The Role of Non-Overlap in Image Registration

Jonas August and Takeo Kanade

Healthcare Robotics Center, The Robotics Institute
Carnegie Mellon University, Pittsburgh, Pennsylvania

Abstract. Here we model the effect of non-overlapping voxels on image registration, and show that a major defect of overlap-only models—their limited capture range—can be alleviated. Theoretically, we introduce a maximum likelihood model that combines histograms of overlapping and non-overlapping voxels into a common joint distribution. The convex problem for the joint distribution is solved via iterative application of replicator equations that converge monotonically. We then focus on rigidly aligning images with unknown translation, where we present a fast FFT-based method for computing joint histograms for all relative translations of an image pair. We then apply this method to standard overlap-only information theoretic registration criteria such as mutual information as well as to our variants that exploit non-overlap. Our experimental results show that global optima correspond to the correct registration generally only when non-overlapping image regions are included.

1 Introduction

This paper addresses a long-standing complaint with intensity-based image registration methods: they generally converge correctly only if given an initial guess within a limited "capture range" of the correct alignment. We are led to ask: even if processing were free but *no* initial guess were given, do current registration criteria select the correct alignment? Unfortunately not, since the global optima of information theoretic registration criteria such as entropy may be far away from the correct result [5, p. R27]. Here we suggest a fix.

Spurious global optima can arise when there is too little overlap of the image pair for reliable estimation of the joint distribution of corresponding voxels. We thus revisit the concept of overlap beginning in §2, where we review a common probabilistic registration model that assumes full overlap and which explains why the joint histogram of the image pair can be used as an estimate of the joint distribution of intensities of corresponding voxels. In §3, we generalize this model to allow for merely partial overlap, and it is here we see terms in the likelihood that depend on the *non-overlapping* voxels. The revised model gives rise to a joint distribution that trades off the joint histogram on the overlapping voxels with univariate histograms from the non-overlapping voxels, unlike [7, 9]. In §4, we solve for this joint distribution using a monotonically-convergent iterative scheme, i.e., where no step size is required.

To solve for the alignment itself, we focus on the case of unknown translation. In §5, we compute the *globally optimal* alignment to within one voxel using a fast FFT-based method for computing joint and non-overlap histograms over all translations. This also makes it practical to visualize various registration criteria over the entire set of transformations, not only those within a local neighborhood of a potential solution. These complex registration landscapes (§6) highlight the difficulties that registration search strategies must confront, and put into question the feasibility of local search for fully automatic (full capture range) image registration. We suggest that global methods not based on local search will be necessary in the absence of a good initial guess. In hindsight, the standard practice of ignoring the non-overlap seems strange since it uses different image data to evaluate competing alignments that differ in overlap. This violates the principle that all hypotheses be compared using the same information.

2 Idealized Registration Configuration: Full Overlap

To introduce our argument and notation, we start with the simpler situation where the effects of overlap are ignored. Let $u : X \to \{1, \ldots, M\}$ and $v : Y \to \{1, \ldots, N\}$ be the two images to be aligned, where region X (resp. Y) is the finite cardinality set of possible voxels (locations) and M (resp. N) is the number of possible intensities for image u (resp. v). Typically, X and Y are the vertices of a finite lattice in 2- or 3-dimensions. Thus $u_x = u(x)$ is the intensity (in the range $\{1, \ldots, M\}$) at voxel $x \in X$ and $v_y = v(y)$ is the intensity (in the range $\{1, \ldots, N\}$) at voxel $y \in Y$.

The goal of intensity-based image registration is to optimally choose that spatial transformation $y = T(x)$ that maps between the two image regions so that u_x and $v_{T(x)}$, the intensities at corresponding voxels x and $T(x)$, are in some sense correlated, suggesting that their joint distribution will be important. We assume that the intensities for pairs of corresponding voxels are independent and identically distributed (IID), i.e., if $x' \neq x$, then $(u_x, v_{T(x)})$ and $(u_{x'}, v_{T(x')})$ are IID, each pair having joint distribution (probability mass function) $p(m, n) = p_{m,n}, m \in \{1, \ldots, M\}, n \in \{1, \ldots, N\}$. Further assuming full overlap, i.e., that the mapping $T : X \to Y$ is one-to-one and onto, the likelihood (joint probability) of the two images is therefore

$$\mathrm{Prob}\{u, v|\text{full overlap}\} = \prod_{x \in X} p(u_x, v_{T(x)}),$$

and the log likelihood is

$$L_{\text{full}} := \sum_{x \in X} \log p(u_x, v_{T(x)}). \tag{1}$$

Recall the identity $\sum_n \delta(k, n) = 1$, where the Kronecker delta function $\delta(k, n)$ is equal to 1 if $k = n$ and is 0 otherwise. We apply this identity twice to obtain

$$L_{\text{full}} = \sum_{x \in X} \left[\sum_m \delta(u_x, m) \right] \left[\sum_n \delta(v_{T(x)}, n) \right] \log p(u_x, v_{T(x)}). \tag{2}$$

By changing the order of summation (permissible because all sums are finite), we can write

$$L_{\text{full}} = \sum_{m,n} a^T_{m,n} \log p_{m,n}, \quad \text{where } a^T_{m,n} := a_{m,n} := \sum_{x \in X} \delta(u_x, m)\delta(v_{T(x)}, n) \quad (3)$$

is the joint histogram (raw, unnormalized counts) of intensity pairs at corresponding voxels for transformation T. Observe that $\sum_{m,n} a_{m,n} = |X|$, the number of voxels in X. To determine the unknown joint distribution p, we solve an optimization problem: maximizing the (log) likelihood. We first show L_{full} is well behaved, and then show the solution is the normalized histogram.

Proposition 1. *L_{full} is a concave function of p.*

Proof. Observe in (3) that L_{full} is a nonnegatively-weighted sum of the concave function log [4]. □

Let S be the **simplex of distributions**[1]

$$\begin{aligned} S := \{\, & p \in \mathbb{R}^{MN} : \\ & p_{m,n} \geq 0, \forall m,n; \qquad \text{[Nonnegativity constraint]} \qquad (4) \\ & \sum_{m,n} p_{m,n} = 1 \}. \qquad \text{[Normalization constraint]} \qquad (5) \end{aligned}$$

Observe that set S is convex.

Proposition 2. *Fix transformation T and suppose $a^T_{m,n} > 0$, for all m, n. Then normalized histogram $p^* = a^T/|X|$ is the global optimum of the convex problem*

$$\max_p L_{\text{full}}(T, p) \text{ subject to } p \in S.$$

Proof. We first ignore the nonnegativity constraint but later check that it is satisfied. Applying the method of Lagrange multipliers to the constrained optimization problem (now with only the normalization equality constraint having the Lagrange multiplier γ), we seek the maximum of $\phi(p, \gamma) = L_{\text{full}} + \gamma(\sum_{m,n} p_{m,n} - 1)$. Recall that the first-order necessary conditions for optimality are obtained by setting to zero the partial derivatives of ϕ with respect to the unknowns. Differentiating w.r.t. $p_{m,n}$, we get $a^T_{m,n}/p^*_{m,n} + \gamma^* = 0$, and therefore $p^*_{m,n} = a^T_{m,n}/\gamma^*$. Differentiating w.r.t. γ we get the normalization constraint (5), and thus $p^*_{m,n} = a^T_{m,n}/\sum_{k,l} a^T_{k,l} = a^T_{m,n}/|X|$. Since $a^T_{m,n}$ is strictly positive, so is $p^*_{m,n}$, and therefore the nonnegativity constraint is not active at p^* and can be ignored. Since L_{full} is concave in p, the unique stationary point p^* is the global maximizer. □

The following consequence of Prop. 2 may be viewed as a justification, first shown in [7], for the use of minimum entropy for (fully overlapping) image registration: the transformation T that minimizes the empirical entropy of distribution $a^T/|X|$ maximizes the likelihood.

Corollary 1. $L^*_{\text{full}}(T) := \max_p L_{\text{full}}(T, p) = -|X| \,\text{entropy}\,(a^T/|X|)$.

[1] Here all distributions are normalized and histograms are unnormalized, unless otherwise stated.

3 Realistic Registration Configuration: Partial Overlap

Now we include the effect of partial overlap of the two images. There are three regions to consider: (a) the voxels that overlap, as before; (b) the voxels in image u that do not map to voxels in image v; and (c) the voxels in image v that do not get mapped to from image u. Even if the only dependencies are between corresponding voxel intensities, as before, what distributions should be used for the non-overlapping regions (b) and (c)? We suggest that no new information about the non-overlapping voxels should be assumed; a non-overlapping voxel is to be treated just the same as an overlapping voxel pair, but where one voxel of the pair was not observed. In other words, the reason why there are no corresponding v-voxels for the non-overlapping u-voxels is that we have limited our region of interest (ROI) for image v, and vice versa. Thus we obtain the probability for intensity u_x at non-overlapping voxel x by marginalizing the joint distribution: sum the joint probability of u_x and v_y over all possible values of the unknown v_y. Specifically, if $p(m,n) = p_{m,n}$ is the joint distribution for intensities $u_x = m$ and $v_y = n$ at corresponding voxels x and y, then the intensity $u_x = m$ at non-overlapping voxel x is distributed according to the marginal distribution $\sum_n p(m,n)$. Similarly, the intensity $v_y = n$ at non-overlapping voxel y is distributed according to the marginal distribution $\sum_m p(m,n)$.

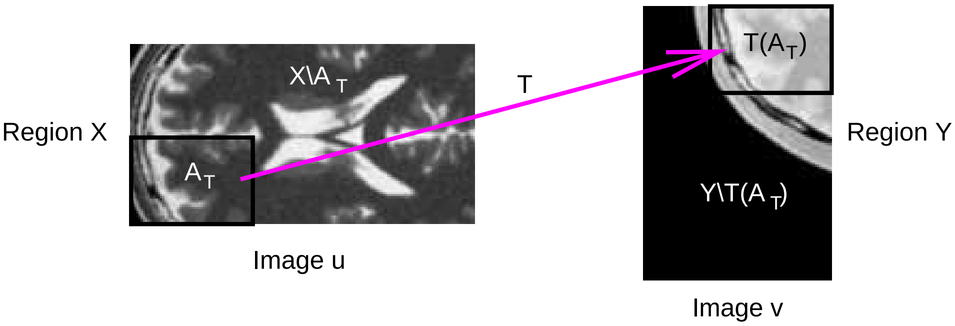

Fig. 1. Partially overlapping images u and v represent different regions of interest in the patient. See text for notation.

When we explicitly consider partial overlap, both the domain of definition and the mapping rule can vary (Fig. 1); thus the alignment transformation is $T : A_T \rightarrow Y$, where domain $A_T \subset X$ is the set of voxels in image u that map to voxels in image v. Note that $T(A_T) \subset Y$ is the set of voxels in image v that get mapped to. Thus the non-overlapping portion of image u is $X \setminus A_T$, i.e., everything in X but A_T; similarly, the non-overlapping part of image $v(y)$ is $Y \setminus T(A_T)$. Again assuming IID distributions, the probability of the image pair

u, v at transformation T is

$$\begin{aligned}&\text{Prob}\{u, v|\text{partial overlap}\}\\ &= \left[\prod_{x\in A_T} p(u_x, v_{T(x)})\right] \times \left[\prod_{x\in X\setminus A_T} \sum_n p(u_x, n)\right] \times \left[\prod_{y\in Y\setminus T(A_T)} \sum_m p(m, v_y)\right].\end{aligned}$$

Again using the Kronecker identity and changing order of summation as in the fully-overlapping case, the log likelihood is

$$\begin{aligned}L_{\text{partial}} = &\sum_{x\in A_T} \log p(u_x, v_{T(x)})\\ &+ \sum_{x\in X\setminus A_T} \log \sum_n p(u_x, n) + \sum_{y\in Y\setminus T(A_T)} \log \sum_m p(m, v_y) \qquad (6)\\ = &\sum_{m,n} a^T_{m,n} \log p_{m,n}\\ &+ \sum_m b^T_m \log \sum_n p_{m,n} + \sum_n c^T_n \log \sum_m p_{m,n}, \qquad (7)\end{aligned}$$

where we define

$$a^T_{m,n} := a_{m,n} := \sum_{x\in A_T} \delta(u_x, m)\delta(v_{T(x)}, n) \qquad (8)$$

$$b^T_m := b_m := \sum_{x\in X\setminus A_T} \delta(u_x, m) \qquad (9)$$

$$c^T_n := c_n := \sum_{y\in Y\setminus T(A_T)} \delta(v_{T(x)}, n), \qquad (10)$$

the (T-dependent) joint histogram for the overlapping region, and the histograms for the non-overlapping regions of image u and v, respectively. Again we can maximize this likelihood L_{partial} to determine the unknown joint distribution p. But unlike §2, clearly some sort of numerical optimization will be needed to compute this $p \in S$: we have to trade off the effects of the overlapping versus the non-overlapping histograms. Fortunately, objective function L_{partial} is well-behaved, leading to a convex problem for p.

Proposition 3. *L_{partial} is a concave function of p.*

Proof. Since log is concave and $\sum_n p_{m,n}$ is affine in p, their composition $\log \sum_n p_{m,n}$ is concave in p; similarly for $\log \sum_m p_{m,n}$ [4]. Thus, L_{partial}, a nonnegatively-weighted sum of concave functions, is concave. □

Before introducing our optimization strategy in §4, we suggest how this optimal p be used.

Proposal 1 (Non-Overlap Imperative) *Given partially overlapping images u and v, to evaluate information-theoretic image comparison measures such as joint entropy and mutual information, use the distribution p that maximizes L_{partial} instead of the overlap-only-based normalized joint histogram.*

4 Replicator Equations for Combining Histograms

Now we present an iterative method for estimating the distribution p that maximizes the log likelihood L_{partial} for partial overlap. We suppress T for now as it will be optimized for after we have optimized for p at each fixed T. Since our problem is to maximize concave L_{partial} over convex set S, we could attempt to exploit the arsenal of convex programming. Instead, we suggest an iterative technique with a simple implementation, where the iteration cost is low (unlike other second-order methods that might apply) and which requires no tuning of parameters at all. Specifically, the **replicator equations** for updated distribution p' are similar to a gradient ascent on the log likelihood, except the gradient multiplicatively—not additively—updates the previous distribution p, and the result is normalized to sum to one to remain in the simplex of distributions:

$$p'_{m,n} := \frac{p_{m,n} L_{m,n}}{\sum_{i,j} p_{i,j} L_{i,j}}, \text{ where } L_{m,n} := \frac{\partial L_{\text{partial}}}{\partial p_{m,n}} = \frac{a_{m,n}}{p_{m,n}} + \frac{b_m}{\sum_j p_{m,j}} + \frac{c_n}{\sum_i p_{i,n}}. \tag{11}$$

Observe that this simplex-preserving multiplicative update method converges in one step to the the result in Prop. 2 if b and c are both zero. More importantly, in contrast to the undesirable instability of (additive) gradient ascent when too large a step size is chosen, each multiplicative update increases the log likelihood *without choosing a step size*.

Definition 1. *Continuous mapping $f : D \to D$ is* **growth transformation** *for objective function $\phi : D \to \mathbb{R}$ if $\phi(f(p)) \geq \phi(p)$, for all $p \in D$.*

The concept of growth transformation was used in papers by Baum and coworkers [2, 3] and Pelillo [6] to characterize the dynamics of replicator equations, which are a particular class of relaxation labeling processes [8], for certain polynomial objective functions ϕ that arise in evolutionary game theory, computer vision and parameter estimation for Markov chains. Although our objective function L_{partial} is non-polynomial, we have obtained the same result.

Proposition 4. *Update (11) is a growth transformation for $L_{\text{partial}} : S \to \mathbb{R}$.*

Explicitly, this states that $L_{\text{partial}}(p') \geq L_{\text{partial}}(p)$, for any distribution $p \in S$ and its update p' from (11): we can depend on the update to monotonically improve the log likelihood. We have proved Prop. 4 using the log-sum and arithmetic-geometric means inequalities [1].

Because the replicator equations describe a growth transformation for L_{partial}, the choice of initial distribution p^0 that starts the iterations is unimportant, but to avoid degeneracies we suggest that all components be non-zero. We use the normalized version of overlap histogram a as the initial condition in our experiments. To maximize L_{partial}, we iteratively apply the replicator equations until a termination condition is satisfied. In our experiments, we simply stopped after completing only two iterations.

5 FFTs for Global Optimization of Translation

Designers of information theoretic objective functions for image registration have not insisted that global optima approximate the true solution, and have instead focused on local optima. Perhaps this bias stems from the seeming intractability of computing the global optimum. To illustrate, even when T is restricted to a translation and $n = |X| \approx 10^5$ to 10^9 is the number of voxels, $O(n)$ operations are required to compute the joint histogram at each of $O(n)$ possible translations, for an apparent total of $O(n^2)$ operations to find the global optimum! These two onerous $O(n)$ are usually [10] reduced to $O(1)$ by (i) using statistical sampling to approximate the joint distribution and (ii) abandoning global optimization entirely for local, greedy search.

Here we introduce a method to allow exact global optimization of translation for information theoretic objectives in only $O(k\, n \log n)$ operations, where $k = MN$ is the number of bins in the joint histogram. This technique applies to both the full overlap and partial overlap likelihoods, as well as to any registration method that that requires computation of the joint distribution, such as entropy, mutual information [10], and normalized mutual information [9]. The trade-off is histogram resolution for image resolution, which is often acceptable because the joint histogram requires crude quantization just to maintain sufficient bin counts for reliability.

The main idea is that the (m, n)-th bin of the joint histogram is the cross-correlation $a^T_{m,n} = \sum_x f(x)g(x + t) =: \mathrm{corr}_{f,g}(t)$ between two binary vectors $f(x) := \delta(u_x, m)$ and $g(y) := \delta(v_y, n)$, where $y = T(x) := x + t$ for translation t. The translation is a 2- or 3-dimensional vector depending on the dimensionality of images u and v. Zero-padding f and g to an appropriate size l^2 or l^3 for 2- or 3-d, resp., we can avoid wrap around artifacts in assuming their periodicity, and thus apply Fourier methods. Specifically, if the discrete Fourier transform of f at frequency vector ω is $\hat{f}(\omega) := \sum_x f(x)e^{-2\pi i\omega\cdot x/l}$, and z^* is the complex conjugate of $z \in \mathbb{C}$, we know that $\widehat{\mathrm{corr}_{f,g}}(\omega) = \hat{f}^*(\omega)\hat{g}(\omega)$. Thus for the (m, n)-th bin, computing $a^T_{m,n}$ over all translations takes $O(n \log n)$ work using the FFT. By performing this over all k bins we can calculate the joint histogram over all translations in $O(k\, n \log n)$ time.

The non-overlap histograms b^T and c^T require a similar approach, because they depend on the non-constant region of overlap A_T. (We cannot simply compute the histograms for each image; we need a histogram for each possible overlap.) For b^T, our computation is based on a cross-correlation between f and a mask (of ones) the size of image v. For c^T, the cross-correlation is between a u-sized mask and g. Each also requires $O(k\, n \log n)$ computations.

Given the overlap and non-overlap histograms, for each translation we can solve the optimization problem for p in §4 with $O(k)$ work, and all translations with $O(k\, n)$ work. Evaluating any of the optimization criteria L_{full}, L_{partial}, entropy, mutual information or normalized mutual information is only $O(k\, n)$ more work and the selection of its global optimum takes $O(n)$ time for a grand total of $O(k\, n \log n)$ operations.

6 Experimental Results

To test the effect of including non-overlapping image portions, we evaluated several registration criteria over all 2-d translations, thus computing a "registration landscape". For joint distribution estimates using overlapping image portions only, the criteria included mutual information, normalized mutual information, and L_{full}. For the non-overlap-based joint distribution computed by optimizing L_{partial} w.r.t. p, these criteria, as well as L_{partial}, were also used to form landscapes. For joint distribution $q = (q_{m,n})$, the formula for mutual information is $\text{MI}\,(q) = \text{entropy}\,(\sum_m q_{m,n}) + \text{entropy}\,(\sum_n q_{m,n}) - \text{entropy}\,(q)$, and for normalized mutual information it is $\text{NMI}\,(q) = [\text{entropy}\,(\sum_m q_{m,n}) + \text{entropy}\,(\sum_n q_{m,n})]/\text{entropy}\,(q)$. All joint histograms had 16 uniformly-spaced bins to which we added 0.1 to avoid degeneracy. Computations (in Numerical Python under GNU/Linux) used up to 0.5GB and took tens of seconds on a 2.4GHz Intel Xeon. We began with a synthetic example where each image started as a common uniform(0,1) noise field, to which independent uniform(0,1) noise was added (Fig. 2). Figs. 3, 4, and 5 show registration results for T1-, T2-, and PD-weighted MRIs of the same brain (images from http://www.bic.mni.mcgill.ca/brainweb). To combat the spatially nonhomogeneous statistics induced by the black background, the images in Fig. 4 and 5 were thresholded at 2 out 256 gray levels, and the supra-threshold mask was regularized by morphologically closing and then opening by a 3-pixel disk, with the background of the resulting masked image indicated by a checkerboard. Histograms were then computed using only non-background pixels. In contrast to the concave dependency of L_{partial} and L_{full} on distribution p, observe that the landscapes are highly non-concave functions of translation T. Thus it will be difficult to significantly increase the capture range of standard methods that locally search for T. Unsurprisingly, local search usually finds only local optima.

References

1. J. August and T. Kanade. The role of non-overlap in image registration. Technical report, The Robotics Institute, Carnegie Mellon University, 2005.
2. L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecoogy. *Bull. Amer. Math. Soc.*, 73:360–363, May 1967.
3. L. E. Baum and G. R. Sell. Growth transformations for functions on manifolds. *Pacific J. Math.*, 27:211–227, 1968.
4. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge UP, 2004.
5. D. L. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46:R1–R45, 2001.
6. M. Pelillo. The dynamics of nonlinear relaxation labeling processes. *J. Math. Imaging and Vision*, 7:309–323, 1997.
7. A. Roche, G. Malandain, and N. Ayache. Unifying maxium likelihood approaches in medical image registration. *Intl. J. Imaging Syst. Technol.*, 11:71–80, 2000.

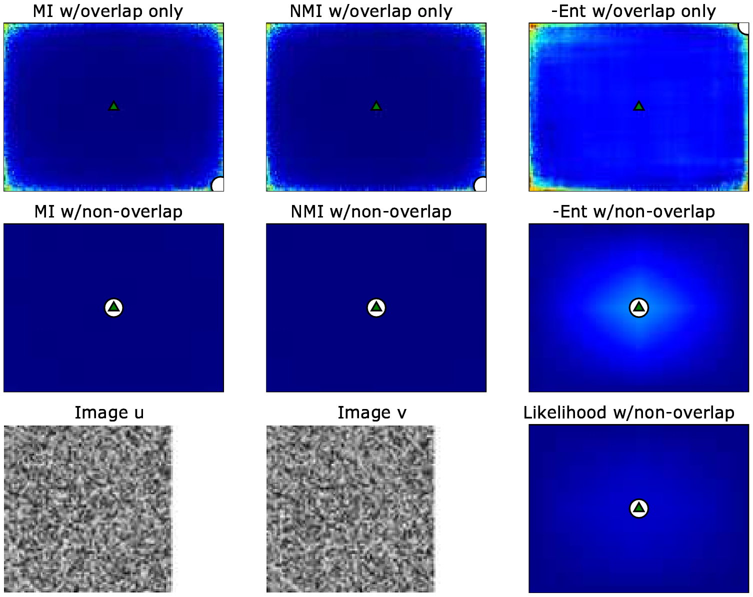

Fig. 2. A correlated pair of uniform noise images (bottom left). Registration landscapes (other images) indicate confidence (via color: blue=low, red=high) in a translation represented by (horizontal,vertical) position of the colored pixel. Top row shows registration landscapes for criteria computed using joint distribution of overlapping voxels only (left to right: mutual information (MI), normalized mutual information (NMI), and L_{full}=weighted negative entropy (-Ent)). The same criteria are shown in middle row, except non-overlapping pixels were also included to compute joint distribution p via minimization of L_{partial} (see §4). Bottom right shows evaluation of optimized $L^*_{\text{partial}}(T)$ at translation T. The white circle and green triangle indicate the computed global optimum of landscape and ground truth translation, respectively. Observe that criteria computed using only the overlap (top) incorrectly have global optima in the corners due to spurious responses from small sample effects, i.e., the image pair overlaps by only a few pixels near the corners of these landscapes. The non-overlapping pixels help make criteria calculations more reliable (middle row and bottom right), so that the global optima are correct. *For all landscapes in this paper, the green triangle hides a spike with a local optimum, but only for the non-overlap-based landscapes is this also a global optimum.*

8. A. Rosenfeld, R. A. Hummel, and S. W. Zucker. Scene labeling by relaxation operations. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(6):420–433, 1976.
9. C. Studholme, D. L. G. Hill, and D. J. Hawkes. An overlap invariant entropy meausre of 3d medical image alignment. *Pattern Recognition*, 32:71–86, 1999.
10. P. Viola and W. M. W. III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

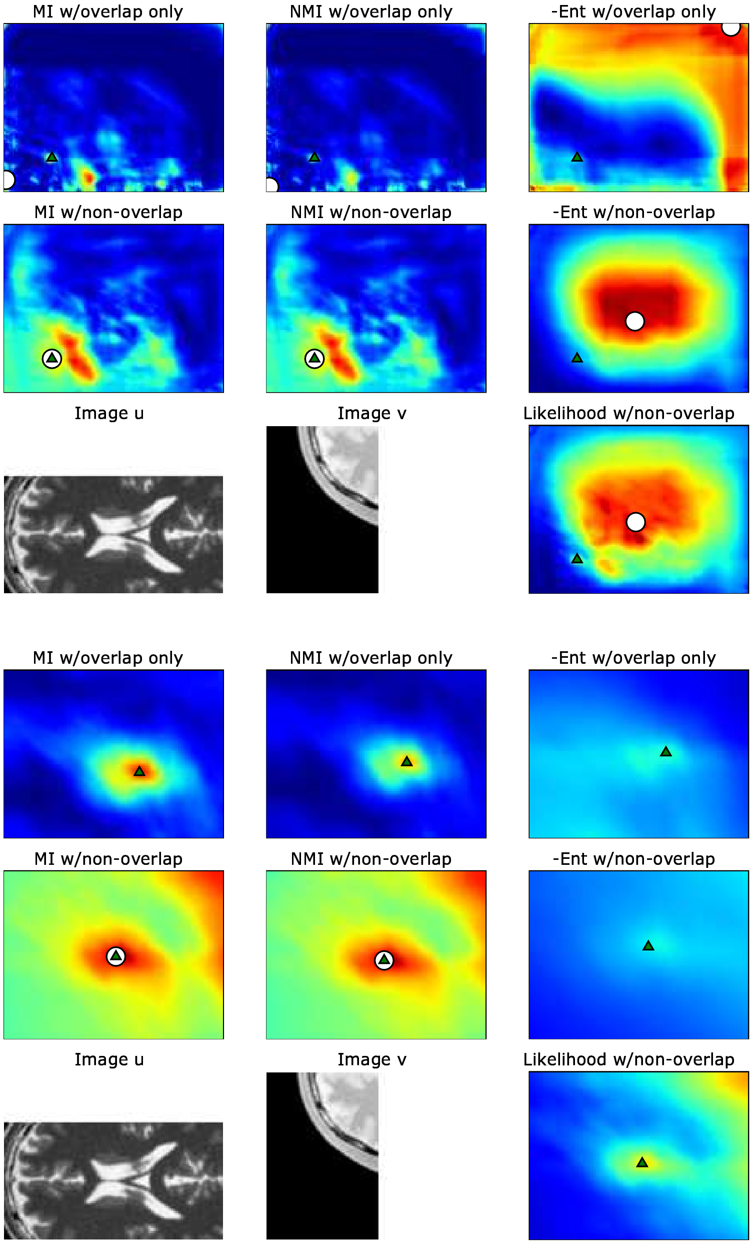

Fig. 3. T2- and PD-weighted MRIs images of human head with restricted ROIs. Top 3x3 image grid shows entire registration landscapes (see Fig. 2 for explanation), while bottom 3x3 grid shows zoom of vicinity of ground truth showing nearby local peak. Most current registration methods use a local search strategy for finding this peak. However, the many spurious peaks, especially in overlap-only landscapes, confound local search unless a close initial guess is provided. The large background in image v is a major violation of the homogeneity/IID assumption. Fig. 5 reduces these artifacts by automatically masking out the background.

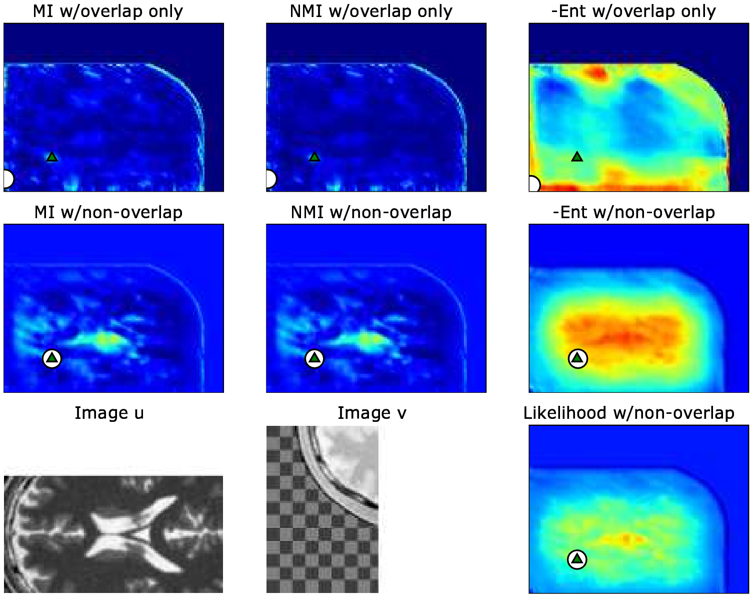

Fig. 4. T2- and PD-weighted MRIs images of human head with restricted ROIs, *with background elimination* via thresholding. Checkerboard indicates background, which was masked out of histogram computations to ensure greater statistical spatial homogeneity. Registration landscapes using non-overlap pixels (middle row and bottom right) have correct global optima; overlap-only landscapes still have only correct local optima.

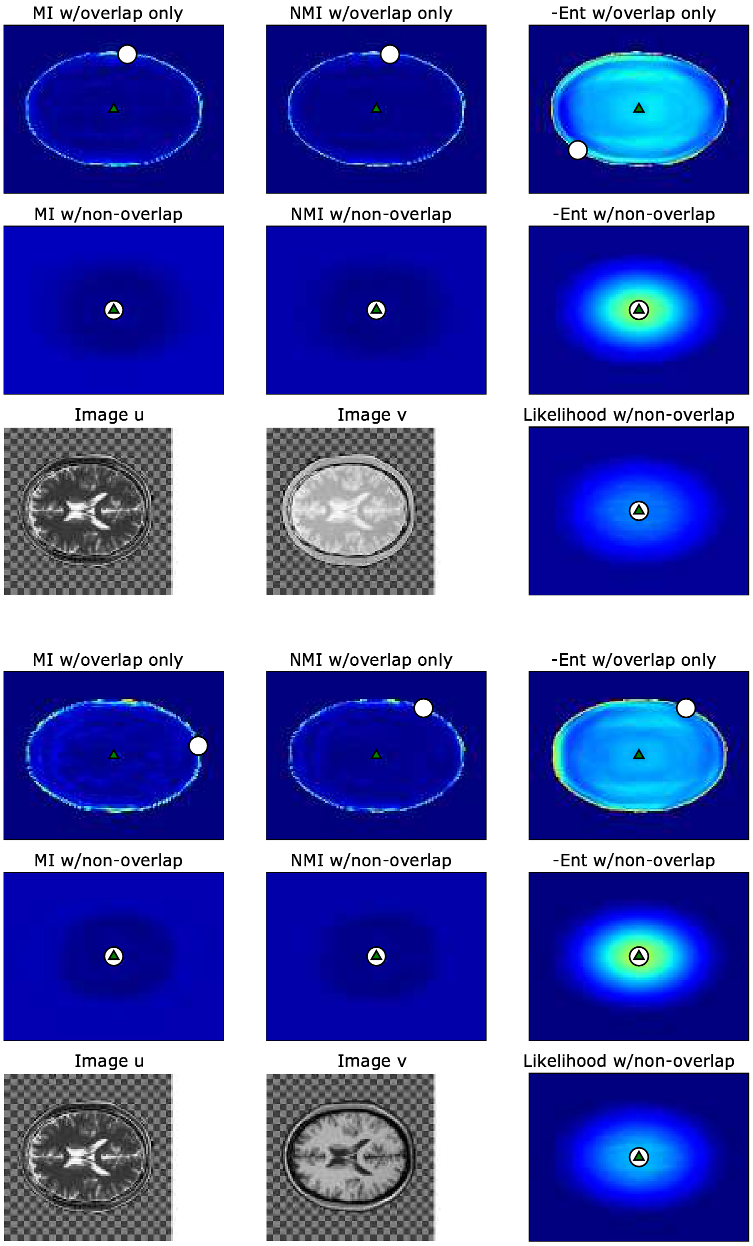

Fig. 5. T2- and PD-weighted MRIs (top 3x3 grid) and T2- and T1-weighted MRIs (bottom 3x3 grid), with full ROI and background-elimination. The spurious global optima for overlap-only landscapes occur on the outer rim of the Minkowski sum of the masks of the two brain regions and are due to small sample effects, similar to the noisy corner/border responses in Fig. 2. Registration landscapes using non-overlap pixels here also have correct global optima, as well as much smoother response in the periphery.