

Bayesian Multimodality Non-rigid Image Registration via Conditional Density Estimation

Jie Zhang and Anand Rangarajan

Department of Computer and Information Science & Engineering
University of Florida, Gainesville, FL, USA

Abstract. We present a Bayesian multimodality non-rigid image registration method. Since the likelihood is unknown in the general multimodality setting, we use a density estimator as a drop in replacement to the true likelihood. The prior is a standard small deformation penalty on the displacement field. Since mutual information-based methods are in widespread use for multimodality registration, we attempt to relate the Bayesian approach to mutual information-based approaches. To this end, we derive a new criterion which when satisfied, guarantees that the displacement field which minimizes the Bayesian maximum *a posteriori* (MAP) objective also maximizes the true mutual information (with a small deformation penalty) as the number of pixels tends to infinity. The criterion imposes an upper bound on the number of configurations of the displacement field. Finally, we compare the results of the Bayesian approach with mutual information, joint entropy and joint probability approaches on synthetic data and simulated T1 and T2 2D MR images.

1 Introduction

Non-rigid multimodality image registration is an impossibly general term covering that aspect of the registration problem wherein the data are from different modalities or imaging protocols and where there is a need to recover deformations. Potentially, the data could arise from diverse domains such as MR, fMRI, CT, PET, SPECT, ultrasound, portal films etc. and the registration may need to be performed across subjects, between pre-operative and post-operative situations and so on. Despite the overarching generality of the term, the emerging consensus in medical imaging is that there is a need to have a general information processing vocabulary to deal with situations where a) imaging data are from different modalities, b) a flexible (non-rigid as opposed to rigid, similarity or affine) registration is required and c) there appears to be a reasonable similarity (usually unspecified) in the intensity patterns of the two modalities to warrant embarking upon this task in the first place.

One particular general methodology that has arisen in recent years—first for rigid registration and increasingly for non-rigid registration—is maximization of mutual information (MI) [7], [12]. Since the main stumbling block in multimodality registration is the inability (or unwillingness?) to model the joint probability between the two imaging sources directly using imaging physics, most recent approaches take the nonparametric density estimation route and model the joint probability using histogramming, Parzen windows and the like. Once the joint

probability is available, the mutual information can be computed under the assumption that the density estimator is the *true* probability distribution. Since this is not true in general, henceforth we refer to these approaches as maximizing the empirical mutual information (EMI).

The Bayesian approach taken in this paper has much in common with EMI and with maximizing the true mutual information (both with an added deformation penalty). Instead of computing the empirical mutual information from the joint probability, we compute the conditional probability using a density estimator and use it as a drop in replacement for the true likelihood, which following tradition, we assume to be unknown. We then ask the following question: What is the relationship between the minimizer of the Bayesian maximum *a posteriori* (MAP) objective function and the maximizer of the true mutual information (with an added deformation penalty)? To answer this question, we derive a criterion—a *sufficient* condition—which when satisfied, *guarantees* that the minimizer of the Bayesian MAP objective function also maximizes the true mutual information in the limit as the number of pixels tends to infinity. The criterion imposes an $O(N^{N^c})$ (with $c < 1$) upper bound on the number of allowed configurations of the displacement field which is trivially satisfied in rigid registration provided the parameters are quantized. The case in the general non-rigid setting is less clear. Since the number of allowed configurations is only implicitly restricted via a deformation penalty rather than being limited by fiat, the implications of this upper bound for Bayesian MAP non-rigid registration *vis-a-vis* the true mutual information remain to be seen.

Since the pioneering contributions of [7], [12], mutual information-based methods have become the *de facto* standard for multimodality rigid registration. Beginning with [4] and [8], mutual information-based methods have slowly started seeing traction for non-rigid registration as well [9], [2], [3]. The work in [6] was one of the first to point out a connection between maximizing the joint probability and mutual information. Also, the work in [3] uses sample averages in place of the true expectation which is similar to this work. Recently, attempts have been made to relate likelihood ratios (with joint densities pitted against products of marginals in a hypothesis testing paradigm) and mutual information [5]. We have not found any previous work which begins with the conditional probability estimate and uses it as a likelihood in an overall Bayesian estimation framework.

2 Bayesian and MI-based registration: Deriving a relationship

2.1 Conditional density estimation

Assume that we have two images $I^{(1)}$ and $I^{(2)}$, and let $I_i^{(k)}$ be the intensity value of image $I^{(k)}$ at location i [pixel position (x, y)], $k = 1, 2$, $i = 1, \dots, N$, where N is the total number of pixels in each image. (While our development is mainly for 2D, extensions to 3D appear straightforward.) The conditional probability of $I^{(1)}$ given $I^{(2)}$ at location i is denoted by $\Pr(I_i^{(1)} | I_i^{(2)})$. Also, for the remainder of the paper, non-rigid warps will be applied solely to $I^{(2)}$ with $I^{(1)}$ held fixed. Instead

of using the original (possibly real-valued) intensity, we will use the following binned intensity value. This is done for a *technical* reason which will become more obvious as we proceed. The binned intensity for each image is

$$B_i^{(k)} = \text{round}[(K^{(k)} - 1) \times \frac{I_i^{(k)}}{\max_{1 \leq j \leq N} \{I_j^{(k)}\}} + 1], \quad k = 1, 2. \quad (1)$$

From (1), we see that the binned intensity values are integers in $\{1, \dots, K^{(k)}\}$. We use (2) to compute the conditional probability at location i :

$$\Pr(B_i^{(1)}|B_i^{(2)}) = \frac{\Pr(B_i^{(1)}, B_i^{(2)})}{\Pr(B_i^{(2)})}. \quad (2)$$

2.2 Bayesian Non-rigid Registration

As mentioned previously, we seek to register image $I^{(2)}$ to image $I^{(1)}$. To achieve this, we set up a displacement vector \mathbf{u}_i at each location i . We denote an entire *configuration* of $\{\mathbf{u}_i, i \in 1, \dots, N\}$ by \mathbf{u} and the set of allowed configurations of \mathbf{u} by Λ . A non-rigid warping of $I^{(2)}$ exists for each configuration \mathbf{u} (with the understanding that interpolation may be necessary to generate a valid warped $I^{(2)}$). From Bayes' rule, we get

$$\Pr(\mathbf{u}|B^{(1)}, B^{(2)}) = \frac{\Pr(B^{(1)}|B^{(2)}, \mathbf{u}) \Pr(\mathbf{u})}{\Pr(B^{(1)}|B^{(2)})} \quad (3)$$

from which we get

$$\log \Pr(\mathbf{u}|B^{(1)}, B^{(2)}) = \log \Pr(B^{(1)}|B^{(2)}, \mathbf{u}) + \log \Pr(\mathbf{u}) - \log \Pr(B^{(1)}|B^{(2)}). \quad (4)$$

Since the probability $\Pr(B^{(1)}|B^{(2)}) = \sum_{\mathbf{u} \in \Lambda} \Pr(B^{(1)}|B^{(2)}, \mathbf{u}) \Pr(\mathbf{u})$ is independent of \mathbf{u} , we have

$$\log \Pr(\mathbf{u}|B^{(1)}, B^{(2)}) \propto \log \Pr(B^{(1)}|B^{(2)}, \mathbf{u}) + \log \Pr(\mathbf{u}). \quad (5)$$

Consequently, from a Bayesian perspective, the non-rigid registration problem becomes

$$\hat{\mathbf{u}} = \arg \min_{\mathbf{u}} E(\mathbf{u}) = \arg \min_{\mathbf{u}} -\log \Pr(B^{(1)}|B^{(2)}, \mathbf{u}) - \log \Pr(\mathbf{u}). \quad (6)$$

We use a standard small deformation smoothness constraint on \mathbf{u} which can be written as $-\log \Pr(\mathbf{u}) \propto \|L\mathbf{u}\|^2$. We assume conditional independence of $B^{(1)}$ given $B^{(2)}$ over the pixel locations. This assumption is clearly not necessary since the image processing/computer vision literature is replete with correlated random field image models (ranging from simplistic to baroque) [13]. However, most random field models require the estimation of further parameters which increases the estimation burden. In addition, EMI-based registration methods have traditionally used very simple density estimation procedures such as histogramming [7] and Parzen windows [12] which sets a clear precedent for us.

With these simplifications in place, we obtain the following Bayesian maximum *a posteriori* (MAP) objective function

$$E_{\text{MAP}}(\mathbf{u}) = -\frac{1}{N} \sum_{i=1}^N \log \Pr(B_i^{(1)}|B_i^{(2)}, \mathbf{u}) + \lambda \|L\mathbf{u}\|^2 \quad (7)$$

where we have normalized the negative log-likelihood in the first term of (7). The parameter λ is a regularization parameter and L is the regularization operator. In the 2D case, we choose

$$\|L\mathbf{u}\|^2 = \int_x \int_y [(\frac{\partial^2 \mathbf{u}}{\partial x^2})^2 + 2(\frac{\partial^2 \mathbf{u}}{\partial x \partial y})^2 + (\frac{\partial^2 \mathbf{u}}{\partial y^2})^2] dx dy \quad (8)$$

which is a standard thin-plate spline [11] small deformation cost.

2.3 Convergence of the Bayesian MAP minimizer in the general setting

In this section, we examine the convergence properties of the minimizer of the Bayesian MAP objective function in (7) as the number of pixels N tends to infinity. In the non-rigid setting, the cardinality of \mathbf{u} scales linearly with N and as we shall see, this complicates the convergence proof.

We begin by assuming that the chosen density estimator converges to the true density as N tends to infinity. This is usually true of histogram and Parzen window estimators. Denote the true density of $B^{(1)}$, $B^{(2)}$ and the pair $(B^{(1)}, B^{(2)})$ by $\bar{\Pr}(B^{(1)})$, $\bar{\Pr}(B^{(2)})$ and $\bar{\Pr}(B^{(1)}, B^{(2)})$ respectively and the corresponding estimated densities by $\hat{\Pr}(B^{(1)})$, $\hat{\Pr}(B^{(2)})$, and $\hat{\Pr}(B^{(1)}, B^{(2)})$. We assume that

$$\lim_{N \rightarrow \infty} \hat{\Pr}(B^{(1)}) = \bar{\Pr}(B^{(1)}), \quad \lim_{N \rightarrow \infty} \hat{\Pr}(B^{(2)}) = \bar{\Pr}(B^{(2)}) \quad (9)$$

and

$$\lim_{N \rightarrow \infty} \hat{\Pr}(B^{(1)}, B^{(2)}) = \bar{\Pr}(B^{(1)}, B^{(2)}). \quad (10)$$

With this notation in place, we can write the true mutual information as

$$MI(\mathbf{u}) = \sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \bar{\Pr}(a, b|\mathbf{u}) \log \bar{\Pr}(a|b, \mathbf{u}) + \text{terms independent of } \mathbf{u} \quad (11)$$

and the empirical mutual information as

$$EMI(\mathbf{u}) = \sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \hat{\Pr}(a, b|\mathbf{u}) \log \hat{\Pr}(a|b, \mathbf{u}) + \text{terms independent of } \mathbf{u}. \quad (12)$$

The objective function we would like to minimize is

$$E_{\text{MI}}(\mathbf{u}) = -MI(\mathbf{u}) + \lambda \|L\mathbf{u}\|^2. \quad (13)$$

Since we use the framework of statistical learning theory [10] throughout this paper, we call $E_{\text{MI}}(\mathbf{u})$ the *expected risk*. This objective function is not computable since the true distribution $\hat{\text{Pr}}$ is unknown and is only approached by our density estimator $\hat{\text{Pr}}$ as N tends to infinity. Instead of minimizing the expected risk, we minimize the Bayesian MAP objective function, a.k.a. the *empirical risk* which is the same as (7):

$$E_{\text{MAP}}(\mathbf{u}) = -LL(\mathbf{u}) + \lambda\|\mathbf{L}\mathbf{u}\|^2 \quad (14)$$

where the log-likelihood $LL(\mathbf{u})$ is defined as

$$LL(\mathbf{u}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^N \log \hat{\text{Pr}}(B_i^{(1)}|B_i^{(2)}, \mathbf{u}). \quad (15)$$

In (15), $\hat{\text{Pr}}(B_i^{(1)}|B_i^{(2)}, \mathbf{u})$ is the estimated distribution from N samples. We are interested in the relationship between the minimizers of (14) and (13).

Let the minimum of the expected risk in (13) be achieved by the displacement field \mathbf{u}_r and the minimum of the empirical risk in (14) be achieved by the displacement field \mathbf{u}_e . The following question is our main concern in this paper:

What is $E_{\text{MI}}(\mathbf{u}_e) - E_{\text{MI}}(\mathbf{u}_r)$, or how close is the value of the expected risk attained by the minimizer \mathbf{u}_e of the Bayesian MAP objective function (empirical risk) to the maximum value of the true mutual information (expected risk) which is attained by \mathbf{u}_r ?

We answer this question by proving the following theorem. It turns out that the theorem requires a specific choice of the density estimator. We pick the estimator used in [13] which is closely related to histogramming. Assuming an i.i.d. density for each pixel, our chosen density estimator is

$$p(I) = \sum_{j=1}^N \delta(I - I_j) \quad (16)$$

where $\delta(\cdot)$ is the Dirac delta function with $\int_{-\infty}^{+\infty} \delta(x)dx = 1$. Note that $p(I)$ in (16) is a density function and not a discrete distribution as required. Since the Dirac delta function is not differentiable, we switch to a continuous approximation

$$\delta(x) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\}. \quad (17)$$

This approximation is increasingly exact as $\sigma \rightarrow 0$ and is continuous and differentiable for $\sigma > 0$. Finally, since we use binned intensities B , we normalize the above delta function approximation to get the final density estimator used in this paper. The joint probability distribution is

$$\hat{\text{Pr}}(B_i^{(1)}, B_i^{(2)}) = \frac{\sum_{j=1}^N \exp\left\{-\frac{(B_i^{(1)} - B_j^{(1)})^2 + (B_i^{(2)} - B_j^{(2)})^2}{2\sigma^2}\right\}}{\sum_{\beta^{(1)}=1}^{K^{(1)}} \sum_{\beta^{(2)}=1}^{K^{(2)}} \sum_{j=1}^N \exp\left\{-\frac{(\beta^{(1)} - B_j^{(1)})^2 + (\beta^{(2)} - B_j^{(2)})^2}{2\sigma^2}\right\}} \quad (18)$$

with similar expressions for the marginals and with the understanding that we are dealing with i.i.d. pixels.

As mentioned previously, $\Lambda = \{\mathbf{u}\}$ is the set of all possible configurations of \mathbf{u} . For the non-rigid registration problem with N -pixel images, $\|\Lambda\|$ —the cardinality of Λ —is bounded from above by N^N , since each pixel can potentially move to any other location. Since the upper bound for $\|\Lambda\|$ scales with N , we let $\|\Lambda\|$ be a function of N :

$$\|\Lambda\| = g(N) \quad (19)$$

Theorem 1: With the probability distributions estimated as in (18), and with the total number of configurations of \mathbf{u} as in (19), the inequality

$$\begin{aligned} \Pr(\sup_{\mathbf{u} \in \Lambda} [\sum_{ab} \bar{\Pr}(a, b|\mathbf{u})[-\log \hat{\Pr}(a|b, \mathbf{u})] + \frac{1}{N} \sum_i \log \hat{\Pr}(B_i^{(1)}|B_i^{(2)}, \mathbf{u})] > \epsilon) \\ \leq g(N) e^{-\frac{2\epsilon^2 N}{\tau(K^{(1)}, K^{(2)}, \sigma)}} \quad (20) \end{aligned}$$

is valid where $\tau(K^{(1)}, K^{(2)}, \sigma) \stackrel{\text{def}}{=} \frac{(K^{(1)})^2 + (K^{(2)})^2}{2\sigma^2} + \log(K^{(1)}K^{(2)})$ and ϵ is any given positive small number.

Proof: An abbreviated proof of Theorem 1 follows. The basic idea is to calculate lower and upper bounds of the estimated distribution $-\log \hat{\Pr}(B_i^{(1)}|B_i^{(2)}, \mathbf{u})$ and then apply Hoeffding's inequality [10] for real-valued bounded functions to get

$$\begin{aligned} \Pr(\sup_{\mathbf{u} \in \Lambda} [\sum_{ab} \bar{\Pr}(a, b|\mathbf{u})[-\log \hat{\Pr}(a|b, \mathbf{u})] + \frac{1}{N} \sum_i \log \hat{\Pr}(B_i^{(1)}|B_i^{(2)}, \mathbf{u})] > \epsilon) \\ \leq \sum_{\mathbf{u} \in \Lambda} \hat{\Pr}[\sum_{ab} \bar{\Pr}(a, b|\mathbf{u})[-\log \hat{\Pr}(a|b, \mathbf{u})] + \frac{1}{N} \sum_i \log \hat{\Pr}(B_i^{(1)}|B_i^{(2)}, \mathbf{u}) > \epsilon] \\ \leq g(N) e^{-\frac{2\epsilon^2 N}{\tau(K^{(1)}, K^{(2)}, \sigma)}} \quad (21) \end{aligned}$$

which is the desired result. From (21), we obtain with probability $1 - \eta$ (where $\eta \stackrel{\text{def}}{=} \exp\{-\frac{2\epsilon^2 N}{\frac{(K^{(1)})^2 + (K^{(2)})^2}{2\sigma^2} + \log(K^{(1)}K^{(2)})}\}$), the inequality

$$\begin{aligned} \sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \bar{\Pr}(a, b|\mathbf{u}_e) [-\log \hat{\Pr}(a|b, \mathbf{u}_e)] - \frac{1}{N} \sum_{i=1}^N [-\log \hat{\Pr}(B_i^{(1)}|B_i^{(2)}, \mathbf{u}_e)] \\ \leq \left(\frac{(K^{(1)})^2 + (K^{(2)})^2}{2\sigma^2} + \log(K^{(1)}K^{(2)}) \right) \sqrt{\frac{\log g(N) - \log \eta}{2N}} \quad (22) \end{aligned}$$

Let

$$E_{\text{EMAP}}(\mathbf{u}) \stackrel{\text{def}}{=} - \sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \bar{\Pr}(a, b|\mathbf{u}) [-\log \hat{\Pr}(a|b, \mathbf{u})] + \lambda \|\mathbf{L}\mathbf{u}\|^2. \quad (23)$$

Adding and subtracting $\lambda \|\mathbf{L}\mathbf{u}_e\|^2$ to both terms on the left side of (22), we get

$$E_{\text{EMAP}}(\mathbf{u}_e) - E_{\text{MAP}}(\mathbf{u}_e) \leq \tau(K^{(1)}, K^{(2)}, \sigma) \sqrt{\frac{\log g(N) - \log \eta}{2N}}. \quad (24)$$

Once again, from Hoeffding's inequality, we have

$$\sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \bar{\Pr}(a, b | \mathbf{u}) [-\log \hat{\Pr}(a|b, \mathbf{u}_r)] \geq \frac{1}{N} \sum_{i=1}^N [-\log \hat{\Pr}(B_i^{(1)} | B_i^{(2)}, \mathbf{u}_r)] - \left[\frac{(K^{(1)})^2 + (K^{(2)})^2}{2\sigma^2} + \log(K^{(1)}K^{(2)}) \right] \sqrt{\frac{-\log \eta}{2N}}. \quad (25)$$

Adding $\lambda \|\mathbf{L}\mathbf{u}_r\|^2$ to both sides of (25) we can state more simply that

$$E_{\text{EMAP}}(\mathbf{u}_r) - E_{\text{MAP}}(\mathbf{u}_r) \geq - \left[\frac{(K^{(1)})^2 + (K^{(2)})^2}{2\sigma^2} + \log(K^{(1)}K^{(2)}) \right] \sqrt{\frac{-\log \eta}{2N}}. \quad (26)$$

Since \mathbf{u}_e is the minimizer of the empirical risk (Bayesian MAP objective function), $E_{\text{MAP}}(\mathbf{u}_r) \geq E_{\text{MAP}}(\mathbf{u}_e)$ and hence

$$0 \leq E_{\text{EMAP}}(\mathbf{u}_e) - E_{\text{EMAP}}(\mathbf{u}_r) \leq [E_{\text{EMAP}}(\mathbf{u}_e) - E_{\text{MAP}}(\mathbf{u}_e)] + [E_{\text{MAP}}(\mathbf{u}_r) - E_{\text{EMAP}}(\mathbf{u}_r)]. \quad (27)$$

Substituting (24) and (26) in (27), we see that

$$0 \leq E_{\text{EMAP}}(\mathbf{u}_e) - E_{\text{EMAP}}(\mathbf{u}_r) \leq \tau(K^{(1)}, K^{(2)}, \sigma) \left[\sqrt{\frac{\log g(N) - \log \eta}{2N}} + \sqrt{\frac{-\log \eta}{2N}} \right]. \quad (28)$$

If we assume that

$$\lim_{N \rightarrow \infty} \frac{\log g(N)}{N} = 0, \quad (29)$$

then we have

$$\lim_{N \rightarrow \infty} |E_{\text{EMAP}}(\mathbf{u}_e) - E_{\text{EMAP}}(\mathbf{u}_r)| = 0. \quad (30)$$

Using the above, we determine $E_{\text{MI}}(\mathbf{u}_e) - E_{\text{MI}}(\mathbf{u}_r)$:

$$E_{\text{MI}}(\mathbf{u}_e) - E_{\text{MI}}(\mathbf{u}_r) = [E_{\text{MI}}(\mathbf{u}_e) - E_{\text{EMAP}}(\mathbf{u}_e)] + [E_{\text{EMAP}}(\mathbf{u}_e) - E_{\text{EMAP}}(\mathbf{u}_r)] + [E_{\text{EMAP}}(\mathbf{u}_r) - E_{\text{MI}}(\mathbf{u}_r)] \quad (31)$$

where

$$E_{\text{MI}}(\mathbf{u}_e) - E_{\text{EMAP}}(\mathbf{u}_e) = \sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \bar{\Pr}(a, b | \mathbf{u}_e) \log \left[\frac{\hat{\Pr}(a|b, \mathbf{u}_e)}{\bar{\Pr}(a|b, \mathbf{u}_e)} \right] \quad (32)$$

and

$$E_{\text{MI}}(\mathbf{u}_r) - E_{\text{EMAP}}(\mathbf{u}_r) = \sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \bar{\Pr}(a, b | \mathbf{u}_r) \log \left[\frac{\hat{\Pr}(a|b, \mathbf{u}_r)}{\bar{\Pr}(a|b, \mathbf{u}_r)} \right]. \quad (33)$$

From our assumptions (9) and (10), we know that

$$\lim_{N \rightarrow \infty} \hat{\Pr}(a|b, \mathbf{u}) = \bar{\Pr}(a|b, \mathbf{u}). \quad (34)$$

Hence

$$\lim_{N \rightarrow \infty} |E_{\text{EMAP}}(\mathbf{u}_e) - E_{\text{EMAP}}(\mathbf{u}_r)| = 0, \text{ and } \lim_{N \rightarrow \infty} |E_{\text{EMAP}}(\mathbf{u}_r) - E_{\text{EMAP}}(\mathbf{u}_r)| = 0. \quad (35)$$

With (30), (31) and (35) we have

$$\lim_{N \rightarrow \infty} |E_{\text{MI}}(\mathbf{u}_e) - E_{\text{MI}}(\mathbf{u}_r)| = 0. \quad (36)$$

We have shown that the minimizers of the expected risk (true mutual information) and the empirical risk (Bayesian MAP objective function) coincide as the number of samples approaches infinity provided that a sufficient condition (29) is met.

It is time that we took a closer look at $g(N)$ which is the cardinality of the number of allowed configurations of \mathbf{u} . In non-rigid registration, the upper bound of $g(N)$ is N^N if we allow each pixel to move to any other location. In sharp contrast, in rigid registration, the upper bound of $g(N)$ is C^6 in 2D where we have assumed 6 free parameters (affine) with each parameter quantized into C (independent of N) bins. It should be obvious that if $g(N) = N^N$, *we cannot conclude* that $\lim_{N \rightarrow \infty} \frac{\log g(N)}{N} = 0$ and therefore that $\lim_{N \rightarrow \infty} |E_{\text{MI}}(\mathbf{u}_e) - E_{\text{MI}}(\mathbf{u}_r)| = 0$. But if we impose an extremely minor restriction on \mathbf{u} to the effect that $g(N) = N^{(N^c)}$, where $0 \leq c < 1$, then we have $\lim_{N \rightarrow \infty} \frac{\log g(N)}{N} = 0$ and the proof goes through resulting in $\lim_{N \rightarrow \infty} |E_{\text{MI}}(\mathbf{u}_e) - E_{\text{MI}}(\mathbf{u}_r)| = 0$. Clearly, from this proof's viewpoint, allowing every pixel to potentially visit any other location would not allow us to state that the Bayesian and the true MI measure minimizers coincide. To satisfy this criterion, we would need to place a small restriction on the set of allowed configurations.

3 Experiments and Results

In this section, we compare four different multimodality registration methods on synthetic data and simulated T1 and T2 2D MR images. The four methods used are i) Bayesian MAP, ii) empirical mutual information, iii) empirical joint probability and iv) empirical joint entropy with a small deformation penalty added to each one. The objective functions corresponding to the four methods are

1. Bayesian MAP: $E_{\text{MAP}}(\mathbf{u}) = -\frac{1}{N} \sum_{i=1}^N \log \hat{\text{Pr}}(B_i^{(1)} | B_i^{(2)}, \mathbf{u}) + \lambda \|L\mathbf{u}\|^2$.
2. Mutual information: $E_{\text{EMI}}(\mathbf{u}) = -\sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \hat{\text{Pr}}(a, b | \mathbf{u}) \log \hat{\text{Pr}}(a | b, \mathbf{u}) + \lambda \|L\mathbf{u}\|^2$.
3. Joint probability: $E_{\text{EJP}}(\mathbf{u}) = -\frac{1}{N} \sum_{i=1}^N \log \hat{\text{Pr}}(B_i^{(1)}, B_i^{(2)} | \mathbf{u}) + \lambda \|L\mathbf{u}\|^2$.
4. Joint entropy: $E_{\text{EJE}}(\mathbf{u}) = -\sum_{a=1}^{K^{(1)}} \sum_{b=1}^{K^{(2)}} \hat{\text{Pr}}(a, b | \mathbf{u}) \log \hat{\text{Pr}}(a, b | \mathbf{u}) + \lambda \|L\mathbf{u}\|^2$.

For each method, we use essentially the same optimization approach:

Multimodality Non-rigid registration

Set $n = 0$.

Perform Gaussian smoothing (with σ_{Smooth}) on the two images.

Begin A: Do A until $|\Delta E| \leq \delta$ or $n \geq T$.

Initialize \mathbf{u} to zero. This is equivalent to an identity transformation.

$\mathbf{u}^{(n+1)} = \mathbf{u}^{(n)} - \alpha^{(n+1)} \frac{\Delta E}{\Delta \mathbf{u}}$. Choose $\alpha > 0$ such that $\Delta E^{(n+1)} < 0$.

Perform bilinear interpolation.

$n \leftarrow n + 1$

End A

In the above, T is an iteration cap, δ is a convergence threshold and α is a standard step-size parameter. We use numerical differentiation (with $\epsilon = 1$) for computing $\frac{\Delta E}{\Delta \mathbf{u}}$. It should be understood that the objective function E in the algorithm is a placeholder for any of the four methods mentioned above. In all the experiments below, $T = 20$, $\delta = 0.01$, $\sigma = 0.1$, $\sigma_{\text{Smooth}} = 0.5$, $K^{(1)} = K^{(2)} = 8$, and $\lambda = 0.05$.

3.1 Experiments on simple multimodality shape images

We first create two images as shown in Figure 1. The circle and square shapes are swapped and the intensities of each shape differ between the two images.

The Bayesian MAP approach was used to register the right image to the left image. A small amount of isotropic Gaussian smoothing was performed prior to registration. At each iteration, we also observe the values of the empirical mutual information (EMI), empirical joint probability (EJP) and empirical joint entropy (EJE). From Figure 2, we see that the likelihood and mutual information plots are very similar whereas the joint probability and negative joint entropy actually decrease which is counter intuitive. (We also observe that the likelihood does not increase monotonically which is perhaps due to a) bilinear interpolation factors and b) the deformation penalty.) The algorithm was able to convert the circle and the square to approximately a square and a circle in about 6 iterations.

3.2 Experiments on simulated T1 and T2 2D MR images

In our next experiment, we chose two frames generated by the powerful Brainweb MR simulator [1]. The 2D T1 and T2 axial images are shown in Figure 3. We used a 3% noise level for the simulation which uses the ICBM protocol. The advantage of using the Brainweb simulator is that the ground truth is known. Any non-rigid deformation applied to a T1 image can be simultaneously applied to its T2 counterpart.

In our experiments, we used a Gaussian radial basis function (GRBF) spline as the non-rigid parameterization. The deformed T1 image and the intensity difference between the original T1 and the deformed T1 image are shown in Figure 3. The registration algorithm attempts to register the deformed T1 image to the original T2 image. The deformed T1 image is gradually *unwarped* during registration. During the execution of the Bayesian MAP algorithm, we also observe the values of the empirical mutual information (EMI), empirical joint probability (EJP) and empirical joint entropy (EJE). The results are shown in Figure 4. In this case, all four curves mostly show an increase which is somewhat different from the behavior in Figure 2. Once again, the MAP and EMI curves are in lockstep as are EJP and EJE. As a comparison between the different algorithms, we executed all four approaches on the same data. The difference images

(between original T1 and unwarped T1) shown in Figure 5 clearly indicate that the MAP and EMI algorithms are superior to the EJP and EJE algorithms.

We performed another experiment on the pair of T1 and T2 simulated 2D MR images. This time, the deformation on the original T1 image was much larger. The original and deformed T1 images are shown on the left in Figure 6. The results of executing the Bayesian MAP algorithm on the deformed T1 and T2 images are shown on the right in Figure 6. Clearly, despite errors near high gradient boundaries (which are mostly caused by interpolation artifacts), the MAP algorithm is up to the task of recovering a reasonably large global deformation.

4 Conclusion

The main contributions of this paper are i) a Bayesian approach to multimodality non-rigid registration, ii) use of an analytical, smoothed histogram-like density, iii) the derivation of an upper bound for the number of configurations of the displacement field which guarantees that the minimizer of the MAP objective approaches that of the *true* mutual information as the number of pixels tends to infinity and iv) experimental confirmation of the Bayesian approach and its close relationship to the empirical mutual information (EMI). We think this is the first time that such a quantitative criterion has been derived to help assess the validity of non-parametric density estimation approaches to multimodality non-rigid registration. The criterion—a sufficient condition—requires that the number of allowed configurations of the displacement field be restricted to $O(N^{N^c})$. The criterion is easily satisfied in the case of quantized, affine parameter configurations and is usually satisfied in practice in the general non-rigid setting. However, more work is needed to fully understand the impact of this criterion on Bayesian MAP non-rigid registration.

While we have elected to use a nonparametric (histogram-like) drop in replacement for the true likelihood, there is no reason why parametric alternatives (such as mixture models [6], Gibbs-Markov models [13] and the like) cannot also be considered. The principal drawback of using a parametric density is that additional parameters have to be estimated. However, if the additional parameters are $O(1)$, then the added estimation burden does not appear to be formidable. These alternatives present appealing avenues for future research.

Acknowledgements

We acknowledge support from NSF IIS 0196457. This paper was inspired by a conversation with Sarang Joshi at IPMI 2001, Davis, CA.

References

1. D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE Trans. Med. Imag.*, 17(3):463–468, 1998.
2. T. Gaens, F. Maes, D. Vandermeulen, and P. Suetens. Non-rigid multimodal image registration using mutual information. In W. Wells, A. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 1099–1106. Springer, 1998.

3. N. Hata, T. Dohi, S. Warfield, W. Wells, R. Kikinis, and F. A. Jolesz. Multimodality deformable registration of pre- and intraoperative images for MRI-guided brain surgery. In W. Wells, A. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 1067–1074. 1998.
4. B. Kim, J. L. Boes, K. A. Frey, and C. R. Meyer. Mutual information for automated unwarping of rat brain autoradiographs. *NeuroImage*, 5:31–40, 1997.
5. J. Kim, J. W. Fisher, A. Tsai, C. Wible, A. S. Willsky, and W. Wells. Incorporating spatial priors into an information theoretic approach for fMRI data analysis. In W. Wells, A. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 62–71. Springer, 2000.
6. M. E. Leventon and W. E. L. Grimson. Multi-modal volume registration using joint intensity distributions. In W. Wells, A. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 1057–1066. Springer, 1998.
7. F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imag.*, 16(2):187–198, 1997.
8. J. B. A. Maintz, H. W. Meijering, and M. A. Viergever. General multimodal elastic registration based on mutual information. In *Medical Imaging—Image Processing (SPIE 3338)*, volume 3338, pages 144–154. SPIE Press, 1998.
9. D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Non-rigid registration using free-form deformations: Application to breast MR images. *IEEE Trans. Med. Imag.*, 18(8):712–721, 1999.
10. V. N. Vapnik. *Statistical learning theory*. John Wiley, New York, 1998.
11. G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, PA, 1990.
12. W. Wells III, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–52, 1996.
13. S. C. Zhu, Y. N. Wu, and D. B. Mumford. Minimax entropy principle and its applications to texture modeling. *Neural Computation*, 9(8):1627–1660, 1997.



Fig. 1. Left and middle: Two simple shape images. Right: Final registration result.

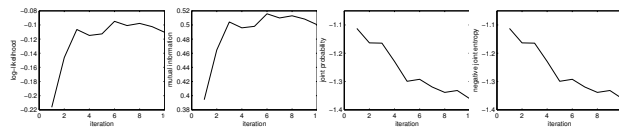


Fig. 2. The changes in i) log-likelihood, ii) mutual information, iii) joint probability and iv) negative joint entropy when minimizing the Bayesian MAP objective function.



Fig. 3. Leftmost: transverse T2 image. Left middle: transverse T1 image. Middle: Deformed T1. Right middle: Intensity difference between original T1 and deformed T1 prior to registration. Right: Unwarped final T1 image

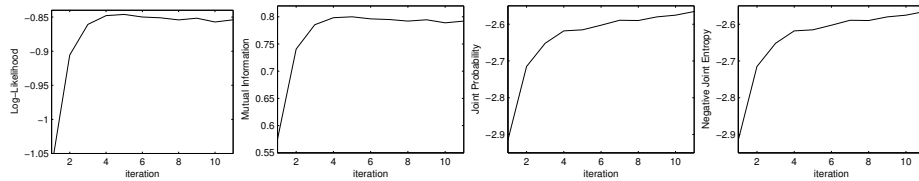


Fig. 4. The changes in i) log-likelihood, ii) mutual information, iii) joint probability and iv) negative joint entropy when minimizing the Bayesian MAP objective function.



Fig. 5. Difference images between original T1 and unwarped T1. Left: MAP. Left middle: EMI. Right middle: EJP. Right: EJE. The SSDs were 609 before registration, and 20.38 (MAP), 20.74 (EMI), 52.49 (EJP) and 52.47 (EJE) after registration.



Fig. 6. Leftmost: Original T1 image, Left middle: deformed T1 image. Middle: Intensity difference between original T1 and deformed T1 before registration. Right middle: Intensity difference between original T1 and unwarped T1 after registration. Right: Unwarped final T1 image. The before and after SSDs were 647 and 59 respectively.