SIMPLER ALTERNATIVES TO INFORMATION THEORETIC SIMILARITY METRICS FOR MULTIMODAL IMAGE ALIGNMENT

Shannon M. Hughes, Ingrid Daubechies

Departments of Electrical Engineering and Mathematics
Princeton University, Princeton, NJ 08544

ABSTRACT

Mutual information (MI) based methods for image registration enjoy great experimental success and are becoming widely used. However, they impose a large computational burden that limits their use; many applications would benefit from a reduction of the computational load.

Although the theoretical justification for these methods draws upon the stochastic concept of mutual information, in practice, such methods actually seek the best alignment by maximizing a number that is (deterministically) computed from the two images. These methods thus optimize a fixed function, the "similarity metric," over different candidate alignments of the two images. Accordingly, we study the important features of the computationally complex MI similarity metric with the goal of distilling them into simpler surrogate functions that are easier to compute.

More precisely, we show that maximizing the MI similarity metric is equivalent to minimizing a certain distance metric between equivalence classes of images, where images f and g are said to be equivalent if there exists a bijection ϕ such that $f(x) = \phi(g(x))$ for all x.

We then show how to design new similarity metrics for image alignment with this property. Although we preserve only this aspect of MI, our new metrics show equal alignment accuracy and similar robustness to noise, while significantly decreasing computation time. We conclude that even the few properties of MI preserved by our method suffice for accurate registration and may in fact be responsible for MI's success.

1. INTRODUCTION

Image registration is the problem of aligning two images of a common subject, e.g., those acquired from different viewpoints or imaging modalities. Accurate registration allows such image data to be compared or fused into a single image and is important to many applications, including medical imaging, computer vision, and military applications.

Shannon Hughes is partially supported by an NSF Graduate Research Fellowship. Ingrid Daubechies is partially supported through NIH grant #347-6019.

Typically, images are aligned by choosing the geometric transformation of one image (within some predefined group of possible transformations) that maximizes a given similarity metric between the two images. Of many such similarity metrics proposed for image alignment, the mutual information based similarity metrics proposed by Viola and Wells [1, 2] are often considered the most accurate [3]. In these metrics, pairs of corresponding pixels from the images are treated as realizations of two random variables, and the similarity metric is an estimate of their mutual information.

There are a variety of methods [4, 5, 3] for estimating the MI metric, in particular the histogram based methods against which we will benchmark. However, none of these methods are particularly fast. Registration applications require many instances of these similarity computations, often on large images, resulting in long running times. Faster methods can therefore have a large impact.

In this paper, we choose to view the MI similarity metric as a deterministic function of two images, which enables us to identify some important properties of this similarity metric, discussed in Section 2. In Section 3, we then create some decidedly simpler functions that share these properties. In Section 4, we compare our new functions with the established similarity metrics and observe similar registration accuracy and significantly decreased computation time.

2. IDENTIFYING FUNDAMENTAL PROPERTIES OF THE MI SIMILARITY METRIC

Mutual information is defined as a function of the joint probability distribution of two random variables. Because one typically does not have access to the random variables of which the two images are viewed as realizations, the MI similarity metric is computed from the two images themselves. More explicitly, the two images f_1 and f_2, both mapping D (labeling the pixel locations) to a subset of $\mathbb{R}$, are viewed as realizations of the random variables X_1 and X_2. As a proxy for the (unavailable) mutual information $I(X_1; X_2)$, one uses a number $\mu(f_1, f_2)$ computed from the set $\{(f_1(y), f_2(y))|y \in D\}$. We now identify a few properties of MI that we expect an acceptable proxy μ to satisfy.

2.1. Invariance under "vertical rearrangements"

Information theoretic quantities such as entropy and mutual information do not depend on the actual values taken by the random variables, but only on their distribution. Put more precisely, if X_1 and X_2 are two random variables and ϕ and ψ are one-to-one functions, then

$$I(\phi(X_1); \psi(X_2)) = I(X_1; X_2). \quad (1)$$

In terms of the images f_1, f_2 that are viewed as realizations of X_1 and X_2, taking $\phi(X_1)$ amounts to taking $\phi(f_1)$. Hence, $\phi(X_1)$ corresponds to a "vertical rearrangement" of the corresponding image f_1, conceptually a restacking of the level sets of the function. Property (1) then says that the MI between two images is unaffected by such "vertical rearrangements" of one or both image functions. Thus, the similarity of two images, as measured by MI metrics, will be high whenever we can make the images alike by applying a one-to-one function to one of them.

This property allows MI methods to succeed in multimodal image registration, where the mappings from underlying physical quantities to image intensity values may reasonably be modelled as bijective, inducing approximately this type of correspondence between the two images.

Hence, our proxy μ for the MI similarity metric should satisfy a similar property. If ϕ and ψ are any two bijective functions from $\mathbb{R}$ to itself, then we require

$$\mu(\phi(f_1), \psi(f_2)) = \mu(f_1, f_2). \quad (2)$$

This in turn implies that μ should be a functional acting on equivalence classes of functions, where we define equivalence (denoted by $\sim$) in terms of the "vertical rearrangement" concept defined above: $f \sim g$ iff there exists a bijection ψ such that

$$\forall x : f(x) = \psi(g(x)). \quad (3)$$

2.2. Maximizing MI as minimizing a distance metric

The following theorem shows that MI generates a distance metric on (equivalence classes of) random variables.

Theorem. *Let H(X) denote the entropy of X. The function*

$$\begin{aligned}\delta(X, Y) &= \tfrac{1}{2}(H(X) + H(Y)) - I(X, Y) \\ &= H(X, Y) - \tfrac{1}{2}(H(X) + H(Y))\end{aligned}$$

is a pseudometric distance between random variables; $\delta(X, Y) = 0$ *is equivalent to* $I(X; Y) = H(X) = H(Y)$.

Proof. To establish that δ is a pseudometric, we have to show that it is symmetric, nonnegative and that it satisfies the triangle inequality. Because we need only show pseudometricity, we allow $\delta(X, Y) = 0$ for some pairs where $X \neq Y$; we do require, however, $\delta(X, X) = 0$.

Clearly, δ is symmetric. Moreover, for all random variables X and Y, we have $I(X; Y) \leq \min\{H(X), H(Y)\} \leq \frac{H(X)+H(Y)}{2}$ and hence $\delta(X, Y) \geq 0 \ \forall X, Y$.

If $\delta(X, Y) = 0$, then $I(X, Y) = \frac{H(X)+H(Y)}{2}$ but since $I(X, Y) \leq \min\{H(X), H(Y)\}$, this means $H(X) = H(Y) = I(X; Y)$. Conversely, $I(X; Y) = H(X) = H(Y)$ implies $\delta(X, Y) = 0$.

Finally, to show the triangle inequality, we first note that

$$\begin{aligned}H(X|Z) \leq \ & H(X, Y|Z) \leq H(Y|Z) + H(X|Y, Z) \\ \leq \ & H(Y|Z) + H(X|Y).\end{aligned}$$

Thus,
$$\begin{aligned}\delta(X, Z) &= H(X, Z) - \tfrac{1}{2}(H(X) + H(Z)) \\ &= H(Z) + H(X|Z) - \tfrac{1}{2}(H(X) + H(Z)) \\ &\leq H(Z) + H(Y|Z) - \tfrac{1}{2}(H(Y) + H(Z)) \\ &\quad + H(Y) + H(X|Y) - \tfrac{1}{2}(H(Y) + H(Z)) \\ &= H(Y, Z) - \tfrac{1}{2}(H(Y) + H(Z)) \\ &\quad + H(X, Y) - \tfrac{1}{2}(H(X) + H(Y)) \\ &= \delta(X, Y) + \delta(Y, Z).\end{aligned}$$

Hence, we have shown that δ is a pseudometric and that $\delta(X, Y) = 0$ iff $I(X, Y) = H(X) = H(Y)$. ☐

Because $\delta(X, Y) = 0$ iff each of X,Y gives complete information about the other, it follows that $\delta(X, Y) = 0$ iff there exists a bijection ϕ from the range of X to the range of Y such that $Y = \phi(X)$. But this is of course the same notion of equivalence we defined earlier in Section 2.1. The pseudometric distance δ induces therefore a true metric on the equivalence classes of random variables defined above.

In the image alignment setting, the spatial transformations to be carried out on the images will not affect the entropies of the two random variables of which the images are viewed as realizations. Maximizing $I(X; Y)$ in this framework is thus equivalent to minimizing $\delta(X, Y)$.

We can therefore choose to minimize δ rather than maximize I. In terms of the proxy function setting discussed above, this means that we shall work with a metric distance d on function equivalence classes, standing in as a proxy for δ, rather than with the proxy μ for I.

3. DESIGN OF ALTERNATIVE METRICS

Based on our mathematical observations in Section 2, we shall design functionals d with the following properties.

First of all, d maps pairs (f, g) of real-valued functions on D to positive real numbers; moreover, d should be a pseudometric distance on this function space, which should be invariant under the equivalence introduced in Section 2. More precisely, we want

$$d(\phi(f), \psi(g)) = d(f, g) \quad (4)$$

for all f,g and for all one-to-one maps ϕ and ψ. Thus, d defines an induced pseudometric distance on the quotient of the function space by the equivalence relation $\sim$ defined by (3). Inspired by the corresponding property of δ, we require that this induced pseudometric is a true metric on the set of equivalence classes of functions.

One way to design such d is to take a standard functional norm $\|\cdot\|$ and to let

$$\tilde{d}(f,g) = \inf_{\xi,\eta:\xi\sim f,\eta\sim g} \|\xi - \eta\|. \tag{5}$$

Clearly, $\tilde{d}$ is a pseudometric distance. Moreover,

$$\tilde{d}(\psi(f),\phi(g)) = \tilde{d}(f,g) \tag{6}$$

for any bijections ψ and ϕ since the sets $\{\xi|\xi \sim f\}$ and $\{\xi|\xi \sim \phi(f)\}$ (and likewise $\{\eta|\eta \sim g\}$ and $\{\eta|\eta \sim \psi(g)\}$) are equivalent. Furthermore, in practice, we will want to impose a condition of type

$$\exists \beta \in \mathbb{R} : \tfrac{1}{\beta} \leq \phi'(x) \leq \beta, \tag{7}$$

with β fixed for all bijections ϕ, so that we do not minimize $\|\phi(f) - \psi(g)\|$ simply by mapping both function ranges to an infinitesimally small region of $\mathbb{R}$. For all f,g with $f \not\sim g$, we will then have

$$\tilde{d}(f,g) = \inf_{\xi,\eta:\xi\sim f,\eta\sim g} \|\xi - \eta\| \neq 0.$$

Thus, $\tilde{d}$ is a distance metric between equivalence classes of functions as desired.

3.1. Computing the alternative metrics efficiently

It turns out that the unique form of these metrics allows them to be simplified so that they may be computed easily.

First of all, we note that (7) implies that for all a,b,

$$\tfrac{1}{\beta}\|a-b\| \leq \|\phi(a) - \phi(b)\| \leq \beta\|a-b\|.$$

Hence, if we define $\hat{d}$ by

$$\hat{d}(f,g) = \inf_{\eta:\eta\sim g} \|f - \eta\|, \tag{8}$$

then since

$$\tfrac{1}{\beta}\|\phi(f) - \psi(g)\| \leq \|f - \phi^{-1}(\psi(g))\| \leq \beta\|\phi(f) - \psi(g)\|$$

for all ϕ and ψ, we have that

$$\tfrac{1}{\beta}\tilde{d}(f,g) \leq \hat{d}(f,g) \leq \beta\tilde{d}(f,g).$$

Thus, $\tilde{d}$ and $\hat{d}$ define equivalent topologies. Below, we shall choose to minimize $\hat{d}$ instead of $\tilde{d}$.

At this point, we must consider the fact that our images are discrete (quantized) functions taking values in some finite, though potentially large, set. Hence, the infimum in our expression for $\hat{d}$ is in fact achieved, i.e. it is a minimum. Minimizing $\hat{d}$ is thus equivalent to finding a bijection ϕ that minimizes $\|f - \phi(g)\|$ and then computing this norm.

Fortunately, this optimal ϕ is directly calculable. In practice, our functions f and g are quantized, taking values in a specific finite set Q, so that ϕ is completely determined by the $\phi(q), q \in Q$. Hence, we can optimize each of these values of $\phi(q)$ independently. To show this, let $\chi_{\{g=q\}}$ denote the characteristic function of the subset of D on which g takes the value q. We note that $\{x|g(x) = q\}$ and $\{x|g(x) = q'\}$ are disjoint whenever $q \neq q'$ and that $\cup_{q\in Q}\{x|g(x) = q\}$ covers D. Hence, each quantized image function f can be written in the form

$$f = \sum_{q\in Q} f\chi_{\{g=q\}}$$

for any other function g. We also have $g = \sum_{q\in Q} q\chi_{\{g=q\}}$, so $\phi(g) = \sum_{q\in Q}\phi(q)\chi_{\{g=q\}}$. Therefore, we have

$$\begin{aligned}\|f - \phi(g)\| &= \|\sum_{q\in Q} f\chi_{\{g=q\}} - \sum_{q\in Q}\phi(q)\chi_{\{g=q\}}\| \\ &= \|\sum_{q\in Q}(f - \phi(q))\chi_{\{g=q\}}\|.\end{aligned}$$

For many norms, including those considered here, there exists $p > 0$ so that $\|\chi_A + \chi_B\|^p = \|\chi_A\|^p + \|\chi_B\|^p$ if $A \cap B = \emptyset$. Hence, we have that

$$\begin{aligned}\hat{d}(f,g)^p &= \inf_\phi \|f - \phi(g)\|^p = \inf_\phi \|\sum_{q\in Q}(f - \phi(q))\chi_{\{g=q\}}\|^p \\ &= \inf_\phi \sum_{q\in Q}\|(f - \phi(q))\chi_{\{g=q\}}\|^p \\ &= \sum_{q\in Q}\inf_{\phi(q)}\|(f - \phi(q))\chi_{\{g=q\}}\|^p \\ &= \sum_{q\in Q}\inf_\alpha \|(f-\alpha)\chi_{\{g=q\}}\|^p.\end{aligned}$$

For many norms, minimizing $\|f - \alpha\|$ w.r.t. a constant α is easy. Furthermore, in practice, $\inf_\alpha \|(f-\alpha)\chi_{\{g=q\}}\|$ may be easy to determine without explicitly finding the best α.

For example, if $\|\cdot\|$ is the L^2 norm, then for each $q \in Q$, $\phi(q) = \frac{1}{|\{x|g(x)=q\}|}\sum_{x\in D:g(x)=q} f(x)$ where $|\cdot|$ denotes set cardinality; if $\|\cdot\|$ is the L^1 norm, then for each $q \in Q$, $\phi(q)$ is the median of the set $\{f(x)|x \in D, g(x) = q\}$.

When aligning images f and g for a set of possible transformations $\mathcal{T}$, we must compare the transformed image $f \circ T$ with g, $\forall T \in \mathcal{T}$. The number of pixels $N(T)$ in the "overlap" $T^{-1}(D) \cap D$ on which both $f \circ T$ and g are defined depends on T, so to compare distances for different T, we normalize by $N(T)$. Our **simplified algorithm for registering images** f and g thus becomes:

1. Determine the level sets of the function g.
2. For $T \in \mathcal{T}$, define

$$\gamma(T) = \tfrac{1}{N(T)}\sum_{q\in Q}\inf_\alpha \|(f\circ T - \alpha)\chi_{\{g=q\}}\|^p.$$

3. Pick (e.g. by gradient descent) the transformation $T^* = \text{argmin}_{T\in\mathcal{T}}\gamma(T)$. The image $f \circ T^*$ is aligned with g.

4. EXPERIMENTAL RESULTS

We now benchmark our metrics against MI similarity metrics. We choose histogram-based methods [2, 4] for our comparison since these methods are among the most popular and computationally efficient image alignment metrics.

As a toy setup of an alignment, we take for our first image, the 256 grayscale "peppers" of size 128×128, with added zero-mean white Gaussian noise of variance 100 (Fig. 1a). For the second image (Fig.1b), we apply a bijective function ϕ (Fig.1c) to the pixel-wise intensity of "peppers", and we again add zero-mean white Gaussian noise of variance 100. In practice, discrepancies between two images, such as those due to occlusions or erased data, may preclude a perfect bijection between pixel intensities, so we also block out a disk of radius 20 in the first image and another of radius 15 in the second to test for robustness to these obstacles.

We then rotate one of the images by up to 10 degrees (with nearest neighbor interpolation) and/or translate it by up to 10 pixels and plot the resulting values of the metric against these motion parameters. Fig.1d shows the result for the histogram-based metric while Figs.1e and 1f show the result for the L^1 and L^2-distance-based metrics respectively. We plot the negative of the MI metric in 1d, so that in all three figures the minimum indicates the best alignment. To compare the robustness to noise, we vary the standard deviation σ of the noise we add to the partially occluded pepper images. Fig.2 plots the cusp depth, and the capture ranges for rotation and translation, averaged over different realizations of the Gaussian noise, as a function of σ.

The L^2 and L^1 based metrics are computed in roughly half the time of the histogram metric (15.4 seconds for Fig. 1d vs. only 8.9 seconds for Fig.1e and 8.5 seconds for Fig.1f, computed with Matlab 7.0.1 on a Dell Latitude D600 laptop with an Intel Pentium 1.8 GHz processor); nevertheless, they achieve comparable results.

5. CONCLUSIONS AND DISCUSSION

We have studied properties of MI-based similarity metrics and identified a few properties of MI-based metrics that a successful image alignment metric should satisfy. We have then created very simple metrics for alignment that have these properties. Although our selected properties do not by any means represent a complete list of properties of MI, they are nevertheless experimentally successful. This suggests that the success of MI-based methods may be mostly due to their minimizing a distance between images that is invariant under "vertical rearrangements."

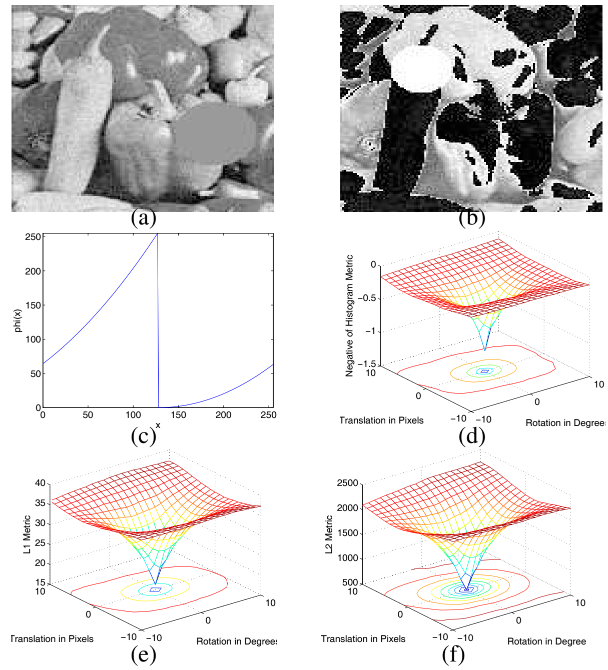

Fig. 1. (a) Original noisy peppers image, with one area blocked out; (b) Transformed noisy peppers image (using the bijection ϕ from $\{0,\ldots,255\}$ to itself plotted in (c)) with one area removed (see text). (d-f) Similarity metric values vs. misalignment between the two peppers images for (d) the histogram based method, (e) our L^1 based distance metric, and (f) our L^2 based distance metric.

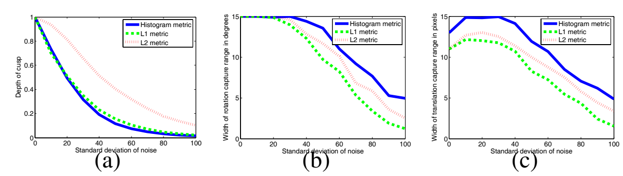

Fig. 2. (a) Cusp depth, (b) Capture range in degrees of rotation, and (c) Capture range in pixels of translation vs. standard deviation of added Gaussian noise.

6. REFERENCES

[1] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *Int. Journal of Computer Vision*, vol. 24, no. 2, pp. 137–54, 1997.

[2] F. Maes et. al., "Multimodality image registration by maximization of mutual information," *IEEE Trans. on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[3] J. Pluim et. al., "Mutual information based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, 2003.

[4] P. Thevenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. on Im. Proc.*, vol. 9, no. 12, pp. 1083–1100, 2000.

[5] B. Ma, A.O. Hero, J.D. Gorman, and O. Michel, "Image registration with minimum spanning tree algorithm," in *IEEE Int. Conf. on Image Proc.*, 2000, vol. I, pp. 481–4.