Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding 112 (2008) 160–172
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Image registration based on kernel-predictability ☆

Héctor Fernando Gómez-García [a,b,*], José L. Marroquín [a], Johan Van Horebeek [a]

[a] Center for Research in Mathematics (CIMAT), Department of Computer Science, Apartado Postal 402, C.P. 36000 Guanajuato, Gto, Mexico
[b] Department of Basic Sciences and Engineering, Universidad del Caribe, C.P. 77528, Cancún Q, Roo, Mexico

Received 8 August 2006; accepted 8 February 2008
Available online 15 February 2008

Abstract

In this work, a new similarity measure between images is presented, which is based on the concept of predictability of random variables evaluated through kernel functions. Image registration is achieved maximizing this measure, analogously to registration methods based on entropy, like mutual information and normalized mutual information. Compared experimentally with these methods in different problems, our proposal exhibits a more robust performance specially for problems involving large transformations and in cases where the registration is done using a small number of samples, such as in nonparametric registration.

Keywords: Multimodal image registration; Parametric and nonparametric transformations; Gini entropy; Information measures

1. Introduction

Due to its wide range of applications, image registration is a problem that has been largely explored (see [8,17,12] and references contained there in). It has become a fundamental task in many important fields such as robot vision and medical image processing, among others. Given a source and a reference image, represented by I_S and I_R, respectively, the registration problem consists in finding a transformation T that applied to I_S aligns it spatially to I_R. Different approaches can be followed to solve the problem; many of them are based on the assumption that the intensity of every point x in the image I_R is conserved in image I_S but at a different spatial position $T(x)$. This means that the equality $I_S[T(x)] = I_R(x)$ holds for every point in I_R (known as the *Optical Flow Constraint*), and there is a huge number of registration methods based on it [11,16,21,22,2]. Hereafter, we denote by I_T the transformed source image, that is $I_S[T(x)] = I_T(x)$.

The optical flow constraint is not always applicable, for example, when registering medical images obtained from different modalities. For this case, registration by the maximization of *Mutual Information* (*MI*) has been widely used because it does not assume a functional relationship between the intensities of the images; instead, it is based on the fact that if aligned, the maximal dependency (information) between the intensities is found.

Given two images, I_T and I_R, their mutual information is defined as:

$$MI(I_T, I_R) = H(I_T) + H(I_R) - H(I_T, I_R) \tag{1}$$

where H is the entropy function of the image intensities. If the space of intensity values is discrete, then the entropy function is written as:

$$H(I) = -\sum_i p_i \log p_i \tag{2}$$

where p_i is the probability to observe the intensity value i; and in case of a continuous space, the entropy is defined as:

$$H(I) = -\int_{-\infty}^{\infty} p(i) \log[p(i)] \mathrm{d}i. \tag{3}$$

☆ The authors were partially supported by Grant 46270 of CONACyT (Consejo Nacional de Ciencia y Tecnología, México).
* Corresponding author. Address: Center for Research in Mathematics (CIMAT), Department of Computer Science, Apartado Postal 402, C.P. 36000 Guanajuato, Gto, Mexico.
E-mail addresses: hector@cimat.mx (H.F. Gómez-García), jlm@cimat.mx (J.L. Marroquín), horebeek@cimat.mx (J. Van Horebeek).

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.02.001

The first applications of *MI* to the image registration problem, were published simultaneously by Viola et al. [23] and Collignon et al. [3], both in the middle of the last decade. Since then, a great number of publications has appeared extending the initial work to problems like nonparametric multimodal image registration [10,4], registration of stereoscopic pairs [7,13] or feature tracking in images [5].

In general, methods based on the maximization of *MI*, start with an initial transformation T^0, leading to a *MI* value MI^0, and using a proper optimization method, a sequence of transformations is generated in such a way that the associated *MI* is increased until convergence. During the optimization process, the increments in *MI* are calculated with the expression:

$$\Delta MI = \Delta H(I_T) + \Delta H(I_R) - \Delta H(I_T, I_R).$$

If the discrete version of the entropy (2) is considered, this is a function of the entries of the probability vector; hence, using a Taylor series expansion, a linear approximation for the increment in entropy is given by:

$$\Delta H = -\sum_i [1 + \log p_i]\Delta p_i.$$

Because the coefficient $[1 + \log p_i]$ is large for small probability values, this increment is highly determined by small features in the images to be registered (which are generally associated with small probability values). This can trap the registration algorithm in local optima when aligning small features, particularly if the small probabilities are not accurately computed. This makes it difficult to apply *MI* in cases where only a limited sampling is available, for example when measuring entropy at a local level in images, which is important in interesting problems like nonparametric image registration, and in the segmentation of motion between frames, where local measurements must be taken in order to learn the local motion models and to have enough spatial definition at the motion interfaces.

Another problem related to the application of *MI*, occurs when working with images with a large background compared to the region of interest, as frequently happens in medical image problems. Under this circumstance the sum of the marginal entropies can become larger than the joint entropy, leading to an increase of *MI*, instead of decreasing it in misregistration. Studholme et al. [20] proposed the use of a normalized version of the *MI* to overcome this disadvantage. This measure is known as *Normalized Mutual Information* (*NMI*):

$$NMI(I_T, I_R) = \frac{H(I_T) + H(I_R)}{H(I_T, I_R)}. \quad (4)$$

In this work we propose a new criteria for the registration of images with different intensity structure (e.g., medical images in different modalities) which uses a new predictability measure for probability distributions, which we call *Kernel-Predictability* (*KP*). *KP*, evaluated in the marginal and joint distributions of two images, is integrated in a similarity measure between images, normalized as (4), and applied to the registration problem. Unlike entropy, the increment of this measure when updated by an iterative optimization method, is mostly determined by the larger entries of the probability vector, which is reflected in a higher robustness in problems where only limited sampling is available. Our proposal is discussed in Sections 2 and 3, and in Section 4 its performance in image registration problems is compared to that obtained under maximization of *MI* and *NMI*. The experimental results show that an important reduction in registration errors is obtained by the use of our method compared to *MI* and *NMI*.

2. Kernel-predictability

In order to introduce our predictability measure for a given distribution F, consider the following guessing game: someone generates a value x_1 from F and we guess x_1 by generating (independently) another value x_2 from F. We denote by $K(x_1, x_2)$ the obtained reward. Repeating this game, we can define the average reward $E[K(X_1, X_2)]$. We suppose that the reward function favors guesses close to the true value, i.e., K is a decreasing function of the distance between x_1 and x_2. Under this assumption it is clear that the less uncertainty is contained in F, the higher will be the average reward.

The above motivates the following measure for a given distribution F:

$$KP(F) = E[K(X_1, X_2)] = \int_{R^d}\int_{R^d} K(x_1, x_2)\mathrm{d}F(x_1)\mathrm{d}F(x_2). \quad (5)$$

This functional measures the predictability of the random variables distributed according to F, weighted by the kernel function K, and we denominate it *kernel-predictability*. It should be noted that *KP* is a predictability measure, so it behaves in an inverse way compared to entropy, which is an uncertainty measure.

For the discrete case, this becomes:

$$KP(\mathbf{p}) = \sum_i \sum_j K_{ij} p_i p_j = \mathbf{p}^{\mathrm{T}}\mathbf{K}\mathbf{p} \quad (6)$$

where the entry (i, j) of the matrix $\mathbf{K}$ equals the reward given for guessing the value x_i when the generated value was x_j, i.e., $K_{ij} = K(x_i, x_j)$. In the past, some measures have been presented that are apparently similar to our proposal. However an important difference must be noted. In [25], a functional like (5) is used to compute the expected distance between two groups of images. In [24,19], similarity measures between images are presented that can be confused with one of the estimators for (5) (discussed below). However, these three measures are evaluated over two different distributions, in contrast to (5), which takes only one distribution for its argument and therefore represents a property of the underlying distribution, such as its entropy or its variance.

We can measure the increment in kernel-predictability, which may be associated to the optimization process as:

$$\Delta KP = 2\sum_i \left(\sum_j K_{ij}p_j\right)\Delta p_i.$$

Note that the increment for every element of the probability vector, p_i, is multiplied by the coefficient $(\sum_j K_{ij}p_j)$; this coefficient equals the ith element of the vector generated by the product of the matrix $\mathbf{K}$ with the distribution vector $\mathbf{p}$. This product just smooths the probability vector $\mathbf{p}$ if we assume that the closer K_{ij} is to the main diagonal, the higher its value. Consequently, $(\sum_j K_{ij}p_j)$ is larger for large p_i values, and the increment in KP is mainly determined by the larger entries of the probability vector, and for that reason, by the most important features in the images to be registered. This is an important difference with respect to entropy.

2.1. Kernel-predictability with Gaussian kernels

Many choices for K are possible; a natural one is the Gaussian kernel, which is defined as:

$$K(x_1,x_2) = \frac{1}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{\|x_1-x_2\|^2}{2\sigma^2}\right) \tag{7}$$

where d is the dimension of the distribution and σ a free parameter.

For an arbitrary continuous distribution F, one can build a nonparametric approximation of its density by means of gaussian windows [6], centered over a set of points $\{a_i\}$ (e.g., independent samples obtained from F):

$$f_X(x) = \frac{1}{N}\sum_{i=1}^{N} f_{a_i,\sigma_2}(x) \tag{8}$$

where $f_{a_i,\sigma_2}(x)$ is the multivariate Gaussian density, $\mathcal{N}(a_i,\sigma_2^2\mathbf{I})$, with a $d\times d$ covariance matrix $\sigma_2^2\mathbf{I}$. Moreover, if one uses a multivariate Gaussian kernel to measure $KP(F)$, using the fact that a convolution of two Gaussians is another Gaussian, one can show that:

$$KP(F) = \frac{1}{N(2\pi(\sigma^2+2\sigma_2^2)^{d/2}} \times \sum_i\sum_j \exp(-||a_i-a_j||^2/2(\sigma^2+2\sigma_2^2)). \tag{9}$$

Note that the higher the spread of the points $\{a_i\}$ in the distribution, the lower will be its KP value. The maximum is reached when all the points in the set $\{a_i\}$ are equal, which represents a single Gaussian distribution. In this case, the value of KP is inversely proportional to the variance σ_2 of this distribution, which implies that the maximum value of KP will be reached when σ_2 is equal to zero, i.e., if one has a degenerate random variable that can only take one fixed value. Note that this will be true for the discrete case and for an arbitrary kernel as well, provided that the elements on the main diagonal of the matrix $\mathbf{K}$ contain the maximal reward value, say K_M (given for an exact prediction). This follows from the next inequality:

$$KP(\mathbf{p}) = \sum_i\sum_j K_{ij}p_ip_j \leqslant K_M\sum_i\sum_j p_ip_j = K_M$$

and from the fact that K_M is the value obtained for such degenerate random variables.

One important difference of KP with respect to entropy also follows from Eq. (9) and is illustrated in Fig. 1. If we move the Gaussian window centered over a_1 towards a_1^*, i.e., if we move a portion of the mass of the distribution to a position where there is practically no overlap with the original distribution, KP will be reduced, since the spread of the set $\{a_i\}$ will increase, and the entropy will increase. However, if one moves a_1 to a point a_1^{**} which is farther to the right, KP will be reduced even more, but the entropy will remain practically constant. This property of the entropy is not an advantage when applied in problems like image registration where the quality of a spatial transformation is measured by the narrowness of the joint distribution of gray tones between a pair of images; in this case, the gradient of KP will contain more information about the location of the optimal transformation.

Constructing the matrix $\mathbf{K}$ in (6) according to the Gaussian kernel (ignoring the normalizing constant for simplic-

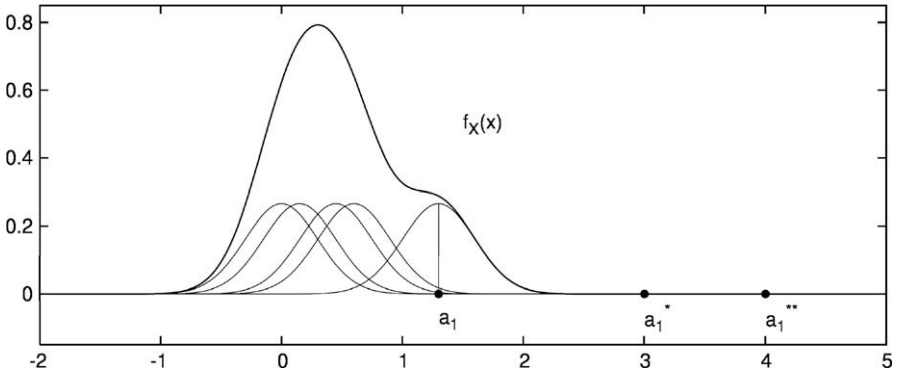

Fig. 1. Moving the gaussian window centered over a_1 towards a_1^* will reduce entropy and KP. Moving a_1 further to the right will reduce even more KP, while entropy remains constant.

ity), generates two interesting cases when evaluating the kernel at extreme values of the amplitude parameter σ. In the first case the Gaussian kernel can be approximated by the Kronecker delta for very small values of σ in the following way:

$$G(x_i, x_j) = \delta_{ij} = \begin{cases} 1 & \text{if} x_i = x_j \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

for this case, $KP(\mathbf{p}) = 1 - Gini(\mathbf{p})$, where *Gini* is the well known *Gini* entropy of Machine Learning [9]. The Gini entropy is maximized, and the associated *KP* minimized, under the uniform distribution.

For large values of σ the Gaussian kernel can be approximated by:

$$G_\sigma(x_1, x_2) \approx 1 - \frac{\|x_1 - x_2\|^2}{2\sigma^2} \tag{11}$$

and for this case, $KP(\mathbf{p}) \approx 1 - \frac{\sum_i Var[(X)_i]}{\sigma^2}$, where $Var[(X)_i]$ is the variance of the ith element of the multivariate random variable X. It can be shown that for univariate distributions with finite domain over the interval $[a, b]$, the distribution with maximal variance, and hence minimal associated *KP*, has a density equally concentrated on its two extreme values, a and b.

Random variables with uniform distribution are more difficult to predict than variables that take only two different values with the same probability, thus we prefer *KP* to behave in a way similar to the Gini entropy; for this reason we choose small values for the width of the Gaussian kernel; in practice for univariate random variables we take σ around 2–10% of their range.

2.2. Estimation of the kernel-predictability

The expression (5) is a *regular statistical functional* of degree two (two refers to the number of arguments of K), and for its estimation three different approaches are available in the literature [14,15]. The estimators are always based on a sampling set composed by n independent and identically distributed random variables, $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, with $X_i \sim F, \forall i$; and are defined as:

$$\widehat{KP}^1 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K(X_i, X_j) \tag{12}$$

$$\widehat{KP}^2 = \frac{4}{n^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K(X_i, X_j) \tag{13}$$

$$\widehat{KP}^3 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(X_i, X_j). \tag{14}$$

For the estimator $\widehat{KP}^1$, the kernel is evaluated over all different pairs of variables in $\mathbf{X}$; in $\widehat{KP}^2$, the set $\mathbf{X}$ is divided in two subsets and the kernel is evaluated at each pair formed by taking one variable from the first set and other variable from the second one; finally, in the third estimator, $\widehat{KP}^3$, the kernel is evaluated in all possible pairs of variables, as with estimator $\widehat{KP}^1$, but it adds the evaluations where the first and second variable coincide. The first two estimators are unbiased. If the kernel K is symmetric then $\widehat{KP}^1$ has the minimal variance among all the unbiased estimators, as shown in [14,15]; $\widehat{KP}^2$ has more variance than $\widehat{KP}^1$ but has a lower computational cost; the estimator $\widehat{KP}^3$ has minimal variance among these three estimators, but is biased. When the sampling set is increased in size, the variances of these estimators tend to the same value and the bias of the estimator $\widehat{KP}^3$ tends to zero.

3. Image registration with kernel-predictability

Application of *KP* to the registration problem can be done considering the joint distribution of the intensities of the images I_R and I_T, that is, $p(I_R, I_T) = p(\mathbf{I}_J(T))$. The intuitive idea is that when $T = T^*$ (the correct aligning transformation), $p(\mathbf{I}_J(T^*))$ should be more concentrated than $p(\mathbf{I}_J(T))$ for $T \neq T^*$, and therefore, $KP[p(\mathbf{I}_J(T^*))] > KP[p(\mathbf{I}_J(T))]$ for $T \neq T^*$. For example, if there exists a deterministic tone transfer function Φ, between I_R and $I_{T^*}, p(\mathbf{I}_J(T^*))$ must be ordered along a ridge-like structure determined by Φ: in this case, the conditional density $p(I_{T^*}|I_R = i) = \delta(I_{T^*} - \Phi(i))$, and any other transformation must redistribute the conditional density at different tone values. It is not enough, however, to consider only the *KP* evaluated over the joint distribution of I_R and I_T, because, for example, it can be maximized under transformations that assign all points in the image I_S to a single point in I_R. Restriction over the solution space can be considered normalizing the joint *KP*, in a way similar to what is done for mutual information [20]. We propose the next similarity measure between images based on *KP*:

$$SKP(I_T, I_R) = \frac{KP[p(\mathbf{I}_J)]}{KP[p(I_T)] + KP[p(I_R)]}. \tag{15}$$

This similarity measure makes a comparison between the predictability of the joint distribution and that of the marginal distributions for the images I_T and I_R. An upper bound for *SKP* in the discrete case is derived in Appendix A. For the particular case where the kernel K used for the evaluation of *KP* is the Kronecker delta, it is possible to show rigorously that SKP reaches its global maximum for $T = T^*$ (see Appendix A). This kernel, however, is not appropriate for practical computations, because in this case the gradient of SKP has very little information about the location of its maximum. In general, at least a local maximum of SKP is obtained for Gaussian kernels as well; to see this, note that for T different, but close to $T^*, KP[p(I_T)] + KP[p(I_R)] \approx KP[p(I_{T^*})] + KP[p(I_R)]$ and $KP[p(\mathbf{I}_J(T^*))] > KP[p(\mathbf{I}_J(T))]$, since $p(\mathbf{I}_J(T))$ is less concentrated than $p(\mathbf{I}_J(T^*))$ (see Eq. (9) and discussion above). In practice, this condition holds also for smooth kernels, for which *KP* behaves very much like the Gaussian case (see Section 2.1).

Registration of the images I_S and I_R is done by searching for the transformation T which maximizes the *SKP* value

between the corresponding I_T image and I_R. The transformation can be classified as *parametric* or *nonparametric*; for each case, a different registration strategy must be followed as detailed below. Assuming it is clear from the context for which images the similarity measure is evaluated, we will write $SKP(T)$ instead of the expression $SKP(I_T, I_R)$.

3.1. Parametric registration

Suppose the transformation T is determined by a vector of m real parameters, $\mathbf{a} = (a_1, a_2, \ldots, a_m)$, and m is considerably smaller than the total number of points in the images to be registered; in this case, we write $T(x; \mathbf{a})$ instead of $T(x)$ (e.g., when registering images under affine or projective transformations). For ease of notation, the intensity values associated to an arbitrary sampled coordinate X_i, can be abbreviated with the expressions: $I_R^i = I_R(X_i)$, $I_T^i = I_S[T(X_i; \mathbf{a})]$, and $\mathbf{I}_J^i = (I_T^i, I_R^i)$. Then, an approximation to (15) using the estimator (13) can be written in the following way:

$$\widehat{SKP}[T(\mathbf{a})] = \frac{\widehat{KP}_J[T(\mathbf{a})]}{\widehat{KP}_T[T(\mathbf{a})] + \widehat{KP}_R} \tag{16}$$

with

$$\widehat{KP}_J[T(\mathbf{a})] = \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j)$$

$$\widehat{KP}_T[T(\mathbf{a})] = \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_{\sigma_M}(I_T^i, I_T^j)$$

$$\widehat{KP}_R = \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_{\sigma_M}(I_R^i, I_R^j),$$

K_{σ_J} is the kernel employed to measure the predictability of the joint distribution of I_T and I_R, and K_{σ_M} for the marginal distributions of I_T and I_R. Note that the constant coefficient in the estimators can be ignored due to normalization.

For example, if Gaussian kernels are used (ignoring the normalizing constants), then:

$$K_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j) = G_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j) = \exp\left\{-\frac{\|\mathbf{I}_J^i - \mathbf{I}_J^j\|^2}{2\sigma_J^2}\right\} \tag{17}$$

$$K_{\sigma_M}(I^i, I^j) = G_{\sigma_M}(I^i, I^j) = \exp\left\{-\frac{(I^i - I^j)^2}{2\sigma_M^2}\right\}. \tag{18}$$

The maximization can be done using stochastic gradient ascent, starting with an initial transformation defined by the vector $\mathbf{a}^0$ and actualizing it with the relation:

$$\mathbf{a}^{t+1} = \mathbf{a}^t + \lambda \nabla_{\mathbf{a}} \widehat{SKP}[T(\mathbf{a}^t)]$$

with:

$$\nabla_{\mathbf{a}} \widehat{SKP}[T(\mathbf{a}^t)] = \frac{1}{\widehat{KP}_T[T(\mathbf{a}^t)] + \widehat{KP}_R} \nabla_{\mathbf{a}} \widehat{KP}_J[T(\mathbf{a}^t)] - \frac{\widehat{KP}_J[T(\mathbf{a}^t)]}{(\widehat{KP}_T[T(\mathbf{a}^t)] + \widehat{KP}_R)^2} \nabla_{\mathbf{a}} \widehat{KP}_T[T(\mathbf{a}^t)] \tag{19}$$

and in particular, when using the kernels (17) and (18), these gradients become:

$$\nabla_{\mathbf{a}} \widehat{KP}_J[T(\mathbf{a}^t)] = -\frac{1}{\sigma_J^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} G_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j)(I_T^i - I_T^j)(\nabla_{\mathbf{a}} I_T^i - \nabla_{\mathbf{a}} I_T^j)$$

$$\nabla_{\mathbf{a}} \widehat{KP}_T[T(\mathbf{a}^t)] = -\frac{1}{\sigma_M^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} G_{\sigma_M}(I_T^i, I_T^j)(I_T^i - I_T^j)(\nabla_{\mathbf{a}} I_T^i - \nabla_{\mathbf{a}} I_T^j).$$

The gradient can be estimated using a different sampling set for every iteration, giving a stochastic behavior to the gradient ascent (as is proposed in [23]) which allows the optimization procedure to escape from local optima; in this sense the use of the estimator (13) is more suitable due to the fact that its higher variance introduces an additional stochastic component. Besides it has the lowest computational cost among the three options.

When working with large transformations, the part of the image I_R in the overlapping region between the two images can vary with T, and the gradient of the similarity must consider this variation. Unfortunately there is no explicit dependence of I_R on the transformation; therefore one must approximate the gradient of the similarity by finite differences. The partial derivative of (16) with respect to any parameter a_i can be evaluated with centered finite differences as:

$$\frac{\partial \widehat{SKP}}{\partial a_i}[T(\mathbf{a}^t)] \approx \frac{\widehat{SKP}[T(\mathbf{a}^t + \epsilon_i \mathbf{e}_i)] - \widehat{SKP}[T(\mathbf{a}^t - \epsilon_i \mathbf{e}_i)]}{2\epsilon_i}, \tag{20}$$

where $\mathbf{e}_i$ is a vector with a one in the ith component and zeros in the rest, and ϵ_i is a small real value. Using this approximation, the similarity must be evaluated twice for each parameter in the transformation and because every evaluation determines a different overlapping region between the images, in order to calculate accurately the gradient, the samples used for estimation must lie in the intersection of all overlapping regions.

The use of (20) for the gradient approximation is advantageous in the case of registrations with large transformations, where the variation of I_R during the process is not negligible; otherwise one can ignore this variation and employ the simpler approach defined in (19). In this paper, the approximation (20) was employed for registration.

3.2. Nonparametric registration

To obtain a nonparametric (dense) field, the registration must find a different translation vector for each point in the

images; in this case, the transformation for every pixel is defined in the following way: $T(x_i) = x_i + u_i$, $i \in \{1, \ldots, N\}$. A large amount of sampling is necessary in order to estimate accurately the complete transformation field, $\mathbf{u} = \{u_1, \ldots, u_N\}$, and the registration by the maximization of our similarity measure can be prohibitive due to its quadratic cost over the sampling size. Instead of maximizing it globally, one can restrict its evaluation to a local level, focusing on a small region around each point in the images; then we can maximize the sum of the local similarities for every point x. For example, if we consider a small squared region defined by the window W_x centered on the point x, then the local similarity will be a function only of the translation vectors associated to the points enclosed by W_x, that is the set $\mathbf{v}_x = \{u_i | i \in W_x\}$. Besides the reduction of the computational cost, evaluating the similarity at a local level can help to avoid the irregularities of the probability distributions of the intensities, which results from large spatial inhomogeneities in the intensity of the images. Also regularization of the field $\mathbf{u}$ must be considered. Therefore, for nonparametric registration, the minimization of the following energy is proposed, which is a combination of a data fidelity term, E_D, and a smoothness term, E_S:

$$E(\mathbf{u}) = E_D(\mathbf{u}) + \lambda E_S(\mathbf{u})$$

where

$$E_D(\mathbf{u}) = \sum_x \{-\widehat{SKP}_{W_x}(\mathbf{v}_x)\} \tag{21}$$

$$E_S(\mathbf{u}) = \sum_x \left\{ \sum_{x' \in N_x} \|u_x - u_{x'}\|^2 \right\} \tag{22}$$

λ is a constant which controls the smoothness of the field, and N_x is a small neighborhood around the point x.

The local similarity is evaluated in the following way:

$$\widehat{SKP}_{W_x}(\mathbf{v}_x) = \frac{\widehat{KP_J}(\mathbf{v}_x)}{\widehat{KP_T}(\mathbf{v}_x) + \widehat{KP_R}(x)} = \frac{\sum_{i,j \in W_x} K_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j)}{\sum_{i,j \in W_x} K_{\sigma_M}(I_T^i, I_T^j) + \sum_{i,j \in W_x} K_{\sigma_M}(I_R^i, I_R^j)}. \tag{23}$$

For this case $I_T^i = I_S(x_i + u_i)$. We have written the $\widehat{KP}_R$ value as a function of the centering point, x, in order to stress its local evaluation. Note that now the estimator (14) is being used; this is due to the fact that when working with small windows, only a few samples are available for the estimation of the similarities, and the smaller variance of (14) allows for a more accurate calculation of the field; the estimator (12) can be used as well with little difference in the results, but the use of estimator (13) should be avoided, mostly for very small windows (e.g., windows with 3×3 pixels).

The minimization is done by gradient descent. When using the Gaussian kernels (17) and (18), the partial derivative of the data fidelity term in Eq. (21) with respect to any translation vector u_l is:

$$\frac{\partial E_D}{\partial u_l} = 2 \sum_{x: l \in W_x} \sum_{i \in W_x} \left\{ \begin{matrix} f_J(x) G_{\sigma_J}(\mathbf{I}_J^l, \mathbf{I}_J^i) - \\ f_M(x) G_{\sigma_M}(I_T^l, I_T^i) \end{matrix} \right\} (I_T^l - I_T^i) \nabla I_S(x_l + u_l) \tag{24}$$

where: $f_J(x) = \frac{1}{\sigma_J^2 [\widehat{KP_T}(\mathbf{v}_x) + \widehat{KP_R}(x)]}$, $f_M(x) = \frac{\widehat{KP_J}(\mathbf{v}_x)}{\sigma_M^2 [\widehat{KP_T}(\mathbf{v}_x) + \widehat{KP_R}(x)]^2}$, and $\nabla I_S(x_l + u_l)$ is the spatial gradient of the image I_S evaluated at the point $(x_l + u_l)$. Note that the first sum runs over every window, W_x, containing the point l, and the second one runs over every point within the window W_x.

Finally, the gradient of the smoothness term is:

$$\frac{\partial E_S}{\partial u_l} = 4 \left(|N_l| u_l - \sum_{l' \in N_l} u_{l'} \right).$$

Image registration by the use of (24) can be time consuming for large windows (e.g., 7×7 pixels or more). Supposing that a local kernel-predictability has been evaluated for a given point x and for a fixed set of vectors $\mathbf{v}_x^0$, then it is possible to make an approximation to evaluate the kernel-predictability for a new set of vectors $\mathbf{v}_x$, making a linear approximation in Taylor series around $\mathbf{v}_x^0$ in the following way:

$$\widehat{KP}(\mathbf{v}_x) \approx \widehat{KP}(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}} \nabla_{\mathbf{v}} \widehat{KP}(\mathbf{v}_x^0). \tag{25}$$

Once the values of $\widehat{KP}(\mathbf{v}_x^0)$ and $\nabla_{\mathbf{v}} \widehat{KP}(\mathbf{v}_x^0)$ are evaluated, the approximation to the kernel-predictability is reduced from $|W|^2$ kernel evaluations, to the calculation of a product of two vectors containing $|W|$ elements without any kernel evaluation. Substituting the linearized approximations for $\widehat{KP}_J(\mathbf{v}_x)$ and $\widehat{KP}_T(\mathbf{v}_x)$ in (23), it can be rewritten as:

$$\widehat{SKP}_{W_x}(\mathbf{v}_x) = \frac{\widehat{KP}_J(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_\mathbf{x}^0)^{\mathrm{T}} \nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0)}{\widehat{KP}_T(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}} \nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0) + \widehat{KP}_R(x)}. \tag{26}$$

Substitution of (26) into the term (21) simplifies the gradient of the data fidelity term to:

$$\frac{\partial E_D}{\partial u_l} = -\sum_{x: l \in W_x} \{ f_J(x) [\nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0)]_l - f_M(x) [\nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0)]_l \} \tag{27}$$

where $[\nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0)]_l$ and $[\nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0)]_l$, are the lth component of the kernel-predictability gradients:

$$[\nabla_{\mathbf{v}} \widehat{KP}_M(\mathbf{v}_x^0)]_l = -\frac{2}{\sigma_M^2} \sum_{i \in W_x} G_{\sigma_M}(I_T^l, I_T^i)(I_T^l - I_T^i) \nabla I_S[x_l + (\mathbf{v}_x^0)_l]$$

$$[\nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0)]_l = -\frac{2}{\sigma_J^2} \sum_{i \in W_x} G_{\sigma_J}(\mathbf{I}_J^l, \mathbf{I}_J^i)(I_T^l - I_T^i) \nabla I_S[x_l + (\mathbf{v}_x^0)_l]$$

and $I_T^i = I_S[x_i + (\mathbf{v}_x^0)_i]$, $f_J(x) = \frac{1}{\widehat{KP}_T(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}} \nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0) + \widehat{KP}_R(x)}$, $f_M(x) = \frac{\widehat{KP}_J(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}} \nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0)}{[\widehat{KP}_T(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}} \nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0) + \widehat{KP}_R(x)]^2}$.

The optimization by gradient descent using (27), requires a periodical reevaluation of the values and gradients of the kernel-predictability, in practice, after every 5–10 iterations. Using this approach an important reduction in the convergence time is reached without loosing too much accuracy.

4. Results

In this section we present some results obtained with the application of our proposal to different image registration problems.

4.1. Parametric registration

In the first set of experiments we compared the performance of our method with respect to registration by maximization of mutual information and normalized mutual information, in affine registration problems. For these measures, two different implementations were considered. The first one, uses the discrete version of the entropy (2), approximating the probability distributions by normalized histograms, and performing the optimization with the simplex method [18]; this implementation is widely used and its advantages over other implementations (in all cases using the discrete version of the entropy) are documented by Zhu and Cochoff [26]. The second implementation is based on the continuous version of the entropy (3), using Parzen windows for the estimation of the probability densities, and following [23] for the entropy estimation; these approximations are:

$$H(I_R) = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_M}(I_R^i - I_R^j)\right\} \tag{28}$$

$$H[I_L(T)] = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_M}(I_T^i - I_T^j)\right\} \tag{29}$$

$$H[I_L(T), I_R] = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_J}(I_J^i - I_J^j)\right\}, \tag{30}$$

where A and B, are two different sets of sampled coordinates in the overlapping region of the images, and G_σ, is the normal density with variance σ^2; the optimization is done using stochastic gradient ascent, approximating the partial derivatives with centered finite differences.

Affine transformations can be applied multiplying a squared matrix $\mathbf{A}$ with a point $\mathbf{p}$ and adding a translation vector $\mathbf{t}$, to generate a transformed point $\mathbf{p}\prime$. The matrix $\mathbf{A}$ is a composition of three simpler transformations: a rotation $\mathbf{R}$, a scaling $\mathbf{S}$, and a shearing $\mathbf{H}$; this is represented by:

$$\mathbf{p}' = \mathbf{A}\mathbf{p} + \mathbf{t} = (\mathbf{RSH})\mathbf{p} + \mathbf{t}.$$

The order of the matrices multiplication is arbitrary, and for bidimensional transformations the exact representation for each matrix is:

$$\mathbf{R} = \begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix}\alpha & 0\\ 0 & \beta\end{pmatrix}$$

$$\mathbf{H} = \begin{pmatrix}1 & \gamma\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ \delta & 1\end{pmatrix}.$$

Five sets, composed of 50 affine transformations each one, were generated assigning random values for the $\phi, \alpha, \beta, \gamma$, and δ parameters, and for the translation vector. These values were sampled uniformly from certain intervals, as is summarized in Table 1.

Different bidimensional images were used for registration (see Fig. 2). Reference images were created applying a change in intensity and affine transformations to the original images (128×128 pixels), and then extracting a square of 90×90 pixels from the center of the transformed images, as is shown in Fig. 2(a)–(c), excepting images 2(d) (217×181 pixels), which correspond to two magnetic resonances obtained by the simulator at the Montreal Neurological Institute [1]; for this case, the floating image is a $T1$-weighted MRI with 9% of noise level and 40% of spatial inhomogeneities in intensity, and the reference images were created applying affine transformations to a corresponding $T2$-weighted image. The intensities of every image pair were scaled between 0 and 100; after that, the change in intensity was applied through the function $I_R = 100(\frac{I_L}{100})^{1.35}$ for images 2(a) and (b) and $I_R = 100(1 - \frac{I_L}{100})^{1.35}$ for 2(c). This process was repeated for every transformation in each set, and the algorithms executed for registering the original images to the reference images. For every registration, two Gaussian pyramids of three levels were constructed by alternatively smoothing (with a Gaussian kernel) and sub-sampling the original source and reference images; then, the registration started with the identity transformation in the coarsest level of the pyramids and the resulting transformation for every level was used as the initial transformation for the subsequent level. The implementation details for the two discrete algorithms were set according to Zhu and Cochoff [26]. For the case of continuous entropy, two different sets of coordinates composed of 50 samples each one were used. A multiple of the identity matrix, $\sigma^2 I$, was used as the covariance matrix in the

Table 1
Composition of the five transformations sets

Set	ϕ (degrees)	α, β	γ, δ	t (pixels for each component)
$S1$	$[-10°, 10°]$	$[0.9, 1.1]$	$[-0.1, 0.1]$	$[-10.0, 10.0]$
$S2$	$[-20°, 20°]$	$[0.8, 1.2]$	$[-0.2, 0.2]$	$[-20.0, 20.0]$
$S3$	$[-30°, 30°]$	$[0.7, 1.3]$	$[-0.3, 0.3]$	$[-30.0, 30.0]$
$S4$	$[-40°, 40°]$	$[0.6, 1.4]$	$[-0.4, 0.4]$	$[-40.0, 40.0]$
$S5$	$[-50°, 50°]$	$[0.5, 1.5]$	$[-0.5, 0.5]$	$[-50.0, 50.0]$

The width of the generating interval for each parameter is progressively augmented.

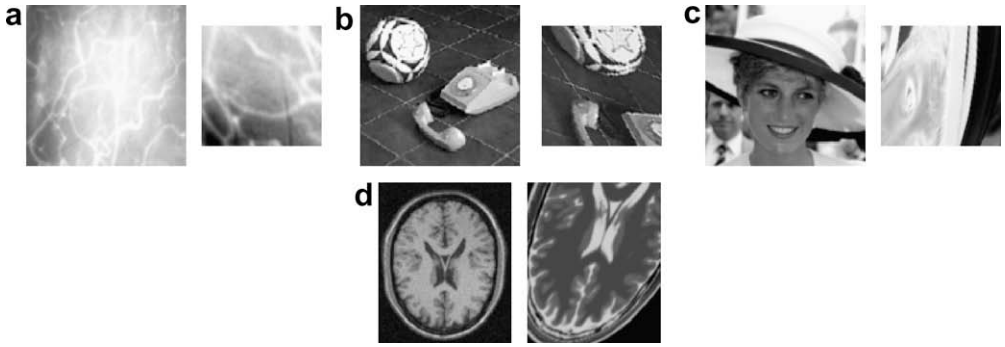

Fig. 2. Images used for registration. For cases (a)–(c), reference images were obtained applying changes in intensity and affine transformations to the original images, and then extracting a subsquare of the center of the transformed images. For case (d), the reference images were generated applying only affine transformations to a MR in modality $T2$.

estimation of the joint entropy of images, and for the marginal entropies the variance was set to the value σ^2; this value was fixed manually, considering a percentage of the dynamic range of the images to be registered. The values used in these experiments were $\sigma = 5\%$ for image 2(a) and $\sigma = 10\%$ for the rest of the images. In the case of *SKP*, estimator (13) was employed, using the same number of samples for estimation as was done with *MI* and *NMI*, and the width of the kernels used were set with the same considerations, except that a fixed value of $\sigma = 8\%$ was used for all registrations. The number of successful registrations for each set and for each algorithm, is plotted in Fig. 3; a registration was considered successful if the mean error between the applied and recovered vector fields was lower than one pixel. It can be noted that, almost in all cases, our method outperformed all versions of registration by mutual information and normalized mutual information, specially for large transformations; and that the algorithms based on the discrete version of the entropy have no robustness when used for registrations with large transformations.

Considering the algorithms based on the continuous estimation of entropy, our method presents another advantage. Due to the quadratic cost of the estimation of both kernel-predictability and entropy, a very important parameter is the number of samples used for registration; the

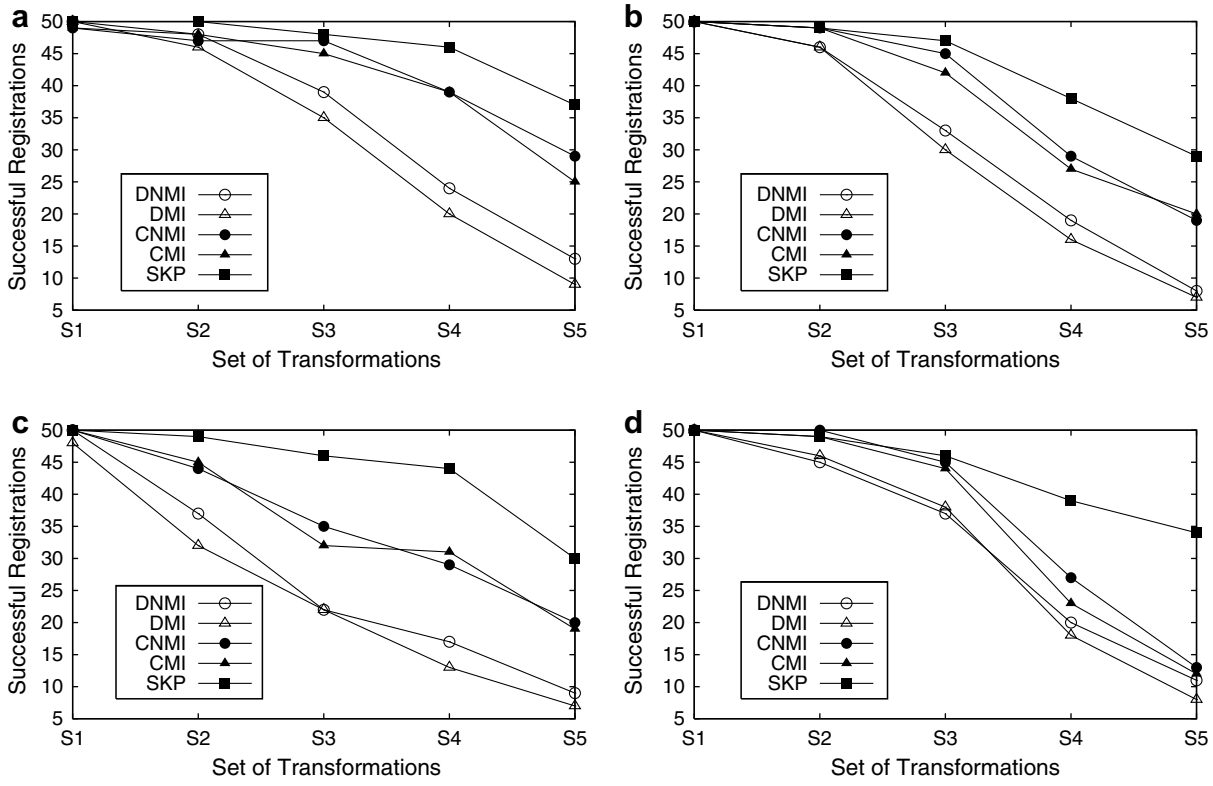

Fig. 3. Successful registrations in function of the complexity of the transformations. The plots show results corresponding to images 2(a)–(d). In the plot *SPK* means "Similarity based on Kernel-Predictability", *CNMI* and *DNMI* refers to "Continuous" and "Discrete Normalized Mutual Information"; finally *CMI* and *DMI*, refers to "Continuous" and "Discrete Mutual Information".

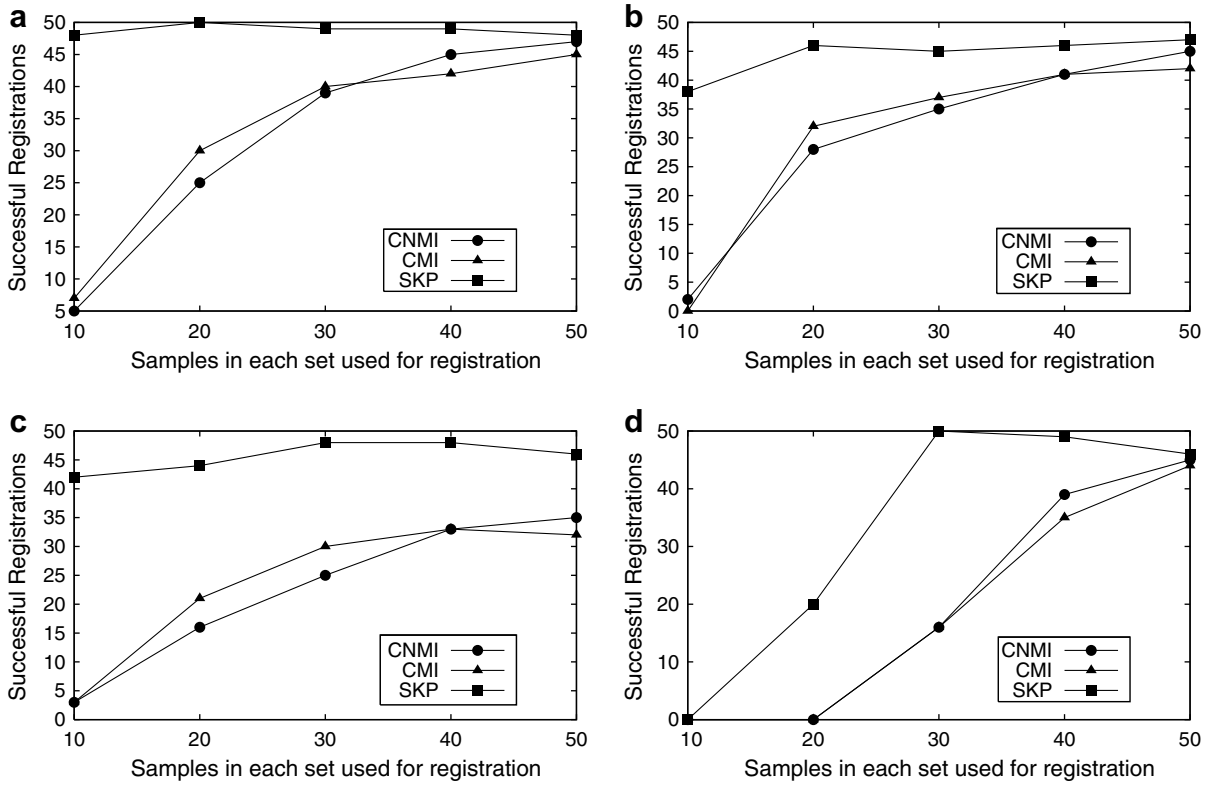

Fig. 4. Successful registrations as a function of the number of samples used for estimation. The plots show results corresponding to images 2(a)–(d). In the plot *SKP* means "Similarity based on Kernel-Predictability", *CNMI* refers to "Continuous Normalized Mutual Information"; finally *CMI* refers to "Continuous Mutual Information".

plots in Fig. 4 show the performance of the three methods when varying this parameter, in this case the set $S3$ of affine transformations (described in Table 1) was used in the four images; as can be seen, our method works considerably well even using a very small sampling for estimation, which is not the case of mutual information and normalized mutual information.

Finally, the performance of our proposal was evaluated under different kernel functions. Registration of the four image pairs (shown in Fig. 2) was repeated for *SKP* using the one-dimensional kernels described in Table 2 for the evaluation of the marginal *KP*'s. The joint *KP* was evaluated in each case, employing a separable kernel generated by the product of the two marginal kernels, that is $K_J(\mathbf{I}_J^i, \mathbf{I}_J^j) = K_M(I_R^i, I_R^j)K_M(I_T^i, I_T^j)$. It can be noted in Fig. 5(a)–(d) that the selection of the kernel for registration by maximization of *SKP* is not a critical factor. Small differences were obtained for different smooth kernels, however a poor performance is obtained in the case of the triangular kernel.

4.2. Nonparametric registration

The robustness of our proposal for working correctly with large transformations and using only few samples, makes it very suitable to be applied in nonparametric registration problems. In order to measure the performance of $\widehat{SKP}$ in these problems, 10 different synthetic transformation fields were generated using two grids with 15×15 nodes of cubic B-spline functions, and assigning random values to every node. Then, for each pixel (x, y) in the image, a translation vector $(u(x,y), v(x,y))$ was defined in the following way:

$$\begin{aligned} u(x,y) &= \sum_{i=1}^{15}\sum_{j=1}^{15} U_{ij}\beta[k_1(x-x_i)]\beta[k_2(y-y_j)] \\ v(x,y) &= \sum_{i=1}^{15}\sum_{j=1}^{15} V_{ij}\beta[k_1(x-x_i)]\beta[k_2(y-y_j)] \end{aligned} \tag{31}$$

where $U_{ij}, V_{ij} \sim U\{-7, 7\}$, for all centering nodes (x_i, y_j), and k_d is the proportion of nodes versus the image dimen-

Table 2
Different kernels used for registration with *SKP*

Gaussian kernel	$K(x_1,x_2) = \exp[-(x_1-x_2)^2/\sigma^2]$
Cauchy kernel	$K(x_1,x_2) = \frac{1}{1+\alpha(x_1-x_2)^2}$
Exponential kernel	$K(x_1,x_2) = \exp(-\lvert x_1-x_2\rvert/\sigma^2)$
Triangular kernel	$K(x_1,x_2) = 1-\alpha\lvert x_1-x_2\rvert$ for $\alpha\lvert x_1-x_2\rvert < 1$ and $K(x_1,x_2) = 0$, otherwise

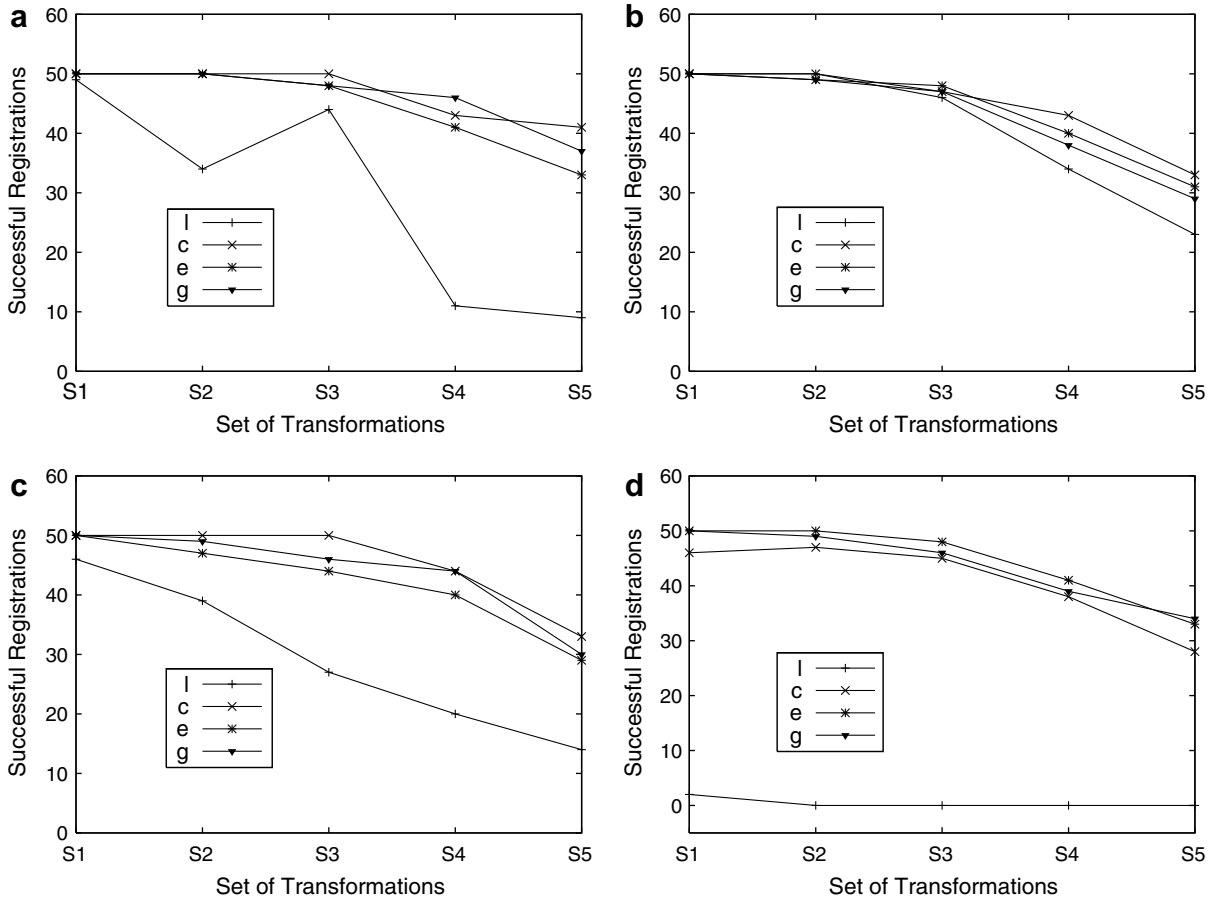

Fig. 5. Registration results for *SKP* using different kernels (described in Table 2). The plot shows registration results using *SKP* with Gaussian (g), Cauchy (c), exponential (e) and triangular kernels (l).

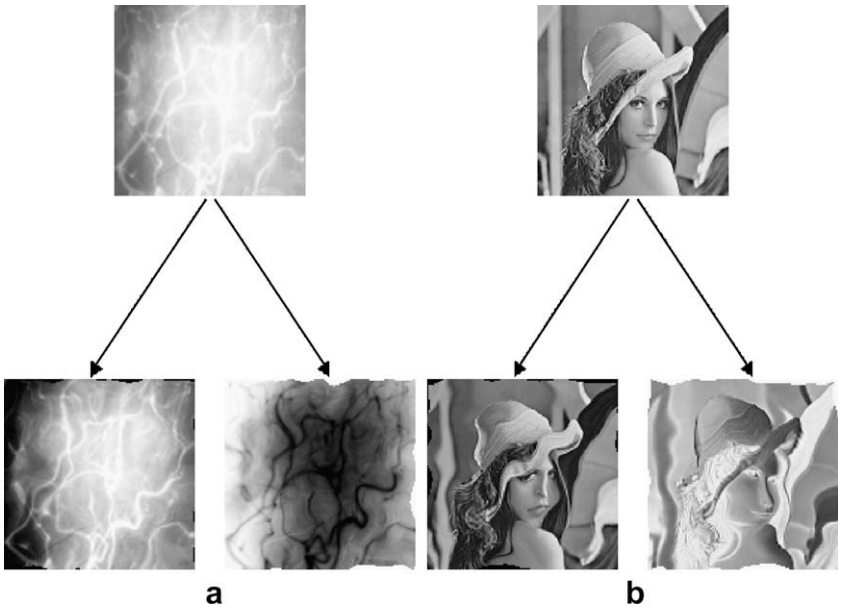

Fig. 6. Images used for nonparametric registration. Reference images were created applying changes in intensity and different synthetic transformation fields to the original images.

sion in the direction d. The cubic B-spline functions used are:

$$\beta(z) = \begin{cases} \frac{2}{3} - |z|^2 + \frac{|z|^3}{2}, & |z| < 1. \\ \frac{(2-|z|)^3}{6}, & 1 \leqslant |z| < 2 \\ 0, & |z| \geqslant 2. \end{cases}$$

The synthetic fields were applied to two images after a change in intensity determined by two different tone transfer functions, $f_1(I) = 100(\frac{I}{100})^{1.35}$ and $f_2(I) = 100(1 - \frac{I}{100})^{1.35}$ for every image, as shown in Fig. 6. Then, our nonparametric registration algorithm was executed to recover the original transformation field and the error measured for each case. The error was calculated as the average length of the difference between the applied and recovered vectors for all pixels. As was done with parametric registration, Gaussian pyramids of three levels were used for the source and reference images; in the coarsest level of the pyramids every vector of the transformation field was initialized to zero and for all the subsequent levels, the transformation was started with the resulting field of the previous level. For comparison, the registration algorithm was run substituting $\widehat{SKP}$ in the term (21) by the corresponding expressions for MI and NMI based on the continuous entropy (Eqs. (28)–(30)). As described in Section 3.2, the similarity measures were evaluated at a local level using small windows placed over each pixel in the images, and windows of different sizes were considered. The results are summarized in Fig. 7(a)–(d); as can be seen, important reductions in the mean error are obtained with our proposal compared to MI and NMI when using small windows for registration, and again, due to the quadratic cost of the estimations over the number of samples, this is reflected in important savings in the execution time (see Fig. 8). To facilitate a qualitative comparison of the errors, the regis-

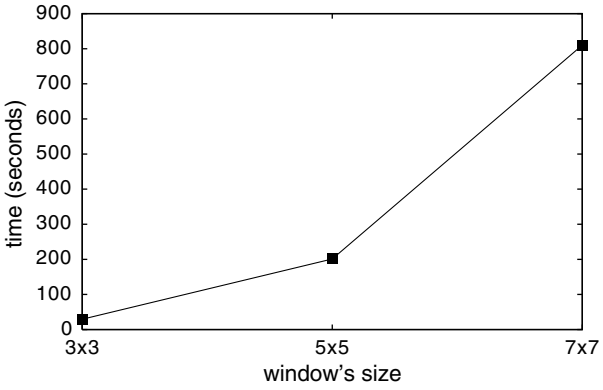

Fig. 8. Execution time for nonparametric registration with SKP as a function of the width of the windows used to measure local similarity. Results are shown for an image of 128×128 pixels. For every window's size 200 iterations of the gradient descent were run in every level of the Gaussian pyramid. The tests were run on a pentium 4, 3.0 GHz, PC.

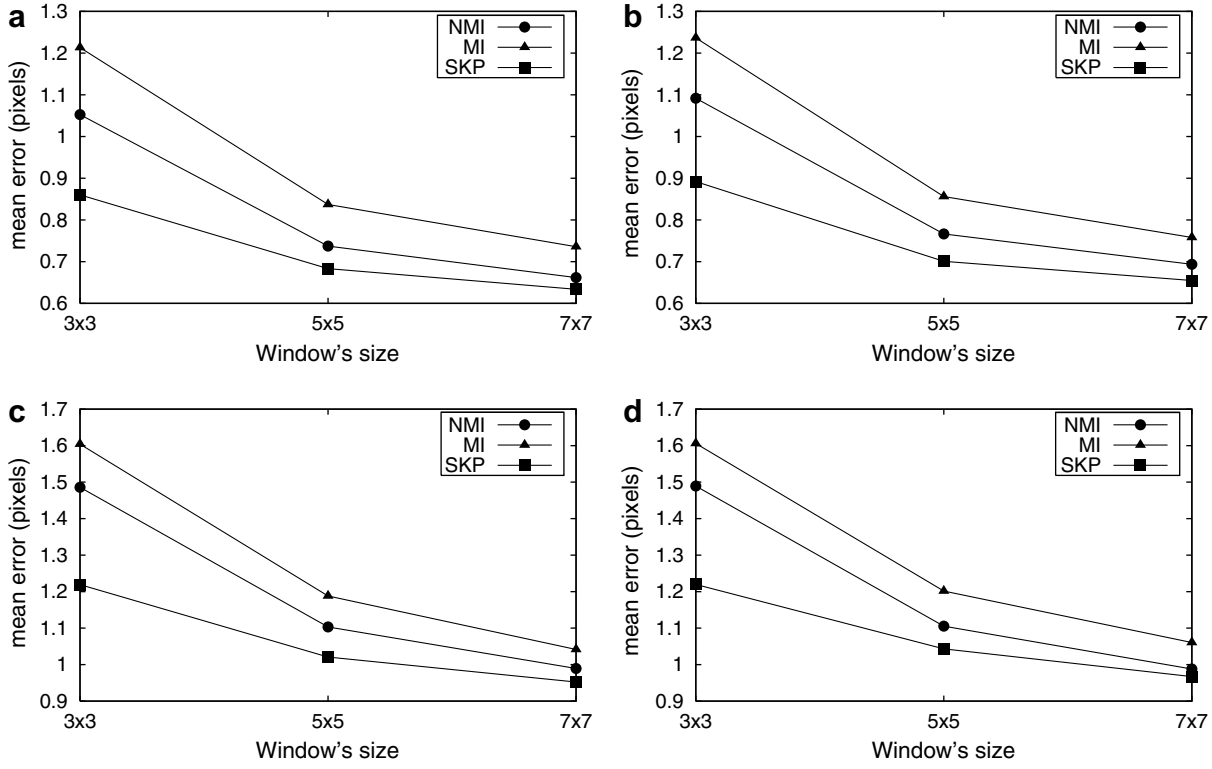

Fig. 7. Mean error in nonparametric registration for different window sizes. The first row shows results for image 6(a) and reference images generated using the tone transfer function $f_1(I) = 100(\frac{I}{100})^{1.35}$ (left plot), and $f_2(I) = 100(1 - \frac{I}{100})^{1.35}$ (right plot). The second row shows the corresponding results for image 6(b).

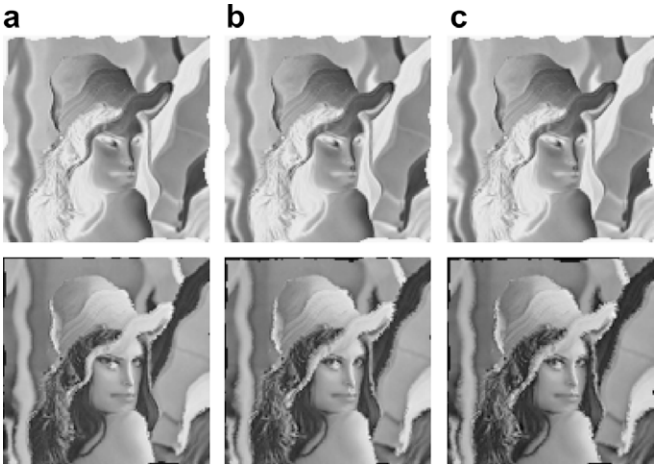

Fig. 9. Registered images for a specific transformation. In the first row, the reference image is shown (the same for each case), and in the second the registered images for *SKP* (left), *NMI* (center) and *MI* (right). The estimation of the deformation field was done locally using windows of 3×3 pixels around every pixel in the images. The respective errors were: 1.23, 1.57 and 1.60 pixels.

tered images by the three methods for a specific transformation are shown in Fig. 9.

5. Conclusions

In this paper, we have proposed the use of a new similarity measure for image registration, based on a novel concept of kernel-predictability for random variables. The performance of our registration method was compared with mutual information and normalized mutual information in different registration situations, including nonparametric registration, and we have shown experimentally that using our method, important reductions in registration errors are obtained, mainly when used for large transformations and in situations where only a small sampling is available. This robustness is due to the fact that the new similarity measure is controlled by the most important features in the images.

Appendix A

An upper bound for the registration measure *SKP* for the discrete case may be found in the following way: suppose one uses a kernel K to measure *KP* for the intensity distributions of a pair of images I, J, which has the property: $K(i,i) = 1 \geqslant K(i,j)$, for $i \neq j$. One may then construct a separable kernel K_2 for measuring *KP* for the joint distribution $p_{IJ}(I,J)$ as:

$$K_2((i_1,j_1),(i_2,j_2)) = K(i_1,i_2)K(j_1,j_2)$$

we now have:

$$\begin{aligned} KP(p_{IJ}(I,J)) &= \sum_{i_1}\sum_{i_2}\sum_{j_1}\sum_{j_2} p_{IJ}(i_1,j_1)p_{IJ}(i_2,j_2) \\ &\quad \times K_2((i_1,j_1),(i_2,j_2)) \\ &= \sum_{i_1}\sum_{i_2} p_I(i_1)p_I(i_2)K(i_1,i_2) \\ &\quad \times \sum_{j_1}\sum_{j_2} p(j_1|i_1)p(j_2|i_2)K(j_1,j_2) \\ &\leqslant \sum_{i_1}\sum_{i_2} p_I(i_1)p_I(i_2)K(i_1,i_2) = KP(p(I)) \end{aligned}$$

In a similar way, one can see that $KP(p_{IJ}(I,J)) \leqslant KP(p(J))$, so that $SKP(I,J) \leqslant \frac{1}{2}$.

Now, consider a reference image I_R and a transformed image I_T, and assume that when the transformation T^*, which correctly aligns both images, is used, one has that the intensities i_R, i_{T^*} are related by a deterministic, invertible tone transfer function Φ, so that $p(i_{T^*}|i_R) = \delta(i_{T^*} - \Phi(i_R))$. Assume also that $K(i,j) = \delta(i-j)$ (a Kronecker delta function). In this case, from the above equation one can see that $KP(p(I_R, I_{T^*})) = KP(p(I_R)) = KP(p(I_{T^*}))$, so that $SKP(I_R, I_{T^*}) = \frac{1}{2}$, which means that *SKP* reaches its global maximum when $T = T^*$.

References

[1] http://www.bic.mni.mcgill.ca/brainweb/.
[2] G. Aubert, R. Deriche, P. Kornprobst, Computing optical flow via variational techniques, SIAM Journal on Applied Mathematics 60 (1) (2000) 156–182.
[3] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, G. Marchal, Automated multi-modality image registration based on information theory, Information Processing in Medical Imaging (1995) 263–274.
[4] E. D'Agostino, F. Maes, D. Vandermeulen, P. Suetens, A viscous fluid model for multimodal image registration using mutual information, MICCAI (2002) 541–548.
[5] N. Dowson, R. Bowden, Metric mixtures for mutual information tracking, ICPR 2 (2004) 752–756.
[6] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973.
[7] E. Geoffrey, Mutual information as a stereo correspondence measure, Technical Report MS-CIS-00-20, University of Pennsylvania, 2000.
[8] L. Gottesfeld, A survey of image registration techniques, ACM Computing Surveys 24 (4) (1992) 325–376.
[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Verlag, New York, 2003.
[10] G. Hermosillo, C. Chefd'hotel, O. Faugeras, Variational methods for multimodal image matching, International Journal of Computer Vision 50 (3) (2002) 329–343.
[11] B. Horn, B.G. Schunck, Determining optical flow, Artificial Intelligence 17 (1981) 185–203.
[12] P.W. Josien, J.B. Pluim, A. Maintz, A. Viergever, Mutual information based registration of medical images: a survey, IEEE Transactions on Medical Imaging 22 (8) (2003) 986–1004.
[13] J. Kim, V. Kolmogorov, R. Zabih, Visual correspondence using energy minimization and mutual information, ICCV (2003) 1033–1040.
[14] A.J. Lee, U-Statistics, Theory and Practice, Marcel Dekker Inc., New York, 1990.
[15] E. Lehmann, Elements of Large Sample Theory, Springer Verlag, New York, 1999.

[16] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, IJCAI81 (1981) 674–679.
[17] A. Maintz, M.A. Viergever, A survey of medical image registration, Medical Image Analysis 2 (1) (1998) 1–36.
[18] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, Cambridge, 1999.
[19] M. Singh, H. Arora, N. Ahuja, Robust registration and tracking using kernel density correlation, CVPRW (2004) 174.
[20] C. Studholme, D.L.G. Hill, D.J. Hawkes, An overlap invariant entropy measure of 3d medical image alignment, Pattern Recognition 32 (1) (1999) 71–86.
[21] R. Szeliski, J. Coughlan, Spline-based image registration, International Journal of Computer Vision 22 (3) (1997) 199–218.
[22] J.-P. Thirion, Image matching as a diffusion process: an analogy with Maxwell's demons, Medical Image Analysis 2 (3) (1998) 243–260.
[23] P. Viola, W. Wells III, Alignment by maximization of mutual information, ICCV (1995) 16–23.
[24] C. Yang, R. Duraiswami, L. Davis, Efficient mean-shift tracking via a new similarity measure, CVPR (2005) 176–183.
[25] S.K. Zhou, R. Chellappa, Probabilistic identity characterization for face recognition, CVPR (2004) 805–812.
[26] Y.M. Zhu, S.M. Cochoff, Influence of implementation parameters on registration of mr and spect brain images by maximization of mutual information, Journal of Nuclear Medicine 43 (2) (2002) 160–166.