Probability Density Estimation using Isocontours and Isosurfaces: Application to Information Theoretic Image Registration

Ajit Rajwade, Arunava Banerjee and Anand Rangarajan,
Department of CISE, University of Florida, Gainesville, USA

***Abstract*—**

We present a new, geometric approach for determining the probability density of the intensity values in an image. We drop the notion of an image as a set of discrete pixels, and assume a piecewise-continuous representation. The probability density can then be regarded as being proportional to the area between two nearby isocontours of the image surface. Our paper extends this idea to joint densities of image pairs. We demonstrate the application of our method to affine registration between two or more images using information theoretic measures such as mutual information. We show cases where our method outperforms existing methods such as simple histograms, histograms with partial volume interpolation, Parzen windows, etc. under fine intensity quantization for affine image registration under significant image noise. Furthermore, we demonstrate results on simultaneous registration of multiple images, as well as for pairs of volume datasets, and show some theoretical properties of our density estimator. Our approach requires the selection of only an image interpolant. The method neither requires any kind of kernel functions (as in Parzen windows) which are unrelated to the structure of the image in itself, nor does it rely on any form of sampling for density estimation.

I. Introduction

Information theoretic tools have for a long time been established as the *de facto* technique for image registration, especially in the domains of medical imaging [22] and remote sensing [3] which deal with a large number of modalities. The groundbreaking work for this was done by Viola and Wells [32], and Maes *et al.* [17] in their widely cited papers. A detailed survey of subsequent research on information theoretic techniques in medical image registration is presented in the works of Pluim *et al.* [22] and Maes *et al.* [18]. A required component of all information theoretic techniques in image registration is a good estimator of the joint entropies of the images being registered. Most techniques employ plug-in entropy estimators, wherein the joint and marginal probability densities of the intensity values in the images are first estimated and these quantities are then used to obtain the entropy. There also exist recent methods which define a new form of entropy using cumulative distributions instead of probability densities (see [25]). Furthermore, there also exist techniques which directly estimate the entropy, without estimating the probability density or distribution as an intermediate step [1]. Below, we present a bird's eye view of these techniques and their limitations. Subsequently, we introduce our method and bring out its salient merits.

The plug-in entropy estimators rely upon techniques for density estimation as a key first step. The most popular density estimator is the simple image histogram. The drawbacks of a histogram are that it yields discontinuous density estimates and requires an optimal choice of the bin width. Too small a bin width leads to noisy, sparse density estimates (variance) whereas too large a bin width introduces oversmoothing (bias). Parzen windows have been widely employed as a differentiable density estimator for several applications in computer vision, including image registration [32]. Here the problem of choosing an optimal bin width translates to the optimal choice of a kernel width and the kernel function itself. The choice of the kernel function is somewhat arbitrary [29] and furthermore the implicit effect of the kernel choice on the structure of the image is an issue that has been widely ignored[1]. The kernel width parameter can be estimated by techniques such as maximum likelihood. Such methods, however, require complicated iterative optimizations, and also a training and validation set. From an image registration standpoint, the joint density between the images undergoes a change in each iteration, which requires re-estimation of the kernel width parameters. This step is an expensive iterative process with a complexity that is quadratic in the number of samples. Methods such as the Fast Gauss transform [33] reduce this cost to some extent but they require a prior clustering step. However, the Fast Gauss transform is only an *approximation* to the true Parzen density estimate, and hence, one needs to analyze the behavior of the approximation error over the iterations if a gradient-based optimizer is used. Also, as per [9] (Section 3.3.2), the ideal width value for minimizing the mean squared error between the true and estimated density is itself dependent upon the second derivative of the (unknown) true density. Yet another drawback of Parzen window based density estimators is the well-known "tail effect" in higher dimensions, due to which a large number of samples will fall in those regions where the Gaussian has very low value [29]. Mixture models have been used for joint density estimation in registration [15], but they are quite inefficient and require choice of the kernel function for the components (usually chosen to be Gaussian) and the number of components. This number again will change across the iterations of the registration process, as the images move with respect to one another. Wavelet based density estimators have also been recently employed in image registration [9] and in conjunction with MI [21]. The problems with a wavelet based method for density estimation include a choice of wavelet function, as well as the selection of the optimal number of levels or coefficients, which again requires iterative optimization.

Direct entropy estimators avoid the intermediate density esti-

[1]Parzen showed in [19] that $\sup|\hat{p}(x) - p(x)| \to 0$, where $\hat{p}$ and p refer to the estimated and true density respectively. However, we stress that this is only an asymptotic result (as the number of samples $N_s \to \infty$) and therefore not directly linked to the nature of the image itself, for all practical purposes.

mation phase. While there exists a plethora of papers in this field (surveyed in [1]), the most popular entropy estimator used in image registration is the approximation of the Renyi entropy as the weight of a minimal spanning tree [16] or a K-nearest neighbor graph [5]. Note that the entropy used here is the Renyi entropy as opposed to the more popular Shannon entropy. Drawbacks of this approach include the computational cost in construction of the data structure in each step of registration (the complexity whereof is quadratic in the number of samples drawn), the somewhat arbitrary choice of the α parameter for the Renyi entropy and the lack of differentiability of the cost function. Some work has been done recently, however, to introduce differentiability in the cost function [27]. A merit of these techniques is the ease of estimation of entropies of high-dimensional feature vectors, with the cost scaling up just linearly with the dimensionality of the feature space.

Recently, a new form of the entropy defined on cumulative distributions, and related cumulative entropic measures such as cross cumulative residual entropy (CCRE) have been introduced in the literature on image registration [25]. The cumulative entropy and the CCRE measure have perfectly compatible discrete and continuous versions (quite unlike the Shannon entropy, though not unlike the Shannon mutual information), and are known to be noise resistant (as they are defined on cumulative distributions and not densities). Our method of density estimation can be easily extended to computing cumulative distributions and CCRE.

All the techniques reviewed here are based on different principles, but have one crucial common point: they treat the image as a set of pixels or samples, which inherently ignores the fact that these samples originate from an underlying continuous (or piece-wise continuous) signal. None of these techniques take into account the ordering between the given pixels of an image. As a result, all these methods can be termed *sample-based*. Furthermore, most of the aforementioned density estimators require a particular kernel, the choice of which is *extrinsic* to the image being analyzed and not necessarily linked even to the noise model. In this paper, we present an entirely different approach in which the density estimate is built directly from a continuous *image* representation (as opposed an arbitrary kernel on the density). Our approach here is based on the earlier work presented in [24], the essence of which is to regard the marginal probability density as the area between two isocontours at infinitesimally close intensity values. A similar approach to density estimation has also been taken in the work of Kadir and Brady [13]. In our work, we have also presented a detailed derivation for the joint density between two or more images, and also extended the work to the 3D case, besides testing it thoroughly on affine image registration for varying noise level and quantization widths. Prior work on image registration using such image based techniques includes [24], [23], [10] and [14]. The work in [10], however, reports results only on template matching with translations, whereas the main focus of [14] is on estimation of densities in vanishingly small circular neighborhoods. The formulae derived are very specific to the shape of the neighborhood. Their paper [14] shows that local mutual information values in small neighborhoods are related to the values of the angles between the local gradient vectors in those neighborhoods. The focus of this method, however is too local in nature, thereby ignoring the robustness that is an integral part of more global density estimates. There also exists some related work by Hadjidemetriou *et al.* [12] in the context of histogram preserving locally continuous image transformations (the so-called Hamiltonian morphisms), which relates histograms to areas between isocontours. The main practical applications discussed in [12] are histograms under weak perspective and paraperspective projections of 3D textured models.

Note that our method, based on finding areas between isocontours, is significantly different from Partial Volume Interpolation (PVI) [17], [30]. PVI uses a continuous image representation to build a joint probability table by assigning fractional votes to multiple intensity pairs when a digital image is warped during registration. The fractional votes are assigned typically using a bilinear or bicubic kernel function in cases of non-alignment with pixel grids after image warping. In essence, the density estimate in PVI still requires histogramming or Parzen windowing.

In this paper, we also present in detail the problems that lead to singularities in the probability density as estimated by the suggested procedure and also suggested principled modifications. The main merit of the proposed geometric technique is the fact that it side-steps the parameter selection problem that affects other density estimators and also does not rely on any form of sampling. The accuracy of our techniques will always *upper bound* all sample-based methods if the image interpolant is known (see Section IV). In fact, the estimate obtained by all sample-based methods will converge to that yielded by our method only in the limit when the number of samples tends to infinity. Empirically, we demonstrate the robustness of our technique to noise, and superior performance in image registration. We conclude with a discussion and clarification of some properties of our method.

II. Marginal and Joint Density Estimation

In this section, we show the derivation of the probability density function (PDF) for the marginal as well as the joint density for a pair of 2D images. We point out practical issues and computational considerations, as well as outline the density derivations for the case of 3D images, as well as multiple images in 2D.

A. Estimating the Marginal Densities in 2D

Consider the 2D gray-scale image intensity to be a continuous, scalar-valued function of the spatial variables, represented as $w = I(x,y)$. Let the total area of the image be denoted by A. Assume a location random variable $Z =< X,Y >$ with a uniform distribution over the image field of view (FOV). Further, assume a new random variable W which is a transformation of the random variable Z and with the transformation given by the gray-scale image intensity function $W = I(X,Y)$. Then the cumulative distribution of W at a certain intensity level α is equal to the ratio of the total area of all regions whose intensity is less than or equal to α to the total area of the image

$$\Pr(W \leq \alpha) = \frac{1}{A} \int \int_{I(x,y)\leq\alpha} dxdy. \tag{1}$$

Now, the probability density of W at α is the derivative of the cumulative distribution in (1). This is equal to the difference in the areas enclosed within two level curves that are separated by an intensity difference of $\Delta\alpha$ (or equivalently, the area enclosed between two level curves of intensity α and $\alpha+\Delta\alpha$), per unit difference, as $\Delta\alpha \to 0$ (see Figure 1). The formal expression for this is

$$p(\alpha) = \frac{1}{A} \lim_{\Delta\alpha\to 0} \frac{\int\int_{I(x,y)\leq\alpha+\Delta\alpha} dxdy - \int\int_{I(x,y)\leq\alpha} dxdy}{\Delta\alpha}. \tag{2}$$

Hence, we have

$$p(\alpha) = \frac{1}{A}\frac{d}{d\alpha}\int\int_{I(x,y)\leq\alpha} dxdy. \tag{3}$$

We can now adopt a change of variables from the spatial coordinates (x,y) to $u(x,y)$ and $I(x,y)$, where u and I are the directions parallel and perpendicular to the level curve of intensity α, respectively. Observe that I points in the direction of the image gradient, or the direction of maximum intensity change. Noting this fact, we now obtain the following:

$$p(\alpha) = \frac{1}{A}\int_{I(x,y)=\alpha} \left| \begin{array}{cc} \frac{\partial x}{\partial I} & \frac{\partial y}{\partial I} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \end{array} \right| du. \tag{4}$$

Note that in Eq. (4), $d\alpha$ and dI have "canceled" each other out, as they both stand for intensity change. Upon a series of algebraic manipulations, we are now left with the following expression for $p(\alpha)$ (with a more detailed derivation to be found in the Appendix):

$$p(\alpha) = \frac{1}{A}\int_{I(x,y)=\alpha} \frac{du}{\sqrt{(\frac{\partial I}{\partial x})^2 + (\frac{\partial I}{\partial y})^2}}. \tag{5}$$

From the above expression, one can make some important observations. Each point on a given level curve contributes a certain measure to the density at that intensity which is inversely proportional to the magnitude of the gradient at that point. In other words, in regions of high intensity gradient, the area between two level curves at nearby intensity levels would be small, as compared to that in regions of lower image gradient (see Figure 1). When the gradient value at a point is zero (owing

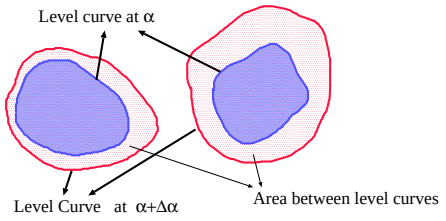


Fig. 1. $p(\alpha) \propto$ area between level curves at α and $\alpha+\Delta\alpha$ (i.e. region with red dots)

to the existence of a peak, a valley, a saddle point or a flat region), the contribution to the density at that point tends to infinity. (The practical repercussions of this situation are discussed later on in the paper. Lastly, the density at an intensity level can be estimated by traversing the level curve(s) at that intensity and integrating the reciprocal of the gradient magnitude. One can obtain an estimate of the density at several intensity levels (at intensity spacing of h from each other) across the entire intensity range of the image.

B. Estimating the Joint Density

Consider two images represented as continuous scalar valued functions $w_1 = I_1(x,y)$ and $w_2 = I_2(x,y)$, whose overlap area is A. As before, assume a location random variable $Z = \{X,Y\}$ with a uniform distribution over the (overlap) field of view. Further, assume two new random variables W_1 and W_2 which are transformations of the random variable Z and with the transformations given by the gray-scale image intensity functions $W_1 = I_1(X,Y)$ and $W_2 = I_2(X,Y)$. Let the set of all regions whose intensity in I_1 is less than or equal to α_1 and whose intensity in I_2 is less than or equal to α_2 be denoted by L. The cumulative distribution $\Pr(W_1 \leq \alpha_1, W_2 \leq \alpha_2)$ at intensity values (α_1,α_2) is equal to the ratio of the total area of L to the total overlap area A. The probability density $p(\alpha_1,\alpha_2)$ in this case is the second partial derivative of the cumulative distribution w.r.t. α_1 and α_2. Consider a pair of level curves from I_1 having intensity values α_1 and $\alpha_1+\Delta\alpha_1$, and another pair from I_2 having intensity α_2 and $\alpha_2+\Delta\alpha_2$. Let us denote the region enclosed between the level curves of I_1 at α_1 and $\alpha_1+\Delta\alpha_1$ as Q_1 and the region enclosed between the level curves of I_2 at α_2 and $\alpha_2+\Delta\alpha_2$ as Q_2. Then $p(\alpha_1,\alpha_2)$ can geometrically be interpreted as the area of $Q_1 \cap Q_2$, divided by $\Delta\alpha_1\Delta\alpha_2$, in the limit as $\Delta\alpha_1$ and $\Delta\alpha_2$ tend to zero. The regions Q_1, Q_2 and also $Q_1 \cap Q_2$ (dark black region) are shown in Figure 2(left). Using a technique very similar to that shown in Eqs. (2)-(4), we obtain the expression for the joint cumulative distribution as follows:

$$\Pr(W_1 \leq \alpha_1, W_2 \leq \alpha_2) = \frac{1}{A}\int\int_L dxdy. \tag{6}$$

By doing a change of variables, we arrive at the following formula:

$$\Pr(W_1 \leq \alpha_1, W_2 \leq \alpha_2) = \frac{1}{A}\int\int_L \left| \begin{array}{cc} \frac{\partial x}{\partial u_1} & \frac{\partial y}{\partial u_1} \\ \frac{\partial x}{\partial u_2} & \frac{\partial y}{\partial u_2} \end{array} \right| du_1 du_2. \tag{7}$$

Here u_1 and u_2 represent directions along the corresponding level curves of the two images I_1 and I_2. Taking the second partial derivative with respect to α_1 and α_2, we get the expression for the joint density:

$$p(\alpha_1,\alpha_2) = \frac{1}{A}\frac{\partial^2}{\partial\alpha_1\partial\alpha_2}\int\int_L \left| \begin{array}{cc} \frac{\partial x}{\partial u_1} & \frac{\partial y}{\partial u_1} \\ \frac{\partial x}{\partial u_2} & \frac{\partial y}{\partial u_2} \end{array} \right| du_1 du_2. \tag{8}$$

It is important to note here again, that the joint density in (8) may not exist because the cumulative may not be differentiable. Geometrically, this occurs if (a) both the images have locally constant intensity, (b) if only one image has locally constant intensity, or (c) if the level sets of the two images are locally parallel. In case (a), we have area-measures and in the other two cases, we have curve-measures. These cases are described in detail in the following section, but for the moment, we shall ignore these degeneracies.

To obtain a complete expression for the PDF in terms of gradients, it would be highly intuitive to follow purely geometric reasoning. One can observe that the joint probability density $p(\alpha_1,\alpha_2)$ is the sum total of "contributions" at every intersection between the level curves of I_1 at α_1 and those of I_2 at α_2. Each contribution is the area of parallelogram ABCD [see Figure 2(right)] at the level curve intersection, as the intensity differences $\Delta\alpha_1$ and $\Delta\alpha_2$ shrink to zero. (We consider a parallelogram here, because we are approximating the level curves locally as straight lines.) Let the coordinates of the point B be $(\tilde{x},\tilde{y})$ and the magnitude of the gradient of I_1 and I_2 at this point be $g_1(\tilde{x},\tilde{y})$ and $g_2(\tilde{x},\tilde{y})$. Also, let $\theta(\tilde{x},\tilde{y})$ be the *acute* angle between the gradients of the two images at B. Observe that the intensity difference between the two level curves of I_1 is $\Delta\alpha_1$. Then, using the definition of gradient, the perpendicular distance between the two level curves of I_1 is given as $\frac{\Delta\alpha_1}{g_1(\tilde{x},\tilde{y})}$. Looking at triangle CDE (wherein CE is perpendicular to the level curves) we can now

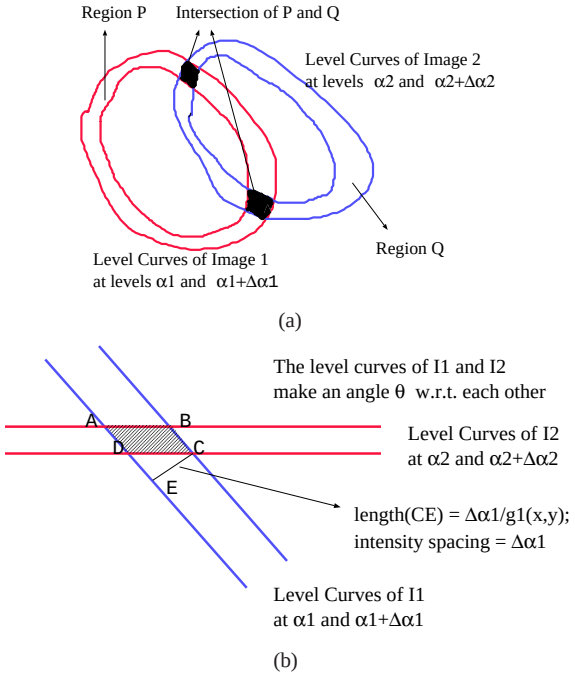


Fig. 2. Left: Intersection of level curves of I_1 and I_2: $p(\alpha_1,\alpha_2) \propto$ area of dark black regions. Right: Parallelogram approximation: PDF contribution = area (*ABCD*)

deduce the length of *CD* (or equivalently that of *AB*). Similarly, we can also find the length *CB*. The two expressions are given by:

$$|AB| = \frac{\Delta\alpha_1}{g_1(\tilde{x},\tilde{y})\sin\theta(\tilde{x},\tilde{y})}, |CB| = \frac{\Delta\alpha_2}{g_2(\tilde{x},\tilde{y})\sin\theta(\tilde{x},\tilde{y})}. \quad (9)$$

Now, the area of the parallelogram is equal to $|AB||CB|\sin\theta(\tilde{x},\tilde{y})$, which evaluates to $\frac{\Delta\alpha_1\Delta\alpha_2}{g_1(\tilde{x},\tilde{y})g_2(\tilde{x},\tilde{y})\sin\theta(\tilde{x},\tilde{y})}$. With this, we finally obtain the following expression for the joint density:

$$p(\alpha_1,\alpha_2) = \frac{1}{A}\sum_C \frac{1}{g_1(x,y)g_2(x,y)\sin\theta(x,y)} \quad (10)$$

where the set C represents the (countable) locus of all points where $I_1(x,y) = \alpha_1$ and $I_2(x,y) = \alpha_2$. It is easy to show through algebraic manipulations that Eqs. (8) and (10) are equivalent formulations of the joint probability density $p(\alpha_1,\alpha_2)$. These results could also have been derived purely by manipulation of Jacobians (as done while deriving marginal densities), and the derivation for the marginals could also have proceeded following geometric intuitions.

The formula derived above tallies beautifully with intuition in the following ways. Firstly, the area of the parallelogram *ABCD* (i.e. the joint density contribution) in regions of high gradient [in either or both image(s)] is smaller as compared to that in the case of regions with lower gradients. Secondly, the area of parallelogram *ABCD* (i.e. the joint density contribution) is the least when the gradients of the two images are orthogonal and maximum when they are parallel or coincident [see Figure 3(a)]. In fact, the joint density tends to infinity in the case where either (or both) gradient(s) is (are) zero, or when the two gradients align, so that $\sin\theta$ is zero. The repercussions of this phenomenon are discussed in the following section.

C. From Densities to Distributions

In the two preceding sub-sections, we observed the divergence

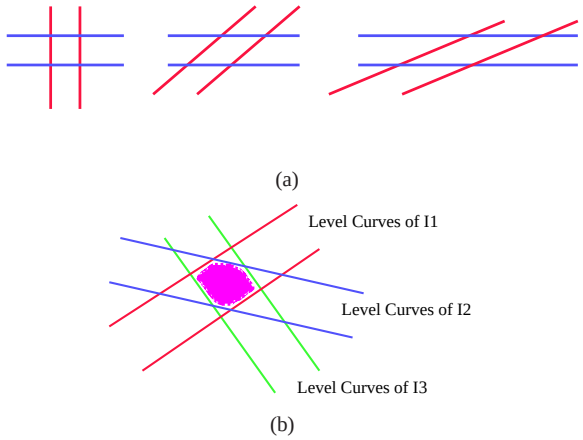
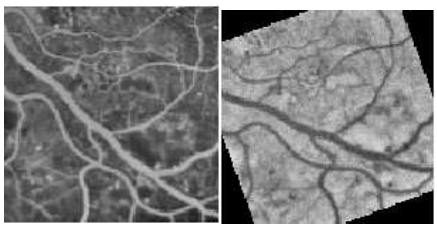


Fig. 3. (a) Area of parallelogram increases as angle between level curves decreases (left to right). Level curves of I_1 and I_2 are shown in red and blue lines respectively. (b) Joint probability contribution in the case of three images.

of the marginal density in regions of zero gradient, or of the joint density in regions where either (or both) image gradient(s) is (are) zero, or when the gradients locally align. The gradient goes to zero in regions of the image that are flat in terms of intensity, and also at peaks, valleys and saddle points on the image surface. We can ignore the latter three cases as they are a finite number of points within a continuum. The probability contribution at a particular intensity in a flat region is proportional to the area of that flat region. Some *ad hoc* approaches could involve simply "weeding out" the flat regions altogether, but that would require the choice of sensitive thresholds. The key thing is to notice that in these regions, *the density does not exist but the probability distribution does*. So, we can switch entirely to probability distributions everywhere by introducing a non-zero lower bound on the "values" of $\Delta\alpha_1$ and $\Delta\alpha_2$. Effectively, this means that we always look at parallelograms representing the intersection between pairs of level curves from the two images, separated by *non-zero* intensity difference, denoted as, say, *h*. Since these parallelograms have finite areas, we have circumvented the situation of choosing thresholds to prevent the values from becoming unbounded, and the probability at α_1,α_2, denoted as $\hat{p}(\alpha_1,\alpha_2)$ is obtained from the areas of such parallelograms. We term this area-based method of density estimation as *AreaProb*. Later on in the paper, we shall show that the switch to distributions is principled and does not reduce our technique to standard histogramming in any manner whatsoever.



(a) (b)

Fig. 4. A retinogram [31] and its rotated negative.

The notion of an image as a continuous entity is one of the pillars of our approach. We adopt a locally linear formulation in this paper, for the sake of simplicity, though the technical

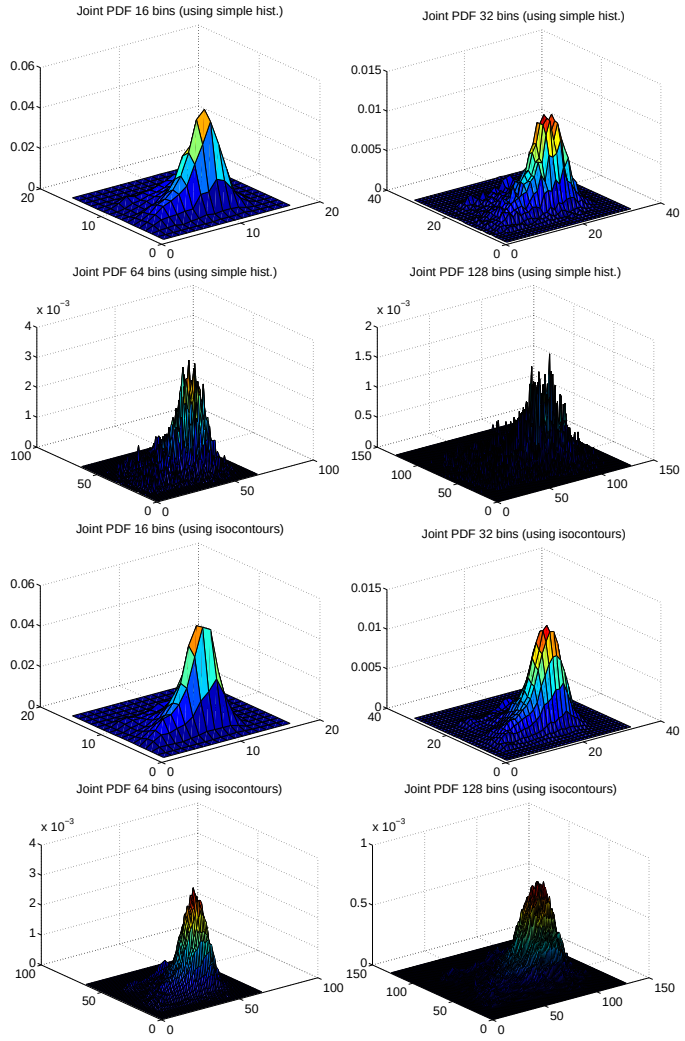


Fig. 5. Joint densities of the retinogram images computed by histograms (top 2 rows) and by our area-based method (bottom 2 rows) using 16, 32, 64 and 128 bins.

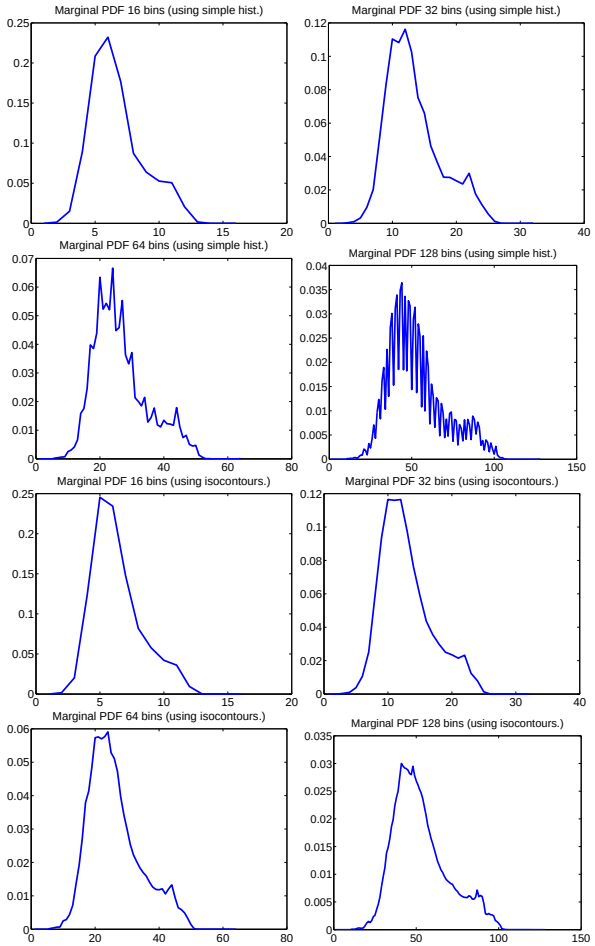


Fig. 6. Marginal densities of the retinogram image computed by histograms (top 2 rows) and our area-based method (bottom 2 rows) using 16, 32, 64 and 128 bins (row-wise order).

contributions of this paper are in no way tied to any specific interpolant. For each image grid point, we estimate the intensity values at its four neighbors within a horizontal or vertical distance of 0.5 pixels. We then divide each square defined by these neighbors into a pair of *triangles*. The intensities within each triangle can be represented as a planar patch, which is given by the equation $z_1 = A_1x + B_1y + C_1$ in I_1. Iso-intensity lines at levels α_1 and $\alpha_1 + h$ within this triangle are represented by the equations $A_1x + B_1y + C_1 = \alpha_1$ and $A_1x + B_1y + C_1 = \alpha_1 + h$ (likewise for the iso-intensity lines of I_2 at intensities α_2 and $\alpha_2 + h$, within a triangle of corresponding location). The contribution from this triangle to the joint probability at (α_1, α_2), i.e. $\hat{p}(\alpha_1, \alpha_2)$ is the area bounded by the two pairs of parallel lines, clipped against the body of the triangle itself, as shown in Figure 7. In the case that the corresponding gradients from the two images are parallel (or coincident), they enclose an infinite area between them, which when clipped against the body of the triangle, yields a closed polygon of finite area, as shown in Figure 7. When both the gradients are zero (which can be considered to be a special case of gradients being parallel), the probability contribution is equal to the area of the entire triangle. In the case where the gradient of only one of the images is zero, the contribution is equal to the area enclosed between the parallel iso-intensity lines of the *other* image, clipped against the body of the triangle (see Figure 7). Observe that though we have to treat pathological regions specially (despite having switched to distributions), we now do not need to select thresholds, nor do we need to deal with a mixture of densities and distributions. The other major advantage is added robustness to noise, as we are now working with probabilities instead of their derivatives, i.e. densities.

The issue that now arises is how the value of h may be chosen. It should be noted that although there is no "optimal" h, our density estimate would convey more and more information as the value of h is reduced (in complete contrast to standard histogramming). In Figure 5, we have shown plots of our joint density estimate and compared it to standard histograms for P equal to 16, 32, 64 and 128 bins in *each* image (i.e. 32^2, 64^2 etc. bins in the joint), which illustrate our point clearly. We found that the standard histograms had a far greater number of empty bins than our density estimator, for the same number of intensity levels. The corresponding marginal discrete distributions for the original retinogram image [31] for 16, 32, 64 and 128 bins are shown in Figure 6.

D. Joint Density Between Multiple Images in 2D

For the simultaneous registration of multiple ($d > 2$) images, the use of a single d-dimensional joint probability has been

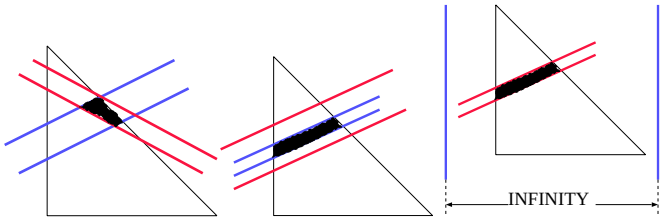


Fig. 7. Left: Probability contribution equal to area of parallelogram between level curves clipped against the triangle, i.e. half-pixel. Middle: Case of parallel gradients. Right: Case when the gradient of one image is zero (blue level lines) and that of the other is non-zero (red level lines). In each case, probability contribution equals area of the dark black region.

advocated in previous literature [2], [35]. Our joint probability derivation can be easily extended to the case of $d > 2$ images by using similar geometric intuition to obtain the polygonal area between d intersecting pairs of level curves [see Figure 3(right) for the case of $d = 3$ images]. Note here that the d-dimensional joint distribution lies essentially in a 2D subspace, as we are dealing with 2D images. A naïve implementation of such a scheme has a complexity of $O(NP^d)$ where P is the number of intensity levels chosen for each image and N is the size of each image. Interestingly, however, this exponential cost can be side-stepped by first computing the at most $(\frac{d(d-1)}{2})P^2$ points of intersection between pairs of level curves from all d images with one another, for every pixel. Secondly, a graph can be created, each of whose nodes is an intersection point. Nodes are linked by edges labeled with the image number (say k^{th} image) if they lie along the same iso-contour of that image. In most cases, each node of the graph will have a degree of four (and in the unlikely case where level curves from all images are concurrent, the maximal degree of a node will be $2d$). Now, this is clearly a planar graph, and hence, by Euler's formula, we have the number of (convex polygonal) faces $\tilde{F} = \frac{d(d-1)}{2} * 4P^2 - \frac{d(d-1)}{2}P^2 + 2 = O(P^2d^2)$, which is quadratic in the number of images. The area of the polygonal faces are contributions to the joint probability distribution. In a practical implementation, there is no requirement to even create the planar graph. Instead, we can implement a simple incremental face-splitting algorithm ([7], section 8.3). In such an implementation, we create a list of faces $\mathscr{F}$ which is updated incrementally. To start with, $\mathscr{F}$ consists of just the triangular face constituting the three vertices of a chosen half-pixel in the image. Next, we consider a single level-line l at a time and split into two any face in $\mathscr{F}$ that l intersects. This procedure is repeated for all level lines (separated by a discrete intensity spacing) of all the d images. The final output is a listing of all polygonal faces $\mathscr{F}$ created by incremental splitting which can be created in just $O(\tilde{F}Pd)$ time. The storage requirement can be made polynomial by observing that for d images, the number of unique intensity tuples will be at most $\tilde{F}N$ in the worst case (as opposed to P^d). Hence all intensity tuples can be efficiently stored and indexed using a hash table.

E. Extensions to 3D

When estimating the probability density from 3D images, the choice of an optimal smoothing parameter is a less critical issue, as a much larger number of samples are available. However, at a theoretical level this still remains a problem, which would worsen in the multiple image case. In 3D, the marginal probability can be interpreted as the total volume sandwiched between two iso-surfaces at neighboring intensity levels. The formula for the marginal density $p(\alpha)$ of a 3D image $w = I(x,y,z)$ is given as follows:

$$p(\alpha) = \frac{1}{V}\frac{d}{d\alpha}\int\int\int_{I(x,y,z)\leq\alpha} dxdydz. \tag{11}$$

Here V is the volume of the image $I(x,y,z)$. We can now adopt a change of variables from the spatial coordinates x, y and z to $u_1(x,y,z)$, $u_2(x,y,z)$ and $I(x,y,z)$, where I is the perpendicular to the level surface (i.e. parallel to the gradient) and u_1 and u_2 are mutually perpendicular directions parallel to the level surface. Noting this fact, we now obtain the following:

$$p(\alpha) = \frac{1}{V}\int\int_{I(x,y,z)=\alpha}\left|\begin{array}{ccc}\frac{\partial x}{\partial I} & \frac{\partial y}{\partial I} & \frac{\partial z}{\partial I}\\ \frac{\partial x}{\partial u_1} & \frac{\partial y}{\partial u_1} & \frac{\partial z}{\partial u_1}\\ \frac{\partial x}{\partial u_2} & \frac{\partial y}{\partial u_2} & \frac{\partial z}{\partial u_2}\end{array}\right| du_1du_2. \tag{12}$$

Upon a series of algebraic manipulations just as before, we are left with the following expression for $p(\alpha)$:

$$p(\alpha) = \frac{1}{V}\int\int_{I(x,y,z)=\alpha}\frac{du_1du_2}{\sqrt{(\frac{\partial I}{\partial x})^2+(\frac{\partial I}{\partial y})^2+(\frac{\partial I}{\partial z})^2}}. \tag{13}$$

For the joint density case, consider two 3D images represented as $w_1 = I_1(x,y,z)$ and $w_2 = I_2(x,y,z)$, whose overlap volume (the field of view) is V. The cumulative distribution $\Pr(W_1 \leq \alpha_1, W_2 \leq \alpha_2)$ at intensity values (α_1,α_2) is equal to the ratio of the total volume of all regions whose intensity in the first image is less than or equal to α_1 and whose intensity in the second image is less than or equal to α_2, to the total image volume. The probability density $p(\alpha_1,\alpha_2)$ is again the second partial derivative of the cumulative distribution. Consider two regions R_1 and R_2, where R_1 is the region trapped between level surfaces of the first image at intensities α_1 and $\alpha_1+\Delta\alpha_1$, and R_2 is defined analogously for the second image. The density is proportional to the volume of the intersection of R_1 and R_2 divided by $\Delta\alpha_1$ and $\Delta\alpha_2$ when the latter two tend to zero. It can be shown through some geometric manipulations that the area of the base of the parallelepiped formed by the isosurfaces is given as $\frac{\Delta\alpha_1\Delta\alpha_2}{|\vec{g_1}\times\vec{g_2}|} = \frac{\Delta\alpha_1\Delta\alpha_2}{|g_1g_2\sin(\theta)|}$, where $\vec{g_1}$ and $\vec{g_2}$ are the gradients of the two images, and θ is the angle between them. Let $\vec{h}$ be a vector which points in the direction of the height of the parallelepiped (parallel to the base normal, i.e. $\vec{g_1}\times\vec{g_2}$), and $d\vec{h}$ be an infinitesimal step in that direction. Then the probability density is given as follows:

$$\begin{aligned} p(\alpha_1,\alpha_2) &= \frac{1}{V}\frac{\partial^2}{\partial\alpha_1\partial\alpha_2}\int\int\int_{V_s} dxdydz \\ &= \frac{1}{V}\frac{\partial^2}{\partial\alpha_1\partial\alpha_2}\int\int\int_{V_s}\frac{d\vec{u_1}d\vec{u_2}d\vec{h}}{|\vec{g_1}\times\vec{g_2}|} = \frac{1}{V}\int_C\frac{d\vec{h}}{|\vec{g_1}\times\vec{g_2}|} \end{aligned} \tag{14}$$

In Eq. (14), $\vec{u_1}$ and $\vec{u_2}$ are directions parallel to the iso-surfaces of the two images, and $\vec{h}$ is their cross-product (and parallel to the line of intersection of the individual planes), while C is the 3D space curve containing the points where I_1 and I_2 have values α_1 and α_2 respectively and $V_s \stackrel{\text{def}}{=} \{(x,y,z) : I_1(x,y,z) \leq \alpha_1, I_2(x,y,z) \leq \alpha_2\}$.

F. Implementation Details for the 3D case

The density formulation for the 3D case suffers from the same problem of divergence to infinity, as in the 2D case. Similar techniques can be employed, this time using level *surfaces* that are separated by finite intensity gaps. To trace the level surfaces,

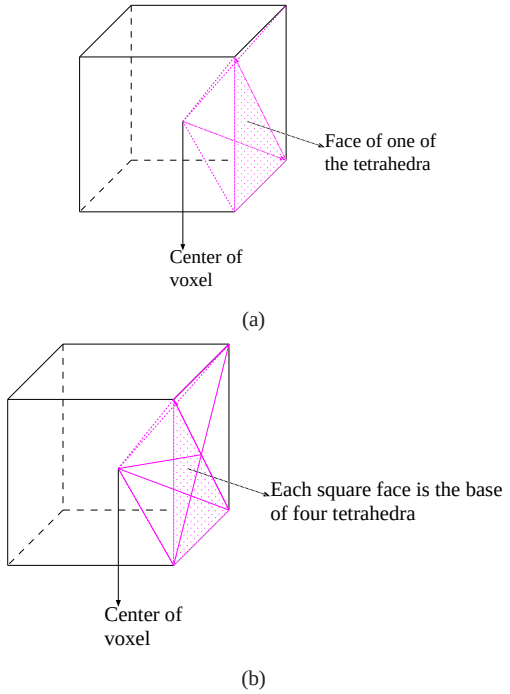


Fig. 8. (a) Splitting a voxel into 12 tetrahedra, two on each of the six faces of the voxel; (b) Splitting a voxel into 24 tetrahedra, four on each of the six faces of the voxel.

each cube-shaped voxel in the 3D image can be divided into 12 tetrahedra. The apex of each tetrahedron is located at the center of the voxel and the base is formed by dividing one of the six square faces of the cube by one of the diagonals of that face [see Figure 8(a)]. Within each triangular face of each such tetrahedron, the intensity can be assumed to be a linear function of location. Note that the intensities in different faces of one and the same tetrahedron can thus be expressed by *different* functions, all of them linear. Hence the isosurfaces at different intensity levels within a single tetrahedron are non-intersecting but not necessarily parallel. These level surfaces at any intensity within a single tetrahedron turn out to be either triangles or quadrilaterals in 3D. This interpolation scheme does have some bias in the choice of the diagonals that divide the individual square faces. A scheme that uses 24 tetrahedra with the apex at the center of the voxel, and four tetrahedra based on every single face, has no bias of this kind [see Figure 8(b)]. However, we still used the former (and faster) scheme as it is simpler and does not noticeably affect the results. Level surfaces are again traced at a finite number of intensity values, separated by equal intensity intervals. The marginal density contributions are obtained as the volumes of convex polyhedra trapped in between consecutive level surfaces clipped against the body of individual tetrahedra. The joint distribution contribution from each voxel is obtained by finding the volume of the convex polyhedron resulting from the intersection of corresponding convex polyhedra from the two images, clipped against the tetrahedra inside the voxel. We refer to this scheme of finding joint densities as *VolumeProb*.

G. Joint Densities by Counting Points and Measuring Lengths

For the specific case of registration of two images in 2D, we present another method of density estimation. This method, which was presented by us earlier in [23], is a biased estimator that does not assume a uniform distribution on location. In this technique, the total number of co-occurrences of intensities α_1 and α_2 from the two images respectively, is obtained by counting the total number of intersections of the corresponding level curves. Each half-pixel can be examined to see whether level curves of the two images at intensities α_1 and α_2 can intersect within the half-pixel. This process is repeated for different (discrete) values from the two images (α_1 and α_2), separated by equal intervals and selected *a priori* (see Figure 9). The co-occurrence counts are then normalized so as to yield a joint probability mass function (PMF). We denote this method as *2DPointProb*. The marginals are obtained by summing up the joint PMF along the respective directions. This method, too, avoids the histogramming binning problem as one has the liberty to choose as many level curves as desired. However, it is a biased density estimator because more points are picked from regions with high image gradient. This is because more level curves (at equi-spaced intensity levels) are packed together in such areas. It can also be regarded as a weighted version of the joint density estimator presented in the previous sub-section, with each point weighted by the gradient magnitudes of the two images at that point as well as the sine of the angle between them. Thus the joint PMF by this method is given as

$$p(\alpha_1,\alpha_2) = \frac{\partial^2}{\partial\alpha_1\partial\alpha_2}\frac{1}{K}\int\int_D g_1(x,y)g_2(x,y)\sin\theta(x,y)dxdy \quad (15)$$

where D denotes the regions where $I_1(x,y) \le \alpha_1, I_2(x,y) \le \alpha_2$ and K is a normalization constant. This simplifies to the following:

$$p(\alpha_1,\alpha_2) = \frac{1}{K}\sum_C 1. \quad (16)$$

Hence, we have $p(\alpha_1,\alpha_2) = \frac{|C|}{K}$, where C is the (countable) set of points where $I_1(x,y) = \alpha_1$ and $I_2(x,y) = \alpha_2$. The marginal (biased) density estimates can be regarded as lengths of the individual isocontours. With this notion in mind, the marginal density estimates are seen to have a close relation with the total variation of an image, which is given by $TV = \int_{I=\alpha} |\nabla I(x,y)| dxdy$ [26]. We clearly have $TV = \int_{I=\alpha} du$, by doing the same change of variables (from x,y to u,I) as in Eqs. (4) and (5), thus giving us the length of the isocontours at any given intensity level.

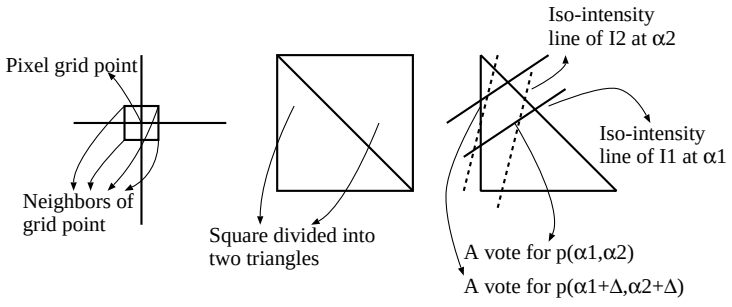


Fig. 9. Counting level curve intersections within a given half-pixel.

In 3D, we consider the segments of intersection of two iso-surfaces and calculate their lengths, which become the PMF contributions. We refer to this as *LengthProb* [see Figure 10(a)].

Both *2DPointProb* and *LengthProb*, however, require us to ignore those regions in which level sets do not exist because the intensity function is flat, or those regions where level sets from the two images are parallel. The case of flat regions in one or both images can be fixed to some extent by slight blurring of

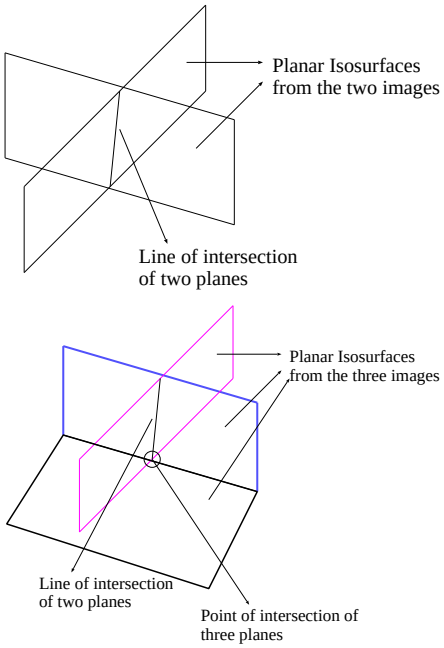


Fig. 10. (a) Segment of intersection of planar isosurfaces from the two images, (b) Point of intersection of planar isosurfaces (each shown in a different color) from the three images

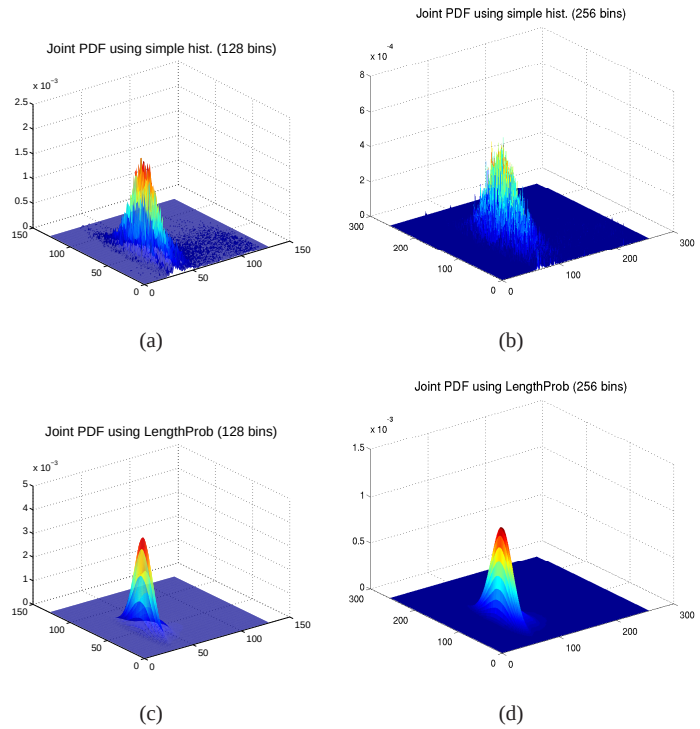


Fig. 11. Joint probability plots using: (a) histograms, 128 bins, (b) histograms, 256 bins, (c) *LengthProb*, 128 bins and (d) *LengthProb*, 256 bins.

Method	2D/3D	Density Contr.	Bias	No. of images
AreaProb	2D	Area	No	Any
VolumeProb	3D	Volume	No	Any
LengthProb	3D	Length	Yes	2 only
2DPointProb	2D	Point count	Yes	2 only
3DPointProb	3D	Point count	Yes	3 only

TABLE I

COMPARISON BETWEEN DIFFERENT METHODS OF DENSITY ESTIMATION W.R.T. NATURE OF DOMAIN, BIAS, SPEED, AND GEOMETRIC NATURE OF DENSITY CONTRIBUTIONS.

the image. The case of aligned gradients is trickier, especially if the two images are in complete registration. However, in the multi-modality case or if the images are noisy/blurred, perfect registration is a rare occurrence, and hence perfect alignment of level surfaces will rarely occur.

To summarize, in both these techniques, *location is treated as a random variable with a distribution that is not uniform, but instead peaked at (biased towards) locations where specific features of the image itself (such as gradients) have large magnitudes or where gradient vectors from the two images are closer towards being perpendicular than parallel.* Such a bias towards high gradients is principled, as these are the more salient regions of the two images. Empirically, we have observed that both these density estimators work quite well on affine registration, and that *LengthProb* is more than 10 times faster than *VolumeProb*. This is because the computation of segments of intersection of planar isosurfaces is much faster than computing polyhedron intersections. Joint PMF plots for histograms and *LengthProb* for 128 bins and 256 bins are shown in Figure 11.

There exists one more major difference between *AreaProb* and *VolumeProb* on one hand, and *LengthProb* or *2DPointProb* on the other. The former two can be easily extended to compute joint density between multiple images (needed for co-registration of multiple images using measures such as modified mutual information (MMI) [2]). All that is required is the intersection of multiple convex polyhedra in 3D or multiple convex polygons in 2D (see Section II-D). However, *2DPointProb* is strictly applicable to the case of the joint PMF between exactly two images in 2D, as the problem of intersection of three or more level curves at *specific* (discrete) intensity levels is over-constrained. In 3D, *LengthProb* also deals with strictly two images only, but one can extend the *LengthProb* scheme to also compute the joint PMF between exactly three images. This can be done by making use of the fact that three planar iso-surfaces intersect in a point (excepting degenerate cases) [see Figure 10(b)]. The joint PMFs between the three images are then computed by counting point intersections. We shall name this method as *3DPointProb*. The differences between all the aforementioned methods: *AreaProb*, *2DPointProb*, *VolumeProb*, *LengthProb* and *3DPointProb* are summarized in Table I for quick reference. It should be noted that *2DPointProb*, *LengthProb* and *3DPointProb* compute PMFs, whereas *AreaProb* and *VolumeProb* compute cumulative measures over finite intervals.

H. Image Entropy and Mutual Information

We are ultimately interested in using the estimated values of $p(\alpha_1, \alpha_2)$ to calculate (Shannon) joint entropy and MI. A major concern is that, in the limit as the bin-width $h \to 0$, the Shannon entropy does *not* approach the continuous entropy, but becomes unbounded [6]. There are two ways to deal with this. Firstly, a normalized version of the joint entropy (NJE) obtained by dividing the Shannon joint entropy (JE) by $\log P$ (where P is the number of bins), could be employed instead of the Shannon joint entropy. As $h \to 0$ and the Shannon entropy

tends toward $+\infty$, NJE would still remain stable, owing to the division by $\log P$, which would also tend toward $+\infty$ (in fact, NJE will have a maximal upper bound of $\frac{\log P^2}{\log P} = 2$, for a uniform joint distribution). Alternatively (and this is the more principled strategy), we observe that unlike the case with Shannon entropy, the continuous MI is indeed the limit of the discrete MI as $h \to 0$ (see [6] for the proof). Now, as P increases, we effectively obtain an increasingly better approximation to the continuous mutual information.

In the multiple image case ($d > 2$), we avoid using a pair-wise sum of MI values between different image pairs, because such a sum ignores the simultaneous joint overlap between multiple images. Instead, we can employ measures such as modified mutual information (MMI) [2], which is defined as the KL divergence between the d-way joint distribution and the product of the marginal distributions, or its normalized version (MNMI) obtained by dividing MMI by the joint entropy. The expressions for MI between two images and MMI for three images are given below:

$$MI(I_1, I_2) = H_1(I_1) + H_2(I_2) - H_{12}(I_1, I_2) \tag{17}$$

which can be explicitly written as

$$MI(I_1, I_2) = \sum_{j_1} \sum_{j_2} p(j_1, j_2) \log \frac{p(j_1, j_2)}{p(j_1) p(j_2)} \tag{18}$$

where the summation indices j_1 and j_2 range over the sets of possibilities of I_1 and I_2 respectively. For three images,

$$MMI(I_1, I_2, I_3) = H_1(I_1) + H_2(I_2) + H_3(I_3) - H_{123}(I_1, I_2, I_3) \tag{19}$$

which has the explicit form

$$MMI(I_1, I_2, I_3) = \sum_{j_1} \sum_{j_2} \sum_{j_3} p(j_1, j_2, j_3) \log \frac{p(j_1, j_2, j_3)}{p(j_1) p(j_2) p(j_3)} \tag{20}$$

where the summation indices j_1, j_2 and j_3 range over the sets of possibilities of I_1, I_2 and I_3 respectively. Though NMI (normalized mutual information) and MNMI are not compatible in the discrete and continuous formulations (unlike MI and MMI), in our experiments, we ignored this fact as we chose very specific intensity levels.

III. Experimental Results

In this section, we describe our experimental results on estimation of PDFs and a comparison between our area-based method for 2D images, versus standard histogramming with sub-pixel sampling. Further, we present results for (a) the case of registration of two images in 2D, (b) the case of registration of multiple images in 2D and (c) the case of registration of two images in 3D.

A. Area-based PDFs versus histograms with several sub-pixel samples

The accuracy of the histogram estimate will no doubt approach the true PDF as the number of samples N_s (drawn from sub-pixel locations) tends to infinity. However, we wish to point out that our method implicitly and efficiently considers every point as a sample, thereby constructing the PDF directly, i.e. the accuracy of what we calculate with the area-based method will always be an upper bound on the accuracy yielded by *any* sample-based approach, *under the assumption that the true interpolant is known to us*. We show here an anecdotal example for the same, in which the number of histogram samples N_s is varied from 5000 to 2×10^9. The L_1 and L_2 norms of the difference between the joint PDF of two 90 x 109 images (down-sampled MR-T1 and MR-T2 slices obtained from Brainweb [4]) as computed by our method and that obtained by the histogram method, as well as the Jensen-Shannon divergence (JSD) between the two joint PDFs, are plotted in the figures below versus $\log N_s$ (see Figure 12). The number of bins used was 128×128 (i.e. $h = 128$). Visually, it was observed that the joint density surfaces begin to appear ever more similar as N_s increases. The timing values for the joint PDF computation are shown in Table II, clearly showing the greater efficiency of our method.

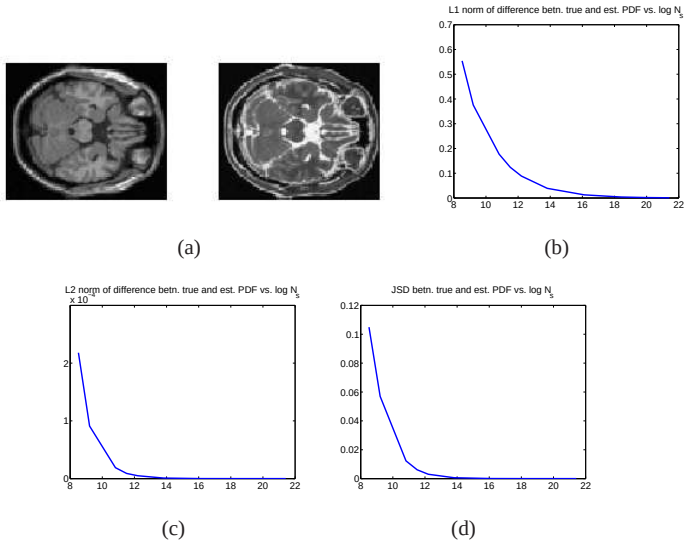


Fig. 12. Plots of the difference between the true joint PDF as computed by the area-based method and the PDF computed by histogramming with N_s sub-pixel samples versus $\log N_s$ using (b) L_1 norm, (c) L_2 norm, and (d) JSD. The relevant images are in sub-figure (a).

Method	Time (secs.)	Diff. with isocontour PDF
Isocontours	5.1	0
Hist. 10^6 samples	1	0.0393
Hist. 10^7 samples	11	0.01265
Hist. 10^8 samples	106	0.0039
Hist. 5×10^8 samples	450	0.00176
Hist. 2×10^9 samples	1927	8.58×10^{-4}

TABLE II

Timing values for computation of joint PDFs and L_1 norm of difference between PDF computed by sampling with that computed using isocontours. Number of bins is 128×128, size of images 122×146.

B. Registration of two images in 2D

For this case, we took pre-registered MR-T1 and MR-T2 slices from Brainweb [4], down-sampled to size 122×146 (see Figure 12) and created a 20° rotated version of the MR-T2 slice. To this rotated version, zero-mean Gaussian noise of different variances was added using the *imnoise* function of MATLAB®. The chosen variances were 0.01, 0.05, 0.1, 0.2, 0.5, 1 and 2. All

these variances are chosen for an intensity range between 0 and 1. To create the probability distributions, we chose bin counts of 16, 32, 64 and 128. For each combination of bin-count and noise, a brute-force search was performed so as to optimally align the synthetically rotated noisy image with the original one, as determined by finding the maximum of MI or NMI between the two images. Six different techniques were used for MI estimation: (1) simple histograms with bilinear interpolation for image warping (referred to as "Simple Hist"), (2) our proposed method using isocontours (referred to as "Isocontours"), (3) histogramming with partial volume interpolation (referred to as "PVI") (4) histogramming with cubic spline interpolation (referred to as "Cubic"), (5) the method *2DPointProb* proposed in [23], and (6) simple histogramming with 10^6 samples taken from sub-pixel locations uniformly randomly followed by usual binning (referred to as "Hist Samples"). These experiments were repeated for 30 noise trials at each noise standard deviation. For each method, the mean and the variance of the error (absolute difference between the predicted alignment and the ground truth alignment) was measured (Figure 13). The same experiments were also performed using a Parzen-window based density estimator using a Gaussian kernel and $\sigma = 5$ (referred to as "Parzen") over 30 trials. In each trial, 10,000 samples were chosen. Out of these, 5000 were chosen as centers for the Gaussian kernel and the rest were used for the sake of entropy computation. The error mean and variance was recorded (see Table III).

Noise Variance	Avg. Error	Std. Dev. of Error
0.05	0.0667	0.44
0.2	0.33	0.8
1	3.6	3
2	4.7	12.51

TABLE III

AVERAGE AND STD. DEV. OF ERROR IN DEGREES (ABSOLUTE DIFFERENCE BETWEEN TRUE AND ESTIMATED ANGLE OF ROTATION) FOR MI USING PARZEN WINDOWS. THE ISOCONTOUR METHOD CONSISTENTLY GAVE BETTER RESULTS THAN PARZEN WINDOWING UNDER HIGHER NOISE.

The adjoining error plots (Figure 13) show results for all these methods for all bins counts, for noise levels of 0.05, 0.2 and 1. The accompanying trajectories (for all methods except histogramming with multiple sub-pixel samples) with MI for bin-counts of 32 and 128 and noise level 0.05, 0.2 and 1.00 are shown as well, for sake of comparison, for one arbitrarily chosen noise trial (Figure 14). From these figures, one can appreciate the superior resistance to noise shown by both our methods, even at very high noise levels, as evidenced both by the shape of the MI and NMI trajectories, as well as the height of the peaks in these trajectories. Amongst the other methods, we noticed that PVI is more stable than simple histogramming with either bilinear or cubic-spline based image warping. In general, the other methods perform better when the number of histogram bins is small, but even there our method yields a smoother MI curve. However, as expected, noise does significantly lower the peak in the MI as well as NMI trajectories in the case of all methods including ours, due to the increase in joint entropy. Though histogramming with 10^6 sub-pixel samples performs well (as seen in Figure 13), our method efficiently and directly (rather than asymptotically) approaches the true PDF and hence the true MI value, under the assumption that we have access to the true interpolant. Parzen windows with the chosen σ value of 5 gave good performance, comparable to our technique, but we wish to re-emphasize that the choice of the parameter was arbitrary and the computation time was much more for Parzen windows.

All the aforementioned techniques were also tested on affine image registration (except for histogramming with multiple sub-pixel samples and Parzen windowing, which were found to be too slow). For the same image as in the previous experiment, an affine-warped version was created using the parameters $\theta = 30^\circ$ = 30, t = -0.3, s = -0.3 and ϕ = 0. During our experiments, we performed a brute force search on the three-dimensional parameter space so as to find the transformation that optimally aligned the second image with the first one. The exact parametrization for the affine transformation is given in [34]. Results were collected for a total of 20 noise trials and the average predicted parameters were recorded as well as the variance of the predictions. For a low noise level of 0.01 or 0.05, we observed that all methods performed well for a quantization up to 64 bins. With 128 bins, all methods except the two we have proposed broke down, i.e. yielded a false optimum of θ around 38°, and s and t around 0.4. For higher noise levels, all methods except ours broke down at a quantization of just 64 bins. The *2DPointProb* technique retained its robustness until a noise level of 1, whereas the area-based technique still produced an optimum of $\theta = 28^\circ$, s = -0.3, t = -0.4 (which is very close to the ideal value). The area-based technique broke down only at an incredibly high noise level of 1.5 or 2. The average and standard deviation of the *estimate* of the parameters θ, s and t, for 32 and 64 bins, for all five methods and for noise levels 0.2 and 1.00 are presented in Tables IV and V. We also performed two-sided Kolmogorov-Smirnov tests [11] for statistical significance on the absolute error between the true and estimated affine transformation parameters for 64 bins and a noise of variance 1. We found that the difference in the error values for MI, as computed using standard histogramming and our isocontour technique, was statistically significant, as ascertained at a level of 0.01.

We also performed experiments on determining the angle of rotation using larger images with varying levels of noise ($\sigma = 0.05, 0.2, 1$). The same Brainweb images, as mentioned before, were used, except that their original size of 183×219 was retained. For a bin count up to 128, all/most methods performed quite well (using a brute-force search) even under high noise. However with a large bin count (256 bins), the noise resistance of our method stood out. The results of this experiment with different methods and under varying noise are presented in Tables VI, VII and VIII.

C. Registration of multiple images in 2D

The images used were pre-registered MR-PD, MR-T1 and MR-T2 slices (from Brainweb) of sizes 90 x 109. The latter two were rotated by $\theta_1 = 20^\circ$ and by $\theta_2 = 30^\circ$ respectively (see Figure 15). For different noise levels and intensity quantizations, a set of experiments was performed to optimally align the latter two images with the former using modified mutual information (MMI) and its normalized version (MNMI) as criteria. These criteria were calculated using our area-based method as well as simple histogramming with bilinear interpolation. The range of

Method	Bins	θ	s	t
MI Hist	32	30, 0	-0.3, 0	-0.3, 0
NMI Hist	32	30, 0	-0.3, 0	-0.3, 0
MI Iso	32	30, 0	-0.3, 0	-0.3, 0
NMI Iso	32	30, 0	-0.3, 0	-0.3, 0
MI PVI	32	30, 0	-0.3, 0	-0.3, 0
NMI PVI	32	30, 0	-0.3, 0	-0.3, 0
MI Spline	32	30.8,0.2	-0.3, 0	-0.3, 0
NMI Spline	32	30.6,0.7	-0.3, 0	-0.3, 0
MI *2DPt.*	32	30, 0	-0.3, 0	-0.3, 0
NMI *2DPt.*	32	30, 0	-0.3, 0	-0.3, 0
MI Hist	64	29.2,49.7	0.4, 0	0.27, 0.07
NMI Hist	64	28.8,44.9	0.4, 0	0.33, 0.04
MI Iso	64	30, 0	-0.3, 0	-0.3,0
NMI Iso	64	30, 0	-0.3, 0	-0.3, 0
MI PVI	64	30, 0	-0.3, 0	-0.3, 0
NMI PVI	64	30, 0	-0.3, 0	-0.3, 0
MI Spline	64	24,21.5	0.4, 0	0.33, 0.04
NMI Spline	64	24.3,20.9	0.4,0	0.33, 0.04
MI *2DPt.*	64	30, 0	-0.3, 0	-0.3, 0
NMI *2DPt.*	64	30, 0	-0.3, 0	-0.3, 0

TABLE IV

AVERAGE VALUE AND VARIANCE OF PARAMETERS θ, s AND t PREDICTED BY VARIOUS METHODS (32 AND 64 BINS, NOISE $\sigma = 0.2$). GROUND TRUTH: $\theta = 30, s = t = -0.3$.

Method	Bins	θ	s	t
MI Hist	32	33.7, 18.1	0.4, 0	0.13,0.08
NMI Hist	32	34.3, 15.9	0.4, 0	0.13, 0.08
MI Iso	32	30,0.06	-0.3, 0	-0.3, 0
NMI Iso	32	30,0.06	-0.3, 0	-0.3, 0
MI PVI	32	28.1, 36.25	0.26, 0.08	0.19, 0.1
NMI PVI	32	28.1, 36.25	0.3, 0.05	0.21,0.08
MI Spline	32	30.3,49.39	0.4, 0	0.09,0.1
NMI Spline	32	31.2,48.02	0.4, 0	0.05,0.1
MI *2DPt.*	32	30.3,0.22	-0.3, 0	-0.3, 0
NMI *2DPt.*	32	30.3,0.22	-0.3, 0	-0.3, 0
MI Hist	64	27.5, 44.65	0.4, 0	0.25,0.08
NMI Hist	64	27,43.86	0.4, 0	0.246, 0.08
MI Iso	64	30.5, 0.12	-0.27, 0.035	-0.28, 0.02
NMI Iso	64	31.2, 0.1	-0.27, 0.058	-0.28, 0.02
MI PVI	64	26.2,36.96	0.4, 0	0.038,0
NMI PVI	64	26.8,41.8	0.4, 0	0.038,0
MI Spline	64	25.9,40.24	0.4, 0	0.3, 0.06
NMI Spline	64	25.7,26.7	0.4, 0	0.3, 0.06
MI *2DPt.*	64	30.5, 0.25	-0.24, 0.0197	-0.23, 0.01
NMI *2DPt.*	64	30.5, 0.25	-0.26, 0.0077	-0.22, 0.02

TABLE V

AVERAGE VALUE AND VARIANCE OF PARAMETERS θ, s AND t PREDICTED BY VARIOUS METHODS (32 AND 64 BINS, NOISE $\sigma = 1$). GROUND TRUTH: $\theta = 30, s = t = -0.3$.

angles was from 1° to 40° in steps of 1°. The estimated values of θ_1 and θ_2 are presented in Table IX.

D. Registration of volume datasets

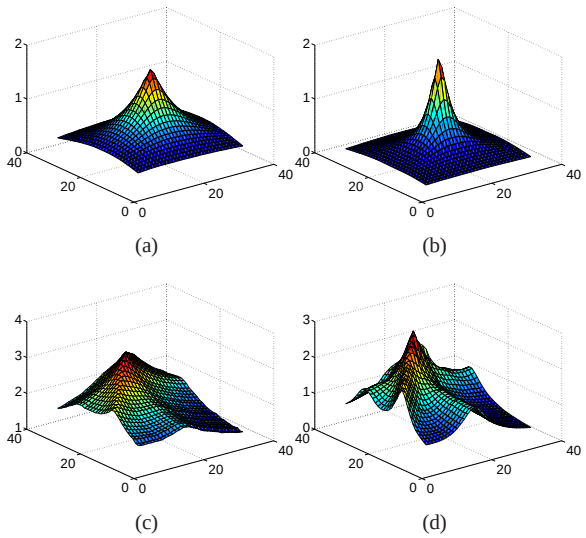


Fig. 16. MI computed using (a) histogramming and (b) *LengthProb* (plotted versus θ_Y and θ_Z); MMI computed using (c) histogramming and (d) *3DPointProb* (plotted versus θ_2 and θ_3).

Experiments were performed on subvolumes of size $41 \times 41 \times 41$ from MR-PD and MR-T2 datasets from the Brainweb simulator [4]. The MR-PD portion was warped by 20° about the Y as well as Z axes. A brute-force search (from 5 to 35° in steps of 1°, with a joint PMF of 64×64 bins) was performed so as to optimally register the MR-T2 volume with the pre-warped MR-PD volume. The PMF was computed both using *LengthProb* as well as using simple histogramming, and used to compute the MI/NMI just as before. The computed values were also plotted against the two angles as indicated in the top row of Figure 16. As the plots indicate, both the techniques yielded the MI peak at the correct point in the θ_Y, θ_Z plane, i.e. at $20°, 20°$. When the same experiments were run using *VolumeProb*, we observed that the joint PMF computation for the same intensity quantization was more than ten times slower. Similar experiments were performed for registration of three volume datasets in 3D, namely $41 \times 41 \times 41$ subvolumes of MR-PD, MR-T1 and MR-T2 datasets from Brainweb. The three datasets were warped through $-2°$, $-21°$ and $-30°$around the X axis. A brute force search was

Method	128 bins	256 bins
MI Hist.	0,0	0.13,0.115
NMI Hist.	0,0	0.067,0.062
MI Iso.	0,0	0,0
NMI Iso.	0,0	0,0
MI PVI	0,0	0,0
NMI PVI	0,0	0,0
MI Spline	0,0	0.33,0.22
NMI Spline	0,0	0.33,0.22
MI *2DPt.*	0,0	0,0
NMI *2DPt.*	0,0	0,0

TABLE VI

AVERAGE ERROR (ABSOLUTE DIFF.) AND VARIANCE IN MEASURING ANGLE OF ROTATION USING MI, NMI CALCULATED WITH DIFFERENT METHODS, NOISE $\sigma = 0.05$.

Method	128 bins	256 bins
MI Hist.	0.07,0.196	0.2,0.293
NMI Hist.	0.07,0.196	0.13,0.25
MI Iso.	0,0	0,0
NMI Iso.	0,0	0,0
MI PVI	0,0	0,0
NMI PVI	0,0	0,0
MI Spline	2.77,10	4.77,10
NMI Spline	2.77,10	18,0.06
MI *2DPt.*	0,0	0,0
NMI *2DPt.*	0,0	0,0

TABLE VII

AVERAGE ERROR (ABSOLUTE DIFF.) AND VARIANCE IN MEASURING ANGLE OF ROTATION USING MI, NMI CALCULATED WITH DIFFERENT METHODS, NOISE $\sigma = 0.2$.

Method	128 bins	256 bins
MI Hist.	1.26,31	27.9,3.1
NMI Hist.	1.2,30	28,3.3
MI Iso.	0,0	0,0
NMI Iso.	0,0	0,0
MI PVI	0,0.26	26.9,14.3
NMI PVI	0,0.26	26.8,14.5
MI Spline	10,0.2	18,0.33
NMI Spline	9.8,0.15	18,0.06
MI *2DPt.*	0.07,0.06	0.07,0.06
NMI *2DPt.*	0.267,0.32	0.07,0.06

TABLE VIII

AVERAGE ERROR (ABSOLUTE DIFF.) AND VARIANCE IN MEASURING ANGLE OF ROTATION USING MI, NMI CALCULATED WITH DIFFERENT METHODS, NOISE $\sigma = 1$.

Noise Variance	Method	32 bins	64 bins
0.05	MMI Hist.	21,30	22,31
0.05	MNMI Hist.	21,30	22,31
0.05	MMI Iso.	20,30	20,30
0.05	MNMI Iso.	20,30	20,30
0.2	MMI Hist.	15,31	40,8
0.2	MNMI Hist.	15,31	40,8
0.2	MMI Iso.	22,29	20,30
0.2	MNMI Iso.	22,29	20,30
1	MMI Hist.	40,9	38,4
1	MNMI Hist.	40,9	34,4
1	MMI Iso.	22,30	35,23
1	MNMI Iso.	22,30	40,3

TABLE IX

THREE IMAGE CASE: ANGLES OF ROTATION USING MMI, MNMI CALCULATED WITH THE ISOCONTOUR METHOD AND SIMPLE HISTOGRAMS, FOR NOISE VARIANCE $\sigma = 0.05, 0.1, 1$ (GROUND TRUTH 20° AND 30°).

Noise Level	Error with *LengthProb*	Error with histograms
0	0.09, 0.02	0.088, 0.009
$\sqrt{50R}$	0.135, 0.029	0.306, 0.08
$\sqrt{100R}$	0.5, 0.36	1.47, 0.646
$\sqrt{150R}$	0.56, 0.402	1.945, 0.56

TABLE X

ERROR (AVERAGE, STD. DEV.) VALIDATED OVER 10 TRIALS WITH *LengthProb* AND HISTOGRAMS FOR 128 BINS. R REFERS TO THE INTENSITY RANGE OF THE IMAGE.

performed so as to optimally register the latter two datasets with the former using MMI as the registration criterion. Joint PMFs of size $64 \times 64 \times 64$ were computed and these were used to compute the MMI between the three images. The MMI peak occurred when the second dataset was warped through $\theta_2 = 19^\circ$ and the third was warped through $\theta_3 = 28^\circ$, which is the correct optimum. The plots of the MI values calculated by simple histogramming and *3DPointProb* versus the two angles are shown in Figure 16 (bottom row) respectively.

The next experiment was designed to check the effect of zero mean Gaussian noise on the accuracy of affine registration of the same datasets used in the first experiment, using histogramming and *LengthProb*. Additive Gaussian noise of variance σ^2 was added to the MR-PD volume. Then, the MR-PD volume was warped by a 4×4 affine transformation matrix (expressed in homogeneous coordinate notation) given as $A = SHR_zR_yR_xT$ where R_z, R_y and R_x represent rotation matrices about the Z, Y and X axes respectively, H is a shear matrix and S represents a diagonal scaling matrix whose diagonal elements are given by 2^{s_x}, 2^{s_y} and 2^{s_z}. (A translation matrix T is included as well. For more information on this parametrization, please see [28].) The MR-T1 volume was then registered with the MR-PD volume using a coordinate descent on all parameters. The actual transformation parameters were chosen to be 7° for all angles of rotation and shearing, and 0.04 for s_x, s_y and s_z. For a smaller number of bins (32), it was observed that both the methods gave good results under low noise and histogramming occasionally performed better. Table X shows the performance of histograms and *LengthProb* for 128 bins, over 10 different noise trials. Summarily, we observed that our method produced superior noise resistance as compared to histogramming when the number of bins was larger. To evaluate the performance on real data, we chose volumes from the Visible Human Dataset[2] (Male). We took subvolumes of MR-PD and MR-T1 volumes of size $101 \times 101 \times 41$ (slices 1110 to 1151). The two volumes were almost in complete registration, so we warped the former using an affine transformation matrix with 5° for all angles of rotation and shearing, and value of 0.04 for s_x, s_y and s_z resulting in a matrix with sum of absolute values 3.6686. A coordinate descent algorithm for 12 parameters was executed on mutual information calculated using *LengthProb* so as to register the MR-T1 dataset with the MR-PD dataset, producing a registration error of 0.319 (see Figure 17).

IV. DISCUSSION

We have presented a new density estimator which is essentially geometric in nature, using continuous image representations and treating the probability density as area sandwiched between

[2]Obtained from the Visible Human Project® (`http://www.nlm.nih.gov/research/visible/getting_data.html`).

isocontours at intensity levels that are infinitesimally apart. We extended the idea to the case of joint density between two images, both in 2D and 3D, as also the case of multiple images in 2D. Empirically, we showed superior noise resistance on registration experiments involving rotations and affine transformations. Furthermore, we also suggested a faster, biased alternative based on counting pixel intersections which performs well, and extended the method to handle volume datasets. The relationship between our techniques and histogramming with multiple sub-pixel samples was also discussed. Here are a few clarifications about our technique in question/answer format.

(1) How does our method compare to histogramming on an up-sampled image?

If an image is up-sampled several times and histogramming is performed on it, there will be more samples for the histogram. At a theoretical level, though, there is still the issue of not being able to relate the number of bins to the available number of samples. Furthermore, it is recommended that the rate of increase in the number of bins be less than the square root of the number of samples for computing the joint density between two images [35], [8]. If there are d images in all, the number of bins ought to be less than $N^{\frac{1}{d}}$, where N is the total number of pixels, or samples to be taken [35], [8]. Consider that this criterion suggested that N samples were enough for a joint density between two images with χ bins. Suppose that we now wished to compute a joint density with χ bins for d images of the same size. This would require the images to be up-sampled by a factor of at least $N^{\frac{d-2}{2}}$, which is exponential in the number of images. Our simple area-based method clearly avoids this problem.

(2) How does one choose the optimal number of bins or the optimal interpolant for our method?

One can choose as many bins as needed for the application, constrained only by availability of appropriate hardware (processor speed or memory). We chose a (piece-wise) linear interpolant for the sake of simplicity, though in principle any other interpolant could be used. It is true that we are making an assumption on the continuity of the intensity function which may be violated in natural images. However, given a good enough resolution of the input image, interpolation across a discontinuity will have a negligible impact on the density as those discontinuities are essentially a measure zero set. One could even incorporate an edge-preserving interpolant [20] by running an anisotropic diffusion to detect the discontinuities and then taking care not to interpolate across the two sides of an edge.

(3) What are the limitations of our technique?

As of now, our technique is not differentiable, which is important for non-rigid registration. Differentiability could be achieved by fitting (say) a spline to the obtained probability tables. However, this again requires smoothing the density estimate in a manner that is not tied to the image geometry. Hence, this goes against the philosophy of our approach. For practical or empirical reasons, however, there is no reason why one should not experiment with this.

(4) Other avenues for future research:

Currently, we do not have a closed form expression for our density estimate. Expressing the marginal and joint densities solely in terms of the parameters of the chosen image interpolant is a challenging theoretical problem. We could also apply our density estimation scheme to images whose pixel values belong to a non-Euclidean manifold, such as unit vectors or covariance matrices, or to data fields that are defined on non-Euclidean surfaces.

APPENDIX

In this section, we derive the expression for the marginal density of the intensity of a single 2D image. We begin with Eq. (4) derived in Section II-A:

$$p(\alpha) = \frac{1}{A} \int_{I(x,y)=\alpha} \left| \begin{array}{cc} \frac{\partial x}{\partial I} & \frac{\partial y}{\partial I} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \end{array} \right| du. \tag{21}$$

Consider the following two expressions that appear while performing a change of variables and applying the chain rule:

$$\left[\begin{array}{cc} dx & dy \end{array} \right] = \left[\begin{array}{cc} dI & du \end{array} \right] \left[\begin{array}{cc} \frac{\partial x}{\partial I} & \frac{\partial y}{\partial I} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \end{array} \right]. \tag{22}$$

$$\left[\begin{array}{cc} dI & du \end{array} \right] = \left[\begin{array}{cc} dx & dy \end{array} \right] \left[\begin{array}{cc} \frac{\partial I}{\partial x} & \frac{\partial u}{\partial x} \\ \frac{\partial I}{\partial y} & \frac{\partial u}{\partial y} \end{array} \right] = \left[\begin{array}{cc} dx & dy \end{array} \right] \left[\begin{array}{cc} I_x & u_x \\ I_y & u_y \end{array} \right]. \tag{23}$$

Taking the inverse in the latter, we have

$$\left[\begin{array}{cc} dx & dy \end{array} \right] = \frac{1}{I_x u_y - I_y u_x} \left[\begin{array}{cc} u_y & -u_x \\ -I_y & I_x \end{array} \right] \left[\begin{array}{cc} dI & du \end{array} \right]. \tag{24}$$

Comparing the individual matrix coefficients, we obtain

$$\left| \begin{array}{cc} \frac{\partial x}{\partial I} & \frac{\partial y}{\partial I} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \end{array} \right| = \frac{I_x u_y - u_x I_y}{(I_x u_y - I_y u_x)^2} = \frac{1}{I_x u_y - I_y u_x}. \tag{25}$$

Now, clearly the unit vector $\vec{u}$ is perpendicular to $\vec{I}$, i.e. we have the following:

$$u_y = \frac{I_x}{\sqrt{I_x^2 + I_y^2}}, \text{ and} \tag{26}$$

$$u_x = \frac{-I_y}{\sqrt{I_x^2 + I_y^2}}. \tag{27}$$

This finally gives us

$$\left| \begin{array}{cc} \frac{\partial x}{\partial I} & \frac{\partial y}{\partial I} \\ \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \end{array} \right| = \frac{1}{\sqrt{I_x^2 + I_y^2}}. \tag{28}$$

Hence the expression in Eq. (5) for the marginal density (i.e. $p(\alpha) = \frac{1}{A} \int_{I(x,y)=\alpha} \frac{du}{\sqrt{I_x^2 + I_y^2}}$) follows.

ACKNOWLEDGMENTS

This work is partially supported by NSF IIS-0307712 and NIH R01NS046812. We would like to thank the anonymous reviewers for thoughtful comments on this paper. The 2D isocontour-based density estimator reported in this paper is available under the terms of the GNU General Public License (GPL version 2) at `http://www.cise.ufl.edu/~anand/isomatch.html`.

REFERENCES

[1] J. Beirlant, E. Dudewicz, L. Györfi, and E. C. van der Meulen. Nonparametric entropy estimation: An overview. *Intl. J. of Mathematical and Statistical Sci.*, 6(1):17–39, June 1997.

[2] J. Boes and C. Meyer. Multi-variate mutual information for registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1679 of *LNCS*, pages 606–612. Springer, 1999.

[3] H. Chen, M. Arora, and P. Varshney. Mutual information-based image registration for remote sensing data. *Journal of Remote Sensing*, 24(18):3701–3706, 2003.

[4] D. L. Collins et al. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3):463–468, 1998.

[5] J. Costa and A. Hero. Entropic graphs for manifold learning. In *IEEE Asilomar Conf. on Signals, Systems and Computers*, volume 1, pages 316–320, 2003.

[6] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, USA, 1991.

[7] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer Verlag, 1997.

[8] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Berlin: Springer Verlag, 1996.

[9] T. Downie and B. Silverman. A wavelet mixture approach to the estimation of image deformation functions. *Sankhya Series B*, 63:181–198, 2001.

[10] N. Dowson, R. Bowden, and T. Kadir. Image template matching using mutual information and NP-Windows. In *Intl. Conf. on Pattern Recognition (ICPR)*, volume 2, pages 1186–1191, 2006.

[11] W. Feller. On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, 19(2):177–189, 1948.

[12] E. Hadjidemetriou, M. Grossberg, and S. Nayar. Histogram preserving image transformations. *Intl. Journal of Computer Vision (IJCV)*, 45(1):5–23, 2001.

[13] T. Kadir and M. Brady. Estimating statistics in arbitrary regions of interest. In *British Machine Vision Conference (BMVC)*, pages 589–598, 2005.

[14] B. Karaçali. Information theoretic deformable registration using local image information. *Intl. Journal of Computer Vision (IJCV)*, 72(3):219–237, 2007.

[15] M. Leventon and W. E. L. Grimson. Multi-modal volume registration using joint intensity distributions. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1496 of *LNCS*, pages 1057–1066. Springer, 1998.

[16] B. Ma, A. Hero, J. Gorman, and O. Michel. Image registration with minimum spanning tree algorithm. In *IEEE Intl. Conf. Image Processing (ICIP)*, volume 1, pages 481–484, 2000.

[17] 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.

[18] F. Maes, D. Vandermeulen, and P. Suetens. Medical image registration using mutual information. *Proceedings of the IEEE*, 91(10):1699–1722, 2003.

[19] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[20] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[21] A. Peter and A. Rangarajan. Maximum likelihood wavelet density estimation with applications to image and shape matching. *IEEE Transactions on Image Processing*, 17(4):458–468, April 2008.

[22] 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.

[23] A. Rajwade, A. Banerjee, and A. Rangarajan. Continuous image repesentations avoid the histogram binning problem in mutual information based image registration. In *IEEE Intl. Symp. Biomedical Imaging (ISBI)*, pages 840–843, 2006.

[24] A. Rajwade, A. Banerjee, and A. Rangarajan. New method of probability density estimation with application to mutual information based image registration. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1769–1776, 2006.

[25] M. Rao, Y. Chen, B. Vemuri, and F. Wang. Cumulative residual entropy: A new measure of information. *IEEE Transactions on Information Theory*, 50(6):1220–1228, 2004.

[26] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[27] M. Sabuncu and P. Ramadge. Gradient based optimization of an EMST image registration function. In *IEEE Intl. Conf. Acoust., Speech, Sig. Proc. (ICASSP)*, volume 2, pages 253–256, 2005.

[28] R. Shekhar and V. Zagrodsky. Mutual information-based rigid and nonrigid registration of ultrasound volumes. *IEEE Transactions on Medical Imaging*, 21(1):9–22, 2002.

[29] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[30] P. Thévenaz and M. Unser. Optimization of mutual information for multiresolution image registration. *IEEE Transactions on Image Processing*, 9(12):2083–2099, 2000.

[31] Tina Is No Acronym (TINA) image database. Available from `http://www.tina-vision.net/ilib.php`. University of Manchester and University of Sheffield, UK.

[32] P. Viola and W. M. Wells. Alignment by maximization of mutual information. *Intl. Journal of Computer Vision (IJCV)*, 24(2):137–154, 1997.

[33] C. Yang, R. Duraiswami, N. Gumerov, and L. Davis. Improved Fast Gauss Transform and efficient kernel density estimation. In *IEEE Intl. Conference on Computer Vision (ICCV)*, volume 1, pages 464–471, 2003.

[34] J. Zhang and A. Rangarajan. Affine image registration using a new information metric. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 848–855, 2004.

[35] J. Zhang and A. Rangarajan. Multimodality image registration using an extensible information metric. In *Information Processing in Medical Imaging (IPMI)*, volume 3565 of *LNCS*, pages 725–737. Springer, 2005.

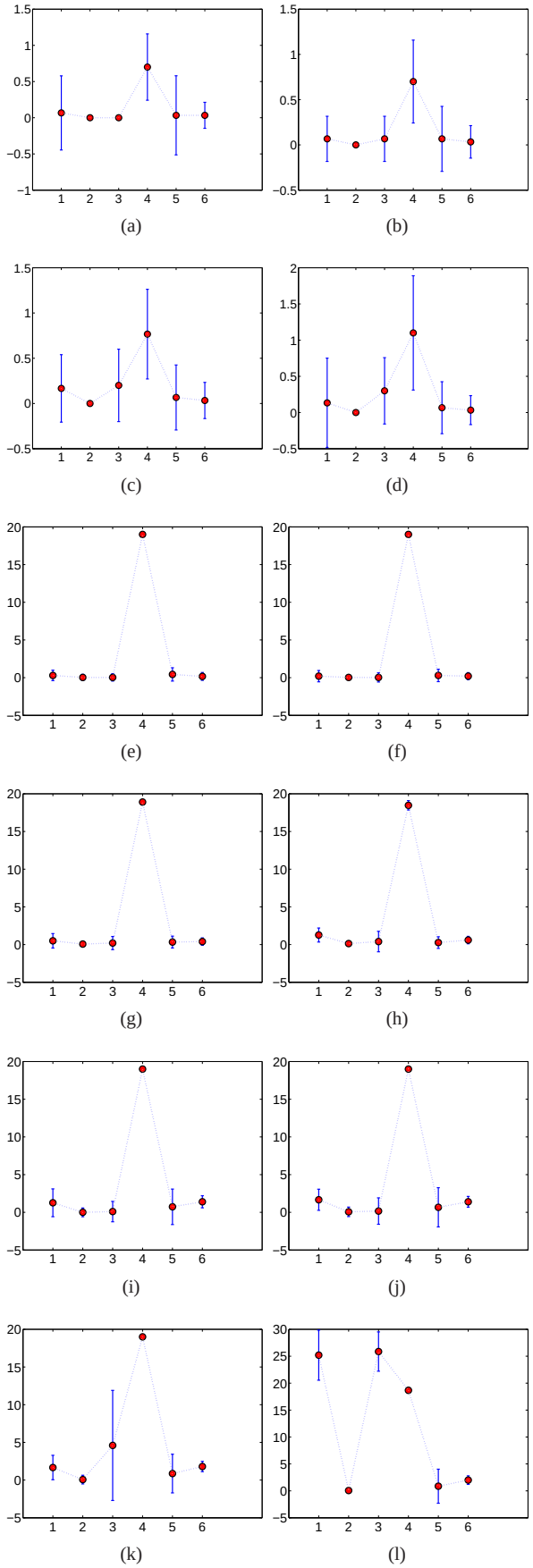


Fig. 13. Graphs showing the average error A (i.e. abs. diff. between the estimated and the true angle of rotation) and error standard deviation S with MI as the criterion for 16, 32, 64, 128 bins (row-wise) with a noise of 0.05 [Top Two Rows: (a) to (d)], with a noise of 0.2 [Middle Two Rows: (e) to (h)] and with a noise of 1 [Bottom Two Rows: (i) to (l)]. Inside each sub-figure, errorbars are plotted for six diff. methods, in the foll. order: Simple Histogramming, Isocontours, PVI, Cubic, *2DPointProb*, Histogramming with 10^6 samples. Errorbars show the values of $A-S$, A, $A+S$. If S is small, only the value of A is shown.

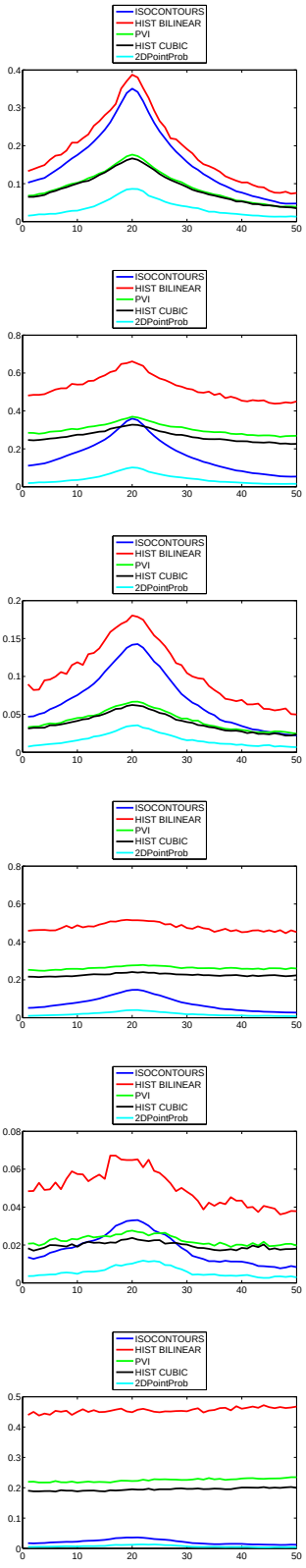
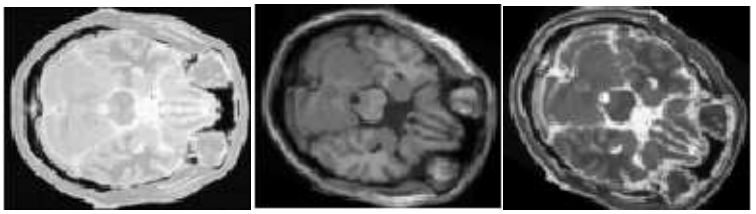


Fig. 14. First two: MI for 32, 128 bins with noise level of 0.05; Third and fourth: with a noise level of 0.2; Fifth and sixth: with a noise level of 1.0. In all plots, dark blue: isocontours, cyan: *2DPointProb*, black: cubic, red: simple histogramming, green: PVI. *(Note: These plots should be viewed in color.)*



(a) (b) (c)

Fig. 15. (a) MR-PD slice, (b) MR-T1 slice rotated by 20 degrees, (c) MR-T2 slice rotated by 30 degrees.

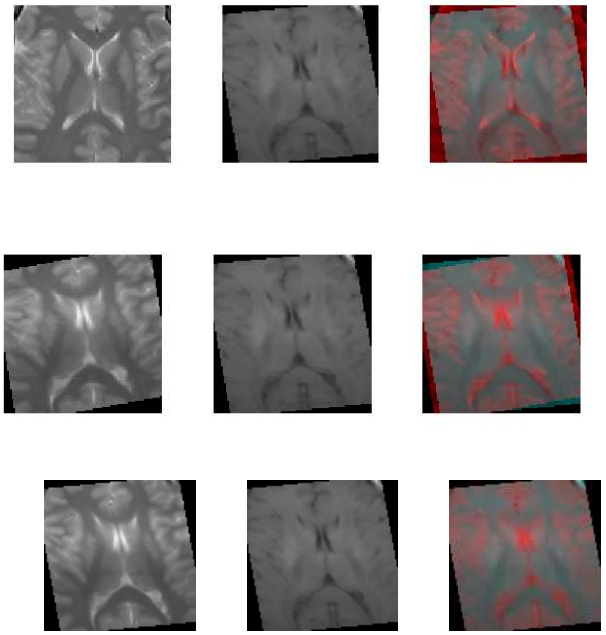


Fig. 17. TOP ROW: original PD image (left), warped T1 image (middle), image overlap before registration (right), MIDDLE ROW: PD image warped using predicted matrix (left), warped T1 image (middle), image overlap after registration (right). BOTTOM ROW: PD image warped using ideal matrix (left), warped T1 image (middle), image overlap after registration in the ideal case (right).