# Image Similarity Using Mutual Information of Regions

Daniel B. Russakoff[1,2], Carlo Tomasi[3],
Torsten Rohlfing[2], and Calvin R. Maurer, Jr.[2]

[1] Department of Computer Science, Stanford University, Stanford CA 94305, USA
daniel.russakoff@cs.stanford.edu
[2] Image Guidance Laboratories, Stanford University, Stanford CA 94305, USA
[3] Department of Computer Science, Duke University, Durham NC 27708, USA

**Abstract.** Mutual information (MI) has emerged in recent years as an effective similarity measure for comparing images. One drawback of MI, however, is that it is calculated on a pixel by pixel basis, meaning that it takes into account only the relationships between corresponding individual pixels and not those of each pixel's respective neighborhood. As a result, much of the spatial information inherent in images is not utilized. In this paper, we propose a novel extension to MI called regional mutual information (RMI). This extension efficiently takes neighborhood regions of corresponding pixels into account. We demonstrate the usefulness of RMI by applying it to a real-world problem in the medical domain—intensity-based 2D-3D registration of X-ray projection images (2D) to a CT image (3D). Using a gold-standard spine image data set, we show that RMI is a more robust similarity meaure for image registration than MI.

## 1 Introduction

### 1.1 Mutual Information

The mutual information (MI) between two variables is a concept with roots in information theory and essentially measures the amount of information that one variable contains about another. Put another way, it is the reduction in uncertainty of one variable given that we know the other [1]. MI was introduced as a similarity measure between images (both 2D and 3D) simultaneously by Viola *et al.* [2] and Maes *et al.* [3]. As a similarity measure, it has a number of advantages. In particular, it assumes no prior functional relationship between the images. Rather, it assumes a statistical relationship that can be captured by analyzing the images' joint entropy. Mutual information is closely related to joint entropy. Specifically, given image $A$ and image $B$, the joint entropy $H(A, B)$ can be calculated as:

$$H(A, B) = -\sum_{a,b} p_{AB}(a, b) \log p_{AB}(a, b)$$

where $p_{AB}$ is the joint probability distribution of pixels associated with images $A$ and $B$. The joint entropy is minimized when there is a one-to-one mapping

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3023, pp. 596–607, 2004.

between the pixels in $A$ and their counterparts in $B$ and it increases as the statistical relationship between $A$ and $B$ weakens. Here, since we are dealing with discrete images, we express all entropies with sums instead of integrals. In general, we must divine the probability distribution associated with each image by binning the values into histograms.

Mutual information considers both the joint entropy $H(A, B)$ and the individual entropies $H(A)$ and $H(B)$ where:

$$H(X) = -\sum_{x} p_X(x) \log p_X(x)$$

MI is defined as:

$$H(A) + H(B) - H(A, B)$$

Intuitively, as Viola notes, maximizing the mutual information between two images seems to try and find the most complex overlapping regions (by maximizing the individual entropies) such that they explain each other well (by minimizing the joint entropy) [2]. As a similarity measure, mutual information has enjoyed a great deal of success, particularly in the medical imaging domain [4]. It is robust to outliers, efficient to calculate, and generally provides smooth cost functions on which to optimize.

There is, however, one important drawback to mutual information as a way of comparing images: it fails to take geometry into account since it considers only pixel values, and not pixel positions.

### 1.2 Previous Work

Since its introduction, there have been a number of extensions to mutual information. In [5], an overlap invariant extension called normalized mutual information is introduced. In [6] and [7], different methods for calculating entropy are used. Each of these extensions, however, continues to ignore important spatial information in images.

More recently, researchers have begun to extend mutual information to include spatial information. In particular, in [8], mutual information is multiplied with a term that compares the local gradients of the two images. In [9], "second order" mutual information is defined. This formulation involves calculating the marginal and joint entropies of a pixel and one neighbor instead of a single pixel. Both cases add some amount of spatial information to the existing framework of mutual information and both cases validate their metrics on the problem of medical image registration. Also, both papers report that the final accuracy of their registrations are essentially the same as when using mutual information. The real improvement of incorporating spatial information lies in the robustness of the measure. A robust similarity measure is one with a smooth, convex landscape with respect to misregistration, specifically, one that does not have too many local extrema on the way to a global optimum. Both [8] and [9] report improved robustness of their metrics over standard mutual information. Each,

however, has important limitations. In [9], only one neighbor at a time is considered which leaves out a great deal of spatial information. In [8], they do not actually extend mutual information, rather MI is multiplied by a different term which accounts for the neighborhood information we are after.

In this paper, we present regional mutual information (RMI), an extension to mutual information that incorporates spatial information in a way that leads to smoother, more robust energy functions than have previously been reported. The paper is organized as follows: Section 2 presents our formulation and justification of RMI and details our algorithm for calculating it. Section 3 presents the results from testing our algorithm vs. MI and those in [8] and [9] on a 2D-3D medical image registration problem using a clinical gold-standard for validation. Finally, Sections 4 and 5 discuss our work and present some conclusions and related future research directions.

## 2 Regional Mutual Information

### 2.1 Formulation

When using mutual information to compare two images, the actual probability distributions associated with each image are not known. Rather, marginal and joint histograms are usually calculated to approximate the respective distributions. One logical way to extend mutual information is to extend the dimensionality of the histograms. For example, using standard mutual information, one set of pixel co-occurrences in the joint distribution is represented by one entry in a two-dimensional joint histogram. What if we were to consider corresponding pixels and their immediate, $3 \times 3$ neighborhoods? This would be an entry into an 18D joint histogram. We could calculate the mutual information exactly as before only this time we would use 9D histograms for the marginal probabilities and an 18D histogram for the joint distribution.

One way to think of a $d$-dimensional histogram is as a set of points in $\Re^d$. Mutual information treats images as distributions of pixel values in 1D. Our formulation of RMI begins by recasting the problem to treat each image as a distribution of multi-dimensional points where each point represents a pixel and its neighborhood as depicted in Figure 1a.

### 2.2 The Curse of Dimensionality

At this point, however, we run into the curse of dimensionality. Essentially, the space where these points reside grows exponentially with each new dimension added. Figure 2 offers a look at why this is a problem. Here, what we've done is created some multi-dimensional distributions with known entropies (e.g. normal, exponential) in $\Re^2$, $\Re^4$, and $\Re^6$ and estimated their entropies with a varying number of samples. The important thing to notice here is that, as dimensionality increases, more and more samples are needed to populate the space to get a reasonable estimate of the entropy. For distributions in $\Re^6$, even 2 million samples

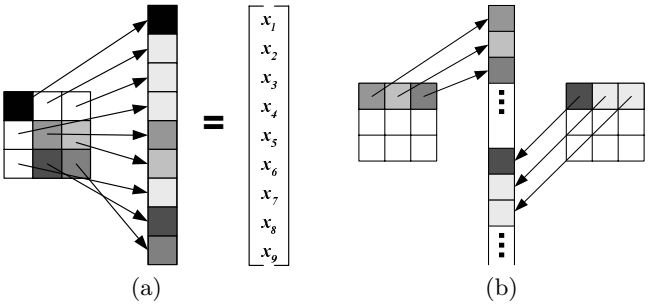


**Fig. 1.** (a) An illustration of the relationship between an image region and its corresponding multi-dimensional point. (b) Corresponding image neighborhoods and the multi-dimensional point representing them in the joint distribution.

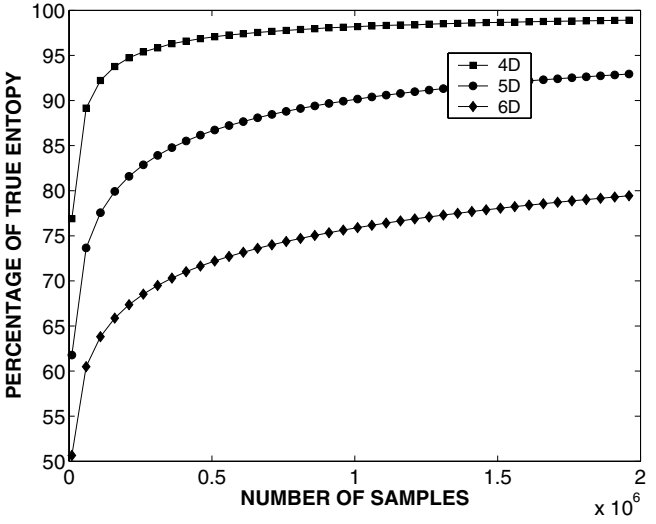


**Fig. 2.** An illustration of the curse of dimensionality as it pertains to sampling and entropy. On the $x$-axis, we have the number of samples and on the $y$-axis, the percentage of the true entropy we get when using those samples to approximate the entropy. As the dimensionality increases linearly, we need an exponentially increasing number of samples to get a reasonable estimate of a distribution's entropy.

are not sufficient to approximate the entropy correctly. Two million samples would correspond roughly to a comparison between images of resolution $1500 \times 1500$, an impractical figure for most real world applications of image similarity.

### 2.3 A Simplifying Assumption

We can take advantage of the fact that the entropy of a discrete distribution is invariant to rotations and translations [10] in order to make our problem more

tractable. Specifically, we can try and rotate and translate our high-dimensional distribution into a space where each dimension is independent. This is a restatement of the independent components analysis (ICA) problem as defined by Bell and Sejnowski [11]. Unfortunately, ICA is an extremely underdetermined problem and computationally too expensive for our purposes.

Instead we make the simplifying assumption that our high-dimensional distribution is approximately normally distributed. If this is the case, all we need to do is transform the points into a space where they are uncorrelated which, given the normal assumption, implies independence. Independence in each dimension allows us to decouple the entropy calculation [1] from one involving a $d$-dimensional distribution to one involving $d$ independent 1-dimensional distributions. Specifically, the entropy of a normally distributed set of points in $\Re^d$ with covariance matrix $\Sigma_d$ is [10]:

$$H_g(\Sigma_d) = \log((2\pi e)^{\frac{d}{2}} \det(\Sigma_d)^{\frac{1}{2}}).$$

This is mathematically equivalent to transforming the points into a new basis $(B_u)$ where each dimension is uncorrelated, projecting the data onto each of the $d$ new axes, and summing the entropies of those $d$ independent 1-dimensional distributions.

In practice, given a high-dimensional distribution represented by a set of data points $P = [p_1, \cdots, p_N]$, we can calculate this approximation to its entropy by first centering the data with respect to its mean, then diagonalizing its covariance matrix, and finally summing the entropies along each dimension. This process is the same as that used by principal components analysis (PCA) [12] and, essentially, what we are doing is summing the entropies along each of the orthogonal principal modes of variation.

### 2.4 Algorithm

Now that we have a method for efficiently calculating the entropy of a high-dimensional histogram, we can use it to calculate the RMI of a pair of images. The algorithm proceeds as follows:

1. Given two images $A$ and $B$, for each corresponding pair of pixels $[A_{ij}, B_{ij}]$, create a vector $\mathbf{v}_{ij}$ (Figure 1b) representing the co-occurrences of the pixels and their neighbors for some specified square radius $r$. This vector is now a point $p_i$ in a $d$-dimensional space where $d = 2(2r + 1)^2$. The image margins can be handled in a number of ways. We chose simply to ignore the pixels along the edges as we assume that they will not have too pronounced of an effect on the final entropy. Given radius $r$ and $m \times n$ images, we now have a distribution of $N = (m - 2r)(n - 2r)$ points represented by a $d \times N$ matrix $P = [p_1, \cdots, p_N]$.
2. Subtract the mean from the points so that they are centered at the origin: $P_0 = P - \frac{1}{N} \sum_i^N p_i$
3. Calculate the covariance of the points: $C = \frac{1}{N} P_0 P_0^T$

4. Estimate the joint entropy as: $H_g(C)$
5. Estimate the marginal entropies as $H_g(C_A)$ and $H_g(C_B)$ where $C_A$ is the $\frac{d}{2} \times \frac{d}{2}$ matrix in the top left of $C$ and $C_B$ is the $\frac{d}{2} \times \frac{d}{2}$ matrix in the bottom right.
6. Calculate the RMI $= H_g(C_A) + H_g(C_B) - H_g(C)$

Asymptotically, RMI has the same performance as standard mutual information. Given $n \times n$ images, the performance of standard mutual information is $O(n^2)$. Similarly, RMI's performance is $O(n^2 d^2)$ which represents the work required to calculate the covariance matrix. Since $d$ remains constant for a given choice of the neighborhood radius, asymptotically, RMI's performance converges to $O(n^2)$ as well.

### 2.5 Justification

In general, 1-dimensional projections of data made up of vectors of independent and identically distributed values tend to be normally distributed [13]. At a high level, this observation is derived by an appeal to the central limit theorem. Unfortunately, in our case, the vectors that make up our data are not independent (pixel values close together are not generally independent of each other) so the central limit theorem does not hold. There are, however, forms of the central limit theorem which allow weak dependence among the data. One such form, proven in [14], holds that the central limit theorem still applies in the case of $m$-dependent variables. A sequence of random variables $(X_1, X_2, \ldots, X_n)$ is $m$-dependent if, for some positive integer $m$, the inequality $s - k > m$ implies that the two sets $(X_1, X_2, \ldots, X_k)$ and $(X_s, X_{s+1}, \ldots, X_n)$ are independent. So, given data made up of $m$-dependent vectors, the central limit theorem does apply which implies that 1-dimensional projections of this data should also tend to be normally distributed.

Now what we must do is demonstrate that our data, though not independent, is at least $m$-dependent which would tend to support our assumption that its projections are normal. Intuitively, $m$-dependence requires that variables far away from each other in a sequence are independent. In our case, each data point represents a sequence of pixels (Figure 1a). Though the pixel values are locally dependent, pixels that are far apart in the sequence are further from each other in the neighborhood they are drawn from and, hence, more likely to be independent. Indeed, as the size of the neighborhood increases, it becomes more and more likely that pixels far enough apart in the sequence representing that neighborhood are independent.

As mentioned earlier, calculating the RMI is equivalent to projecting the data onto each of the axes of the new, uncorrelated basis $(B_u)$ and summing up the entropies. To test the validity of our assumption that these projections are generally normal, we generated 200 random pairs of medical images (amorphous silicon detector X-ray images) of the type we would expect to see in a typical registration problem. We then calculated RMI as usual, projected the data onto each of the axes of $B_u$, and used a Kolmogorov-Smirnov normality test ($\alpha = 0.01$)

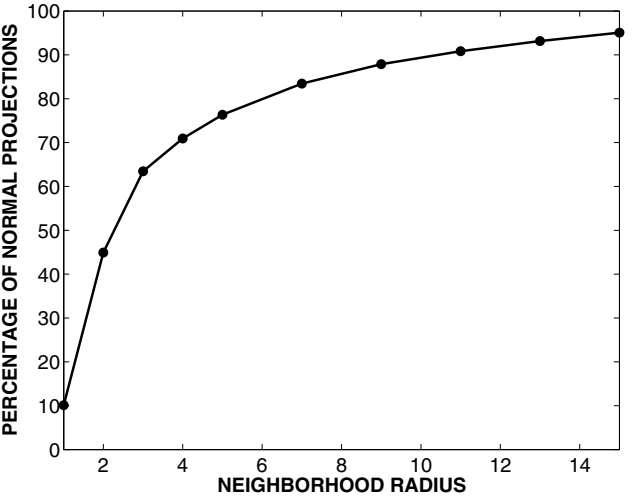


**Fig. 3.** For 200 random X-ray images, we plot the mean percentage of projections onto $B_u$ which are normal vs. the radius of the neighborhood used. As predicted by the $m$-dependent central limit theorem, the percentage of projections that can be considered normally distributed increases as the radius increases.

to determine which of the projections could be considered normal. For each pair of images, we performed this test for neighborhood radii ranging from 1 to 15. The results can be seen in Figure 3 and suggest a tradeoff between the time to calculate RMI and the accuracy of our assumption. Experimentally, we have found that even a radius as low as $r = 2$ works quite well in practice.

## 3 Results

### 3.1 Validation

As we have mentioned, mutual information as an image similarity measure has enjoyed a large degree of success in medical image registration applications. We chose to validate RMI on just such a problem, specifically 2D-3D medical image registration. The 2D-3D registration problem involves taking one or more X-ray projection (2D) images of a patient's anatomy and using these projections to determine the rigid transformation $\mathbf{T}$ (rotation and translation) that aligns the coordinate system of a CT (3D) image with that of the X-ray projection images and an operating room. 2D-3D registration is an important primitive in applications such as image-guided spine surgery [15,16] and radiosurgery [17,18].

We validate RMI as a means of performing 2D-3D intensity-based registration using the same experimental set-up and clinical gold-standard as in [19]. We use archived clinical data from the CyberKnife Stereotactic Radiosurgery System (Accuray, Inc., Sunnyvale, CA) which includes a preoperative contrast 3D CT

**Table 1.** 2D-3D Spine Image Target Registration Error

| Similarity Measure | TRE (mm) Mean | Unsuccessful Registrations |
|---|---|---|
| RMI | 1.25 | 8% |
| Pluim, *et al.* | 1.31 | 15% |
| Mutual Information | 1.33 | 18% |
| Rueckert, *et al.* | 1.43 | 30% |

scan, 2 orthogonal intraoperative amorphous silicon detector 2D X-ray images, and a built-in gold-standard calculated using bone-implanted fiducial markers. We compared RMI not only to mutual information (implemented using histograms as per [3]), but also to our own implementations of the aforementioned similarity measures with spatial information from Pluim, *et al.* [8] and Rueckert, *et al.* [9]. There have been a number of other similarity measures used for this problem [20] including cross-correlation and gradient-correlation, two measures that also take neighborhood relationships into account. However, following [21], we focus on mutual information as, on real data, it has been shown to be more accurate.

### 3.2 Experiments

Each experiment involved an initial transformation generated by perturbing the gold-standard reference transformation by adding randomly generated rotations and translations. The initial transformations were characterized by computing the target registration error (TRE) [22] for the transformation and grouped into eight initial TRE intervals: 0–2, 2–4, 4–6, 6–8, 8–10, 10–12, 12–14, and 14–16 mm. For each of 6 patients (3 with cervical lesions and 3 with thoracic lesions) and each similarity measure, 160 registrations were performed, 20 in each of the six misregistration intervals. The TRE value was computed for each registration transformation as the difference between the positions of a target mapped by the evaluated transformation and the gold-standard transformation. The TRE values were computed for each voxel inside a rectangular box bounding the vertebra closest to the lesion and then averaged. The registrations were characterized as either "successful" if the TRE $< 2.5$ mm or "unsuccessful" if the TRE $\geq 2.5$ mm. The results are listed in Table 1. Here we see that, while slightly more accurate, the real win from using RMI lies in its robustness, or its success rate with respect to misregistration. We take a more detailed look at robustness in Figure 4 where we plot percentage of successful registrations vs. initial TRE for all 4 similarity measures. While most of the measures perform well when started close to the initial solution, as initial TRE increases, RMI performs much better. By the time the initial TRE is in the range of 14–16 mm, RMI performs almost 50% better than the next best measure. In terms of CPU time, the difference between standard mutual information and RMI is quite small. On average registrations with mutual information took 101 sec., while those with RMI took 174 sec.

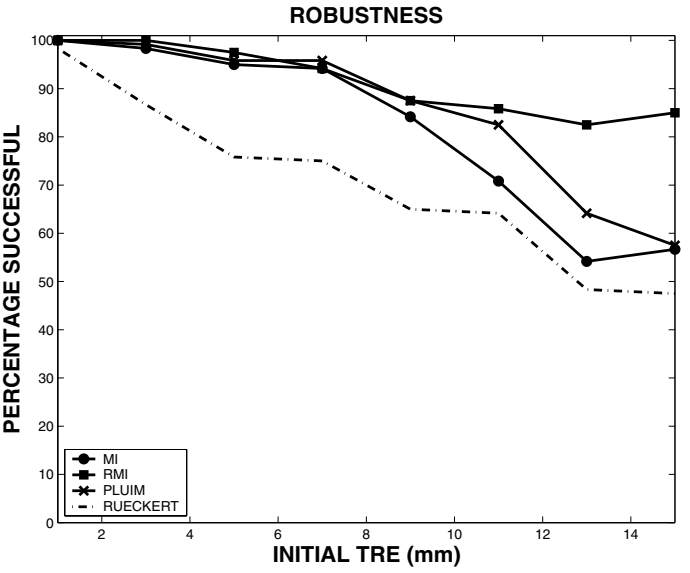


**Fig. 4.** Percentage of successful registrations for initial transformations with different initial TRE values. Each data point represents a 2 mm range of initial TRE values centered at the $x$-coordinate.

### 3.3 Noisy CT Data

Intensity-based 2D-3D registration requires an accurate CT scan. Usually this means that the patients are required to hold their breath during the scan so that breathing artifacts don't introduce noise into the data. Not all patients, however, are able to hold their breath for a sufficient period of time. Those that don't usually leave CTs that are noisy enough to severely affect the performance of an image-based registration algorithm. We performed the same experiments as above on two patients whose CT scans contained severe breathing artifacts. The results, seen in Figure 5, show the same basic trends as those from Figure 4 but are much more dramatic. For initial transformations close to the gold-standard, RMI still succeeds 100% of the time and its performance drops off much more gradually than the other three measures.

## 4 Discussion

To get a closer look at why RMI is more robust than MI, we analyzed a specific situation from the data-set above where MI often fails. In particular, instead of considering 6 independent parameters, we looked only at misregistration with respect to the $x$-axis and plotted RMI with varying neighborhood sizes (up to $r = 4$) vs. MI. The results, shown in Figure 6, illustrate one of the advantages of RMI. MI clearly shows a strong local maximum to the right of the global

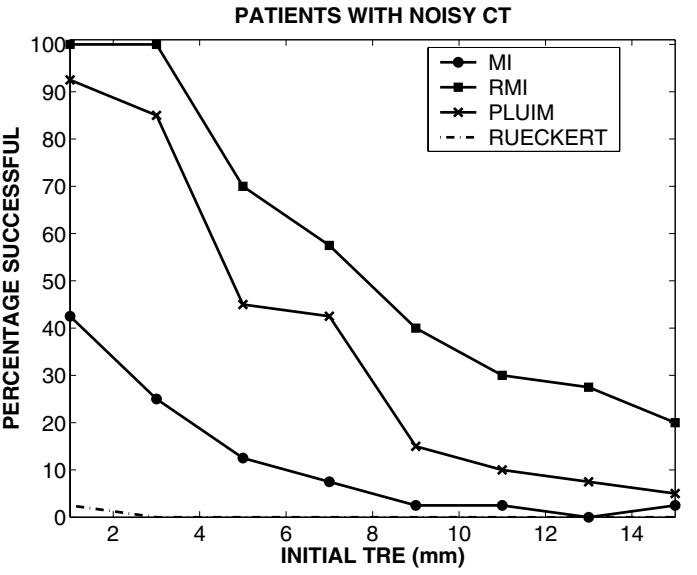


**Fig. 5.** Percentage of successful registrations for initial transformations with different initial TRE values. Each data point represents a 2 mm range of initial TRE values centered at the $x$-coordinate.

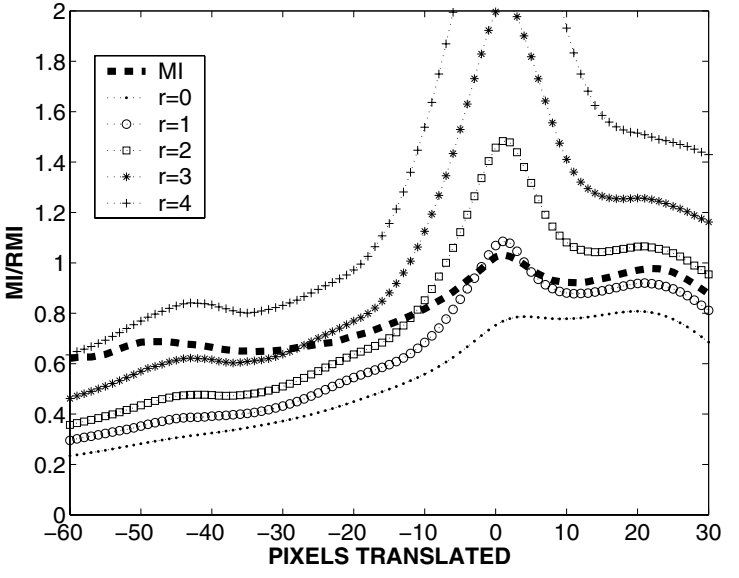


**Fig. 6.** Plot of mutual information as a function of misregistration along the $x$-axis. Also includes plots of RMI with neighborhoods of varying sizes ($r = 0$, $r = 1$, $r = 2$, $r = 3$, $r = 4$).

maximum which could have a large affect on the success of the optimization. As we begin to consider RMI with larger and larger neighborhoods, more spatial information is brought into the metric and we get a stronger peak at the global optimum and a smoother, more accurate similarity landscape away from it.

## 5 Conclusions

Mutual information as an image similarity measure has enjoyed a great deal of success in a variety of fields, medical image registration in particular. We have extended mutual information in a principled way to arrive at RMI which incorporates spatial information inherent in images. We have demonstrated RMI's improved robustness as a similarity measure and validated its use on real, clinical data from the medical imaging domain. In the future, we hope to validate RMI more extensively on larger clinical data sets. In addition, we're interested in applying RMI to non-rigid registration of clinical data. We would also like to apply RMI to more problems from other domains such as the creation of image mosaics, the querying of image databases, and the tracking of human motion.

## References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, New York (1991)
2. Viola, P., Wells III, W.M.: Alignment by maximization of mutual information. International Journal of Computer Vision **24** (1997) 137–154
3. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. IEEE Transactions on Medical Imaging **16** (1997) 187–198
4. Pluim, J., Maintz, J., Viergever, M.: Mutual information based registration of medical images: a survey. IEEE Transactions on Medical Imaging **22** (2003) 986–1004
5. Studholme, C., Hill, D.L.G., Hawkes, D.J.: An overlap invariant entropy measure of 3D medical image alignment. Pattern Recognit. **32** (1999) 71–86
6. Rodriguez-Carranza, C., Loew, M.: A weighted and deterministic entropy measure for image registration using mutual information. Medical Imaging 1998: Image Processing **Proc. SPIE 3338** (1998) 155–166
7. Ma, B., Hero, A.O., Gorman, J., Michel, O.: Image registration with minimal spanning tree algorithm. In: IEEE Int. Conf. on Image Processing, Vancouver, BC (2000)
8. Pluim, J.P.W., Maintz, J.B.A., Viergever, M.A.: Image registration by maximization of combined mutual information and gradient information. IEEE Trans. Med. Imaging **19** (2000) 809–814
9. Rueckert, D., Clarkson, M.J., Hill, D.L.G., Hawkes, D.J.: Non-rigid registration using higher-order mutual information. Medical Imaging 2000: Image Processing **Proc. SPIE 3979** (2000) 438–447
10. Shannon, C.: A mathematical theory of communication. The Bell System Technical Journal **27** (1948) 379–423

11. Bell, A., Sejnowski, T.: An information-maximization approach to blind separation and blind deconvolution. Neural Computation **7** (1995) 1129–1159
12. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley and Sons Inc. (2001)
13. Diaconis, P., Freedman, D.: Asymptotics of graphical projection pursuit. Annals of Statistics **12** (1984) 793–815
14. Hoeffding, W., Robbins, H.: The central limit theorem for dependent random variables. Duke Mathematical Journal **15** (1948) 773–780
15. Lavallée, S., Troccaz, J., Sautot, P., Mazier, B., Cinquin, P., Merloz, P., Chirossel, J.P.: Computer-assisted spinal surgery using anatomy-based registration. In Taylor, R.H., Lavallée, S., Burdea, G., Mösges, R., eds.: Computer-Integrated Surgery: Technology and Clinical Applications. MIT Press, Cambridge, MA (1996) 425–449
16. Weese, J., Penney, G.P., Buzug, T.M., Hill, D.L.G., Hawkes, D.J.: Voxel-based 2-D/3-D registration of fluoroscopy images and CT scans for image-guided surgery. IEEE Trans. Inform. Technol. Biomedicine **1** (1997) 284–293
17. Murphy, M.J., Adler, Jr., J.R., Bodduluri, M., Dooley, J., Forster, K., Hai, J., Le, Q., Luxton, G., Martin, D., Poen, J.: Image-guided radiosurgery for the spine and pancreas. Comput. Aided Surg. **5** (2000) 278–288
18. Ryu, S.I., Chang, S.D., Kim, D.H., Murphy, M.J., Le, Q.T., Martin, D.P., Adler, Jr., J.R.: Image-guided hypo-fractionated stereotactic radiosurgery to spinal lesions. Neurosurgery **49** (2001) 838–846
19. Russakoff, D.B., Rohlfing, T., Maurer, Jr., C.R.: Fast intensity-based 2D-3D image registration of clinical data using light fields. Proc. 9th IEEE Int. Conf. Computer Vision (ICCV 2003) (2003) 416–423
20. Penney, G.P., Weese, J., Little, J.A., Desmedt, P., Hill, D.L.G., Hawkes, D.J.: A comparison of similarity measures for use in 2D-3D medical image registration. IEEE Trans. Med. Imaging **17** (1998) 586–595
21. Russakoff, D.B., Rohlfing, T., Ho, A., Kim, D.H., Shahidi, R., Adler, Jr., J.R., Maurer, Jr., C.R.: Evaluation of intensity-based 2D-3D spine image registration using clinical gold-standard data. In Gee, J.C., Maintz, J.B.A., Vannier, M.W., eds.: Proc. Second Int. Workshop on Biomedical Image Registration (WBIR 2003). Lecture Notes in Computer Science 2717. Springer-Verlag, Berlin (2003) 151–160
22. Maurer, Jr., C.R., Maciunas, R.J., Fitzpatrick, J.M.: Registration of head CT images to physical space using a weighted combination of points and surfaces. IEEE Trans. Med. Imaging **17** (1998) 753–761