A New Method of Probability Density Estimation with Application to Mutual Information Based Image Registration

Ajit Rajwade, Arunava Banerjee and Anand Rangarajan,
Department of CISE,
University of Florida, USA.
{avr,arunava,anand}@cise.ufl.edu*

Abstract

We present a new, robust and computationally efficient method for estimating the probability density of the intensity values in an image. Our approach makes use of a continuous representation of the image and develops a relation between probability density at a particular intensity value and image gradients along the level sets at that value. Unlike traditional sample-based methods such as histograms, minimum spanning trees (MSTs), Parzen windows or mixture models, our technique expressly accounts for the relative ordering of the intensity values at different image locations and exploits the geometry of the image surface. Moreover, our method avoids the histogram binning problem and requires no critical parameter tuning. We extend the method to compute the joint density between two or more images. We apply our density estimation technique to the task of affine registration of 2D images using mutual information and show good results under high noise.

1. Introduction

Ever since the pioneering work of Viola and Wells in [9] and Maes *et al* in [7], mutual information based image registration techniques have gained popularity, particularly in the field of medical imaging. The process of estimating the probability density function (both marginal and joint) of the intensity values in the images to be registered, lies at the core of all MI-based techniques. Current density estimation techniques include histogramming, Parzen windows and Gaussian mixture models (GMMs). Despite their simplicity and popularity, histogram-based methods suffer from the binning problem, due to the absence of a principled method to estimate the "optimal" number of bins in the marginal and joint histograms, or to relate the number of bins to a particular image *size*. A smaller than optimal number of bins is known to yield an over-smoothed density estimate, whereas an excess in the same number produces an estimate that is highly sparse and prone to noise. Furthermore, histograms are not differentiable. Continuous histograms (obtained by fitting say, a spline between the values in the chosen bins) do bring in differentiability, but they do not overcome the binning problem, as the shape of the final density will vary depending on how many bins were chosen to start with. Parzen windowing [9] does not suffer from the binning problem, but it requires careful selection of the σ parameter of the kernel, as well as the kernel function itself. The σ parameter can be estimated by maximum likelihood methods, but this process is computationally demanding, especially because the value of the parameter changes across the iterations of the registration process. Furthermore, there is the problem of maintaining consistency between the σ values of the marginal and joint densities. Lastly, the process of calculation of density estimates at M points, using Gaussians centered at M other points, has a complexity of $O(M^2)$, which is inefficient for large M. Methods such as the Fast Gauss Transform (FGT) [10] produce an approximation which can be computed in linear time, but they require a prior clustering step. Also, one needs to take into account the growth of the approximation error across the iterations of the registration process.

GMMs have been used for MI based registration in [4]. They also do avoid the binning problem, but they are a computationally inefficient density estimator. They require the estimation of a large number of parameters (the means and covariance matrices of the component Gaussians and their relative weights) and the optimization is highly prone to local minima. Also, one is confronted with the issue of choosing an "optimal" number of mixture components. This number, again, may change across the different iterations of the image registration process (for the joint density).

Another popular method in MI-based image registration is to estimate entropy directly, by-passing the actual density estimation process, as has been done by Ma et al. [6].

*We acknowledge support from NSF 0307712 and NIH 2 R01 NS046812-04A2.

They create a minimal spanning tree to estimate the joint and marginal entropies of a set of samples drawn from a pair of images. However, the entropy involved here is the Renyi entropy (as against the Shannon entropy which was used in [9] and [7]). The construction of the MST itself has a time complexity of $O(E \log E)$ where E is the number of edges in the fully connected graph in which each image pixel is a vertex. This renders the method computationally expensive as the MST has to be created at every step of the registration process. Therefore, one needs to resort to some form of thresholding in order to reduce the complexity of the graph. Alternative entropic graphs such as k-nearest neighbor graphs [1] could also be used to compute Renyi entropy, but they have a quadratic time complexity in the number of nodes.

A point to be noted is that none of the above-mentioned techniques explicitly take into account the geometry of the "surface" of the image, and thus ignore the relative ordering of the intensity values at different image locations. In the work presented here, we drop the notion of an image as a discrete set of pixels, and treat it as a continuous entity. We then proceed to relate image gradients to probability density. For a single image, we see that the cumulative distribution function at a particular intensity value α is equal to the ratio of the total area of all regions in the image whose intensity is less than or equal to α, to the total area of the image. (Note that the boundary of such regions could be level curves of the image or the boundary of the image itself). The derivative of this ratio w.r.t. the intensity change yields the probability density at that intensity value. We also present a method to ascertain the joint density of a pair of images, by looking at the area of intersection of a pair of level curves (at nearby intensity levels with infinitesimally small intensity difference) in the first image, with a similar pair from the second. Next, we also determine the probability distributions by considering successive level curves with a non-zero intensity difference. With this theoretical development, we estimate image entropy and mutual information, and use these calculated values for the task of registering 2D images. We empirically show the robustness of our technique and the smoothness of the information-theoretic optimization functions w.r.t. the transformation. A point to underline is that our technique requires *no* setting of critical parameters, and neither does it rely on any form of sampling. To the best of our knowledge, the only work other than ours to adopt such a geometric approach to density estimation is that by Kadir and Brady in [3][1]. However unlike [3], we explicitly take into account the effect of singularities in the density estimate and present a solution (see Sections (2.3) and (2.4)), and also apply the technique to MI-based registration.

The paper is organized as follows. In Section (2), we present the complete theoretical treatment of our method, followed by a discussion of some practical issues. Section (3) presents in detail the experimental results. A few salient features of the technique as well as directions for future work are discussed in Section (4).

[1]This was brought to our notice after the acceptance of this paper.

2. Theory

In this section, we describe our method for estimating the marginal image density, followed by the joint density given a pair of images.

2.1. Estimating the Marginal Densities

Consider the 2D gray-scale image intensity to be a continuous, scalar-valued function of the spatial variables, represented as $z = I(x, y)$. Consider a random experiment whose outcome is a location in the image. Let the probabilty distribution associated with the experiment be uniform on location. This distribution on location induces a corresponding probability distribution on intensity. The cumulative distribution 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 (denoted as $\mathcal{A}$). This can be written as follows:

$$\Pr(z < \alpha) = \frac{1}{\mathcal{A}} \iint_{z<\alpha} dxdy. \tag{1}$$

Now, the probability density at α is the derivative of the cumulative distribution. 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:

$$p(\alpha) = \frac{1}{\mathcal{A}} \lim_{\Delta\alpha \to 0} \frac{\iint_{z<\alpha+\Delta\alpha} dxdy - \iint_{z<\alpha} dxdy}{\Delta\alpha}. \tag{2}$$

Hence, we have

$$p(\alpha) = \frac{1}{\mathcal{A}} \frac{d}{d\alpha} \iint_{z=\alpha} dxdy. \tag{3}$$

We can now adopt a change of variables from the spatial coordinates x and 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 parallel to the image gradient, or the direction of maximum intensity change. Noting this fact, we now obtain the following:

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

Note that in this equation, $d\alpha$ 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 $p(\alpha)$:

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

From the above expression, one can make two important observations. Firstly, each point on a given level curve contributes a certain measure to the density at that intensity. The density contribution for a given intensity value at a location (x, y) 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 to the existence of a peak, a valley or a saddle point), the density at that point tends to infinity. (The practical repercussions of this situation are discussed later on in the paper). Secondly, 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. Such an estimate does not suffer from the binning problem, as here, an explicit relation between intensity and the spatial coordinates has been exploited. Therefore, as h becomes smaller and smaller (i.e. as the number of bins N increases), the density estimate becomes increasingly accurate. *Au contraire*, a naive histogram calculation leaves most of the bins empty. This scenario is clearly illustrated in Figure (2), where we plot histogram envelopes of the face image on the left side of Figure (7), comparing our method to the standard histogram for 32, 64, 128, and 256 bins, given a 100 by 100 image of 256 intensity levels. Yet another fact to note is that we are adopting a continuous representation of the image. This is quite unlike the standard histogram which assumes an image to be discrete and flat within each pixel, and in which each pixel contributes to one and only one bin.

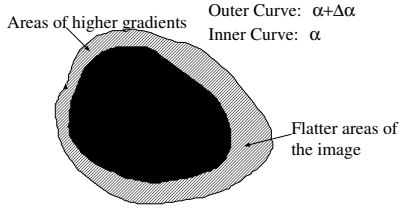
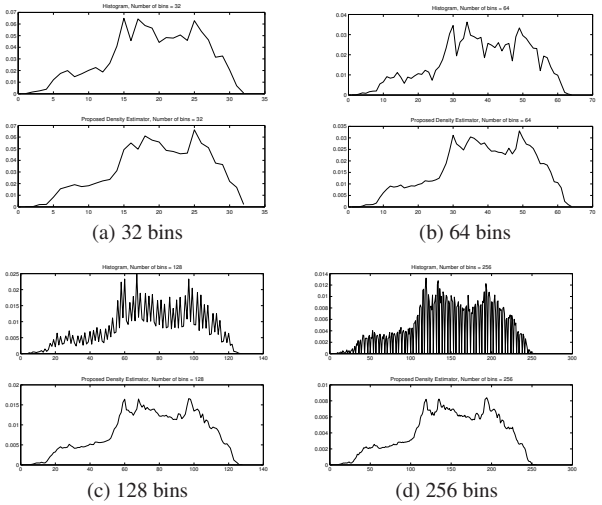


Figure 1. $p(\alpha) \propto$ area of hatched region

2.2. Estimating the Joint Density

Consider two images represented as $z_1 = I_1(x, y)$ and $z_2 = I_2(x, y)$, whose overlap area is $\mathcal{A}$. Their cumulative



(a) 32 bins (b) 64 bins (c) 128 bins (d) 256 bins

Figure 2. Comparison of conventional 1D histograms (TOP half of each sub-figure) to the proposed density estimator (BOTTOM half of each sub-figure)

distribution at intensity values (α_1, α_2) is equal to the ratio of the total area 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, to the total overlap area. The probability density $p(\alpha_1, \alpha_2)$ in this case is the second order derivative of the cumulative distribution. 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 $\mathcal{P}$ and the region enclosed between the level curves of I_2 at α_2 and $\alpha_2 + \Delta\alpha_2$ as $\mathcal{Q}$. Then $p(\alpha_1, \alpha_2)$ can geometrically be interpreted as the area of $\mathcal{P} \bigcap \mathcal{Q}$, divided by $\Delta\alpha_1 \Delta\alpha_2$, in the limit as $\Delta\alpha_1$ and $\Delta\alpha_2$ tend to zero. The regions $\mathcal{P}$, $\mathcal{Q}$ and also $\mathcal{P} \bigcap \mathcal{Q}$ (hatched region) are shown in Figure (3). Using a technique very similar to that shown in equations (2) to (4), we obtain the joint probability density as follows:

$$p(\alpha_1, \alpha_2) = \frac{1}{\mathcal{A}} \frac{\partial^2}{\partial\alpha_1 \partial\alpha_2} \iint_C du_1 du_2 \left| \begin{array}{cc} \frac{\partial x}{\partial u_1} & \frac{\partial y}{\partial u_1} \\ \frac{\partial x}{\partial u_2} & \frac{\partial y}{\partial u_2} \end{array} \right| \quad (6)$$

where C represents the locus of all points where $z_1 = \alpha_1$ and $z_2 = \alpha_2$. Here u_1 and u_2 represent directions along the corresponding level curves of the two images I_1 and I_2. 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 nothing but the area of parallelogram ABCD (see Figure (4)) 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 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 DE is perpendicular to the level curves) we can now deduce that the length CD (or equivalently AB) is given as

$$|AB| = \frac{\Delta\alpha_1}{g_1(\tilde{x}, \tilde{y}) \sin\theta(\tilde{x}, \tilde{y})}. \tag{7}$$

Similarly, the length CB is given by

$$|CB| = \frac{\Delta\alpha_2}{g_2(\tilde{x}, \tilde{y}) \sin\theta(\tilde{x}, \tilde{y})}. \tag{8}$$

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:

$$p(\alpha_1, \alpha_2) = \frac{1}{\mathcal{A}} \iint_C \frac{du_1 du_2}{g_1(x, y)g_2(x, y) \sin\theta(x, y)}. \tag{9}$$

It is easy to show through algebraic manipulations that equations (6) and (9) are equivalent formulations of the joint probability density $p(\alpha_1, \alpha_2)$. Thus, we see that $p(\alpha_1, \alpha_2)$ can be computed by summing up the values of $\frac{1}{g_1(x,y)g_2(x,y) \sin\theta(x,y)}$ (i.e. the density contribution) at all points where the level curve of I_1 at α_1 and that of I_2 at α_2 intersect. These results could also have been derived following an algebraic method, i.e. by manipulation of Jacobians, as was done while deriving the expression for the marginal densities. Furthermore, the derivation for the marginals could also have proceeded following geometric intuitions.

It is worth mentioning that the formula derived above tallies beautifully with intuition in the following ways. Firstly, the area of the parallelogram $ABCD$ (and hence the joint density) 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 $ABCD$ (and hence the joint density) is maximum in the case where the gradients of the two images are parallel or completely align, and the least when they are orthogonal (see Figure (5)). Lastly, the determinant of the Jacobian in equation (6) is equal to the area of the parallelogram in Figure (4), which is equal to the cross product of the vectors AB and CB along the corresponding level curves. This treatment could be easily extended to the case of joint density between $d > 2$ images, by using the concept of the wedge product or using similar geometric intuition to obtain the area between d intersecting pairs of level curves (see Figure (6) for the case of three images). 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.

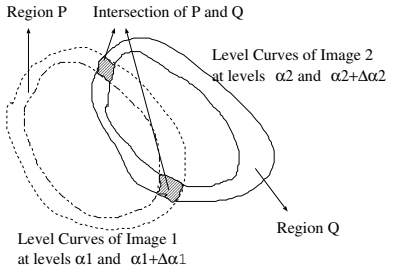


Figure 3. Intersection of level curves of I_1 and I_2: $p(\alpha_1, \alpha_2) \propto$ area of hatched region.

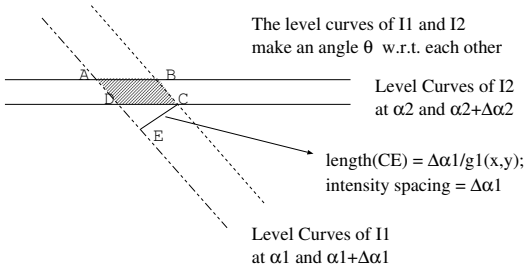


Figure 4. Parallelogram approximation: pdf contribution $\propto$ area($ABCD$)

Figure 5. Area of parallelogram increases as angle between level curves decreases (top to bottom). Level curves of I_1 and I_2 are shown in black and red lines respectively.

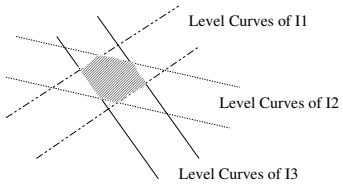


Figure 6. Joint probability contribution in the case of three images.

2.3. Practical Issues

In the two preceding sub-sections, we observed the divergence of the marginal density in regions of zero gradient.

We also noticed the divergent behavior of the joint density in regions where either (or both) image gradient(s) is (are) zero, or when the gradients completely align. A practical solution would be to observe that in such regions, the cumulative distribution is actually non-differentiable, and that it might be beneficial to calculate probability distributions (as opposed to densities) in such areas. 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. In any region that is flat, there exists one and only one intensity value. The contribution to the probability at a particular intensity in a flat region is proportional to the total area of that flat region. More *ad hoc* approaches could involve simply "weeding out" the flat regions altogether. All such methods require the choice of appropriate thresholds to distinguish between flat and non-flat regions. The nature of the density surface is highly dependent on the threshold values chosen. Too high a threshold leads to loss of useful image data (owing to the fact that non-flat regions are implicitly being flattened), whereas too low a threshold allows the flatter regions of the image to completely dominate regions of high gradient in terms of density contribution. This has serious ramifications in the computation of entropy, which is required for image registration. As such there is no principled method to obtain an "optimal" threshold. Also, following this path would cause us to deal with a mixture of densities and distributions.

2.4. Work-around: Probability Distributions

A robust work-around to solve this conundrum, is to 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. Note that, throughout this paper, we denote probability distributions using the operator $\hat{p}$ and densities using the operator p. 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. Also, note that we deal with these issues in detail unlike the work in [3].

The notion of an image as a continuous entity is one of the pillars of our approach. While any continuous image representation would hold good, we adopt a locally linear formulation in this paper, for the sake of simplicity. 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. The values A_1, B_1 and C_1 can be calculated by solving three simultaneous equations. Iso-intensity lines at levels α_1 and $\alpha_1 + h$ within this triangle are represented by the equation $A_1x + B_1y + C_1 = \alpha_1$ and $A_1x + B_1y + C_1 = \alpha_1 + h$. Similar equations exist for the 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 (9(a)). In the case

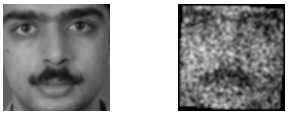
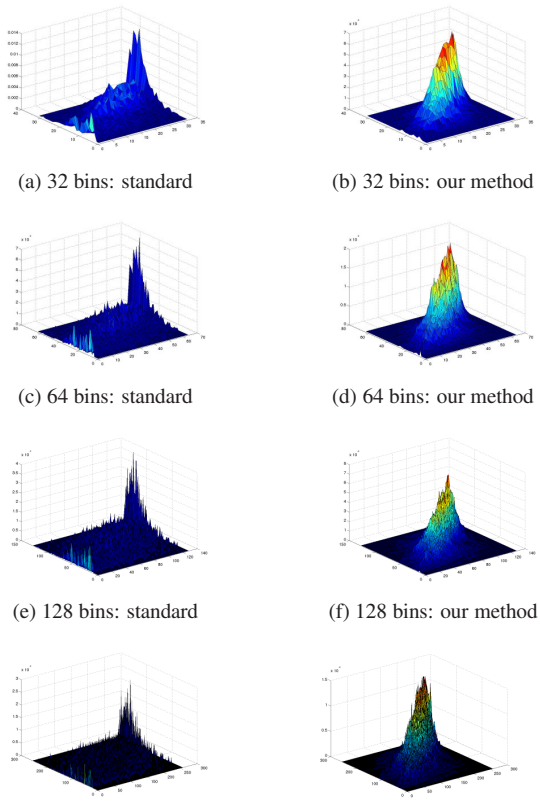


Figure 7. A face image and its noisy, rotated version



(a) 32 bins: standard (b) 32 bins: our method

(c) 64 bins: standard (d) 64 bins: our method

(e) 128 bins: standard (f) 128 bins: our method

(g) 256 bins: standard (h) 256 bins: our method

Figure 8. Comparison of standard joint histograms to our new joint density estimator

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 (9(b)). 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 (9(c))). Observe that we still have to treat flat regions and regions where the gradients from the two images align, in a special manner, even though we have switched to distributions. The basic difference is that now we neither have 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 get more and more accurate as the value of h is reduced. This, again, is in complete contrast to standard histogramming, as has been mentioned before. In Figure (8), we have shown plots of our joint density estimate and compared it to standard histograms for N equal to 32, 64, 128 and 256 bins in *each* image (i.e. 32^2, 64^2 etc. bins in the joint), which illustrate our point clearly. On an average, we saw that the standard histograms had a far greater number of empty bins than our density estimator, for the same number of intensity levels.

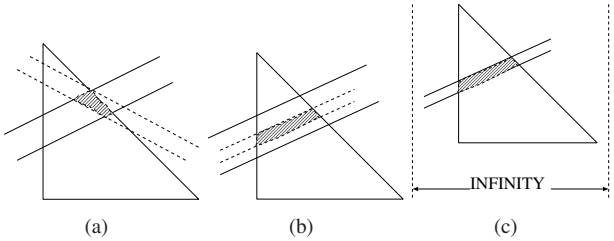


Figure 9. (a) Probability contribution $\propto$ area of parallelogram between level curves clipped against the triangle, i.e. half-pixel. (b) Case of parallel gradients. (c) Case when the gradient of one image is zero (dotted level lines) and that of the other is non-zero (solid level lines). In each case, probability contribution $\propto$ area of the hatched region.

2.5. Image Entropy

We are ultimately interested in using the developed theory to calculate MI which requires us to calculate the (Shannon) joint entropy of the images, which in turn is calculated from the probability distributions $\hat{p}(\alpha_1, \alpha_2)$ as described in the previous section. A major concern would be that, in the limit as $h \to 0$, the Shannon entropy does *not* approach the continuous entropy, but becomes unbounded [2]. There are two ways to deal with this situation. Firstly, a normalized version of the joint entropy (NJE) obtained by dividing the Shannon joint entropy (JE) by $\log N$, could be employed instead of the Shannon joint entropy itself. As $h \to 0$ and the Shannon entropy tends toward $+\infty$, NJE would still remain relatively stable, owing to the division by $\log N$, which would also tend toward $+\infty$. In fact, any pair of images that has a uniform joint probability distribution, as calculated by our method, will have the maximum possible joint entropy value, which is $\log N^2$. From this, it is clear that for such an image pair, NJE will therefore have an upper bound of 2, and that no image pair can have an NJE value greater than 2. Alternatively (and this is the more principled strategy), we observe that unlike the case with Shannon entropy, the continuous mutual information is indeed the limit of the discrete mutual information as $h \to 0$ (see [2] for an elegant proof). With this observation in mind, we need not concern ourselves with the unbounded nature of Shannon entropy in the limit. In fact, as N increases, we effectively obtain an increasingly better approximation to the continuous mutual information.

3. Experiments

We now proceed to explore our algorithm from the point of view of image registration. In our first experiment, we considered the simplest case of a single rotation between a pair of images. The aim was to iteratively rotate one of the images in a brute-force manner so that it was optimally registered with the other based on some criterion. The following four measures were calculated as required criteria: Joint Entropy between the two images, i.e. JE, MI, normalized MI (NMI) and the measure ρ defined in [11]. Of all these four measures, MI is the only measure whose continuous version is the limit of the discrete version. We calculated JE, NMI and ρ only because we know that h here is non-zero. JE (needed while computing all the other three measures) was calculated from the joint probability estimated as described in Section (2). The marginal probabilities were computed by summing up the joint probability matrix row-wise or column-wise (analogous to integration). Not only was this more efficient than following the procedure in Section (2.1), but this also helped ensure a consistency between the joints and the marginals. The marginal probabilities thus estimated were used in the marginal entropy calculation.

The first experiment was performed on the face image shown in Figure (7) and its -15 degree rotated version. A noise of variance 0.1 was added[2] to the latter and it was then blurred slightly. We then sought to register the original

[2]using the "imnoise" function of MATLAB

clean face image with its noisy rotated version by performing a brute force search (between -25 to 0 degrees) for the angle of rotation, so as to minimize JE and ρ, or to maximize MI and NMI. For each angle, the values of the aforementioned four entropic measures were calculated, using the standard histograms as well as our method (both with 128 bins). A trajectory of all these quantities was then plotted to visualize the nature of their variation w.r.t. the rotation (see Figure(11)). This process was repeated exactly as stated with the noise level raised to 0.8 and the same amount of blurring (see Figure (12)). From these graphs, especially Figure (12), one can appreciate the superior noise resistance of our method, due to the smoother trajectories of JE and MI. NMI and ρ had similar trajectories which have been omitted to save space. Smoothness of objective function is of paramount importance if brute force search is to be abandoned for more efficient search mechanisms[3]. Moreover, our method predicted the transformation parameters more accurately as seen in Table (1).

The standard histogram methods will no doubt perform better when the number of bins is reduced. However, we wish to emphasize that there is no way of correctly predicting the number of bins in standard histograms. Also, in practical registration systems, situations could arise where a large number of bins is essential for accurate rigid/affine registration, and even more so in non-rigid settings. One such example is the registration of depth maps to color images of the same object, where very small depth changes do correspond to significant changes in intensity. Also, see Section (4) for further comparison between standard histograms and the proposed strategy and the fundamental point of departure between these two methods.

The second experiment was aimed at demonstrating the use of our method for affine image registration. For this, we chose a synthetically generated MR-PD slice and an MR-T2 slice, both obtained from the BrainWeb simulator [5]. Both slices were initially in complete alignment with one another. The T2 slice was given an affine transformation, with an in-plane rotation of $\theta = -20$ degrees, a scaling in both directions by a factor of $s = -0.3$ and $t = -0.3$ respectively, and a translation in the X and Y directions by $t_x = 2$ and $t_y = 2$ pixels respectively (see [11] for details of the affine matrix). In our experiments, the angle ϕ was set to 0 for simplicity. Following the transformation, the T2 slice was treated with zero mean Gaussian noise of variance 0.1. The images are shown in Figure (10). A multi-resolution brute force search was performed for the optimal parameters, within an angular range of $[-24, -12]$, a translation of $[-3, 3]$ and a scale range of $[-0.5, 0.5]$. The affine transformation was applied to the PD slice so as to optimally

[3]Indeed, our method is not tied to the brute-force search, and could work with any optimization technique. Brute-force search was employed only to do a fair comparison with standard histogram based MI.

Metric	Predicted Angle	
-	Our Method.	Std. Hist.
JE	$-18°$	$-25°$
MI	$-15°$	$-21°$
NMI	$-15°$	$-21°$
ρ	$-15°$	$-25°$

Table 1. Angle predicted using JE, MI, NMI and ρ, ground truth = -15, noise $\sigma = 0.8$

align it with the T2 slice, by seeking the maximum of the MI value. With our method, the estimated parameters (i.e. the maximum of MI) were $\theta = -18$ degrees, a translation of 3 pixels along both X and Y, and a scale of -0.2 along both X and Y. With the standard histograms, the maximum of MI occurred at an angle of $\theta = -12$ degrees, a scale of $s = 0.5$ and $t = 0.4$, and translations of 0 and 3 pixels. For both methods, the number of bins used was 128. Clearly our method outperformed the standard histogram.

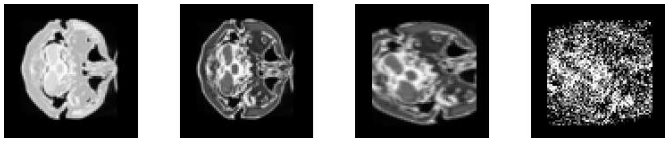
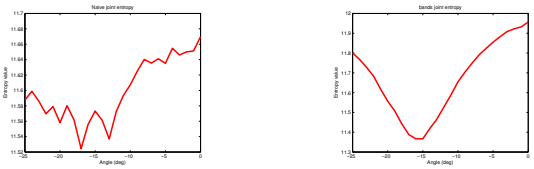


Figure 10. (a) An MR-PD slice (b) An MR-T2 slice (c) MR-T2 slice, synthetically warped (d) Warped MR-T2 slice with noise



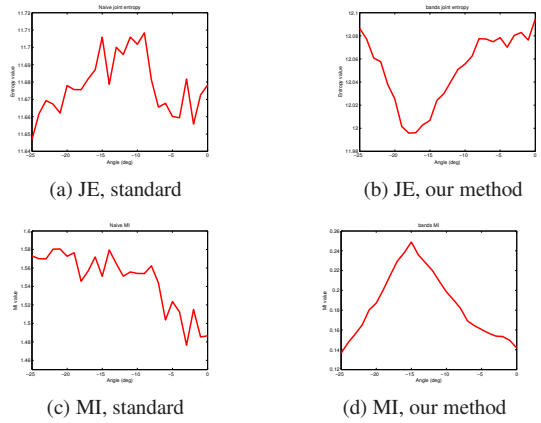
(a) JE, standard (b) JE, our method

(c) MI, standard (d) MI, our method

Figure 11. Comparison of the trajectory of JE and MI w.r.t. rotation, computed with standard histograms as well our method, noise $\sigma = 0.1$

4. Discussion

In this paper, we have presented a new procedure for estimation of the probability density of image intensity, which has its foundations in the notion of images as continuous functions of the spatial coordinates. Our method directly relates probability density to image gradients. The adopted notion enables us to solve the so-called binning problem completely, while calculating both the marginals as well as



(a) JE, standard (b) JE, our method

(c) MI, standard (d) MI, our method

Figure 12. Comparison of the trajectory of JE and MI w.r.t. rotation, computed with standard histograms as well our method, noise $\sigma = 0.8$

the joint probabilities. Our method divides an image into piecewise linear patches (two in each pixel). Each triangle can contribute to the joint probability at N^2 pairs of intensity levels in the worst case. Therefore, the approach has a computational complexity of $O(SN^2)$ for registration of two images with S pixels and N chosen intensity levels per image. As such, it is more efficient than the Parzen window estimator which is quadratic in the number of samples, or the MST method which requires an $O(E \log E)$ creation of the spanning tree, E itself being quadratic in S. The technique we have proposed requires no parameter tuning or choosing of any kernel function centered around randomly drawn (i.i.d.) samples, unlike the existing methods. Rather, *every* point in the image contributes to the density estimate in our technique. Furthermore, our technique innately incorporates spatial information into the density estimate, unlike histograms, Parzen windows or GMMs, all of which are highly global. This essentially means that given a digital image (for which we use a continuous representation), the ordering of the pixel values is exploited in our method and changes in the ordering would affect our density estimate. Other methods ignore such information. A further merit of our method is its superior resistance to noise as compared to standard histograms, as has been demonstrated empirically in the experimental section. The reason for this is that noise causes votes in standard histograms to erratically switch over from one bin to another. In our case, the effects of noise are spread out over several bins without any discontinuous switching.

In this paper, we have preferred to remain within the ambit of Shannon entropy (and related measures) as it is the most widely used entropy formulation. As such, it is trivial to calculate the Renyi entropy from our distribution estimates. It is also trivial to calculate cumulative distributions and use the highly robust cross-cumulative residual entropy (CCRE) [8] for registration. The main difference between this paper and [8] is that the latter presents yet another sampling-based technique that does not exploit the relative positioning of intensities.

Our future work would involve application of the new density estimator to non-rigid image registration. On the theoretical front, we note that our technique in its present form is not differentiable, which is essential for finding the analytic derivatives required for the efficient (and accurate) implementation of gradient-based search methods. Though we do have a continuous formulation already as described in Section (2), further work is required to deal with flat regions and aligned gradients, without sacrificing differentiability (which is the unfortunate consequence of switching to distributions). Furthermore, although our approach finds immediate application in group-wise registration of multiple images, the overall computational cost would be exponential in the number of images. These issues pose interesting challenges for future research.

References

[1] J. Costa and A. Hero. Entropic graphs for manifold learning. In *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 317–320, 2003.

[2] T. Cover and J. Thomas. *Elements of information theory*. Wiley-Interscience, New York, USA, 1991.

[3] T. Kadir and M. Brady. Estimating statistics in arbitrary regions of interest. In *British Machine Vision Conference*, 2005.

[4] M. Leventon and E. Grimson. Multi-modal volume registration using joint intensity distributions. In *MICCAI*, pages 1057–1066, 1998.

[5] D. Louis Collins et al. Design and construction of a realistic digital brain phantom. *IEEE Trans. Med. Imaging*, 17(3):463–468, 1998.

[6] B. Ma, A. Hero, J. Gorman, and O. Michel. Image registration with minimum spanning tree algorithm. In *ICIP*, pages 481–484, 2000.

[7] 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.

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

[9] P. Viola and W. M. Wells. Alignment by maximization of mutual information. *Int. J. Comput. Vision*, 24(2):137–154, 1997.

[10] C. Yang, R. Duraiswami, N. Gumerov, and L. Davis. Improved Fast Gauss Transform and efficient kernel density estimation. In *ICCV*, pages 464–471, 2003.

[11] J. Zhang and A. Rangarajan. Affine image registration using a new information metric. In *CVPR (1)*, pages 848–855, 2004.