An Introduction to Statistical Methods of Medical Image Registration

Lilla Zöllei, John Fisher, William Wells

ABSTRACT After defining the medical image registration problem, we provide a short introduction to a select group of multi-modal image alignment approaches. More precisely, we choose four widely-used statistical methods applied in registration scenarios for analysis and comparison. We clarify the implicit and explicit assumptions made by each, aiming to yield a better understanding of their relative strengths and weaknesses. We also introduce a figural representation of the methods in order to provide an intuitive way of illustrating their similarities and differences.

1 Introduction

Registration of medical image data sets is the problem of identifying a set of geometric transformations which map the coordinate system of one data set to that of the others. Depending on the nature of the input modalities, we distinguish between uni-modal and multi-modal cases, according to whether the images being registered are of the same type. The multi-modal registration scenario is more challenging as corresponding anatomical structures will have differing intensity properties. In our analysis, we focus on the multi-modal case.

When designing a registration framework, one needs to decide on the nature of the transformations that will be used to bring images into agreement. For example, rigid transformations are generally sufficient in the case of bony structures while non-rigid mappings are mainly utilized for soft tissue matching. One must also evaluate the quality of alignment given an estimate of the aligning transformation. *Objective functions* or *similarity measures* are special-purpose functions that are designed to provide these essential numerical scores. The goal of a registration problem can then be interpreted as the optimization of such functions over the set of possible transformations. In general, these problems correspond to multi-dimensional non-convex optimization problems where we cannot automatically bracket the solution (as we would in case of a 1D line-search). Thus an initial estimate of the aligning transformation is needed before the search

begins.

In the past few decades there have been numerous types of objective functions proposed for solving the registration problem. Among these, there exist a variety of methods that are based on sound statistical principles. These include various maximum likelihood [4, 10], maximum mutual information [5, 11], minimum Kullback-Leibler divergence [1], minimum joint entropy [9] and maximum correlation ratio [8] methods. We are primarily interested in these, and in our discussion we select four of these registration approaches for further analysis. We explore the relative strengths and weaknesses of the selected methods, we clarify the type of explicit and implicit assumptions they make and demonstrate their use of prior information. By such an analysis and some graphical representations of the solution manifold for each method, we hope to facilitate a deeper and more intuitive understanding of these formulations.

In the past, similar or more detailed overview studies of the registration problem have been reported. Roche et al. [8], for example, have described the modeling assumptions in *uni-modal* registration applications and a general maximum likelihood framework for a certain set of multi-modal registration approaches, and we have described a *unified information theoretic* framework for analyzing multi-modal registration algorithms [13, 14].

2 The Similarity Measures

In our analysis, we discuss four objective criteria that rely on clear statistical principles: maximum likelihood (ML), approximate maximum likelihood (MLa), Kullback-Leibler divergence (KL) and mutual information (MI). While not an exhaustive list, these similarity measures are representative of a significant group of currently used registration algorithms. Many registration approaches either directly employ or approximate one of these measures.

While the analysis presented here carries straightforwardly to registration of multiple data sets, for simplicity, we focus on 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

$$v_o(x) = v(T^*(x)) \quad \text{or} \quad v(x) = v_o\left((T^*)^{-1}(x)\right), \tag{1.1}$$

where $T^* : \Re^M \rightarrow \Re^M$ is a bijective mapping corresponding to an unknown *relative* transformation. The goal of registration is to find an estimate of an

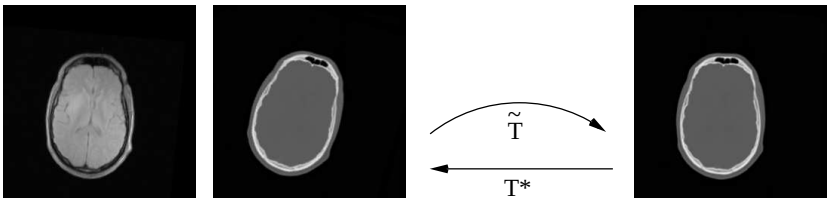

FIGURE 1. An 2D example of the registration problem. The *observed* input images are $u(x)$, an MRI slice, and $v_o(x)$, a CT slice. $v(x)$ is the CT slice that is in correct alignment with the MRI slice. The unknown transformation that relates the observed data to the aligned image is T^*. The goal of the registration algorithm is to make $\tilde{T}$ be the best estimate of $(T^*)^{-1}$.

aligning transformation $\tilde{T} \approx (T^*)^{-1}$ which optimizes some objective function of the observed data sets.[1] Figure 1 demonstrates the key components of the registration problem via a 2D example.

Throughout our analysis (and consistent with practice) spatial samples x_i are modeled as independent random draws of a uniformly distributed random variable X whose support is the domain of $u(x)$. Consequently, all the analyzed methods assume that

(IID-i) observed intensities $v_o(x_i)$ and $u(x_i)$ can be viewed as independent and identically distributed (*i.i.d.*) random variables, despite spatial dependencies present within the data.

This is a simple consequence of the property that a *function* of an *i.i.d.* random variable is itself an *i.i.d.* random variable under very general conditions.

2.1 Maximum Likelihood

The maximum likelihood (ML) method of parameter estimation has served as the basis for many registration algorithms. Its popularity in parameter estimation can be explained by the fact that as the sample size increases, ML becomes the smallest variance unbiased estimator. As we will see, practical issues generally preclude a direct ML approach. Analysis of the method is however useful for comparison purposes. Given that the input images are related by an unknown transformation T^* (see Figure 1), we parameterize the observed data samples (a sequence of joint measurements drawn *i.i.d*)

[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.

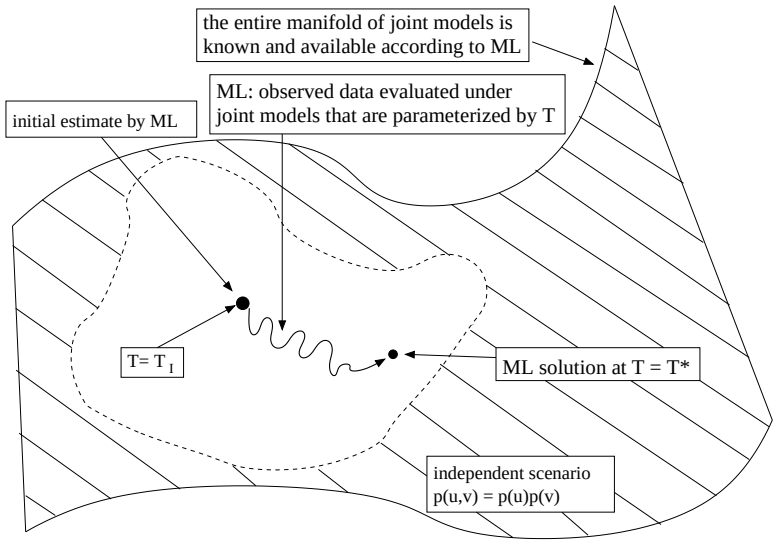

FIGURE 2. Joint density manifold of the registration search space parameterized by T. According to the classical ML approach, the entire manifold of joint models is known and available for the optimization task. The solution is defined at the location which maximizes the likelihood of the observed sample pairs. Here T_I has been chosen as an initial estimate for the search.

as

$$\begin{aligned} \mathcal{Y}_{T^*} &= \{[u, v_{T^*}]_1, \ldots, [u, v_{T^*}]_N\} \\ &= \{[u(x_1), v(T^*(x_1))], \ldots, [u(x_N), v(T^*(x_N))]\} \\ &= \{[u(x_1), v_o(x_1)], \ldots, [u(x_N), v_o(x_N)]\}. \end{aligned}$$

According to the ML criterion, we obtain estimates by varying some parameters of a probabilistic model that is being evaluated on a set of observed data. In the case of our registration problem, the optimal geometrical transformation that *explains* the observations according to the ML criterion satisfies the (normalized) log-likelihood criterion:

$$T_{\mathrm{ML}} = \arg\max_T \mathcal{L}_T(\mathcal{Y}_{T^*}) \tag{1.2}$$

$$= \arg\max_T \frac{1}{N} \sum_i \log\left(p\left([u, v_{T^*}]_i; T\right)\right). \tag{1.3}$$

$\mathcal{L}_T(.)$ in Equation (1.2) indicates that we are evaluating a model parameterized by the transformation T.

This formulation of the registration problem implicitly assumes that

(ML-i) as T approaches T^*, Equation (1.3) is non-decreasing.

An important distinction between currently used registration methods and the classical ML approach is that the former optimize the objective criterion by transforming the joint observations ($[u, v_{T^*}]_i$). In contrast, a classical ML approach optimizes the objective function by changing the parameters of the joint *density* model under which we evaluate the observations (as a function of transformation T), leaving the observations static throughout the search process. Below, we will indicate these differences via notional graphs of the solution paths of the selected methods. In Figure 2, according to the ML approach, the entire search space of joint models (parameterized by transformation T) is considered to be known and available. We make the initial estimate of this example be $T = T_I$ (the identity transformation) and the solution lies at transformation T^* that maximizes the likelihood function with respect to the currently observed images. Thus the initial guess by ML is modified in order to satisfy the criterion.

This framework highlights two practical obstacles to a direct ML approach. The optimization of Equation (1.2) requires the solution of a system of non-linear equations for which no direct global solution typically exists. Finding a globally optimal solution would likely require that $p(u, v; T)$ be pre-computed over all relative transformations T (see Figure 2). An alternative is to use an optimization procedure that searches for a local optimum, which would require the ability to produce $p(u, v; T)$ on demand, as we search. The first approach may be impractical due to computational and memory limitations. While the second approach may be feasible, as far as we know, it has not been tested or used. The second obstacle is that there are configurations of the data for which a considerable set of transformations form an equivalence class under the ML criterion. As the relative transformations away from the solution $T = T^*$ become large, we observe empirically that the joint models tend toward statistical independence. In addition, they may tend towards the same independent model (more on this appears in Section 2.4, below). In this situation, the ML criterion will lose traction for such large transformations. (In Figure 2, such models are located outside of the dashed outline.) As we shall see, MI-based approaches can be interpreted as moving away from these models.

2.2 Approximate Maximum Likelihood

As mentioned above, the optimization of Equation (1.2) is generally a very difficult problem. Suppose, however, that we have a model of the joint density of our data sets at one particular parameter setting, specifically when the multi-modal images are registered. We can estimate this model from other *registered* data sets and evaluate new observations under the resulting model. This idea was first suggested by Leventon and Grimson and we refer to it as an approximate maximum likelihood registration approach (MLa) [4]. (A similar approach has been discussed more recently in [12].)

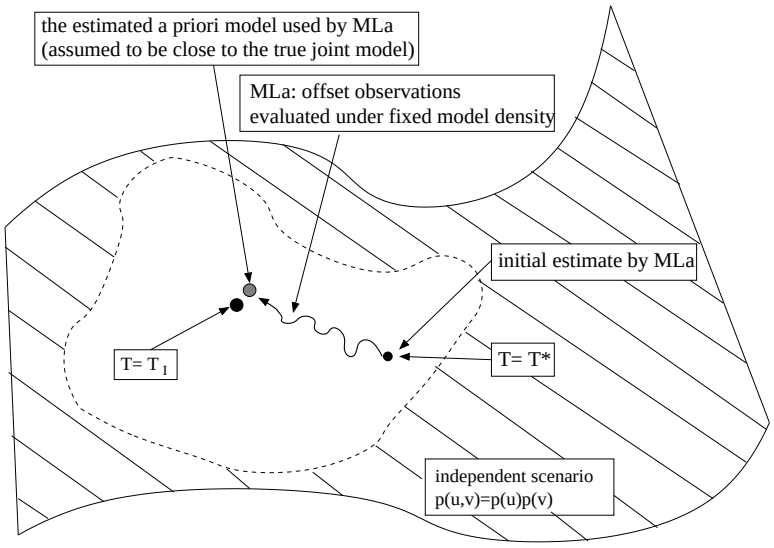

FIGURE 3. The approximate ML method (MLa) searches over the set of joint data sets offset by T. The goal is to maximize a criterion that is similar to likelihood with respect to a fixed model.

The approach makes two strong modeling assumptions:

(MLa-i) It is feasible to estimate or learn a joint probability model over the data modalities of interest at the correct alignment[2], and

(MLa-ii) the resulting model accurately captures the statistical properties of other unseen image pairs (of the same anatomy and with the same modality pairing as the training set).

We denote the estimated joint density model as

$$p^{\circ}(u,v) \approx p(u,v;T_I).$$

As with all of the remaining methods, the MLa approach transforms the observations prior to evaluating the objective criterion. We denote the transformed observations as

$$\begin{aligned}
\mathcal{Y}_T &= \left\{ \left[u(x_1), v_{\circ}(\hat{T}(x_1)) \right], \ldots, \left[u(x_N), v_{\circ}(\hat{T}(x_N)) \right] \right\} \\
&= \left\{ \left[u(x_1), v(T^* \circ \hat{T}(x_1)) \right], \ldots, \left[u(x_N), v(T^* \circ \hat{T}(x_N)) \right] \right\} \\
&= \{ [u(x_1), v(T(x_1))], \ldots, [u(x_N), v(T(x_N))] \} \\
&= \{ [u, v_T]_1, \ldots, [u, v_T]_N \}. \qquad (1.4)
\end{aligned}$$

[2] Assuming manual or other types of ground truth results are available from previous registration experiments.

We emphasize that the transformation $T = \left(T^* \circ \hat{T}\right)$ in this particular notation refers to the relative transformations on $v(x)$ rather than on the observed image of $v_\circ(x)$. In practice, it is $\hat{T}$ that we apply to the observed image, so optimization is performed over $\hat{T}$ through $v_\circ(\hat{T}(x))$. This is equivalent to implicit optimization over T through the relation $v(T(x)) = v_\circ(T^* \circ \hat{T}(x))$. While we express results on the implicit transformation, there are simple relationships which allow results to be expressed in terms of either T or $\hat{T}$.

The MLa approach estimates T to be the transformation that maximizes a criterion that is similar to the likelihood criterion:

$$T_{\text{MLa}} = \arg\max_T \mathcal{L}_{T_I}(\mathcal{Y}_T) \tag{1.5}$$

$$= \arg\max_T \frac{1}{N} \sum_i \log\left(p\left([u, v_T]_i ; T_I\right)\right). \tag{1.6}$$

Notice that, according to this approach the joint observations $([u, v_T]_i)$ are varied as a function of T and the model density p° is held static. It is under this particular fixed probability model that all the transformed inputs are evaluated. In Figure 3, we indicate the path of the MLa approach by tracing a sample search path. Beginning with the initial estimate, the algorithm searches over transformations to maximize the likelihood-like criterion with respect to the previously constructed, static density model.

The MLa method also makes an implicit assumption when solving the registration problem. It assumes that:

(MLa-iii) as $\hat{T}$ approaches $(T^*)^{-1}$, or equivalently as $(T^* \circ \hat{T})$ approaches T_I, Equation (1.6) is non-decreasing.

In general, one cannot guarantee the validity of this assumption. Theoretically, there might exist some counter-intuitive scenarios for which this implicit hypothesis would fail. The existence of these is explained by the information theoretic phenomenon of *typicality* [2]. A more detailed discussion of this issue is not in the scope of this chapter; it is described in an information-theoretic framework in [14].

This obstacle, in the context of multi-modal registration, may explain some shortcomings of the MLa approach that were observed empirically by Chung *et al.* [1]. It motivates their registration approach, which is described in the next section.

2.3 Kullback-Leibler Divergence

Chung *et al.* suggested the use of KL divergence as a registration measure in order to align digital-subtraction angiography (DSA) and MR angiography

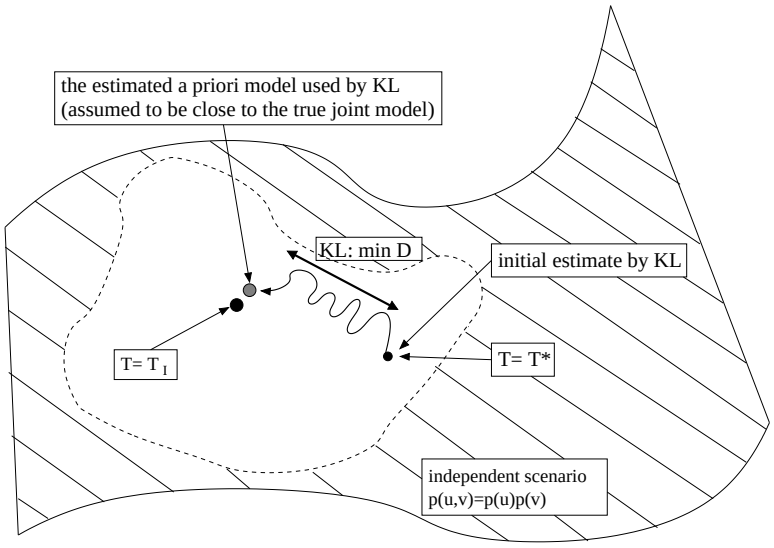

FIGURE 4. According to the KL framework, at each point on the manifold a joint density is estimated from the offset data pairs. The aligning transformation is located where the KL distance (D) is minimized between that current estimate and a previously defined fixed model.

(MRA) data sets [1]. Using the same modelling assumption as in MLa (i.e. a model of the joint intensity data can be estimated from a set of registered data sets), they optimize an objective function based on a KL divergence term, that is, the distance between the joint density at the current transformation estimate and the fixed model is to be minimized:

$$T_{\mathrm{KL}} \approx \arg\min_{T} D\left(\hat{p}(u,v;T) \| p^{\circ}(u,v)\right),$$

where p^o is constructed as in the MLa approach from correctly registered data sets and $\hat{p}(u,v;T)$ is a probability model estimated from the transformed sets of observed pixel intensities $\{u(x_i), v(T(x_i))\}$ (or $\{u(x_i), v_o(\hat{T}(x_i))\}$ as discussed above). Whereas the previous methods utilize a likelihood function of the observed data sets, here numerical or Monte Carlo integration is used in order to calculate the KL divergence terms directly.

Consequently, in addition to assumptions MLa-i and MLa-ii, this approach makes the following hypothesis:

(KL-i) There is a reliable method for estimating $p(u,v;T)$ from transformed observations, and

(KL-ii) the KL divergence $D\left(p(u,v;T)||p^{\circ}(u,v)\right)$ can be accurately estimated via numerical or Monte Carlo integration of

$$\int\int \hat{p}\left(u,v;T\right) \log\left(\frac{\hat{p}\left(u,v;T\right)}{p^{\circ}\left(u,v\right)}\right) dudv \tag{1.7}$$

by substituting $\hat{p}\,(u, v; T)$ for $p\,(u, v; T)$ in the KL divergence integral.

The KL method has been demonstrated to be more robust with respect to, or less dependent on, the size of the sampling region (the area from which the joint sample pairs are drawn from) than the MLa (or the MI) approaches [1]. This robustness is demonstrated empirically [1] and can be partly explained by *typicality*, as discussed in the preceding section.

Provided that both of the KL assumptions are valid (the density estimate and the integration methods are accurate), the KL divergence estimate is non-increasing as $\hat{T}$ approaches $(T^*)^{-1}$. This is supported by empirical comparisons in which KL did not exhibit some of the undesirable local extrema encountered in the MLa method[1]. Additionally, the authors emphasize that even though the estimated models represent a strong assumption, sufficient model distributions can be constructed even if manual alignment is unavailable. For example, the joint probability distribution could be estimated from segmented data for corresponding structures.

In relation to the previous methods, both the samples $([u, v_T]_i)$ *and* the evaluation density $(\hat{p}(u, v; T))$ are being varied as a function of the transformation T, while the algorithm approaches the static joint probability density model $(p^{\circ}(u, v))$ constructed prior to the alignment procedure. Instead of evaluating the joint characteristics of the transformed input data sets under the model distribution, the KL approach re-estimates the joint model $(\hat{p}(u, v; T))$ at every iteration and uses that when evaluating the observations. In Figure 4, the KL method is shown to approach the solution by minimizing the KL distance between the model and the current estimate.

2.4 Mutual Information and Joint Entropy

As has been amply documented in the literature [5, 6, 7, 11], Mutual Information (MI) is a popular information theoretic objective criterion. It estimates the transformation parameter T by maximizing the mutual information (or the statistical dependence) between the input image data sets:

$$T_{\mathrm{MI}} = \arg\max_{T} I\,(u; v_T)\,.$$

One way to define the MI term is to use marginal and joint entropy measures. By definition, given random variables A and B, it is the sum of their marginal entropies minus their joint:

$$I(A, B) = H(A) + H(B) - H(A, B).$$

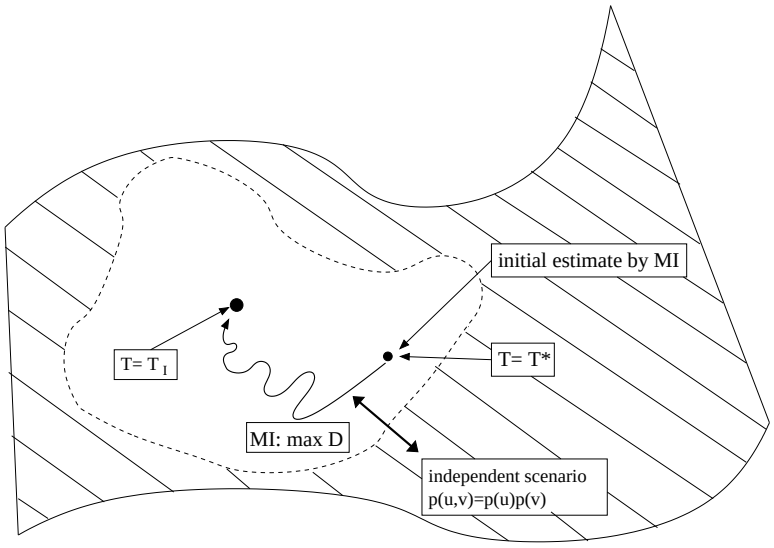

FIGURE 5. According to MI, the solution is located maximum KL distance away from the worst-case, independent scenario, where the joint density is defined as the product of its marginals: $p(u, v; T) = p(u)p(v; T)$.

In the multi-modal alignment scenario that translates to

$$I(u; v_T) = H(p(u)) + H(p(v; T)) - H(p(u, v; T)). \tag{1.8}$$

If T is restricted to the class of symplectic transformations (i.e. volume preserving), then $H(p(u))$ and $H(p(v; T))$ are invariant to T. In that case, maximization of MI is equivalent to minimization of the joint entropy term, $H(p(u, v; T))$, the presumption being that this quantity is minimized when $\hat{T} = (T^*)^{-1}$. The minimization of the joint entropy term has also been widely used in the registration community.

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

$$I\left(u, v_T\right) = D\left(p(u, v; T) \| p(u)p(v; T)\right).$$

That is, mutual information is the KL divergence between the observed joint density term and the product of its marginals. Accordingly, the implicit assumption of MI-based methods is that:

(MI-i) as $(T^* \circ \hat{T})$ diverges from T_I (as we are getting farther away from the ideal registration pose) the joint intensities look less statistically dependent, tending towards statistical independence.

This allows us to write the MI optimization problem as maximizing the divergence from the current density estimate to the scenario where the

images are completely independent:

$$T_{\mathrm{MI}} \approx \arg\max_{T} D\left(\hat{p}(u,v;T) \| \hat{p}(u)\hat{p}(v;T)\right).$$

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 known model point according to KL distance, the aim is to move farthest away from the condition of statistical independence among the images, in the KL sense. This behavior is illustrated in Figure 5.

Numerous variations on the mutual information metric have been introduced; for instance, one making it invariant to image overlap (normalized mutual information [9]) and another enhancing its robustness using additional image gradient information (gradient-augmented mutual information [6]). In this report, we do not list and analyze these, given that they operate with similar underlying principles.

3 Conclusion

We have provided a brief comparison of four well-known and widely used multi-modal image registration methods. We illustrated the underlying assumptions which distinguish them, and specifically, we clarified the assumed behavior of joint intensity statistics as a function of transformation parameters. Considering the collection of approaches discussed, we see that the ML approach has not actually been used, in practice. The related MLa method and the KL divergence method exploit prior information in the form of static joint density estimates over previously registered data. Subsequently, both make similar implicit assumptions regarding the behavior of joint intensity statistics as the transformation estimate approaches the ideal alignment. In contrast, the MI approach makes no use of specific prior joint statistics – instead, it simply moves away from the general class of statistically independent models. Figure 6 serves as a visual guide to summarize how the different methods approach the solution.

Acknowledgment

This work has been supported by NIH grant #R21CA89449 and #5P41RR13218 by NSF ERC grant (JHU EEC #9731748), by the Whiteman Fellowship and The Harvard Center for Neurodegeneration and Repair.

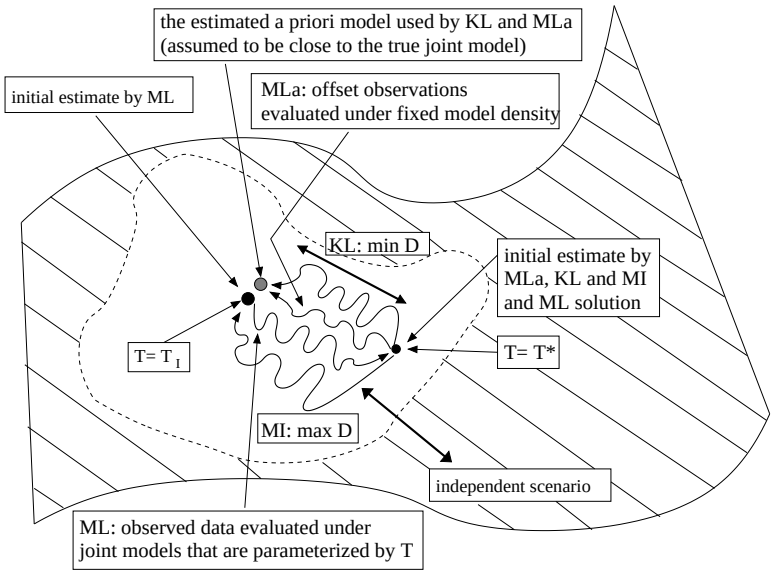

FIGURE 6. Manifold of the registration search space parameterized by transformation T. The illustration shows how each of the examined methods (ML, MLa, KL and MI) search through the settings in order to obtain the best estimate of the aligning transformation. Note that the ML method transforms the model to agree with the observed data, while the rest of the methods operate by transforming the observed data.

4 References

[1] A. Chung, W. Wells III, A. Norbash, and W. 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. Cover and J. 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. 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] 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.

[6] J. Pluim, J. Maintz, and M. 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.

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

[8] A. Roche, G. Malandain, and X. P. 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.

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

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

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

[12] Y. Zhu and S. Cochoff. Likelihood maximization approach to image registration. *IEEE Transactions on Image Processing*, 11(12):1417–1426, 2002.

[13] L. Zollei, J. Fisher, and W. Wells. A unified statistical and information theoretic framework for multi-modal image registration. In *Proceedings of IPMI*, volume 2732 of *Lecture Notes in Computer Science*, pages 366–377. Springer, 2003.

[14] L. Zollei, J. Fisher, and W. Wells. A unified statistical and information theoretic framework for multi-modal image registration. Technical report, MIT, 2004.