

# Spatial Bias in Multi-Atlas Based Segmentation

Hongzhi Wang

Paul A. Yushkevich

Penn Image Computing and Science Lab, Department of Radiology, University of Pennsylvania

## Abstract

*Multi-atlas segmentation has been widely applied in medical image analysis. With deformable registration, this technique realizes label transfer from pre-labeled atlases to unknown images. When deformable registration produces error, label fusion that combines results produced by multiple atlases is an effective way for reducing segmentation errors. Among the existing label fusion strategies, similarity-weighted voting strategies with spatially varying weight distributions have been particularly successful. We show that, weighted voting based label fusion produces a spatial bias that under-segments structures with convex shapes. The bias can be approximated as applying spatial convolution to the ground truth spatial label probability maps, where the convolution kernel combines the distribution of residual registration errors and the function producing similarity-based voting weights. To reduce this bias, we apply a standard spatial deconvolution to the spatial probability maps obtained from weighted voting. In a brain image segmentation experiment, we demonstrate the spatial bias and show that our technique substantially reduces this spatial bias.*

## 1. Introduction

Atlas-based segmentation has been widely applied in medical image analysis. This technique applies example-based knowledge representation, where the knowledge for segmenting a structure of interest is represented by a pre-labeled atlas. Through establishing one-to-one correspondence between a target image and an atlas image by image-based deformable registration, the segmentation label can be transferred to the target image from the atlas.

Segmentation errors produced by atlas-based segmentation are mostly due to registration errors. One effective way to reduce such error is to use multiple atlases. When multiple atlases are available, each atlas produces one candidate segmentation for the target image and the final segmentation is obtained through label fusion that integrates the results obtained from referring to different atlases. Recently, the label fusion technique has been applied in computer vision for segmenting natural images as well [19, 12].

Many label fusion methods are based on weighted voting, where each atlas contributes to the final solution according to a weight. For instance, majority voting [15, 9] applies equal weights to every atlas. The STAPLE algorithm [23, 14] is related to majority voting but is more advanced by taking the segmentation qualities into consideration for label fusion. One key limitation of these two methods is that they make decisions purely based segmentations and completely ignores the information conveyed by images. As recent studies have shown that, as a good indicator of registration accuracy, image similarity between the atlas and target should be included to improve voting weight assignment for more accurate label fusion. Among image similarity-based weighted voting methods, those that derive weights from local image similarity, and thus allow the weights to vary spatially, have been most successful in practice [1, 10, 16, 3, 25, 22].

Our key contribution is to identify and describe the pattern of a spatial bias produced by weighted voting based label fusion. Under mild assumptions, we show that due to the errors in deformable registration, label fusion via image similarity based weighted voting can be modeled as applying a spatial convolution operation to the ground truth spatial label posterior probability maps. The convolution kernel can be approximated by a function that combines the distribution of residual registration errors and the function transferring image similarities into voting weights. Due to this spatial bias, weighted voting based label fusion tends to produce segmentations underestimating the volumes of structures with convex shapes.

To reduce this bias, we apply standard spatial deconvolution to the fused label posterior maps after properly estimating the convolution kernel. In a brain image segmentation application that segments the hippocampus from magnetic resonance images (MRI), we demonstrate the spatial bias produced by majority voting and one recent similarity-based local weighted voting approach. The spatial bias is prominently reduced by the deconvolution method.

To improve multi-atlas segmentation, an alternative solution is to improve the accuracy of image-based registration. However, since it is almost impossible to produce error free image-based registration, the spatial bias induced by regis-

tration error is inevitable. Hence, our results have utility regardless of how accurate the applied registration algorithm is. In our experiment, we used Symmetric Normalization (SyN) [2], which was a top performer in a recent evaluation study [11] comparing 14 freely available deformable registration algorithms, and the spatial bias could still be clearly observed.

## 2. Label fusion by weighted voting

In this section, we briefly describe the weighted voting techniques used in label fusion. Let  $T_F$  be a target image to be segmented and  $A^1 = (A_F^1, A_S^1), \dots, A^n = (A_F^n, A_S^n)$  be  $n$  atlases, warped to the space of the target image by deformable registration.  $A_F^i$  and  $A_S^i$  denote the  $i_{th}$  warped atlas image and manual segmentation. Each  $A_S^i$  is a candidate segmentation for the target image. Label fusion combines these candidate segmentations to produce the final solution.

Many label fusion methods are based on weighted voting. For instance, the combined votes for label  $l$  are:

$$\hat{p}(l|x, T_F) = \sum_{i=1}^n w_x^i p(l|x, A^i) \quad (1)$$

where  $x$  indexes through image locations.  $\hat{p}(l|x, T_F)$  is the estimated label posterior for the target image.  $p(l|x, A^i)$  is the probability that  $A^i$  votes for label  $l$  at  $x$ , with  $\sum_{l \in \{1, \dots, L\}} p(l|x, A^i) = 1$ .  $L$  is the total number of labels. Typically, for deterministic atlases that have one unique label for every location,  $p(l|x, A^i)$  is 1 if  $l = A_S^i(x)$  and 0 otherwise.  $w_x^i$  is a local weight assigned to the  $i_{th}$  atlas, with  $\sum_{i=1}^n w_x^i = 1$ . The voting weights are typically determined based on the quality of registration produced by each atlas such that more accurately registered atlases are weighted more heavily in producing the final solution. To estimate this weight, similarity metrics typically employed by image-based registration such as sum of squared distance (SSD) and normalized cross correlation (NCC) can be applied. For instance, when SSD and a Gaussian weighting model are used [16]<sup>1</sup>, the weights can be estimated by:

$$w_x^i = \frac{1}{Z(x)} \exp \left( - \sum_{y \in \mathcal{N}(x)} [A_F^i(y) - T_F(y)]^2 / \sigma \right) \quad (2)$$

where  $\sigma$  is a model parameter.  $\mathcal{N}(x)$  defines a neighborhood around  $x$  and  $Z(x)$  is a normalization constant. In our experiment, we use a  $(2r+1) \times (2r+1) \times (2r+1)$  cube-shaped neighborhood specified by the radius  $r$ . For robust image matching, instead of using the raw image intensities,

<sup>1</sup>[16] proposes a general method with multiple specific implementations. Here, we refer to the ‘‘local weighted voting’’ implementation, i.e. equation (6) and (10) in [16].

the intensity vector obtained from each local image intensity patch is normalized to have zero mean and a constant norm 1.

With the estimated posterior probability map for each label, the final solution is determined by selecting the label with the highest posterior at each voxel.

**Refining label fusion by local patch search.** As recently shown in [21, 4], the performance of atlas-based segmentation can be improved by applying a local search technique. This method also uses image similarities over local image patches as indicators of registration accuracy and remedies registration errors by searching for the correspondence that gives the most similar appearance matching, within a small neighborhood around the registered correspondence in each warped atlas. The locally searched optimal correspondence is:

$$\xi^i(x) = \arg \min_{x' \in \mathcal{N}'(x)} \|A_F^i(\mathcal{N}(x')) - T_F(\mathcal{N}(x))\|^2 \quad (3)$$

$\xi^i(x)$  is the location in  $i_{th}$  atlas with the best image matching for location  $x$  in the target image within the neighborhood  $\mathcal{N}'(x)$ . Again, we use a cubic neighborhood definition, specified by a radius  $r_s$ . Given the set of local search correspondence maps  $\{\xi^i\}$ , we estimate the label posterior by  $\hat{p}(l|x, T_F) \propto \sum_{i=1}^n w_{\xi^i(x)}^i p(l|\xi^i(x), A^i)$ .

## 3. Spatial bias induced by registration errors

In this section, we describe a spatial bias in weighted voting based label fusion and develop solutions to reduce this bias.

### 3.1. Residual spatial displacement errors resulted from image-based registration

Errors produced by atlas-based segmentation are mainly due to registration errors, i.e. the correspondence computed from registration is incorrect. Hence, characterizing the influence of the residual registration error is critical for understanding the bias in label fusion.

Let  $y^i$  be the residual spatial displacement error in the warped atlas  $A^i$  such that the correct correspondence for location  $x$  in  $A^i$  is  $x - y_x^i$  in  $T_F$ . Hence, we have:

$$p(l|x, A^i) = p(l|x - y_x^i, T_F) + \epsilon(x) \quad (4)$$

where  $\epsilon$  is a random error caused by effects such as partial volume, errors and random variation in atlas segmentations. We use  $L_1$  norm to quantify the residual spatial displacement error, i.e.  $\|y_x^i\|_1 = \sum_{j=1}^3 |y_x^i(j)|$  for three-dimensional images.

The residual spatial displacement error can be modeled by a random variable, characterized by a distribution  $p_D$ . When image-based registration is reliably performed, the expected residual spatial displacement error is small.

### 3.2. Bias induced by residual registration errors

As shown in recent studies, deriving voting weights from image similarities over local patches is one of the most effective weighted voting strategies. Our study starts with investigating how well the residual registration error correlates with the local image similarity.

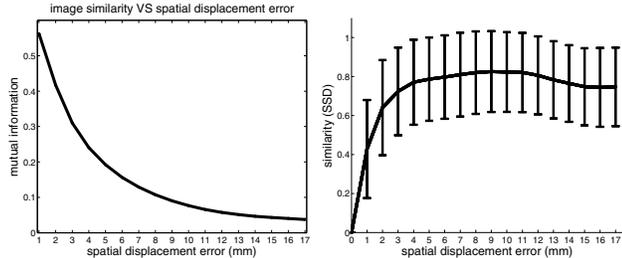


Figure 1. Relationship between spatial displacement errors and local image similarities. The plotted results are averaged over statistics extracted from 20 brain MR images used in our experiments. In this test, for each voxel within a manually label hippocampus in one image, we computed the SSD between the normalized image patch ( $r=2$ ) extracted at this voxel and normalized image patches extracted at voxels within its neighborhood. The left figure shows the mutual information between spatial errors, i.e. the distance between the two voxels, and local image similarities at different spatial errors. Given a spatial error specified in the X-axis, the mutual information is computed using all pairs of local patches with distances no greater than the error. To quantify the entropy of local image similarities, we uniformly divide the image similarity value range  $[0,4]$  into 100 bins. The right figure plots the average local image similarity at different spatial errors (bars at  $\pm 1$  s.d.). Note that when spatial errors are small (i.e.  $\leq 4$ mm), spatial errors strongly correlate to local image similarities. This strong correlation quickly diminishes as the spatial error increases.

Strong image similarities over small patches between two registered images do not necessarily indicate small spatial displacement errors. However, as shown in Fig. 1, when spatial errors are small enough, image similarities over local patches are strongly correlated with spatial displacement errors, with high mutual information and less ambiguous relation between spatial errors and image similarities. Since most residual spatial errors are small after image-based registration, we can approximately represent the voting weight estimated from the local image similarity as a function of the residual registration error. Hence, we have:

$$w_x^i \approx W(I(y_x^i)) \quad (5)$$

where  $I$  approximates the local image similarity given the spatial error and  $W$  is the weighting function that transfers local image similarities into voting weights. In fact, approximating spatial displacement errors by local image similarities has been applied by most image-based registration and similarity-based local weighted voting methods. An example of  $I$  returns the average image similarity or the most

likely image similarity given a spatial error. Since an ideal weighting function assigns smaller weights to poorly registered atlases,  $W(I(y))$  should be a decreasing function of  $\|y\|_1$ . However, we do not make this assumption.

By substituting (4) and (5) into (1), we have:

$$\hat{p}(l|x, T_F) \propto \sum_{i=1}^n W(I(y_x^i)) [p(l|x - y_x^i, T_F) + \epsilon_x^i] \quad (6)$$

$$\approx \int_y W(I(y)) p(l|x - y, T_F) p_D(y) dy + \epsilon \quad (7)$$

$$= [w_D \otimes p(l|T_F)](x) + \epsilon \quad (8)$$

where  $\otimes$  is discrete convolution and  $w_D(y) = W(I(y))p_D(y)$  is the convolution kernel, approximating the expectation of the voting weight given the residual registration error. (7) is obtained because the residual registration error produced by each atlas is a random sample from the residual registration error distribution  $p_D$ .

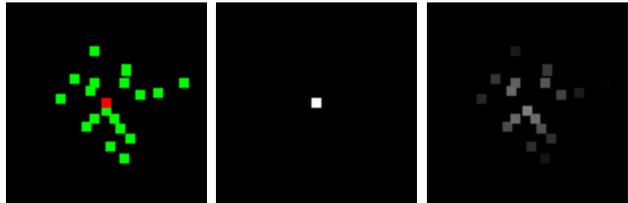


Figure 2. Illustration of the spatial bias. Left image: the red dot is the target voxel to be labeled. Green dots represent warped corresponding voxels from different atlases. The distance,  $d$ , between the red dot and a green dot quantifies a registration error. The label posterior map on the right is produced by assigning voting weights  $\propto e^{-d}$ , which can be approximately modeled as spreading the ground truth label posterior map (middle) to the neighborhood.

Due to the error in deformable registration, the estimated label posterior probability map produced by image similarity-based weighted voting can be modeled as applying spatial smoothing to the ground truth label posterior probability map (see Fig. 2 for an illustration). The convolution kernel is determined by the registration quality,  $p_D$ , and the weighting function  $W(I)$ . Hence, when different registration algorithms or different weighting functions are used, the convolution kernel may be different. In a special case, majority voting applies a constant function for  $W(I)$ . Hence, the resulting convolution kernel is identical to  $p_D$ . In the remaining of this paper, we focus our study on two weighted voting methods: majority voting and the Gaussian-based local weighted voting method (2).

It is well known that spatial smoothing has stronger impacts near boundaries, resulting in under/overestimated label posteriors at edges for convex/concave shaped structures. Without correcting this spatial bias, weighted voting tends to produce less accurate results for thin (or high-curvature shaped) structures. Fig. 4 and Fig. 5 demonstrate

such bias produced by majority voting and by Gaussian-based local weighted voting.

To reduce this spatial bias, we propose to estimate the convolution kernel  $w_D$  and apply standard deconvolution to restore the unbiased spatial label posterior maps from the estimation obtained via weighted voting.

### 3.3. Characterizing the convolution kernel

The convolution kernel,  $w_D$ , captures the relationship between the residual registration error and the expected voting weight received in weighted voting. Accurately estimating the kernel is difficult due to the difficulty in quantifying the residual registration errors. To obtain information on how to characterize the convolution kernel, we conducted empirical studies to measure the relationship between the lower bound of the residual registration error and the expected voting weight produced by Gaussian weighting<sup>2</sup>.

To simplify our study and implementation, without enforcing any priors, we assume that the smoothing kernel is isotropic, i.e.  $w_D(y) = w_D(y')$  if  $\|y\|_1 = \|y'\|_1$ . The assumption is particularly suitable for brain images, where most regions have homogenous intensities. Although the smoothing kernels may be less isotropic around regions with strong edges, the kernels are approximately isotropic averaging over large regions.

We conducted studies on the set of atlases used in our hippocampus segmentation experiments (data description in section 4). For one atlas  $A = (A_F, A_S)$ , we register another atlas to it and let  $B = (B_F, B_S)$  be the corresponding warped atlases obtained from the registration. Under the assumption that the manual segmentations for both atlases are correct and correct correspondence exists, the manual segmentations provide information about the registration error. For example, if  $A_S(x) \neq B_S(x)$  then the correspondence at  $x$  is incorrect. Although retrieving accurate residual registration errors is still difficult, the shortest distance from  $x$  to the regions with the same label of  $B_S(x)$  in  $A$  defines a lower bound for the residual registration error at  $x$ <sup>3</sup>. Similarly, the shortest distance from  $x$  to the regions with the same label of  $A_S(x)$  in  $B$  defines another lower bound for the residual registration error at  $x$ . A more accurate lower bound can be determined by selecting the maximal value from the two estimations.

Approximating the residual registration error by its lower bound, we collected the statistics of the joint occurrence of the residual registration error and the image similarity-based Gaussian weight by registering each pair of the atlases. With one pair of registered atlases ( $A^i, A^j$ ), the accumulative voting weights received at one residual registration

<sup>2</sup>Note that majority voting is a special case of Gaussian-based weighted voting, with  $\sigma \rightarrow \infty$ .

<sup>3</sup>The longest distance from  $x$  to regions with the same label of  $B_S(x)$  in  $A$  defines an upper bound for the registration error at  $x$ .

error  $R$  are:

$$\begin{aligned} \sum_{\{x\|y_x\|_1=R\}} w_x &\propto \int_{\{y\|y\|_1=R\}} w_D(y) dy \\ &= w_D(R) \int_{\{y\|y\|_1=R\}} 1 dy \propto w_D(R) 4\pi R^2 \end{aligned} \quad (9)$$

where  $w_x = e^{-\|A_F^i(\mathcal{N}(x)) - A_F^j(\mathcal{N}(x))\|^2/\sigma}$  is the Gaussian weight assigned at location  $x$  and  $y_x$  defines the lower bound of the spatial displacement error at  $x$ .  $4\pi R^2$  in (9) is derived because  $w_D$  is a 3-dimensional kernel.

To estimate the expected weight  $w_D(R)$  at the residual registration error  $R$ , we normalize the accumulative weights received at  $\|y\|_1 = R$  by dividing by  $R^2$ . Averaging over all pairs of the atlases, we obtain the expected voting weights with respect to the lower bound of residual registration errors. Fig. 3 shows empirical estimations obtained from using 20 atlases and a Gaussian weighting function with  $r = 2, \sigma = 0.05, 0.1$ , respectively.

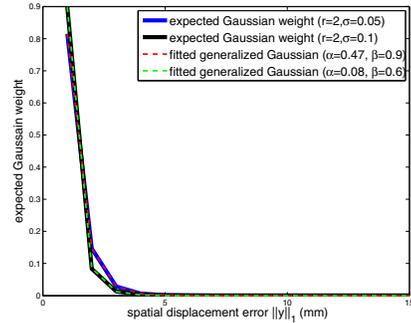


Figure 3. Empirical relation between the lower bound of the residual spatial displacement error and the expected voting weight produced by Gaussian weighting with  $r = 2, \sigma = 0.05, 0.1$ . See text for details. The weights are normalized such that  $\sum_R w_D(R) = 1$ . The empirical functions (solid curves) can be accurately fitted by generalized Gaussian functions (dashed curves).

In our study, the expected Gaussian weight and the lower bound of the residual registration error show an exponential relationship. Hence, we apply a generalized Gaussian function to model this empirical relationship. We have:

$$w_D(y) \sim e^{-(\|y\|_1/\alpha)^\beta} \quad (10)$$

where  $\alpha$  and  $\beta$  are model parameters, which can be determined through least square fitting and gradient descent optimization. As Fig. 3 shows that generalized Gaussians fit very well to the empirical functions.

### 3.4. Estimating the convolution kernel

We found that applying the convolution kernel estimated using the lower bound of the residual registration error to reduce the spatial bias already improved the image segmentation accuracy produced by Gaussian-based weighted voting.

However, since the lower bound may be significantly different from the actual residual registration error and the above empirical estimation does not consider the local search technique used for remedying registration errors, room is still left for further improving the kernel estimation accuracy. Assuming that the optimal convolution kernel can be modeled by a generalized Gaussian function as well, we estimate the optimal convolution kernel using the set of atlases in a leave-one-out fashion as follows.

First, the spatial label posterior maps of each atlas is estimated using weighted voting from the remaining atlases. We determine the optimal parameters of the generalized Gaussian kernel by separately applying deconvolution using a range of parameter values specifying the kernel. The parameters producing the most accurate segmentation for all atlases are selected. Since in our empirical studies, the fitted generalized Gaussian function using the lower bound of the residual registration error all have  $0 \leq \alpha, \beta \leq 1$ , we estimate the convolution kernel by selecting parameters from a discrete range  $\alpha, \beta \in \{0.02, 0.04, \dots, 1\}$ .

#### 4. Experiments

In this section, we apply our method to segment the hippocampus using T1-weighted MRI. The hippocampus plays an important role in memory function. Macroscopic changes in brain anatomy, detected and quantified by MRI, consistently have been shown to be predictive of Alzheimers disease (AD) pathology and sensitive to AD progression [17]. Accordingly, automatic hippocampus segmentation from MRI has been widely studied.

**Imaging data.** We use the data in the Alzheimer’s Disease Neuroimaging Initiative (ADNI, www.loni.ucla.edu/ADNI). Our study is conducted using 3 T MRI and only includes data from mild cognitive impairment (MCI) patients and controls. Overall, the data set contains 139 images (57 controls and 82 MCI patients). The images are acquired sagittally, with  $1 \times 1$  mm in-plane resolution and 1.2 mm slice thickness.

To obtain reference segmentation, we first applied a landmark-guided atlas-based segmentation method [13] to produce the initial segmentation for each image. Each fully-labeled hippocampus was manually edited by a trained human rater following a previously validated protocol [8].

**Experiment setup.** For cross-validation evaluation, we randomly selected 20 images to be the atlases and another 20 images for testing. Each atlas was registered to each test image, as well as to each other atlas. Global registration was performed using the FSL FLIRT tool [18] with six degrees of freedom and using the default parameters (normalized mutual information similarity metric; search range

	CTL left	MCI left	Cohen’s $d$
MV	1996 $\pm$ 339	1705 $\pm$ 355	1.3903
MVDecon	2177 $\pm$ 300	1880 $\pm$ 348	1.6313
LWGaussian	2181 $\pm$ 285	1770 $\pm$ 367	1.9625
GauDecon	2273 $\pm$ 285	1859 $\pm$ 368	1.9920
manual	2297 $\pm$ 321	1835 $\pm$ 379	1.9634
	CTL right	MCI right	Cohen’s $d$
MV	1921 $\pm$ 297	1645 $\pm$ 312	1.5096
MVDecon	2102 $\pm$ 289	1825 $\pm$ 319	1.6088
LWGaussian	2102 $\pm$ 341	1672 $\pm$ 362	1.8073
GauDecon	2194 $\pm$ 338	1760 $\pm$ 368	1.8369
manual	2230 $\pm$ 390	1762 $\pm$ 416	1.7425

Table 1. Hippocampal volume ( $\text{mm}^3$ ) measured by different methods for control and MCI cohorts. The results for left side and right side hippocampi are given separately and are averaged over 5 cross-validation experiments, which together include test images from 41 control subjects and 59 MCI subjects. The third row shows the corresponding Cohen’s  $d$  effect size. The hippocampus volume is normalized by intracranial volume for computing the Cohen’s effect size.

from -5 to 5 in x, y and z). Deformable registration was performed using the ANTS Symmetric Normalization (SyN) algorithm [2], with the cross-correlation similarity metric (with radius 2) and a Gaussian regularizer with  $\sigma = 3$ . The cross-validation experiment was repeated 5 times. In each experiment, a different set of atlases and test images were randomly selected from the ADNI dataset. The results reported below are averaged over the 5 experiments.

To test our method, first, majority voting (MV) and the Gaussian-based local weighted voting label fusion (2) (LWGaussian) were separately applied to produce the initial spatial label posterior maps for each of the test image. The parameters for LWGaussian are  $r = 2$ ,  $r_s = 3$ ,  $\sigma = 0.1$ . We then applied spatial deconvolution with the kernel estimated using the atlases for each label fusion method to reduce the bias in the estimated label posterior maps.

As discussed above, the optimal parameters of the generalized Gaussian function used for deconvolution were obtained by evaluating a range of values ( $\alpha, \beta \in \{0.02, 0.04, \dots, 1\}$ ) using the atlases in a leave-one-out strategy as well. For deconvolution, we applied Wiener deconvolution [6] implemented in MATLAB, the *deconvwnr* function with a constant noise-to-signal power ratio 0.2.

**Results.** Table 1 presents the hippocampal volumes in control and MCI cohort obtained using different segmentation methods. As expected, both majority voting and similarity-based local weighted voting underestimated the hippocampal volume than the reference segmentation. The average absolute volume difference rates,

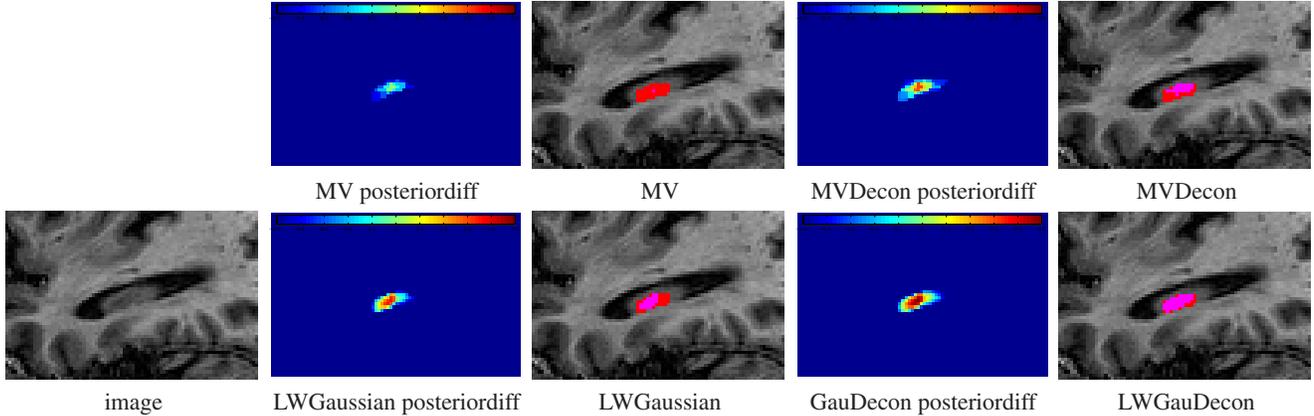


Figure 4. An example of applying deconvolution for label fusion showing in a sagittal cross-section view. From left to right: image, spatial label posterior of the hippocampus subtracted by the posterior of the background and segmentation of the hippocampus obtained from majority voting (upper) and Gaussian-based local weighted voting (lower), spatial label posterior of the hippocampus subtracted by that of the background and segmentation obtained after applying deconvolution with a generalized Gaussian kernel, ( $\alpha = 0.04, \beta = 0.34$ ) for MV and ( $\alpha = 0.1, \beta = 0.4$ ) for LWGaussian, to the spatial label posterior maps produced by the corresponding label fusion method. Red: manual segmentation; Blue: automatic segmentation; Pink: overlap between manual and automatic segmentation.

$\frac{|\text{auto} - \text{reference}|}{|\text{reference}|}$ , produced by the two methods are 13.6% and 6.1%, respectively. Applying spatial deconvolution produced more accurate volume measurements for both label fusion methods.

The corresponding Cohen’s  $d$  effect size [7] is also shown in Table 2 (computed as the difference of the sample means of the two cohort, divided by the pooled sample standard deviation). Larger values of Cohen’s  $d$  indicate greater effect, i.e., greater ability to tell cohorts apart based on hippocampal volume. The hippocampal volumes obtained from applying spatial deconvolution also have slightly more significant differences between the two population groups than those produced by the corresponding label fusion method, as indicated by the increased effect sizes. Since volume differences produced by automatic segmentation methods are all proportional to that of manual segmentation, the hippocampus volume measured using different methods show similar separability between the two population groups. Automatic algorithms yield slightly greater effect sizes than manual segmentation, likely due to reduced variance in volume estimation.

method	left	right
MV	$0.878 \pm 0.044$	$0.871 \pm 0.036$
MVDecon	$0.890 \pm 0.030$	$0.878 \pm 0.039$
LWGaussian	$0.900 \pm 0.026$	$0.890 \pm 0.030$
GauDecon	$0.904 \pm 0.024$	$0.896 \pm 0.029$

Table 2. The performance (Dice) produced by each method. The results for the left and right side hippocampi are shown separately.

Table 2 shows the results in terms of Dice ratio

$[5] \left( \frac{2|A \cap B|}{|A| + |B|} \right)$  produced by each method. Our method improved the performance of majority voting and LWGaussian by  $\sim 1\%$  and  $\sim 0.5\%$  Dice ratio, respectively. Our improvement is statistically significant, with  $p < 0.01$  on the paired Student’s t-test for each cross-validation experiment.

Fig. 4 shows one example produced by majority voting and LWGaussian and by subsequently applying spatial deconvolution. For this example, both majority voting and LWGaussian produced smaller volume for the hippocampus. Applying deconvolution effectively reduced such bias.

**Visualizing the spatial bias.** To visualize and quantify the spatial bias produced by majority voting and LWGaussian across all subjects, we normalize the different hippocampus segmentations into a common coordinate space. Normalization is performed using a structure-specific shape-based normalization approach [24]. A geometrical model, known as the continuous medial representation (cm-rep), is fitted to each reference segmentation of the hippocampus, and the parametrization of this geometrical model is used to establish a one-to-one correspondence between the space inside and near the reference segmentation and a common reference space provided by a single template manual segmentation. Using these correspondence maps, we transfer both the reference segmentations and the automatic segmentations from subject space into the template space. We emphasize that since the same mapping is applied to both reference and automatic segmentations, the differences between these segmentations are maintained by the mapping. Averaging over all subjects across all cross-validation experiments, we computed the spatial label distribution in the template space for the reference segmenta-

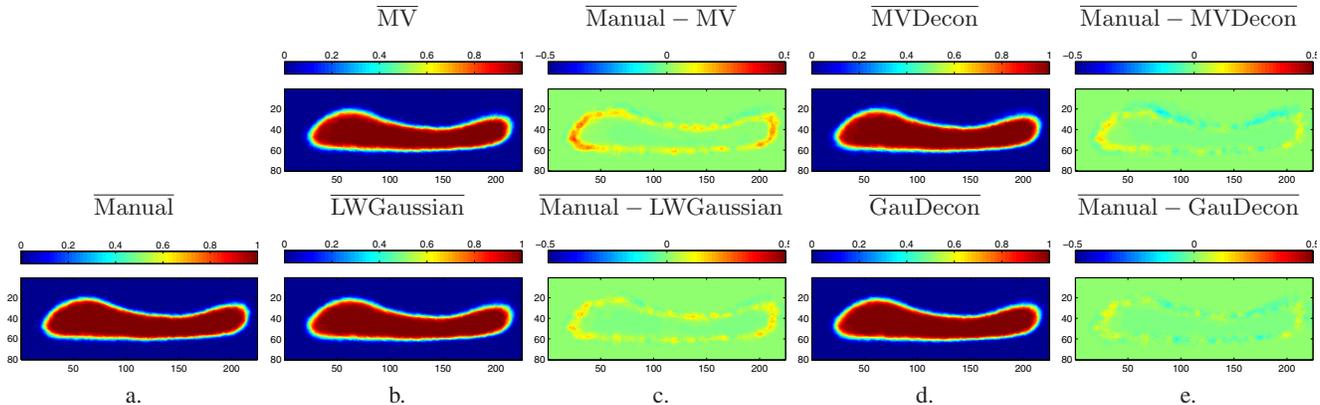


Figure 5. The spatial patterns of disagreement between automatic and reference segmentations of the hippocampus, plotted after normalization to a common reference space. All plots show a sagittal cross-section of the 3D reference space. (a). The mean of the normalized reference segmentations. (b). The mean of the segmentations produced by majority voting (upper) and LWGaussian (lower), mapped into the reference space using the same transformations as the corresponding reference segmentations. (c). Mean signed difference between reference and automatic segmentations. The spatial bias resulted from applying a spatial smoothing operation to the reference segmentation is clearly shown. (d). The mean of the segmentations obtained by applying deconvolution to the results produced by the corresponding label fusion method. (e). Mean signed difference between reference segmentations and the results obtained from applying deconvolution with the estimated kernels. The spatial bias is prominently reduced.

tions and the automatic segmentations. These distributions are shown in Fig. 5. Note that a cm-rep model fitted to a reference segmentation does not overlap it perfectly. Hence, the mean spatial label distribution of the reference segmentations in the template space is not a binary image.

The plot of mean signed difference between the normalized reference and automatic segmentations in Fig. 5(c) clearly shows the strongest spatial bias of underestimation of the figure produced by majority voting and LWGaussian in the anterior and posterior regions of the hippocampus, which are the parts in the hippocampus with the largest curvatures. Such bias is prominently reduced after applying deconvolution with the estimated kernels, shown in Fig. 5(e).

Note that the quality of image-based registration usually varies at different locations. For regions with more distinctive image features, the image-based registration usually can be more reliably conducted than regions with less distinct image features. As a result, the approximated convolution kernel for one weighted voting method may vary at different locations as well. As shown in Fig. 5(e), applying a single kernel to the entire image reduced most spatial bias, while the applied kernel is inadequate to reduce all the spatial bias. Hence, by spatially adapting the applied deconvolution kernel, one may further improve the performance.

## 5. Discussion and Conclusions

We described and demonstrated a spatial bias in similarity-based weighted-voting label fusion that may underestimate the volumes of convex structures. Due to registration errors, weighted voting imposes an effect that can be

modeled as applying a spatial convolution operation to the ground truth label posterior maps. The convolution kernel combines the distribution of residual spatial displacement errors resulted from registration and the weighting function used to transfer local image similarities into voting weights. To reduce this spatial bias, we proposed to apply standard spatial deconvolution to the label posterior maps obtained from weighted voting.

In a hippocampus segmentation application, we demonstrated the spatial bias produced by majority voting and Gaussian-based local weighted voting label fusion. We also showed that the convolution kernel resulted from Gaussian local weighted voting can be modeled by generalized Gaussian functions. Fitting the deconvolution kernel using atlases in a leave-one-out fashion, we showed that applying spatial deconvolution effectively reduced the spatial bias produced by majority voting and Gaussian weighted voting.

### Relation to other bias reduction work in multi-atlas segmentation.

In addition to the spatial bias described in this paper, weighted voting may produce other types of bias. For example, another source of bias is due to the redundancies in the atlases [22, 21]. For a simple example, suppose that a single atlas is duplicated multiple times in the atlas set. If voting weights are derived only from atlas-target similarity, the total contribution of the repeated atlas to the final solution will increase in proportion to the number of times the atlas is repeated, biasing the solution towards the repeated atlases. Since the two types of bias are complementary to each other, it is possible to apply spatial deconvolution on

the spatial label posterior maps produced by [22, 21] to reduce both types of bias in weighted voting.

As shown in [20], machine-learning techniques can be useful for reducing systematic errors produced by a segmentation approach. Our work suggests that when multi-atlas segmentation with similarity-based weighted voting label fusion is applied as the host method, the estimated spatial label posterior maps contain meaningful information for learning how to correct the bias.

## Acknowledgement

This work was supported by the Penn-Pfizer Alliance grant 10295 (PY) and the NIH grants K25 AG027785 (PY) and R01 AG037376 (PY)

## References

- [1] X. Artaechevarria, A. Munoz-Barrutia, and C. O. de Solorzano. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *IEEE TMI*, 28(8):1266–1277, 2009. 1
- [2] B. Avants, C. Epstein, M. Grossman, and J. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12:26–41, 2008. 2, 5
- [3] D. Collins and J. Pruessner. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage*, 52(4):1355–1366, 2010. 1
- [4] P. Coupe, J. Manjon, V. Fonov, J. Pruessner, N. Robles, and D. Collins. Nonlocal patch-based label fusion for hippocampus segmentation. In *MICCAI*, 2010. 2
- [5] L. Dice. Measure of the amount of ecological association between species. *Ecology*, 26:297–302, 1945. 6
- [6] R. Gonzalez, R. Woods, and S. Eddins. *Digital Image Processing Using Matlab*. Prentice Hall, 2003. 5
- [7] J. Hartung, G. Knapp, and B. K. Sinha. *Statistical Meta-Analysis with Application*. Wiley, 2008. 6
- [8] D. Hasboun, M. Chantome, A. Zouaoui, M. Sahel, M. Deladoueille, N. Sourour, M. Duymes, M. Baulac, C. Marsault, and D. Dormont. MR determination of hippocampal volume: Comparison of three methods. *Am J Neuroradiol*, 17:1091–1098, 1996. 5
- [9] R. Heckemann, J. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33:115–126, 2006. 1
- [10] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. Viergever, and B. van Ginneken. Multi-atlas-based segmentation with local decision fusion-application to cardiac and aortic segmentation in CT scans. *IEEE Trans. on MI*, 28(7):1000–1010, 2009. 1
- [11] A. Klein. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786–802, 2009. 2
- [12] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009. 1
- [13] J. Pluta, B. Avants, S. Glynn, S. Awate, J. Gee, and J. Detre. Appearance and incomplete label matching for diffeomorphic template based hippocampus segmentation. *Hippocampus*, 19:565–571, 2009. 5
- [14] T. Rohlfing, R. Brandt, R. Menzel, and C. Maurer. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428–1442, 2004. 1
- [15] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. M. Jr. Quo vadis, atlas-based segmentation? *The Handbook of Medical Image Analysis Volume III: Registration Models*, pages 435–486, 2005. 1
- [16] M. Sabuncu, B. Yeo, K. V. Leemput, B. Fischl, and P. Goland. A generative model for image segmentation based on label fusion. *IEEE TMI*, 29(10):1714–1720, 2010. 1, 2
- [17] R. Scahill, J. Schott, J. Stevens, M. Rossor, and N. Fox. Mapping the evolution of regional atrophy in Alzheimer’s Disease: unbiased analysis of fluidregistered serial MRI. *Proc. Natl. Acad. Sci. U. S. A.*, 99(7):4703–4707, 2002. 5
- [18] S. Smith, M. Jenkinson, M. Woolrich, C. Beckmann, T. B. aand H. JohansenBerg, P. Bannister, M. Luca, I. Drobnjak, D. Flitney, R. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. Stefano, J. Brady, and P. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23(Suppl 1):S208S219, 2004. 5
- [19] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *IEEE Trans. on PAMI*, 30(8):1483–1489, 2008. 1
- [20] H. Wang, S. Das, J. W. Suh, M. Altinay, J. Pluta, C. Craige, B. Avants, and P. Yushkevich. A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, 55(3):968–985, 2011. 8
- [21] H. Wang, J. W. Suh, S. Das, J. Pluta, M. Altinay, and P. Yushkevich. Regression-based label fusion for multi-atlas segmentation. In *CVPR*, 2011. 2, 7, 8
- [22] H. Wang, J. W. Suh, J. Pluta, M. Altinay, and P. Yushkevich. Optimal weights for multi-atlas label fusion. In *IPMI*, 2011. 1, 7, 8
- [23] S. Warfield, K. Zou, and W. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE TMI*, 23(7):903–921, 2004. 1
- [24] P. Yushkevich, J. Detre, D. Mechanic-Hamilton, M. Fernandez-Seara, K. Tang, A. Hoang, M. Korczykowski, H. Zhang, and J. Gee. Hippocampus-specific fMRI group activation analysis using the continuous medial representation. *NeuroImage*, 35(4):1516–1530, 2007. 6
- [25] P. Yushkevich, H. Wang, J. Pluta, S. Das, C. Craige, B. Avants, M. Weiner, and S. Mueller. Nearly automatic segmentation of hippocampal subfields in in vivo focal T2-weighted MRI. *NeuroImage*, 53(4):1208–1224, 2010. 1