerable reduction of computing time.

Further research is needed in order to better investigate issues such as interpolation and maximization. This may improve the accuracy and robustness of the method. Future work must also consist of evaluating and validating the method on a larger database.

Acknowledgments

The images and the standard transformations we used in section 7 were provided as part of the project, "Evaluation of Retrospective Image Registration", National Institutes of Health, Project Number 1 R01 NS33926-01, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN.

Many thanks to Frederik Maes, Jean-Pierre Nadal, and Christophoros Nikou for fruitful discussion and to Janet Bertot for the proofreading of this article.

A Mathematical definition of random variables

We recall the main definitions and properties of random variables. The term *measure* will be used in its mathematical sense, which must not be confused with the notion of similarity measure.

A.1 Measurable space

Let Ω be a set. A *tribe* of Ω is a family $\mathcal{T}$ of subsets of Ω so that, $\Omega \in \mathcal{T}$, and $\mathcal{T}$ is stable by differences and countable unions. If provided with a tribe, Ω is said to be *measurable* and denoted $(\Omega, \mathcal{T})$.

A.2 Measurable mapping

Now let $(\Omega, \mathcal{T})$ and $(\Omega', \mathcal{T}')$ be two measurable spaces. A mapping $f : (\Omega, \mathcal{T}) \to (\Omega', \mathcal{T}')$ is *measurable* if:

$$f^{-1}(\mathcal{T}') \subset \mathcal{T}.$$

A.3 Measure

A *measure* is any mapping μ from a tribe $\mathcal{T}$ into $\overline{\mathbb{R}}_+$ that is not identically equal to $+\infty$ and which verifies the following property: for all series $(A_n)_{n\in\mathbb{N}} \in \mathcal{T}^{\mathbb{N}}$ such that $\forall (p,q) \quad A_p \cap \mathcal{A}_q = \emptyset$,

$$\mu\left(\cup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

A measure, Pr, is called a probability if it is normalized ($\Pr(\Omega) = 1$). A measurable space $(\Omega, \mathcal{T})$ is called a *probability space* when provided with a probability. It is denoted $(\Omega, \mathcal{T}, \Pr)$.

A.4 Random variable

These preliminary definitions allow us to define random variables. Given a probability space $(\Omega, \mathcal{T}, \Pr)$ (the *state* space) and a measurable space $(\mathcal{A}, \mathcal{T}')$ (the *sample* space), a random variable is a *measurable* mapping from Ω onto $\mathcal{A}$.

In this work, we apply this definition to images in the following way. An image X is a mapping from a finite grid of voxels Ω onto a finite alphabet of intensities $\mathcal{A}$. Typically, $\mathcal{A} = \{0, 1, 2, \ldots, 255\}$. Choosing the maximal tribe $\mathcal{T}$ and the uniform probability Pr on Ω, and the maximal tribe $\mathcal{T}'$ on $\mathcal{A}$, X automatically verifies the condition of measurability so that it can be considered as a random variable from Ω onto $\mathcal{A}$. Notice that according to this definition, images are *discrete* random variables. A continuous approach has been proposed in [22], using Parzen density estimates.

A.5 Integration formula

Given a random variable X, its expectation is defined by:

$$\forall \phi : \mathcal{A} \to \mathcal{A}, \quad E\left[\phi(X)\right] = \int_\Omega \phi\left[X(\omega)\right] \, dPr.$$

Let us now suppose that $\mathcal{A} = \mathbb{R}$. Then, the probability density function (pdf) of X is defined as the only function $p : \mathbb{R} \to \mathbb{R}^+$ such that:

$$\forall B, \text{ borelian}, \quad \Pr\left[X^{-1}(B)\right] \stackrel{def}{=} \int_\Omega X^{-1}(\omega) \, dPr = \int_B p(x) \, dx.$$

One shows the following integration formula. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a measurable function. We have:

$$E\left[\phi(X)\right] = \int_\Omega \phi\left[X(\omega)\right] \, dPr = \int_{\mathbb{R}} \phi(x) \, p(x) \, dx.$$

B Optimal estimation in L^2

We prove that the conditional expectation is the best estimate in the sense of the L^2 norm. Let X and Y be two L^2 random variables with joint pdf $p(x,y)$, and respective marginal pdf's $p(x)$ and $p(y)$. Let H_x be the space of functions which are square integrable with respect to the density measure $p(x)$:

$$H_x = \left\{ \phi : \mathbb{R} \to \mathbb{R} \;\middle|\; \int_{\mathbb{R}} \phi^2(x)\, p(x)\, dx \; < +\infty \right\}.$$

H_x is an Hilbert space provided with the dot product:

$$\langle \phi, \psi \rangle = \int_{\mathbb{R}} \phi(x)\, \psi(x)\, p(x)\, dx.$$

Notice that H_x is isomorphic to the space L_x we have introduced in section 3. We want to minimize on H_x:

$$J(\phi) = \|Y - \phi(X)\|_2^2 = \int_{\mathbb{R}} \int_{\mathbb{R}} [y - \phi(x)]^2 \; p(x,y)\, dxdy. \tag{12}$$

J is clearly convex. Therefore, minimizing eq (12) is equivalent to cancelling the gradient of J, i.e. find ϕ^* such that $\nabla J(\phi^*) = 0$. We express the directional derivative of J with respect to ϕ:

$$\begin{aligned}
\frac{dJ}{d\phi}(\phi, f) &= -2 \int_{\mathbb{R}} \int_{\mathbb{R}} [y - \phi(x)] \; f(x)\, p(x,y)\, dxdy, \\
&= -2 \int_{\mathbb{R}} \left[\left(\int_{\mathbb{R}} y\, p(y|x)\, dy \right) - \phi(x) \right] \; f(x)\, p(x)\, dx, \\
&= 2 \left\langle \phi - \int_{\mathbb{R}} y\, p(y|x)\, dy \;,\; f \right\rangle_{H_x}.
\end{aligned}$$

Thus, the gradient of J is:

$$\nabla J(\phi) = 2 \left(\phi - \int_{\mathbb{R}} y\, p(y|x)\, dy \right).$$

It is null if, and only if,

$$\phi(x) = \int_{\mathbb{R}} y\, p(y|x)\, dy = E(Y|x).$$

C Unbiased estimation

Given two variables X and Y, an estimate $\phi(X)$ of Y is unbiased if:

$$E[\phi(X)] = E(Y).$$

Let us prove that the conditional expectation is unbiased. By definition,

$$E(Y|X) = \int_{\mathbb{R}} y\, p(y|x)\, dy.$$

Thus,

$$\begin{aligned}
E[E(Y|X)] &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y\, p(y|x)\, dy \right) \; p(x)\, dx, \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} y\, p(x,y)\, dydx, \\
&= E(Y).
\end{aligned}$$

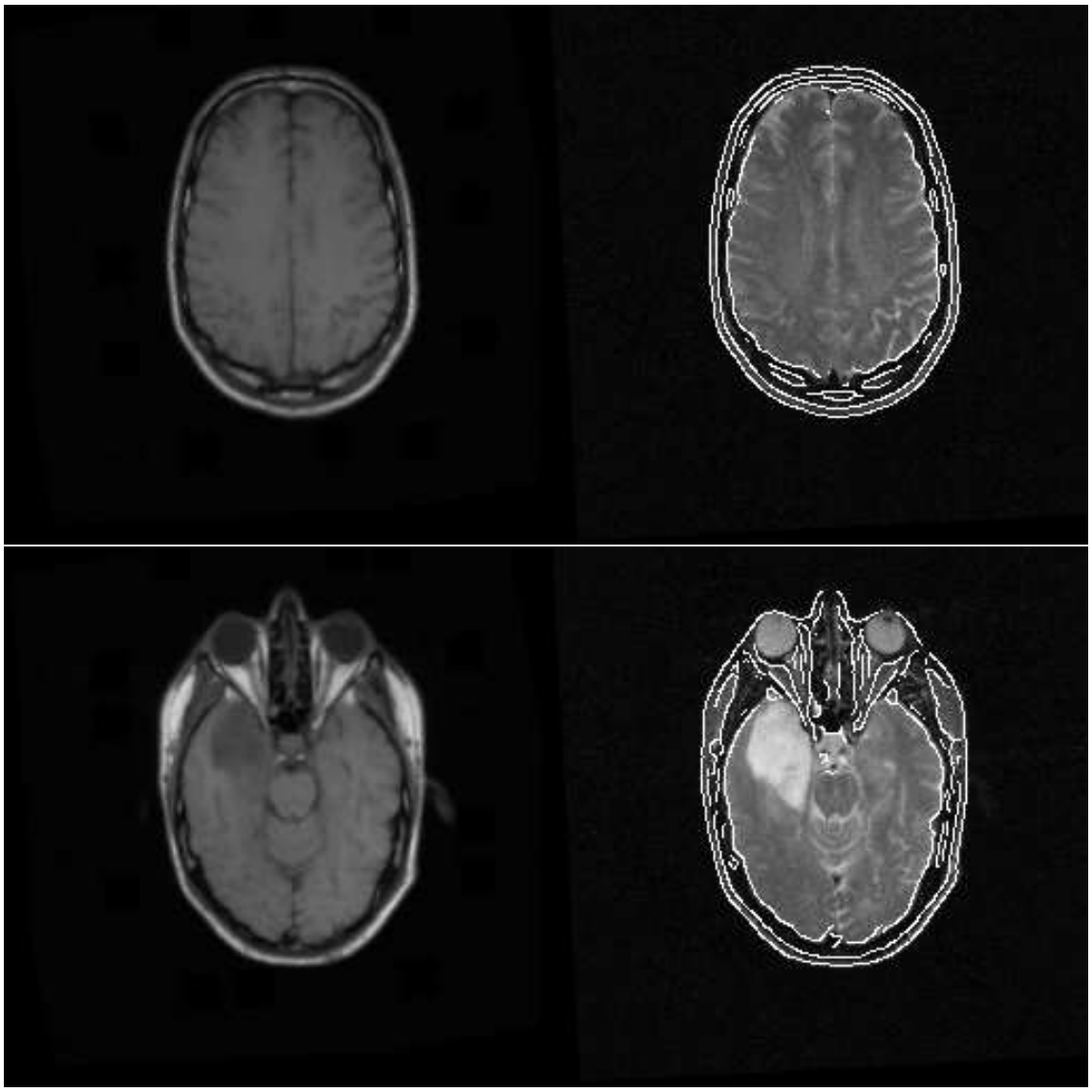


Figure 17: *MR-T1 to MR-T2 registration by maximization of CR (with* $(4 \times 4 \times 1)$ *subsampling of the MR-T1 image). Left, MR-T2 image. Right, registered MR-T1 image with MR-T2 contours superimposed.*

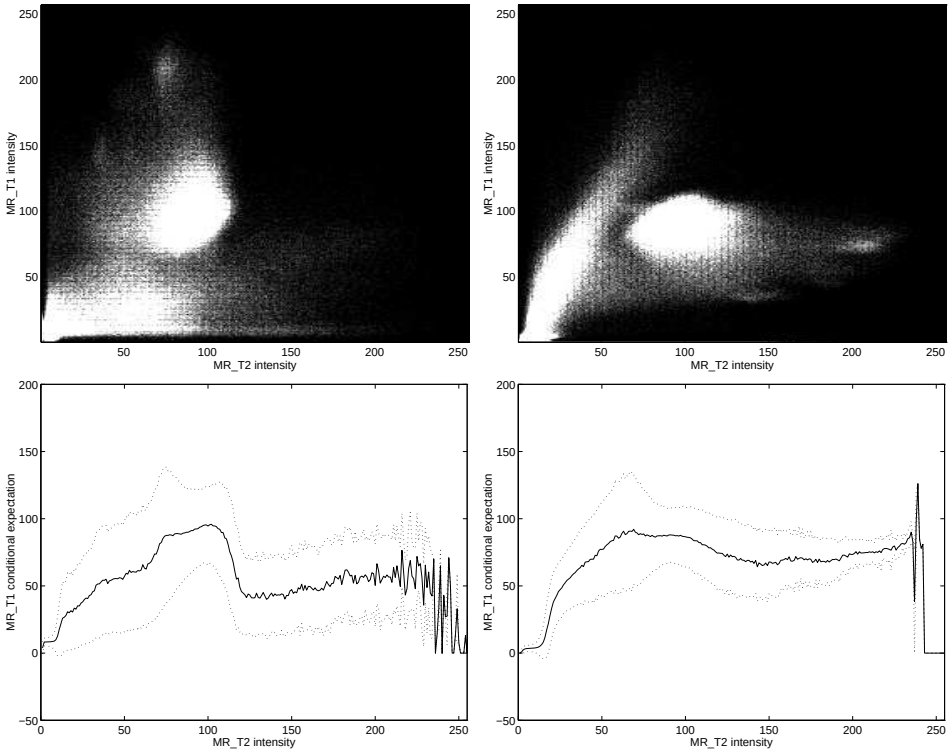


Figure 18: *Top, joint pdf of MR-T2 and MR-T1 brain images. Left, slight misalignment (CR≈ 0.74). Right, aligned case (CR≈ 0.88). Bottom, plots of the corresponding conditional expectation of MR-T1 in terms of MR-T2 (solid line), surrounded by conditional variance plots (dot lines).*

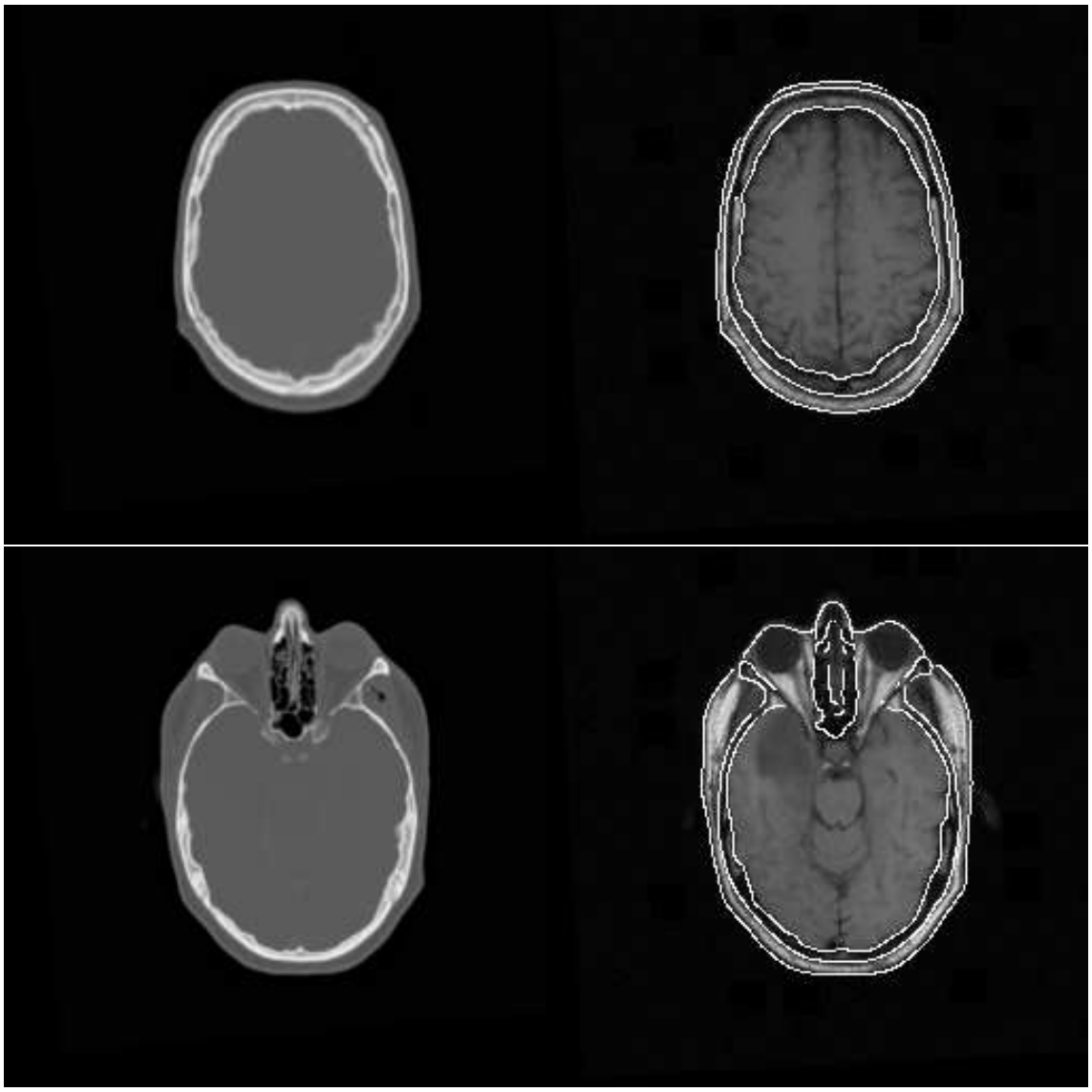


Figure 19: *CT to MR-T1 registration by maximization of CR (with* $(4 \times 4 \times 1)$ *subsampling of the CT image). Left, registered CT image. Right, MR-T1 images with CT contours superimposed.*

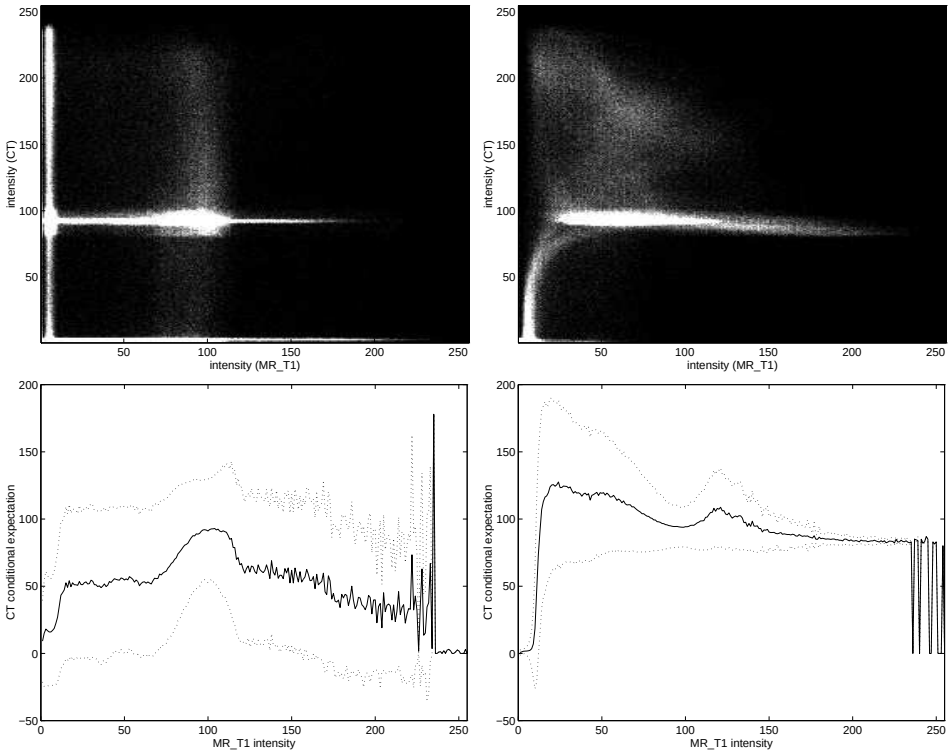


Figure 20: *Top, joint pdf of MR-T1 and CT brain images. Left, large misalignment (CR≈ 0.24). Right, aligned case (CR≈ 0.84). Bottom, plots of the corresponding conditional expectation of CT in terms of MR-T1 (solid line), surrounded by conditional variance plots (dot lines).*

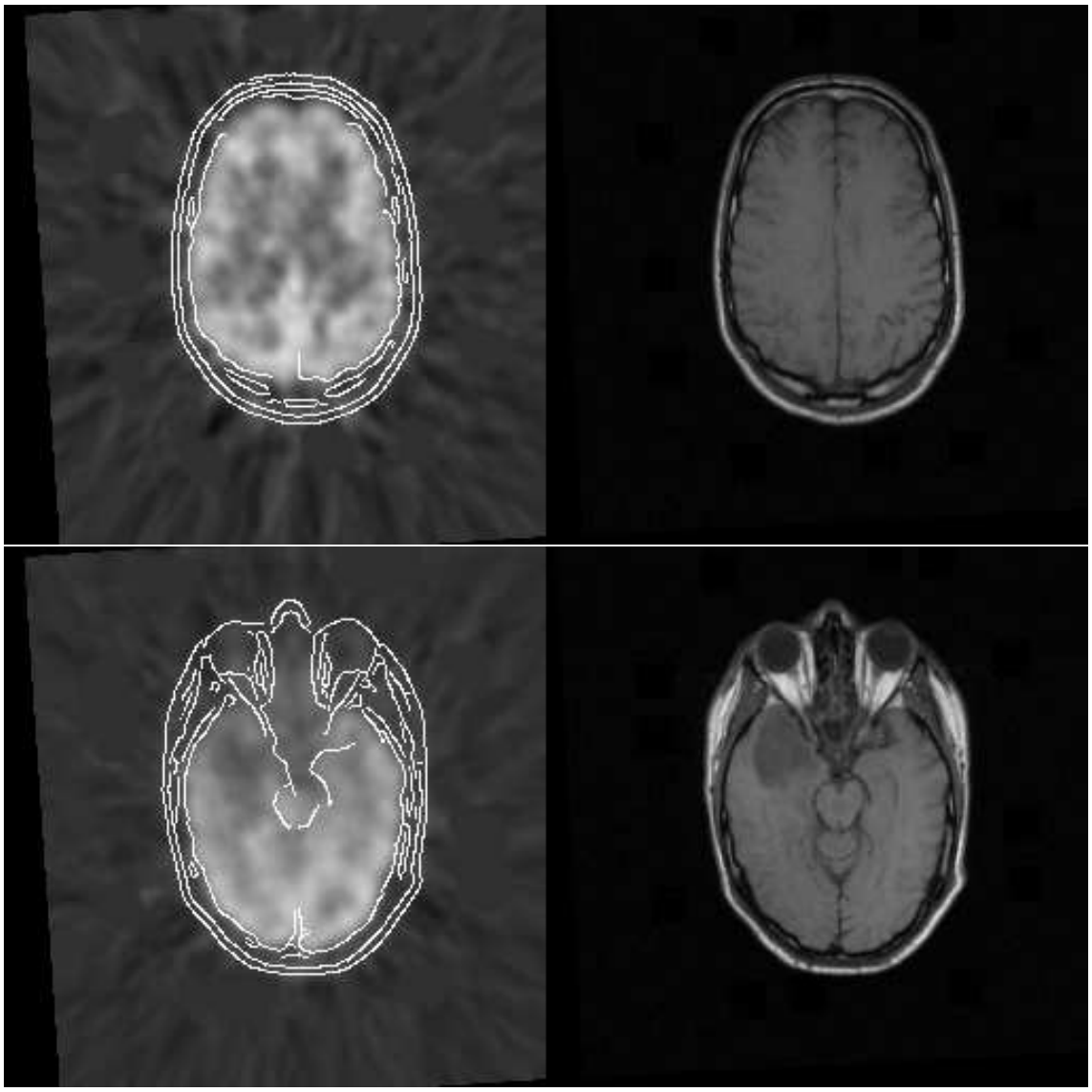


Figure 21: *PET to MR-T1 registration by maximization of CR (no subsampling). Left, registered PET image with MR-T1 contours superimposed. Right, MR-T1 image.*

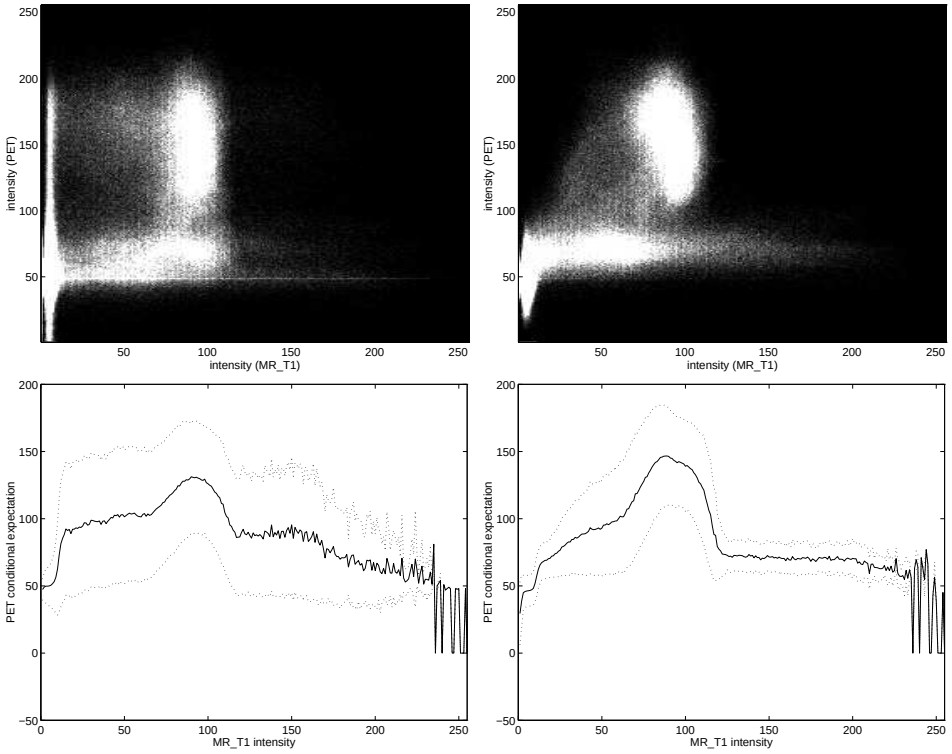


Figure 22: *Top, joint pdf of MR-T1 and PET brain images. Left, unaligned case (CR≈ 0.48). Right, aligned case (CR≈ 0.76). Bottom, plots of the corresponding conditional expectation of PET in terms of MR-T1 (solid line), surrounded by conditional variance plots (dot lines).*

References

[1] N. Ayache. *Artificial Vision for Mobile Robots-Stereo-Vision and Multisensor Perception.* MIT Press, Cambridge, MA, 1991.

[2] R. E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley Pub. Comp., 1987.

[3] M. Bro-Nielsen. Rigid Registration of CT, MR and Cryosection Images Using a GLCM Framework. *CVRMed-MRCAS'97*, pages 171–180, March 1997.

[4] L. G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[5] M. E. Brummer, R. M. Mersereau, R. L. Eisner, and R. R. J. Lewine. Automatic detection of brain contours in MRI data sets. *IEEE Trans. on Medical Imaging*, 12(2):153–166, June 1993.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, inc., 1991.

[7] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Jonh Wiley & Sons, 1973.

[8] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality Image Registration by Maximization of Mutual Information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[9] J. B. A. Maintz and M. A. Viergever. A survey of medical image registration. *MedIA*, 2(1):1–36, 1998.

[10] G. Malandain, S. Fernández-Vidal, and J.C. Rocchisani. Improving registration of 3-D images using a mechanical based method . *ECCV'94*, pages 131–136, May 1994.

[11] C. R. Meyer, J. L. Boes, B. Kim, P. H. Bland, K. R. Zasadny, P. V. Kison, K. Koral, K. A. Frey, and R. L. Wahl. Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate warped geometric deformations. *Medical Image Analysis*, 1(3):195–206, 1996/7.

[12] C. Nikou, F. Heitz, J.-P. Armspach, and I.-J. Namer. Single and multimodal subvoxel registration of dissimilar medical images using robust similarity measures. *SPIE Conference on Medical Imaging*, February 1998.

[13] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, Inc., third edition, 1991.

[14] X. Pennec and J.P. Thirion. A framework for uncertainty and validation of 3D registration methods based on points and frames. *Int. Journal of Computer Vision*, 25(3):203–229, 1997.

[15] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipies in C.* Cambrige Univ. Press, 2nd edition, 1992.

[16] G. Saporta. *Probabilités, analyse des données et statistique.* Editions Technip, Paris, 1990.

[17] C. Studholme, D. L. G. Hill, and D. J. Hawkes. Automated 3-D registration of MR and CT images of the head. *Medical Image Analysis*, 1(2):163–175, 1996.

[18] G. Subsol, J. P. Thirion, and N. Ayache. A General Scheme for Automatically Building 3D Morphometric Anatomical Atlases: application to a Skull Atlas. *Medical Image Analysis*, 1998. Accepted for publication.

[19] J.P. Thirion. Fast non-rigid matching of 3-D medical images. *Medical Robotics and Computer Aided Surgery*, pages 47–54, November 1995.

[20] P. A. van den Elsen, E.-J. D. Pol, T. S. Sumanaweera, P. F. Hemler, S. Napel, and J. R. Adler. Grey value correlation techniques for automatic matching of CT and MR brain and spine images. *Proc. Visualization in Biomed. Comp.*, 2359:227–237, October 1994.

[21] P.A. van den Elsen, E.J.D. Pol, and M.A. Viergever. Medical image matching - a review with classification. *IEEE Engineering in Medicine and Biology*, 12(4):26–39, march 1993.

[22] P. Viola. *Alignment by Maximization of Mutual Information.* PhD thesis, M.I.T. Artificial Intelligence Laboratory, 1995. also A.I.T.R. No. 1548, available at ftp://publications.ai.mit.edu.

[23] P. Viola and W. M. Wells. Alignment by Maximization of Mutual Information. *Intern. J. of Comp. Vision*, 24(2):137–154, 1997.

[24] W. M. Wells, P. Viola, H. Atsumi, and S. Nakajima. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–51, 1996.

[25] R. P. Woods, S. R. Cherry, and J. C. Mazziotta. Rapid Automated Algorithm for Aligning and Reslicing PET Images. *Journal of Comp. Assist. Tomography*, 16(4):620–633, 1992.

[26] R. P. Woods, J. C. Mazziotta, and S. R. Cherry. MRI-PET Registration with Automated Algorithm. *Journal of Comp. Assist. Tomography*, 17(4):536–546, 1993.

Unité de recherche INRIA Sophia Antipolis

2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique

615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)

Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)

Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot St Martin (France)

Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Éditeur

INRIA - Domaine de Voluceau - Rocquencourt, B.P. 105 - 78153 Le Chesnay Cedex (France)

http://www.inria.fr

ISSN 0249-6399