A Unified Statistical and Information Theoretic Framework for Multi-Modal Image Registration

Lilla Zollei, John Fisher and William Wells

Abstract

We formulate and interpret several multi-modal registration methods in the context of a unified statistical and information theoretic framework. A unified interpretation clarifies the implicit assumptions of each method yielding a better understanding of their relative strengths and weaknesses. Additionally, we discuss a generative statistical model from which we derive a novel analysis tool, the *auto-information function*, as a means of assessing and exploiting the common spatial dependencies inherent in multi-modal imagery. We analytically derive useful properties of the *auto-information* as well as verify them empirically on multi-modal imagery. Among the useful aspects of the *auto-information function* is that it can be computed from imaging modalities independently and it allows one to decompose the search space of registration problems.

1 Introduction

Registration of multi-modal data sets is the problem of identifying a geometric transformation (or a set of transformations) which maps the coordinate system of one data set to that of another (or others). Objective functions or similarity measures are special functions that evaluate the current quality of alignment. The goal of a registration problem can be interpreted as the optimization of such a function. There already exist a variety of registration methods whose objective functions are based on sound statistical principles. These include various maximum likelihood [4, 12], maximum mutual information [6, 13], minimum Kullback-Leibler divergence [1], minimum joint entropy [11] and maximum correlation ratio [9] methods. However, the relationship of these approaches to each other from the standpoint of explicit/implicit assumptions, use of prior information, performance in a given context, and failure modes has not received a great deal of attention. (One account on modeling assumptions in uni-modal registration techniques and a general maximum likelihood framework for a certain set of multi-modal registration approaches is presented in [10].) Additionally, while the various objective criteria may be well understood, their relationship to an underlying generative statistical model is often left unspecified. Our motivation here is three-fold. First, we formulate and interpret several registration algorithms in the context of a unified statistical and information theoretic framework which illuminates the similarities and differences between the various methods. Second, a unified statistical interpretation clarifies the implicit assumptions of each method yielding a better understanding of their relative strengths and weaknesses. Third, we discuss a generative statistical model from which we derive a novel analysis tool, the *auto-informaion function*, as a means of assessing and exploiting the common spatial dependencies inherent in multi-modal imagery. Currently, few, if any, of the commonly used registration algorithms exploit spatial dependencies except perhaps in an indirect way. Consequently, we devote significant discussion to the auto-information function, providing both theoretical and empirical analysis.

2 Unified View of Maximum-Likelihood, Mutual Information, and Kullback-Leibler Divergence

For simplicity, we consider the case of two *registered* data sets, $u(x)$ and $v(x)$ sampled on $x \in \Re^M$. These data sets represent, for example, two imaging modalities of the same underlying anatomy in an M-dimensional space. In practice, we observe $u(x)$ and $v_o(x)$ in which the latter is related to $v(x)$ by

$$v_o(x) = v(T^*(x)) \quad \text{or} \quad v(x) = v_o\left((T^*)^{-1}(x)\right),$$

where $T^* : \Re^M \to \Re^M$ is a bijective mapping corresponding to the unknown ground truth alignment transformation. The goal of registration is to find a

transformation estimate $\hat{T} \approx T^*$ (or equivalently its inverse) which optimizes some objective function of the observed data sets.[1]

We now discuss six objective criteria within a common statistical framework: maximum likelihood, approximate maximum likelihood, Kullback-Leibler divergence, iterated generalized likelihood, correlation ratio, and mutual information. We selected these similarity measures to include in our work as they form (though not completely exhaustively) a solid reference to a large group of currently used registration algorithms. Throughout our analysis, spatial samples x_i are modeled as random draws of an independent and identically distributed ($i.i.d.$) random variable X. Consequently, observed pixel / voxel intensities $v_o(x_i)$ and $u(x_i)$ are modeled as $i.i.d.$ random variables as well.

2.1 Maximum Likelihood

We begin our discussion with the classical maximum likelihood (ML) method of parameter estimation. In order to apply this method to image registration we must presume that we can model the joint densities of pixel intensities as a function of transformation parameters. Consequently, we can construct the joint probability density space $p(u, v; T)$. For the actual observations that we aim to align, the joint probability density function can be written as

$$u(x_i), v_o(x_i) \sim p\left(u, v_o\right) = p\left(u, v; T^*\right). \tag{1}$$

Thus we can write the ML estimate of the registration transformation as

$$T_{\mathrm{ML}} \quad = \quad \arg \max_T \sum_{i=1}^{N} \log p(u(x_i), v_o(x_i); T),$$

where N indicates the number of samples analyzed. It is important to note, in contrast to subsequent methods, that the joint observations remain static while the joint *density* under which we evaluate the observations is varied as a function of T.

There is a fundamental link between ML estimation and information theoretic quantities. Specifically, under the $i.i.d.$ assumption for fixed T and T^*,

$$\begin{aligned} T_{\mathrm{ML}} &\approx \arg \max_T - \left[D\left(p\left(u, v; T^*\right) \| p\left(u, v; T\right)\right) + H\left(p\left(u, v; T^*\right)\right)\right] \\ &= \arg \min_T \left[D\left(p\left(u, v; T^*\right) \| p\left(u, v; T\right)\right)\right], \end{aligned} \tag{2}$$

where $H(p)$ is the entropy of the distribution p and $D(p\|q)$ is the Kullback-Liebler (KL) divergence [3] between the distributions p and q. A detailed derivation of this relationship is included in the Appendix. Consequently, the ML estimate (when it is unique) is the one which minimizes the KL divergence

---

[1]Technically speaking, $u(x)$ may have undergone some transformation as well, but without loss of generality we assume it has not. If there were some canonical coordinate frame (e.g. an anatomical atlas) by which to register the data sets one might consider transformations on $u(x)$ as well.

between the ideal $p(u, v; T^*)$ and the modeled $p(u, v; T)$ distributions.

As a practical matter, one generally cannot model the joint density of observations as a function of *all* relative transformations T. Furthermore, even if such a model were available, as the relative transformation becomes "large" it is reasonable to assume that joint observations become independent (i.e. $p(u, v) = p(u)p(v)$). The utility of classical ML decreases greatly for such situations as a large set of transformations become equally likely. (In contrast, mutual information-based similarity measures define the solution of the registration problem to be as far away as possible from the space of such unlikely settings.)

2.2 Approximate Maximum Likelihood

While obtaining a joint density model over all relative transformations is perhaps impractical, suppose we have a model of the joint density of our data sets *when they are registered* which we will denote $p^\circ(u, v)$. Such a density is utilized in an approximate maximum likelihood registration framework (MLa) [4] which estimates T^* as

$$T_{\mathrm{MLa}} \quad = \quad \arg \max_T \sum_{i=1}^{N} \log p^\circ \left(u(x_i), v_o(T(x_i)) \right).$$

For practical reasons (e.g. one might be able to obtain reasonable density models of joint pixel intensities from previously registered data) and in contrast to the classical ML method, the joint observations are varied as a function of T while the density, p°, under which they are evaluated is held static.

Similarly to the relationship presented in the previous section, one can show that

$$T_{\mathrm{MLa}} \quad \approx \quad \arg \min_T \left[D \left(p^o \left(u, v_o(T) \right) \| p^o \left(u, v \right) \right) + H \left(p^o \left(u, v_o(T) \right) \right) \right] \tag{3}$$

$$= \quad \arg \min_T \left[D \left(p^o \left(u, v(T^* \circ T) \right) \| p^o \left(u, v \right) \right) + H \left(p^o \left(u, v(T^* \circ T) \right) \right) \right]. \tag{4}$$

Contrary to Eq.(2), we see that according to this formulation, both the KL-divergence and the entropy terms vary as a function of T, thus it is the sum of the two that needs to be optimized. The implicit assumption of the approximate maximum likelihood method is that as $T^* \circ T$ approaches T_I (the identity transformation), Eq.(4) is non-increasing.

In general, one cannot guarantee the validity of that hypothesis. The reason for this argument is related to the information theoretic notion of typicality [2]. Informally, typicality states that, with probability approaching unity, N independent draws from a density p with a corresponding entropy $H(p)$ have a likelihood very close to $-NH(p)$. Furthermore, N independent draws from a density q with corresponding entropy $H(q)$ evaluated under p have a likelihood very close to $-N(H(q) + D(q\|p))$ of which Eq. (4) is an application. Perhaps counter-intuitively, one can construct a density q such that typical draws from

q are *more likely* under p than typical draws from p. The same observation was empirically demonstrated in [1], which, in part, motivates the registration method that we introduce in the next section.

2.3 Kullback-Leibler Divergence

While one cannot guarantee that the full expression to be optimized in Eq.(4) is non-increasing as $T^* \circ T$ approaches T_I, the KL divergence term in it *does* satisfy such a requirement. Chung *et al* [1] suggest that one estimate T^* as

$$T_{\mathrm{KL}} = \arg \min_T D \left(\hat{p}\left(u, v(T^* \circ T); T\right) \| p^o\left(u, v\right) \right),$$

where $p^o\left(u, v\right)$ is constructed as in [4] from correctly registered data sets and $\hat{p}\left(u, v(T^* \circ T); T\right)$ is estimated from transformed sets of observed joint pixel intensities $\{u(x_i), v(T^* \circ T(x_i))\}$. The authors demonstrate empirically that this objective criterion, as expected, did not exhibit some of the undesirable local extrema encountered in the MLa method.

In relation to the previous methods, both the samples *and* the evaluation densities are being varied as a function of the transformation T while the algorithm is to approach the static joint probability density model constructed prior to the alignment procedure.

2.4 Iterated Generalized Maximum Likelihood

The objective function of another registration technique, which we refer to as *iterated generalized maximum likelihood* (MLit), can also be characterized in our framework. The alignment measure described in [12] defines an iterated maximum *a posteriori* (MAP) approach building on conditional probability densities and on prior knowledge about the distribution of the candidate transformations. Although the prior term carries important information about the transformation space, in this analysis we focus on the likelihood term of the problem formulation. In this case, the optimization goal of the method can be written as

$$T_{\mathrm{MLit}} \quad = \quad \arg \max_T \sum_{i=1}^{N} \log p(u(x_i), v_o(x_i)|T; T).$$

At this point, this criterion closely resembles the MLa formulation. However, instead of assuming that the joint probability density function of the input modalities is available for the correct alignment, the MLit method carries out the estimation of such a model online. At every iteration the joint probability model is re-estimated and at time t of such a process, the best alignment transformation can be defined as:

$$(T_{\mathrm{MLit}})_t \quad = \quad \arg \max_T \sum_{i=1}^{N} \log \hat{p}_{T_{t-1}}(u(x_i), v_o(x_i)|T; T_{t-1}),$$

or in our unified information theoretic framework as

$$(T_{\mathrm{MLit}})_t = \arg \min_T \left[D(\hat{p}_T(u, v_o(T)|T; T) || \hat{p}_{T_{t-1}}(u, v_o(T)|T; T_{t-1})) + \right.$$
$$\left. + H(\hat{p}_T(u, v_o(T)|T; T)) \right]. \qquad (5)$$

In Eq.(5), $\hat{p}_{T_{t-1}}$ refers to the joint probability function estimated with respect to the best current estimate (T_{t-1}) of the aligning transformation. (That estimate is defined at the previous, $(t-1)$th iteration.)

Although, experimentally, good registration results have been reported by applying the MLit technique, we need to investigate the key assumption on which its performance relies. The algorithm presumes that using the best current estimate of the transformation (T_{t-1}) it is possible to find an even more likely aligning transformation (given that the optimal alignment setting has not yet been recovered). Or in other words, the likelihood of samples drawn from density p_T but evaluated under density $p_{T_{t-1}}$ could be greater than the likelihood of samples both drawn from and evaluated under density $p_{T_{t-1}}$:

$$\mathcal{L}_{T_{t-1}}(T) = \int p_T \log p_{T_{t-1}} du \quad > \quad \mathcal{L}_{T_{t-1}}(T_{t-1}) = \int p_{T_{t-1}} \log p_{T_{t-1}} du \qquad (6)$$

A rigorous description of the model under which this assumption holds is still under investigation. However, we demonstrate below that if the MLit approach (maximizing the likelihood criterion with respect to the old transformation estimate) does converge, it converges to the minimum of the joint entropy measure. For local search scenarios, we support our argument by the fact that the gradient of the likelihood function evaluated under the old transformation estimate (Eq.(7)) is equivalent to the negative of the entropy measure gradient evaluated at T_{t-1} (Eq.(8)). In the following equations T^i represents the ith component of transformation T and $\mathcal{L}$ denotes the likelihood function.

$$\mathcal{L}_{T_{t-1}}(T) = \int p_T \log p_{T_{t-1}} du$$
$$\frac{\partial}{\partial T^i} \mathcal{L}_{T_{t-1}}(T) = \int \frac{\partial p_T}{\partial T^i} \log p_{T_{t-1}} du$$
$$\nabla_T \mathcal{L}_{T_{t-1}}(T) = \int \nabla_T p_T \log p_{T_{t-1}} du \qquad (7)$$

$$H(T) = -\int p_T \log p_T du$$

$$\frac{\partial}{\partial T^i} H(T) = -\int \left(\frac{\partial p_T}{\partial T^i} \log p_T + \frac{\partial p_T}{\partial T^i} \right) du$$

$$\frac{\partial}{\partial T^i} H(T) = -\int \left(\frac{\partial p_T}{\partial T^i} \log p_T \right) du - \int \left(\frac{\partial p_T}{\partial T^i} \right) du$$

$$\frac{\partial}{\partial T^i} H(T) = -\int \frac{\partial p_T}{\partial T^i} \log p_T du$$

$$\nabla_T H(T) = -\int \nabla_T p_T \log p_T du$$

$$\nabla_T H(T_{t-1}) = -\int \nabla_T p_T \log p_{T_{t-1}} du \qquad (8)$$

More generally, we conjecture that the following conditions are sufficient to make the assumptions by the MLit method hold globally: the minimum entropy configuration corresponds to the correct alignment of the input images, the marginal densities of the input images do not vary with respect to the various transformations applied to them, the joint entropy of the inputs decreases as the new transformation is applied and finally both of the joint densities can be written as the convex combination of the same two densities (one being the ideal joint probability density at the solution and the other corresponding to the product of marginals, the independent scenario).

The iterated generalized maximum likelihood method applies a non-parametric approach to best approximate the ideal joint density function while estimating the best aligning transformation. It successively identifies a transformation that further maximizes the likelihood criterion. Thus if the conjecture holds, this method gradually drives towards a moving as opposed to a static model used in the MLa and the KL-divergence framework.

2.5 Correlation Ratio

When defining correlation ratio [9] as a similarity metric, one makes the assumption that there is a functional relationship between the input images at the correct registration position. Describing that relationship via an intensity function f as $u(x_k) = f(v_o(T(x_k))) + \epsilon_k \quad \forall k$, where ϵ_k refers to additive stationary Gaussian noise, correlation ratio is defined as

$$\eta^2(u|v_o) = 1 - \frac{\text{Var}(u - \hat{f}(v_o))}{\text{Var}(u)}. \qquad (9)$$

This similarity metric can also be explained in the maximum likelihood framework [10]. The joint probability density function of interest is expressed in a product form $P(u, v_o; T) = P(v_o)P(u|v_o; T)$, and as $P(v_o)$ does not depend on

the transformation, it is the $P(u|v_o; T)$ term that is to be optimized with respect to T. Instead of experimentally defining (and fixing) the model joint density function at the correct registration pose, the optimal probability density function is estimated online. But contrary to the MLit method, here, a parametric model is used. For a particular transformation T, the metric to be optimized with respect to parameters Θ is then $P(u|v_o; T, \Theta)$.

Finding the correct alignment of the input images is formulated as a coupled optimization task: $P(u|v_o; T, \Theta)$ is to be maximized both with respect to T and Θ. The necessary alternating optimization steps are equivalent to the optimization of Eq.(9), due the following exponential relationship ([10]):

$$\eta^2(u|v_o(T)) = 1 - \frac{1}{k}e^{2U(T)/N}, \quad k = 2\pi \text{Var}(u).$$

In our unified statistical framework, we can define the correlation ratio function as:

$$T_{\text{CR}} = \arg\min_T \left[D(\hat{p}_\Theta(u|v_o(T)) || \hat{p}_{\Theta^*}(u|v_o(T))) + H(\hat{p}_\Theta(u|v_o(T))) \right], \quad (10)$$

where $\Theta^* = \arg\max_\Theta p(u|v_o, T_{t-1}, \Theta)$ for a particular transformation T_{t-1}.

The objective function formulation using correlation ratio (Eq.(10)) is also closely related to MLa (Eq.(3)), however, they are distinctly different. While in the former we face two separate, in the latter we address a single optimization task. The approximate maximum likelihood method also makes the assumption that a static model of the joint density function of the input modalities is adequate to describe all input data sets (of corresponding modalities); in contrast, according to the correlation ratio approach, the joint density function is not the same for all registered data sets and it needs to be estimated separately for each alignment scenario. In fact, with the re-estimation requirement, it is possible to obtain a more accurate density model per case, but the sequential optimization of two individual functions could also get attracted to less favorable local solutions. Similarly to the MLit method, correlation ratio is also attracted to a moving point in the solution space. It, however, applies a parametric framework.

2.6 Maximum Mutual Information and Joint Entropy

As has been amply documented in the literature [6, 7, 8, 13], mutual information (MI) is a popular information theoretic objective criterion which estimates the transformation parameter T as

$$T_{\text{MI}} = \arg\max_T I\left(u; v_o(T)\right) = \arg\max_T I\left(u; v(T^* \circ T)\right),$$

where MI is defined to be a function of marginal and joint entropy terms

$$I(u; v\left(T^* \circ T\right)) = H(p(u)) + H(p(v\left(T^* \circ T\right))) - H(p(u, v\left(T^* \circ T\right))). \ (11)$$

If T is restricted to the class of symplectic transformations (i.e. volume preserving), then $H(p(u))$ and $H(p(v(T^* \circ T)))$ are invariant to T. In that case,

maximization of MI is equivalent to minimization of the joint entropy term, $H(p(u, v\,(T^* \circ T)))$, the presumption being that this quantity is minimized when $T_{\mathrm{MI}} = (T^*)^{-1}$.

MI can also be expressed as a KL divergence measure [3]

$$I\,(u, v\,(T^* \circ T)) = D\,(p(u, v\,(T^* \circ T)) \| p(u) p(v\,(T^* \circ T))),$$

that is, mutual information is the KL divergence between the observed joint density term and the product of the marginals. Accordingly, the implicit assumption of MI methods is that as $T^* \circ T$ diverges from T_I or in other words as we are getting farther away from the ideal registration pose, the joint intensities look increasingly independent. This allows us to write the MI optimization problem as maximizing the distance from the scenario when the input images are completely independent:

$$T_{\mathrm{MI}} = \arg \max_{T} D\,(\hat{p}(u, v(T^* \circ T); T) \| \hat{p}(u) \hat{p}(v(T^* \circ T); T)).$$

As in the KL divergence alignment approach, both the samples and the evaluation densities are being simultaneously varied as a function of the transformation T. However, instead of approaching a model point in the solution space, the aim is to move farthest away from the worst case scenario.

Recently, numerous variations on the mutual information metric have been introduced; for instance, one making it invariant to image overlap (normalized mutual information [11]) and another enhancing its robustness using additional image gradient information (gradient-augmented mutual information [7]). In this report, we do not list and analyze all of these given that they operate with similar underlying principles.

Considering the collection of approaches discussed, we see that the MLa and KL divergence methods exploit prior information in the form of joint density estimates over previously registered data. Subsequently, both make similar implicit assumptions regarding the behavior of joint intensity statistics as $T^* \circ T$ approaches T_I closer to the ideal alignment. In contrast, the correlation ratio, the iterated generalized maximum likelihood method and the MI approaches make no use of prior joint statistics – estimating these instead during the search process. While the former two still try to model the correct density function at alignment, the MI approach just assumes (implicitly) that as $T^* \circ T$ approaches T_I, the joint intensity statistics become increasingly dependent, again, as measured by a KL divergence term. In light of this, we now define the *auto-information function* as an empirical analysis tool for exploring aspects of these assumptions.

3   Auto-, Cross-Information Functions

We define the *auto-* and *cross-information* functions. The functions measure statistical dependence, indexed over transformation parameters, much as the well-known *auto-correlation* function measures the degree of second-order correlation as a function of displacement. Given two different image modalities, u and v, we simply define the auto- and cross-information functions as:

$$
\begin{aligned}
R_u^I(T) &= I(u(x); u(T(x))) \text{ and} \\
R_{u,v}^I(T) &= I(u(x); v(T(x))),
\end{aligned}
$$

where $I(u; v)$ is the mutual information measure already introduced in Eq.(11). Analysis of such functions, in particular the auto-information metric (which can be computed *prior* to registration on the individual input images), may provide guidance for commonly used coarse-to-fine search strategies. Additionally, further spatial properties might also be inferred from the auto-information function which lead to better and faster converging registration algorithms.

This new approach can be described in the context of the following latent variable model

$$
p(u, v, l) = p_l(l_1, \cdots, l_N) \prod_i p_{u|l}(u_i|l_i) \, p_{v|l}(v_i|l_i),
$$

where the sets $\{u_1, \cdots, u_N\}$ and $\{v_1, \cdots, v_N\}$ represent observations (e.g. pixels or voxels) of two different image modalities at corresponding coordinate system locations and $\{l_1, \cdots, l_N\}$ a set of latent variables which describe tissue properties (e.g. label types). The model simply asserts the independence of the observations *conditioned* on the latent variables and it does not specify the joint properties of $\{l_1, \cdots, l_N\}$. A partial or a full description of the latter could also be incorporated. A graphical model[2] depicting the same problem formulation is shown on Fig. 1. Each of the algorithms cited in the previous sections corresponds to a hypothesis over this statistical model differing only in which aspects of the graph are specified or assumed *a priori*.

The proposed approach has two notable consequences. First, spatial dependencies in the observations arise directly from known or assumed spatial dependencies in the latent variables. Second, bounds on the spatial dependencies (modulo the unknown transformation) can be *estimated* from the individual imaging modalities. In particular, it is easily derived that the auto-information functions of induced images lower bound that of the underlying latent anatomy and the cross-information values for the pairs of corresponding image elements is always greater than or equal to that of non-corresponding ones. For proofs of these claims, see the Appendix.

$$
I(u_j; u_k), I(v_j; v_k) \leq I(l_j; l_k) \text{ and } I(u_j; v_j) \geq I(u_j; v_k) \quad \forall \, j, k = 1, ..., N. \quad (12)
$$

---

[2]A similar representation incorporating voxel positions has been recently introduced for elastic image registration via conditional probability computations [5].
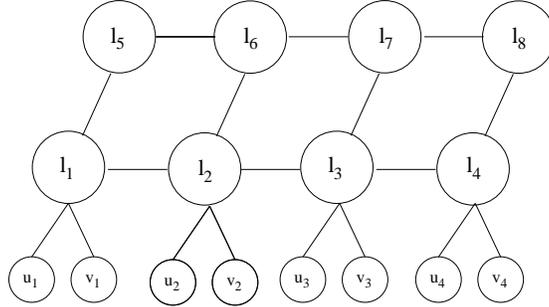
Figure 1: Example of a latent anatomy model

With such inequalities we guarantee local extrema for the MI objective function. More importantly, Eq. 12 shows that under the latent variable model, MI as an objective criterion is guaranteed to have a local maximum about the point of correct registration. To our knowledge, while this property has been empirically observed and exploited, no sets of conditions have been established such that it could be rigorously proven.

3.1 Function properties

We introduce two key properties of the auto-information function: an identity equation and the transformation decoupling. They describe how transformations as a whole and their components individually influence the autoinformation function map if applied to the input image prior to the processing. Their utility is then demonstrated in the experiments section, in a uni- and a multi-modal framework.

Auto-Information Identity

We can define the following identity between the auto-information functions of two datasets (v and v_o) that are related via transformation T^* as $v_o(x) = v(T^*(x))$:

$$
\begin{aligned}
R_{v_0}^I(T) &= I(v_0(x); v_0(T(x))) = I(v(T^*(x)); v(T^* \circ T(x))) \\
&= I(v(y); v(T^* \circ T \circ (T^*)^{-1}(y))) = R_v^I(T^* \circ T \circ (T^*)^{-1}) \\
&= R_v^I(T'),
\end{aligned}
\tag{13}
$$

where T' is a similarity transformation of T by T^*. In other words, the auto-information function of a transformed image (v_o) can be computed from the auto-information function of its non-perturbed counterpart. This property is potentially very useful when examining how the auto-information function changes with respect to an initial transformation applied to the input image and we show an example of it in our experiments section.

Decoupling the transformation components

We demonstrate a way to decouple transformation components when searching for alignment between the input images (or similarity between their autoinformation function maps). This means, that the components of a transformation T^* relating the input images can be searched for separately, reducing the parameter space that needs to be traversed at any given time. In the realm of shearless affine transformations, each operation is composed of a scaling, a rotation and a displacement component. After adopting a convention for the composition order of these operations, we can write any such transformation as $t(s, r, d) = D(d) \circ R(r) \circ S(s)$. Then the transformation $T' = T^* \circ T \circ (T^*)^{-1}$ in identity Eq.(13) can be rewritten as:

$$T' = D(d^*) \circ R(r^*) \circ S(s^*) \circ D(d) \circ R(r) \circ S(s) \circ S(s^*)^{-1} \circ R(r^*)^{-1} \circ D(d^*)^{-1}.$$

If we now investigate the different subspaces of the auto-information map, we notice their unique dependence on certain components of the transformation. First, take a look at the displacement-only subspace, where $T(s, r, d) = D(d)$ at the map creation step. Then

$$
\begin{aligned}
T' &= T^* \circ T \circ (T^*)^{-1} \\
&= D(d^*) \circ R(r^*) \circ \underbrace{S(s^*) \circ D(d) \circ S(s^*)^{-1}}_{D(d')} \circ R(r^*)^{-1} \circ D(d^*)^{-1} \quad (14) \\
&= D(d^*) \circ \underbrace{R(r^*) \circ D(d') \circ R(r^*)^{-1}}_{D(d'')} \circ D(d^*)^{-1} \quad (15) \\
&= D(d^*) \circ D(d'') \circ D(d^*)^{-1} \quad (16) \\
&= D(d'') \\
&= R(r^*) \circ D(d') \circ R(r^*)^{-1} \\
&= R(r^*) \circ S(s^*) \circ D(d) \circ S(s^*)^{-1} \circ R(r^*)^{-1} \\
&= (R(r^*) \circ S(s^*)) \circ D(d) \circ (R(r^*) \circ S(s^*))^{-1}. \quad (17)
\end{aligned}
$$

In Eq. (14) the composition of a scaling, displacement and the inverse of the scaling operation corresponds to a simple displacement, $D(d')$. As the composition of a rotation, displacement and the inverse of the rotation operation is just another displacement, $D(d'')$ in Eq. (15), and displacement operations commute, the $D(d^*)$ terms cancel out in step Eq. (16). Thus the displacement-only subspace of the auto-information map (Eq.(17)) is invariant to displacement $D(d^*)$ component of T^*. Accordingly, we can search for the unknown $(R(r^*) \circ S(s^*))$ composition, by comparing the observed and modeled subspace maps, without considering any potential displacement element of the aligning transformation.

Similarly, let's consider the rotation-only subspace of the map, where $T(s, r, d) = R(r)$,

$$
\begin{aligned}
T' &= T^* \circ T \circ (T^*)^{-1} \\
&= D(d^*) \circ R(r^*) \circ S(s^*) \circ R(r) \circ S(s^*)^{-1} \circ R(r^*)^{-1} \circ D(d^*)^{-1}.
\end{aligned}
$$

By using the $(R(r^*) \circ S(s^*))$ estimate from the previous analysis, we can recover $D(d^*)$.

Finally, in the scaling-only subspace of the auto-information function map, where $T(s, r, d) = S(s)$, is

$$
\begin{aligned}
T' &= T^* \circ T \circ (T^*)^{-1} \\
&= D(d^*) \circ R(r^*) \circ S(s^*) \circ S(s) \circ S(s^*)^{-1} \circ R(r^*)^{-1} \circ D(d^*)^{-1} \\
&= D(d^*) \circ R(r^*) \circ S(s) \circ R(r^*)^{-1} \circ D(d^*)^{-1}.
\end{aligned}
$$

Thus the search strategy can be completed in the following way. Knowing the $D(d^*)$ estimate, compute $R(r^*)$ and then from $(R(r^*) \circ S(s^*))$ and $R(r^*)$, express $S(s^*)$. Such a sequential reduction in search space can facilitate a reduced computational cost in optimization. For a more restricted class of transformations, for example for 2D rigid-body motion, the parameter search could even be done in parallel. However, in general (for higher dimensions and for affine transformations), the individual searches have to be executed one after the other.

3.2 Experiments

In this section, we describe a set of experiments that were constructed to demonstrate the nature of the auto-information function and to give some insight for what kind of applications it might be useful.

We carried out experiments using both simulated and medical image datasets. For the below evaluation, we worked with images in 2D and defined the rotation to be carried out around the center point of the target image. Note also, that prior to running our experiments, we introduced a preprocessing step. We increased and zero-padded the background region of the images in order to ensure that no transformations result in cropped datasets. (This property is required to fully satisfy our theoretical assumptions when defining, for example, the identity relationship. In the future, we intend to investigate how this step restricts our experimental results and whether in the case of higher dimensional data sets it is still a reasonable criterion.)

The two pairs of medical images that we used for our experiments consisted of a pair of corresponding Proton Density (PD) and T2-weighted (T2) acquisitions and a pair of corresponding MRI and CT images of the head. (See these images on Fig. 2.)

Smoothing

We aimed to demonstrate how a smoothing operation would affect the nature of the auto-information function map. Thus we computed the 3D auto-information map for both an image and a smoothed version of it (created by a Gaussian filter with window size 5). As expected, after the smoothing operator was applied to the data, the auto-information map became significantly flatter and less peaky.
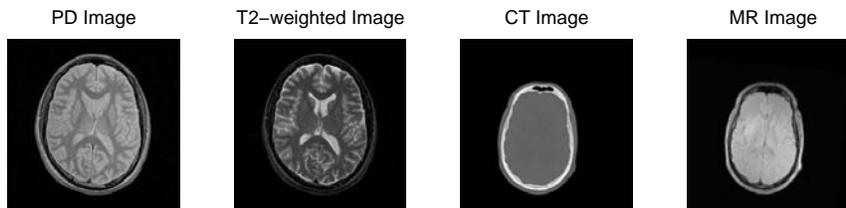
Figure 2: Medical image slices used for our experiments. Left-to-right: Corresponding Proton Density and T2-weighted images; Corresponding CT and MRI acquisitions.

While the initial map has a sharp peak at the zero offset pose and quickly decreasing lobes, in the case of the smoothed image that transition is much more gradual. An example showing the auto-information map slices, in the case of the original and the smoothed PD images is shown on Fig. (3). We plan to further analyze the relationship of such maps to identify parameters for hierarchical search mechanisms.

Changes due to an Initial Pose Difference

Just examining the auto-information map of the input images does not reveal much in the way of underlying structure embedded in the images. (See Fig. 3 (a), (b)). Therefore, we also examined the changes in the auto-information function maps due to an initial transformation applied to the input images. We created a map of the input images and a map of their transformed versions. (The transformation that we applied was the same in the case of both of the input images and it was comprised of both a displacement and a rotational component.) Comparing Fig. 3 (c)-(e) and (d)-(f), we note that there is a distinctive pattern of difference in the maps of one modality due to the initial transformation (the effect of the rotation, for example, is well visible on the slices). Although the delicate changes in the structure of these maps could be predicted/approximated by using the identity formula in Eq. (13), they are difficult to interpret at the first sight. Therefore, we display the difference images of the maps of the input with no initial transformation and that of the transformed image for both of the modalities. The results, (Fig. 3 (g) and (h)), computed on both the CT and MRI images, convey more information about the effects resulting from the transformation. This type of display also allows us to note that the difference maps of the two image modalities are almost identical. This is an essential observation which allows us to conclude that a fixed transformation applied to multi-modal images of the same underlying anatomy results in the same type of changes in the auto-information surfaces. This empirical observation gives indication of the utility of the auto-information function in the context of multi-modal registration and it encouraged us to apply the predictions established by the identity equation Eq.(13) not only in a uni- but also in a multi-modal setting.
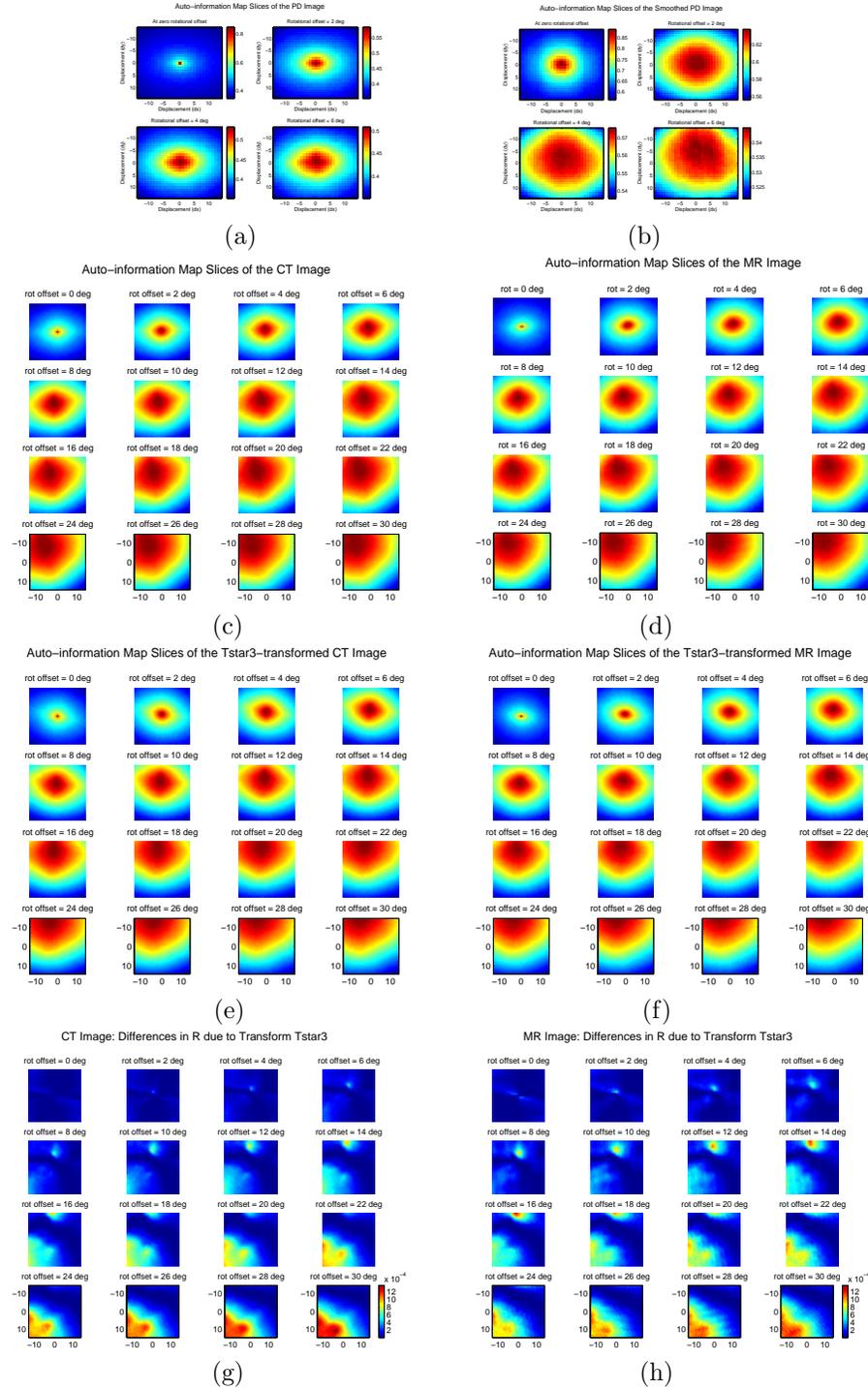
Figure 3: Auto-Information map slices of the (a) PD, (b) smoothed PD, (c) CT, (d) MR, (e) the transformed CT and (f) the transformed MR images. Squared difference maps between the auto-information map of the (g) CT and the transformed CT images and of the (h) MRI and the transformed MRI images. Note the similarities between the image slices of (g) and (h). The slices, each a map of translation, in all cases correspond to various rotational offsets in the auto-information map volume. (Top-to-bottom, left-to-right: the rotational offset is 0,2,...,30 degrees)
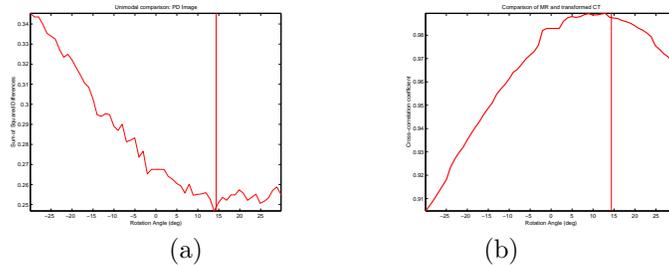
Figure 4: Rotation angle search in the displacement-only subspace: (a) Unimodal search using the PD image – minimizing sum of squared errors (b) Multimodal search using the MRI and CT images – maximizing cross-correlation coefficient. The ground truth solution in both figures is indicated by a vertical line.

Simple registration examples

In a set of preliminary 2D experiments we examined the displacement-only subspace of the auto-information map. We were interested in determining how accurately we could recover a rotational component of a rigid-body transformation applied to one of the input images both in a uni- and a multi-modal scenario. In the former, we were to align a PD image to a transformed version of itself, while in the latter an MRI slice to a similarly perturbed CT image. In both cases we used the identity relationship from Eq.(13) to model the subspaces of interest. We optimized the sum of squared differences and the cross-correlation coefficient, respectively, between the true and the modeled auto-information subspace maps. We decided to apply such simple similarity metrics for the optimization task as the subspaces to be compared were both composed of the same type of information, the autoinformation values (as opposed to, for example, intensities of different modalities). In Fig. 4, we show the results of these experiments. In the uni-modal scenario, the registration result closely matches the ground truth rotation angle (indicated by a vertical line on the graph). In the CT-MRI experiment the search solution was slightly off. This can be explained by the fact that the identity relationship used for modeling the zero rotational subspace in both sets of experiments is accurate only for the uni-modal setup. In the case of multiple modalities it is merely a close approximation. However, the results could still provide valuable global initialization for subsequent local searches.

4 Conclusion

We provided a unified statistical and information theoretic framework for comparing several well known multi-modal image registration methods. The consequence of which was to illustrate the underlying assumptions which distinguish them. Specifically, our investigation served to clarify the assumed behavior of

joint intensity statistics as a function of transformation parameters. This motivated the introduction of a latent variable generative model from which we were able to derive several interesting properties of the statistical dependencies across modalities. Significantly, we provided the first rigorous proof, to our knowledge, of the existence of a local maxima for the mutual information criterion about the point of correct registration in the context of the latent variable model.

We also introduced the auto- and cross-information functions which characterize the joint intensity statistics as a function of the relative transformation between images within and across modalities. Several properties of the auto-information function, which can be computed from each modality independently, were derived analytically and verified empirically. A significant aspect of the auto-information function is that it facilitates decoupling of the transformation parameters in the search space. Furthermore, our empirical results on anatomical data shows that the auto-information functions across modalities exhibit striking similarities. We conjecture that this property can be exploited in multi-modal registration methods currently in development. Further theoretical and empirical analysis of the properties of the auto- and cross-information functions are the subject of future research.

References

[1] A.C.S. Chung, W.M.W. Wells III, A. Norbash, and W.E.L. Grimson. Multi-modal image registration by minimizing kullback-leibler distance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 2 of *Lecture Notes in Computer Science*, pages 525–532. Springer, 2002.

[2] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[3] Kullback and Solomon. *Information Theory and Statistics*. John Wiley and Sons, New York, 1959.

[4] M. Leventon and W.E.L. Grimson. Multi-modal volume registration using joint intensity distributions. In *First International Conference on Medical Image Computing and Computer-Assisted Intervention*, Lecture Notes in Computer Science. Springer, 1998.

[5] A.M.C. Machado, M.F.M. Campos, and J.C. Gee. Bayesian model for intensity mapping in magnetic resonance image registration. *Journal of Electronic Imaging*, 12(1):31–39, Jan 2003.

[6] 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.

[7] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Image registration by maximization of combined mutual information and gradient information.

In *Proceedings of MICCAI*, Lecture Notes in Computer Science, pages 567–578. Springer, 2000.

[8] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[9] A. Roche, G. Malandain, and X. Pennec ad N. Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In *Proceedings of MICCAI*, volume 1496 of *Lecture Notes in Computer Science*, pages 1115–1124. Springer, 1998.

[10] A. Roche, G. Malandain, and N. Ayache. Unifying maximum likelihood approaches in medical image registration. *International Journal of Imaging Systems and Technology*, 11(7180):71–80, 2000.

[11] C. Studholme, D.L.G. Hill, and D.J. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[12] Samson Timoner. *Compact Representations for Fast Nonrigid Registration of Medical Images*. PhD thesis, Massachusetts Institute of Technology, 2003.

[13] W.M. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1:35–52, 1996.

A Maximum Likelihood and Information Theory

In this section we will demonstrate how the relationship between the Maximum Likelihood (ML) formulation and information theoretic quantities can be obtained (Eq. (2)):

$$T_{ML} \quad \approx \quad \arg\min_T \left[D\left(p\left(u,v;T^*\right) \| p\left(u,v;T\right) \right) + H\left(p\left(u,v;T^*\right) \right) \right].$$

The information theoretic Kullback-Leibler (KL) divergence [3] is a nonnegative quantity which can be defined as

$$D(p\|q) = E_p\left\{ \log\left(p/q\right) \right\} = \int p(x) \log\left(p(x)/q(x)\right) dx,$$

where p and q stand for probability densities. If we now apply this definition to determine the KL-divergence between the observed and the modeled joint

density functions,

$$
\begin{aligned}
D\left(p\left(u, v; T^*\right) \| p\left(u, v; T\right)\right) &= \\
&= \int p(u, v; T^*) \log \frac{p(u, v; T^*)}{p(u, v; T)} \, dudv \\
&= \int p(u, v; T^*) \log p(u, v; T^*) dudv - \int p(u, v; T^*) \log p(u, v; T) \, dudv \\
&= -H\left(p\left(u, v; T^*\right)\right) - \int p(u, v; T^*) \log p(u, v; T) \, dudv,
\end{aligned} \tag{18}
$$

where $H(p)$ is the entropy of the distribution p.

On the other hand, maximizing the sum of likelihood functions is equivalent to maximizing the normalized version of the same sum. Thus given Eq.(1) and the fact that the observations are *i.i.d.* draws, we can apply the weak law of large numbers

$$
\begin{aligned}
\max \sum_{i=1}^{N} \log p(u(x_i), v(T^*(x_i)); T) &= \max \frac{1}{N} \sum_{i=1}^{N} \log p(u(x_i), v(T^*(x_i)); T) \\
&\approx \max \mathrm{E}\left[\log p(u, v_o; T)\right].
\end{aligned} \tag{19}
$$

From Eq. (18),

$$
\begin{aligned}
-\left[D\left(p\left(u, v; T^*\right) \| p\left(u, v; T\right)\right) + H\left(p\left(u, v; T^*\right)\right)\right] &= \\
&= \int p(u, v; T^*) \log p(u, v; T) \, dudv = \mathrm{E}\left[\log p(u, v_o; T)\right].
\end{aligned} \tag{20}
$$

Equations (19) and (20) finally allow us to conclude that

$$
\begin{aligned}
T_{ML} &\approx \arg \max_{T} - \left[D\left(p\left(u, v; T^*\right) \| p\left(u, v; T\right)\right) + H\left(p\left(u, v; T^*\right)\right)\right] \\
&= \arg \min_{T} \left[D\left(p\left(u, v; T^*\right) \| p\left(u, v; T\right)\right) + H\left(p\left(u, v; T^*\right)\right)\right].
\end{aligned}
$$

B Bounds on Spatial Dependencies

Both of the relationships presented in (12)

$$
\begin{aligned}
I(u_j; u_k), I(v_j; v_k) &\leq I(l_j; l_k) \tag{21} \\
I(u_j; v_j) &\geq I(u_j; v_k) \quad \forall j, k = 1, ..., N, \tag{22}
\end{aligned}
$$

can be derived from the Data Processing Inequality theorem [2]. Accordingly, if X, Y and Z are random variables forming a Markov chain $(X \rightarrow Y \rightarrow Z)$, then $I(X; Y) \geq I(X; Z)$, i.e. no processing of Y can increase the information that Y contains about X.

Proof I

The relationship between the random variables appearing in inequality (21), $v_j \leftarrow l_j - l_k \rightarrow v_k$ (see Fig. 1), can be rewritten in two different forms using Bayes rule: $v_j \leftarrow l_j \leftarrow l_k \leftarrow v_k$ and $v_j \rightarrow l_j \rightarrow l_k \rightarrow v_k$. Using these formulations and applying the Data Processing Inequality theorem, we obtain:

$$I(v_k; l_k) \geq I(v_k; l_j) \geq I(v_k; v_j) \text{ and } I(l_k; l_j) \geq I(l_k; v_j)$$
$$I(v_j; l_j) \geq I(v_j; l_k) \geq I(v_j; v_k) \text{ and } I(l_j; l_k) \geq I(l_j; v_k)$$

Given $I(X; Y) = I(Y; X)$, we can establish $I(l_j; l_k) \geq I(v_j; v_k) \quad \forall\, j, k$.

Proof II

In a similar manner as above, we can obtain the following inequalities for u_j, v_j, l_j, l_k, v_k (see again Fig. 1):

$$I(v_j; l_j) \geq I(u_j; v_j) \text{ and } I(v_k; l_k) \geq I(v_k; l_j) \geq I(v_k; u_j).$$

Applying Bayes rule, we can establish the following relationships:
$v_j \leftarrow l_j \leftarrow l_k \leftarrow u_k$ and $v_j \leftarrow l_j \leftarrow u_j$. As we assume that $I(v_k; l_k) = I(v_j; l_j)$, we need to consider two scenarios: (a) $l_k \rightarrow l_j$ indicates a lossless relationship and (b) $l_k \rightarrow l_j$ indicates a lossy connection. In the former case, $I(u_j; v_k) = I(u_j; v_j)$, and in the latter $I(u_j; v_k) < I(u_j; v_j)$. Therefore, we can conclude that $I(u_j; v_j) \geq I(u_j; v_k)$, which was stated in inequality (22).