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

Lilla Zöllei[1], John W. Fisher III[1], and William M. Wells III[1,2]

[1] Massachusetts Institute of Technology,
Artificial Intelligence Laboratory, Cambridge, MA 02139, USA
{lzollei, fisher, sw}@ai.mit.edu
[2] Department of Radiology, Harvard Medical School and
Brigham and Women's Hospital, Boston, MA 02115, USA
sw@bwh.harvard.edu

Abstract

We formulate and interpret several 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 multiple 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). There exist a variety of registration methods whose objective functions are based on sound statistical principles. These include maximum likelihood [4], maximum mutual information [6, 9], minimum KL divergence [1] and minimum joint entropy [8] 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. 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-information 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 $x \in \Re^N$. These data sets represent, for example, two imaging modalities of the same underlying anatomy. In practice, we observe $u(x)$ and $v_o(x)$ in which the latter is related to $v(x)$ by

$$v_o(x) = v(T^*(x)) \tag{1}$$

$$v(x) = v_o\left(\left(T^*\right)^{-1}(x)\right), \tag{2}$$

where $T^* : \Re^N \rightarrow \Re^N$ is a bijective mapping. The goal of registration is to find $\hat{T} \approx T^*$ (or equivalently its inverse) which maximizes some objective criterion of the observed data sets.[3]

We now discuss four objective criteria within a common statistical framework. 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 with the classical maximum likelihood (ML) method of parameter estimation. In order to apply the method to image registration we must presume that we can model the joint densities of pixel intensities as a function of transformation parameters. That is

$$u(x_i), v_o(x_i) \sim p\left(U, V; T^*\right), \tag{3}$$

and the ML estimate of the transformation on $v(x)$ is

$$T_{ML} = \arg\max_T \sum_{i=1}^{N} \log p(u(x_i), v(T^*(x_i)); T), \tag{4}$$

[3] 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.

where N is the number of samples. 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^*,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log p(u(x_i), v_o(x_i); T) = \\ -\left(H\left(p\left(u,v;T^*\right)\right) + D\left(p\left(u,v;T^*\right) \| p\left(u,v;T\right)\right)\right), \quad (5)$$

where $H(p)$ is the entropy of the distribution p and $D(p\|q)$ is the Kullback-Leibler (KL) divergence [3] between the distributions p and q. KL divergence is a nonnegative quantity defined as $D(p||q) = E_p \left\{\log\left(p/q\right)\right\} = \int p(x) \log\left(p(x)/q(x)\right) dx$. Eq.(5) follows from Eq.(3) (the observations are *i.i.d.* draws), subsequently the (normalized) summation of Eq.(4) is equivalent to an expectation by the weak law of large numbers. Consequently, the ML estimate (when it is unique) is the one which minimizes the KL divergence between the true and hypothesized 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. $p(u,v) = p(u)p(v)$ - which is an essential assumption exploited by mutual information approaches). The utility of classical ML decreases greatly for such situations as a large set of transformations become equally likely.

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 $p^\circ\left(u,v\right) = p(u,v;T_I)$ where T_I indicates the identity transformation. Such a density is utilized in the approximate maximum likelihood method (MLa) [4] which estimates T as

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

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 under which they are evaluated is held static.

Similar to previous statements, one can show that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log p^\circ(u(x_i), v(T^* \circ T(x_i))) = \\ -\left(H\left(p\left(u,v;T^* \circ T\right)\right) + D\left(p\left(u,v;T^* \circ T\right) \| p\left(u,v;T_I\right)\right)\right). \quad (6)$$

As compared to Eq.(5) we see that both terms vary as a function of T. In general, one cannot guarantee that the combination of terms will be minimized when $T^* \circ T = T_I$. This 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. (6) 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 implicit assumption of the approximate maximum likelihood method is therefore that as $T^* \circ T$ approaches T_I Eq.(6) is nondecreasing. In [1] it was shown empirically that this assumption does not always hold which, in part, motivates the registration method suggested in that work.

2.3 Kullback-Leibler Divergence

While one cannot guarantee that Eq.(6) is nondecreasing as $T^* \circ T$ approaches T_I, the second term of Eq.(6) *is* nondecreasing as $T^* \circ T$ approaches T_I. Consequently, Chung *et al* [1] suggest that one estimate T as

$$\begin{aligned} T_{KL} &= \arg\min_T \sum_{u,v} \hat{p}\left(u, v; T^* \circ T\right) \log \frac{\hat{p}\left(u, v; T^* \circ T\right)}{p^o\left(u, v\right)} &(7) \\ &\approx \arg\min_T D\left(\hat{p}\left(u, v; T^* \circ T\right) \| p\left(u, v; T_I\right)\right), &(8) \end{aligned}$$

where $p\left(u, v; T_I\right)$ is estimated as in [4] from registered data sets and $\hat{p}\left(u, v; T^* \circ T\right)$ is estimated from transformed sets of observed joint pixel intensities $\{u(x_i), v_o(T(x_i))\}$. In relation to the previous methods, both the samples *and* the evaluation densities are being varied as a function of the transformation T. In [1] it was demonstrated empirically that this objective criterion, as expected, did not exhibit some of the incorrect/undesirable local extrema encountered in the MLa method.

2.4 Maximum Mutual Information and Joint Entropy

As has been amply documented in the literature [6, 7, 9], mutual information (MI) is a popular information theoretic objective criterion which estimates the transformation parameter T as

$$T_{\mathrm{MI}} = \arg\max_T I\left(u; v_o(T)\right) = \arg\max_T I\left(u; v(T^* \circ T)\right), \tag{9}$$

where MI is a function of marginal and joint entropy terms as

$$I(u; v_o\left(T\right)) = H(p(u)) + H(p(v_o\left(T\right))) - H(p(u, v_o\left(T\right))). \tag{10}$$

Again by typicality (or by the Weak Law of Large Numbers), this expression can be approximated as

$$\begin{aligned} I(u; v_o(T)) \approx & -\frac{1}{N}\sum_{i=1}^{N} \log \hat{p}\left(u(x_i)\right) - \frac{1}{N}\sum_{i=1}^{N} \log \hat{p}\left(v_o(T(x_i))\right) \\ & + \frac{1}{N}\sum_{i=1}^{N} \log \hat{p}\left(u(x_i), v_o(T(x_i))\right)\right), \end{aligned} \quad (11)$$

where x_i are *i.i.d.* draws of the spatial variable X and $\hat{p}(.)$ are density estimates. There exist variants in the literature which approximate mutual information by other means, but for our purposes we will consider them all to be equivalent.

If T is restricted to the class of symplectic transformations (i.e. volume preserving) then $H(u)$ and $H(v_o(T))$ are invariant to T. In that case, maximization of MI is equivalent to minimization of the joint entropy term, $H(u, v_o(T))$, the presumption being that the joint entropy is minimized when $T_{\mathrm{MI}} = (T^*)^{-1}$. As in the KL divergence approach, both the samples and the evaluation densities are being simultaneously varied as a function of the transformation T.

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

$$I\left(u, v_o(T)\right) = D\left(p(u, v; T^* \circ T) \| p(u)p(v; T^* \circ T)\right), \quad (12)$$

that is, mutual information is the KL divergence between the observed joint density and the product of its marginals. The implicit assumption of MI methods is that as $T^* \circ T$ diverges from T_I, joint intensities look increasingly independent.

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. In contrast, the MI approach makes no use of prior joint statistics – estimating these instead during the search process. On the other hand, MI approaches, implicitly assume 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:

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

where $I(u;v)$ is the mutual information measure already defined in Eq.(10). Analysis of such functions, in particular the auto-information function which can be computed *prior* to registration, 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

$$p\left(u,v,l\right) = p_l\left(l_1,\cdots,l_N\right)\prod_i p_{u|l}\left(u_i|l_i\right)p_{v|l}\left(v_i|l_i\right), \tag{13}$$

where the sets $\{u_1,\cdots,u_N\}$ and $\{v_1,\cdots,v_N\}$ represent observations of two different image modalities and $\{l_1,\cdots,l_N\}$ a set of latent variables which describe tissue properties (e.g. label types). The joint properties of $\{l_1,\cdots,l_N\}$ may be only partially specified. 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 model simply asserts the independence of the observations *conditioned* on the latent variables. An example is shown in the graphical model [4] of Fig. 1.

The proposed formulation 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:

$$I(u_j;u_k), I(v_j;v_k) \leq I(l_j;l_k) \text{ and } I(u_j;v_j) \geq I(u_j;v_k). \quad \forall\, j,k = 1,...,N \tag{14}$$

Consequently, the auto-information functions of induced images lower bound that of the underlying latent anatomy and we guarantee local extrema for the MI objective function given that the auto-information values for the pairs of corresponding image elements is always greater than or equal to that of non-corresponding ones. (For proofs, see the Appendix.) More importantly, Eq. 14 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, no sets of conditions have been established such that it could be rigorously proven.

3.1 Auto-Information Identity

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

$$\begin{aligned} R^I_{v_0}(T) &= I(v_0(x); v_0(T(x))) = I(v(T^*(x)); v(T^* \circ T(x))) \\ &= I(v(y); v(T^* \circ T \circ (T^*)^{-1}(y))) = R^I_v(T^* \circ T \circ (T^*)^{-1}) = R^I_v(T'), \end{aligned} \tag{15}$$

[4] A similar representation incorporating voxel positions has been recently introduced for elastic image registration via conditional probability computations [5].

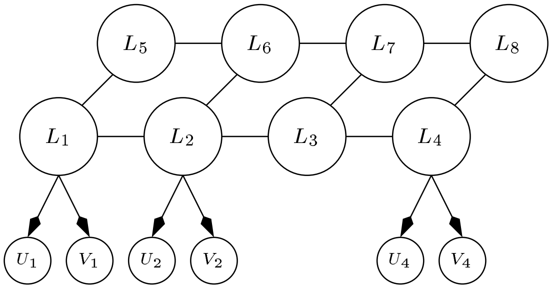


Fig. 1. Example of a latent anatomy model

where T' is a similarity transform 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 the initial input image. This property is potentially very useful when examining how the auto-information function changes with respect to an initial transformation.

3.2 Experiments

In this section, we describe several experiments that were constructed to demonstrate certain key properties of the auto-information function and to give some insight for which applications it might be useful.

We carried out experiments using both simulated and medical image datasets. To date, the experiments have been carried out in 2D and the nature of the transformations was restricted to rigid-body movements (displacement and rotation). We defined the rotation to be carried out around the center point of the input image. Note also, that prior to running our experiments, we introduced a preprocessing step. We increased and padded with zero the background region of the images in order to ensure that no cropping takes place as a result of transformations. (This property is required to fully satisfy our assumptions defining, for example, the identity relationship).

We used two pairs of medical images for our experiments. One pair consisted of a Proton Density and a T2-weighted acquisition and the other of a corresponding MRI and CT image of the head. (See these images on Fig. 2.)

Identity In order to experimentally verify the relationship established in Eq. (15), we compared the auto-information maps of initially transformed datasets to the same maps that were estimated by the identity. Up to numerical precision, the identity holds, the summed squared difference values are zero.

Smoothing With another set of experiments we aimed to demonstrate how the smoothing operation affects the auto-information function maps. We computed the 3D auto-information map for both the image and a smoothed version of it (created by a Gaussian filter with window size 5). As expected, the auto-information map became significantly flatter and less peaky after the smoothing operator was applied to the data. While the initial map has a sharp peak at the

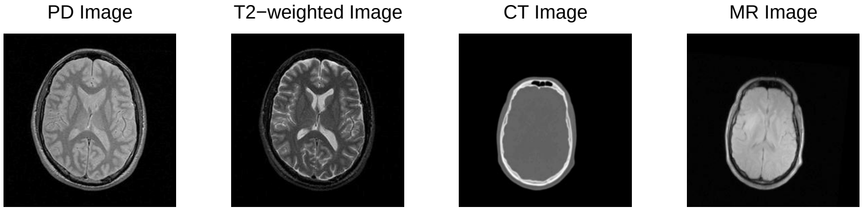


Fig. 2. Medical input images used for our experiments. Left-to-right: Corresponding Proton Density and T2-weighted images; Corresponding CT and MRI acquisitions.

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).

Changes due to an Initial Pose Difference 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 image. We created a map of the input image and a map of its transformed version. (The transformation that we applied is further referred to as T_3^* and it is comprised of both a displacement and a rotational component.) Comparing Fig. 3 (c),(d),(e) and (f), we note that there is a distinctive pattern of difference in the maps due to the initial transformation applied to the input (the effect of the rotation, for example, is well visible on the slices). However it is difficult to interpret these at the first sight. Therefore, we displayed the difference images of the maps of the input with no initial transformation and that of the transformed image. 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. We observe that the two difference maps are almost identical, which allows us to conclude that a fixed transformation applied to multi-modal images of the same underlying object results in the same type of changes in the auto-information surfaces. This empirical observation is encouraging in that it gives indication of the utility of the auto-information function in the context of registration.

Decoupling the transformation components In this section, we demonstrate a way to decouple the transformation components when searching for alignment (or the initial pose) in a registration scenario. It turns out that one can use the autoinformation function to decouple the components of transformation T and search for them separately. (Compare auto-information map slices in Fig. 3 (c) and (e), for example.)

The decoupling observation is explained as follows. If T^* is a composition of a displacement and a rotational component, then it can be written as a rotation operation followed by displacement: $T^*(r^*, d^*) = D(d^*) \circ R(r^*)$. Then consider the identity in Eq. (15); if we rewrite the transformation composition of $T_{comp} =$

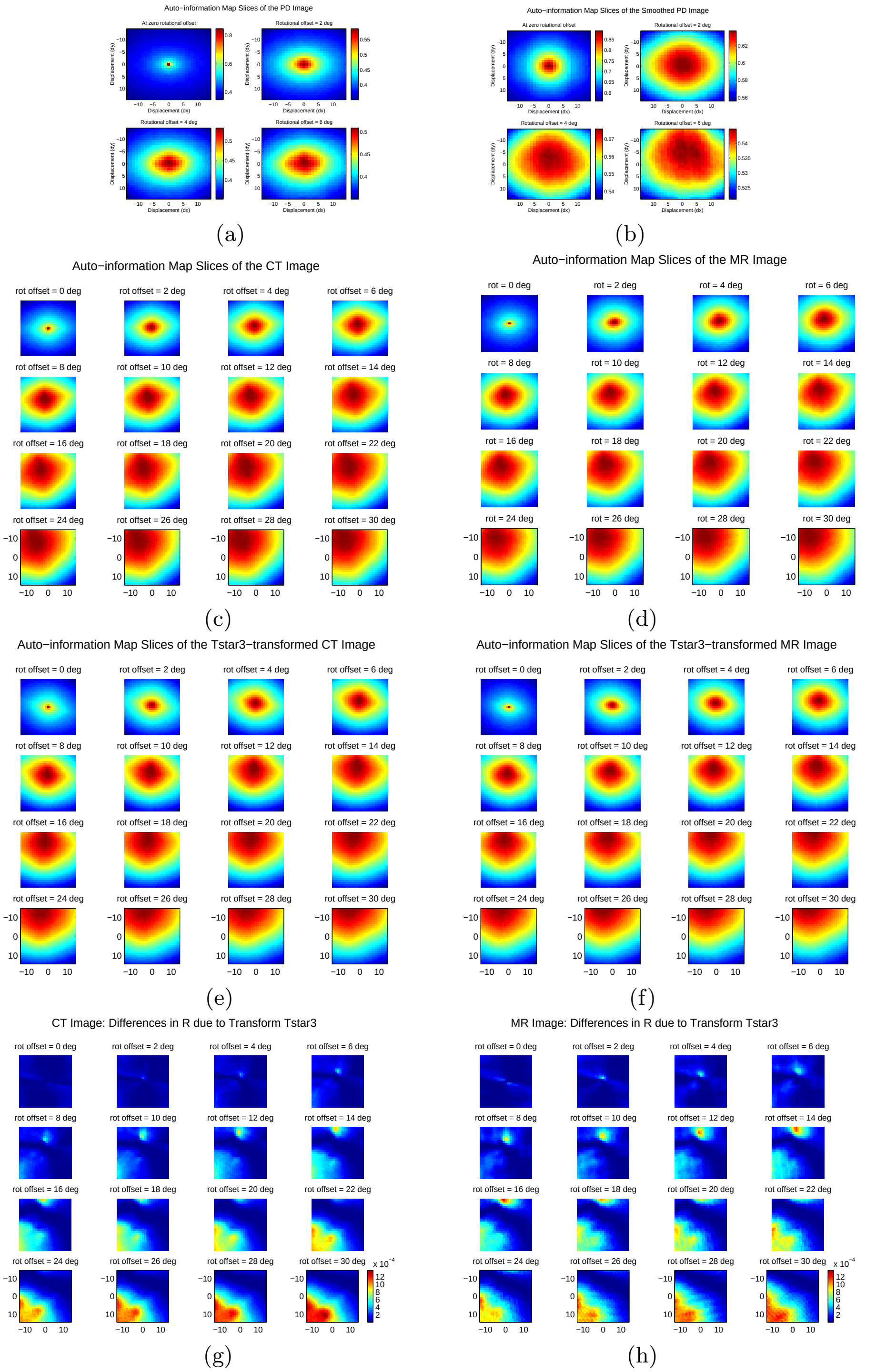


Fig. 3. Auto-Information map slices of the (a) PD, (b) smoothed PD, (c) CT, (d) MR, (e) the transformed CT and (f) the transformed MR images. Squared difference maps between the auto-information map of the (g) CT and the T_3^*-transformed CT images and of the (h) MRI and the T_3^*-transformed MRI images. 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)

$T^* \circ T \circ (T^*)^{-1}$ with the above expression for T^*, we get: $T_{comp} = D(d^*) \circ R(r^*) \circ T \circ R\left((r^*)^{-1}\right) \circ D\left((d^*)^{-1}\right)$. Also after replacing T with $T(r,d) = D(d) \circ R(r)$:

$$T_{comp} = D(d^*) \circ R(r^*) \circ D(d) \circ R(r) \circ R\left((r^*)^{-1}\right) \circ D\left((d^*)^{-1}\right).$$

Now, if we only examine the auto-information map in the displacement dimensions of T, i.e.: $T(r,d) = D(d)$, we would compute the transformation

$$T_{comp} = D(d^*) \circ R(r^*) \circ D(d) \circ R\left((r^*)^{-1}\right) \circ D\left((d^*)^{-1}\right). \tag{16}$$

As the composition of a rotation, displacement and the inverse of the rotation operation is just another displacement, $D(d')$, and displacement operations commute, the $D(d^*)$ terms cancel out:

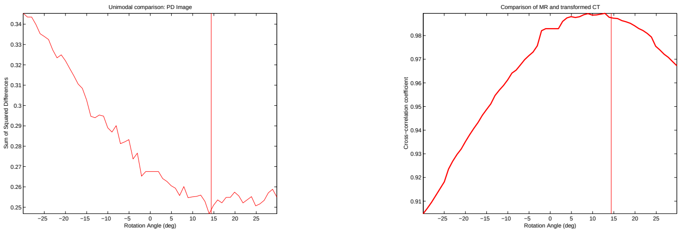
$$T_{comp} = D(d^*) \circ D(d') \circ D\left((d^*)^{-1}\right) = D(d') = R(r^*) \circ D(d) \circ R\left((r^*)^{-1}\right).$$

Thus the zero-rotation subspace of the autoinformation function is invariant to displacement. Accordingly, we can search for the unknown rotational component, by comparing subspace maps, without considering any potential displacement component of the aligning transformation. Such a reduction in search space facilitates a reduced computational cost in optimization.

In a set of preliminary experiments, we looked at the zero rotation subspace of the auto-information map and searched for the rotational component of T^* in both a uni- and a multi-modal scenario. In Fig. 4, we show the results for these cases. In the former, we aligned a PD image to a transformed version of itself, while in the latter the MRI slice to the CT image. We optimized the sum of squared differences and the cross-correlation coefficient, respectively of the auto-information subspace maps, in order to estimate the best transformation component. We decided to apply these simple similarity measures as the surfaces to be compared were composed of the same type of measures, the autoinformation values (as opposed to intensities of different modalities, for example). Both of the registration results closely matched the ground truth rotation angle.

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 this 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.



(a) (b)

Fig. 4. Decoupled rotation angle search: (a) Unimodal search using the PD image – minimizing sum of squared errors (b) Multi-modal search using the MRI and CT images – maximizing cross-correlation coefficient. The ground truth solution in both images is indicated with the vertical line.

We also introduced the auto- and cross- information functions which characterize the joint intensity statistics as a function of the relative transformations 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. One aspect of the auto-information function is that it facilitates decoupling of rotation and displacement parameters in the search space. Furthermore, our empirical results on anatomical data showed that the auto-information functions across modalities exhibit striking similarities which we conjecture 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.

Acknowledgement

This work has been supported by NIH grant R21CA89449, by NSF ERC grant (JHU Agreement #8810-274), by the Whiteman Fellowship and The Harvard Center for Neurodegeneration and Repair.

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*, 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 2000*, pages 567–578.
8. 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.
9. W.M. Wells III, P. Viola, and R. Kikinis. Multi-modal volume registration by maximization of mutual information [medical imaging]. In *Proceedings of 2nd International Symposium on Medical Robotics and Computer Assisted Surgery*, pages 358,55–62, 1995.

Appendix

Both of the relationships in Eq. (14) result from extending 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 in the first inequality of Eq. (14), $v_j \leftarrow l_j - l_k \rightarrow v_k$, can be rewritten in two different forms using Bayes rule: $v_j \leftarrow l_j \leftarrow l_k \leftarrow v_k$ and $v_j \rightarrow l_j \rightarrow l_k \rightarrow v_k$. Given these and applying the Data Processing Inequality theorem, we arrive at the following:

$$I(v_k;l_k) \geq I(v_k;l_j) \geq I(v_k;v_j) \text{ and } I(l_k;l_j) \geq I(l_k;v_j) \tag{17}$$

$$I(v_j;l_j) \geq I(v_j;l_k) \geq I(v_j;v_k) \text{ and } I(l_j;l_k) \geq I(l_j;v_k) \tag{18}$$

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:

$$I(v_j;l_j) \geq I(u_j;v_j) \text{ and } I(v_k;l_k) \geq I(v_k;l_j) \geq I(v_k;u_j). \tag{19}$$

Again, using 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) if $l_k \rightarrow l_j$ indicates a lossless relationship, then $I(u_j;v_k) = I(u_j;v_j)$, (b) if $l_k \rightarrow l_j$ indicates a lossy connection, then $I(u_j;v_k) < I(u_j;v_j)$. Therefore, we can conclude that $I(u_j;v_j) \geq I(u_j;v_k)$.