Combination Strategies in Multi-Atlas Image Segmentation: Application to Brain MR Data

Xabier Artaechevarria*, Arrate Muñoz-Barrutia, *Member, IEEE*, and Carlos Ortiz-de-Solórzano, *Senior Member, IEEE*

***Abstract*—It has been shown that employing multiple atlas images improves segmentation accuracy in atlas-based medical image segmentation. Each atlas image is registered to the target image independently and the calculated transformation is applied to the segmentation of the atlas image to obtain a segmented version of the target image. Several independent candidate segmentations result from the process, which must be somehow combined into a single final segmentation. Majority voting is the generally used rule to fuse the segmentations, but more sophisticated methods have also been proposed. In this paper, we show that the use of global weights to ponderate candidate segmentations has a major limitation. As a means to improve segmentation accuracy, we propose the generalized local weighting voting method. Namely, the fusion weights adapt voxel-by-voxel according to a local estimation of segmentation performance. Using digital phantoms and MR images of the human brain, we demonstrate that the performance of each combination technique depends on the gray level contrast characteristics of the segmented region, and that no fusion method yields better results than the others for all the regions. In particular, we show that local combination strategies outperform global methods in segmenting high-contrast structures, while global techniques are less sensitive to noise when contrast between neighboring structures is low. We conclude that, in order to achieve the highest overall segmentation accuracy, the best combination method for each particular structure must be selected.**

***Index Terms*—Atlas-based segmentation, classifier combination, combination of segmentations, majority voting, weighted voting.**

I. Introduction

THE PRINCIPLES of atlas-based segmentation have been successfully applied to a wide variety of image modalities and segmentation tasks [1]–[5]. This approach has a major advantage when compared to other segmentation algorithms, such as level sets [6] or watersheds [7]. Namely, it allows introducing *a priori* knowledge about the shape and the distribution of the segmented structures in a simple way, by using a presegmented image as a reference that guides the segmentation.

In principle, a single atlas image can be used for segmenting new images. However, it has been recently shown that using multiple atlas images can yield better results [8]–[11]. Information from several reference images can be combined into an average atlas [12], [13] or, if probability values for each particular location are included, into a so-called *probabilistic atlas* [1], [2], [4], [9]. In doing so, these atlases try to comprise all the variability of a given population. However, it has been recently suggested that, to gain full advantage of having multiple atlas images at hand, they must be registered to the target image independently and the resulting segmentations combined [10].

An analogy can be made between the combination of segmentations derived independently from multiple atlas images and the combination of multiple independent classifiers in a generic classification problem [14]. In this analogy, each transformed atlas image can be regarded as a classifier, which assigns a label value to each voxel of the target image. The training process can be assimilated to the registration between the atlas image and the target image. It has been widely proven in the pattern recognition field that combining multiple classifiers can yield more robust and accurate results than using single classifiers [15], [16], this fact being the main motivation for multi-atlas approaches.

The most widely used combination strategy in the literature is majority voting, also named majority rule, decision fusion or label voting. This approach weights each candidate segmentation equally and assigns to each voxel the label that most segmentations agree on [10], [11]. Another popular approach is called simultaneous truth and performance level estimation (STAPLE), which uses an expectation-maximization (EM) approach to reach the best possible final segmentation [17], [18]. STAPLE estimates the performance of each classifier iteratively and weights it accordingly. The two different methods presented in [18] are extensions of the one in [17] for images with multiple segmented structures. Shape-based averaging represents another way of combining segmentations [19] which is based on Euclidean distance maps computed for all structures in each candidate segmentation. The method was shown to keep structure regularity and contiguity better than majority voting. Another possibility is atlas selection: instead of combining segmentations, methods can be devised to select atlases *a priori* (before registration) or *a posteriori* (after registration) [10]. In [20], a number of atlases is selected for combination based on mutual information [21]. In [22], atlas selection is done on a structure basis for the segmentation of brain magnetic resonance

Manuscript received January 02, 2009; revised January 22, 2009. First published February 18, 2009; current version published July 29, 2009. The work of X. Artaechevarria was supported by a predoctoral scholarship of the Basque Government. The work of A. Muñoz-Barrutia was supported by a Ramón y Cajal Fellowship of the Spanish Ministry of Science and Technology. The work of C. Ortiz-de-Solórzano was supported in part by the Spanish Ministry of Science and Education (MCYT TEC2005-04732), in part by the EU Marie Curie Program (MIRG-CT-2005-028342), and in part by the Ramón y Cajal Fellowship. This work was supported by the Government of Navarra and the Spanish Ministry of Health under Project FIS-PI070751. *Asterisk indicates corresponding author.*

*X. Artaechevarria is with the Cancer Imaging Laboratory, Center for Applied Medical Research, University of Navarra, 31008 Pamplona, Spain (e-mail: xabiarta@unav.es)

A. Muñoz-Barrutia and C. Ortiz-de-Solórzano are with the Cancer Imaging Laboratory, Center for Applied Medical Research, University of Navarra, 31008 Pamplona, Spain.

Digital Object Identifier 10.1109/TMI.2009.2014372

0278-0062/$26.00 © 2009 IEEE

(MR) images. For each given structure, the atlas image with highest local mutual information is selected.

Recent works have contributed significantly to the field of atlas-based medical image segmentation, but we believe that some aspects still remain unexplored. Namely, we think that in order to achieve highest possible overall segmentation accuracy, a better understanding of the different combination strategies is required. In this paper, we propose a scheme that divides the combination strategies in two major groups: global strategies, which estimate segmentation accuracy with a single value for the whole image, and local strategies, which evaluate the segmentations at each voxel and perform weighted voting accordingly.

We test different global and local combination methods on digital phantoms and on publicly available MR images of human brains. We show that no combination algorithm is better than the others consistently for all images and regions within the images. We study in which kind of structures local strategies do better than global methods, and conclude that the optimum solution is to select the combination method according to the gray level contrast between each particular region and its neighbors.

The outline of the article is the following. In Section II, different combination strategies are examined. First, global weighting approaches are listed. Then, we show their main limitations and introduce the local combination strategies. In Sections III and IV, we show results from experiments on phantom images and on brain MR data with multiple segmented structures, comparing different combination strategies. The paper ends with a discussion (Section V) and a final conclusion (Section VI).

II. Combination Methods

Notation

For simplicity and uniformity, we adopt the basic notation used in [18]. Consider a segmentation of an image with potentially L different classes that belong to a label set $\Lambda = \{1, 2, \ldots, L\}$. A 3-D atlas image A_k is a mapping from coordinates to labels $A_k : \mathbb{R}^3 \longrightarrow \Lambda$. An atlas-based classifier is defined by a set of atlas images, $A_k,\ k = 1, \ldots, K$ and coordinate transformations that map coordinates from the target image to the atlas images $F_{T \longrightarrow A_k} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3,\ k = 1, \ldots, K$. Using a given transformation $F_{T \longrightarrow A_k}$ to transform the segmentation of the atlas image A_k^s, an estimated segmentation of the target image $\hat{T}_k^s$ is obtained

$$\hat{T}_k^s = A_k^s \circ F_{T \longrightarrow A_k}. \tag{1}$$

Each $\hat{T}_k^s,\ k = 1, \ldots, K$ is a candidate segmentation which must be combined into a final estimated segmentation $\hat{T}^s$. Following [18], we now regard the segmentation task as the classification of N unordered samples. Thus, when candidate k assigns sample x to class i, we write the output of the classifier as

$$e_k(x) = i. \tag{2}$$

Note that in this context the spatial location of the voxels is irrelevant. The set of all samples that belong to class i is denoted by C_i. The outputs from L independent classifiers can be combined to generate a single ensemble response, $E(x)$, which is the output of the combination strategy. The aim when building an ensemble classifier is to achieve a higher probability of correctly classifying the voxels of the image than that obtained by using an individual classifier. The combined classifier output $E(x)$ for a sample x should be the class that maximizes the probability, given all classifier decisions $e_1(x)$ through $e_K(x)$ and some arbitrary classifier performance model $\mathbf{P}$ [18]

$$E(x) = \arg\max_i \ P(x \in C_i | e_1(x), \ldots, e_K(x), \mathbf{P}). \tag{3}$$

A. Global Combination Strategies

This group of combination strategies assigns a global weight w_k to each segmentation of the target image $\hat{T}_k^s$. The weight is calculated by a global evaluation of segmentation accuracy, and its value is used to ponderate the whole target image segmentation derived from the atlas image. Note that although the weight is global for each candidate segmentation, decisions on the combination of the segmentations are taken voxel-by-voxel.

1) Majority Voting: This is the most simple combination method. It is generally used when there is no *a priori* knowledge about the accuracy of each classifier. It assigns to each voxel the label that most segmentations agree on. Thus, the ensemble response $E_{\mathrm{MV}}(x)$ of majority voting can be expressed as

$$E_{\mathrm{MV}}(x) = \max[f_1(x), \ldots, f_L(x)] \tag{4}$$

where $f_i(x) = \sum_{k=1}^{K} w_{k,i}(x)$ for $i = 1, \ldots, L$ and

$$w_{k,i}(x) = \begin{cases} 1, & \text{if } i = e_k(x) \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

This technique labels a voxel correctly if a majority among the K segmentations agree on the correct value.

2) Weighted Voting: The expression for majority voting, given by (4), can be generalized to assign arbitrary weights to each candidate segmentation. Weighted voting is commonly used in pattern recognition, as it allows the maximization of classification accuracy when the accuracy of individual the classifiers is known [16]. However, these results can not be directly translated into the multi-atlas segmentation field, because segmentation accuracies can not be known *a priori*. Thus, a different approach must be adopted. The ensemble response $E_{\mathrm{WV}}(x)$ would respond to (4), but weights would have to be set differently. One option is to set them as

$$w_{k,i}(x) = \begin{cases} m^p, & \text{if } i = e_k(x) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where m is a similarity measure between the registered atlas image and the target image and p is an associated gain exponent. If the similarity measure is not sensitive enough, the differences between the weights might not be relevant. In those cases, it might be necessary to increase the value of p.

The selection of a similarity measure m is not straightforward. We previously proposed using the mutual information [21] between the target image and the registered atlas image

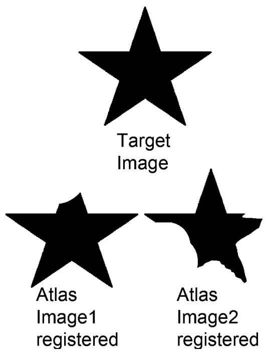


Fig. 1. Example to show limitation of global candidate segmentation combination strategies. Atlas images 1 and 2 have been registered to target image. Image 1 is generally better registered, except for the upper arm of the star. However, global strategies can not evaluate registration performance locally. Therefore, they can not take advantage of this fact to obtain a better fused segmentation.

as weight, i.e., $m^p = [I(A_i^s \circ F_{T \longrightarrow A_i}, T)]^p$ [23]. The rationale behind that choice was that high mutual information values normally imply a better registration, and using this information for the weighting should yield better results. It must be noted that, generally, no strong correlation between mutual information (or any other image similarity measure) and registration accuracy exists [24]. Nevertheless, our tests on MR images of mouse brains resulted in higher segmented region overlap and smaller distances between segmented surfaces than majority voting [23]. It must also be noted that, although statistically significant, differences were small.

More generally, Roche *et al.* [25] studied the assumptions that are implicitly made when using different similarity measures in image registration. They show that the mean square distance can be used when no relative intensity difference between images is expected, while variants of the correlation coefficient should be preferred in the case of an affine relationship between intensities. Finally, if only a statistical relationship can be assumed, mutual information is the best option. These three similarity measures are tested and compared in our experiments, since they are likely to be good predictors of accuracy in registration.

B. Limitation of Global Combination Strategies

Global fusion strategies have shown their potential and perform generally better than single atlas approaches. Nevertheless, they have a major limiting drawback. Namely, weights or performance estimates are the same for all the voxels of the segmentation. We will show how this can negatively affect the overall segmentation performance using a simple example. Let us have a star-shaped target image and two different atlas images, numbered 1 and 2, as shown in Fig. 1. After registration, the first atlas image fits the target image almost perfectly in all parts, except for one of the arms (upper arm). In contrast, the second atlas image fits the target image correctly in that arm, while registration fails in the rest of the image.

As weights are assigned globally, with these approaches it is impossible to take advantage of the fact that registration between atlas image two and the target image succeeded in a particular point, even if it was inaccurate in the rest of the image. A combination strategy that would account for local registration failures could therefore achieve higher segmentation accuracy.

C. Local Combination Strategies

1) Generalized Local Weighted Voting: A logical solution that overcomes the limitation shown in the previous subsection, is to adapt weights locally. Instead of assigning the weight to all the voxels of the segmentation, each voxel can have a different weight value.

In this context, we propose a local segmentation fusion method that we denominate *Generalized Local Weighted Voting*. The general expression for the weights is given by

$$w_{k,i}(x) = \begin{cases} [m(s,r)]^p, & \text{if } i = e_k(x) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $m(s,r)$ can be any local similarity measure that assigns a distance between two regions of the registered atlas image and the target image. The distance is measured over a neighborhood region of shape s, radius r using a gain p. The neighborhood can be of arbitrary shape, for instance cubical or spherical.

As pointed out in Section II-BII, it must be noted that similarity measures do not perfectly correlate with registration accuracy. For instance, high information content areas, such as edges or high contrast areas, are likely to result in regions with high mutual information, regardless of the registration accuracy.

Assigning a different weight to each voxel significantly increases the number of degrees of freedom of the problem. This arises the need for a regularization method to ensure a smooth variation of the weights, thus avoiding the effect of image noise. When local weights are computed, an *intrinsic regularization* occurs due to the voxel-by-voxel shifting of the neighborhood region. For instance, assuming a $2-D$ image and a square region of 10×10 pixels, two neighboring pixels share 180 out of the 200 pixels (100 from the target image, 100 from the registered atlas image) used to compute the local weight. Sharing this large percentage of voxels between neighboring pixels causes a smooth variation of the weights along the image, regardless of the particular similarity measure employed. Increasing the radius of the local neighborhood further ensures the smoothness of the weights, but such computed weights are less local and this generally results in worse perfomance. More detailed effects of increasing the neighborhood radius r are explored in Sections III and IV.

On the other hand, the computational burden can be reduced by limiting the number of voxels in which segmentations must be evaluated. If, for a given voxel, all candidate segmentations agree on a certain label, this label is directly assigned without computing the weights. This approach was adopted in all experiments shown in this paper.

2) STAPLE: This fusion strategy weights each voxel according to the estimated performance of the disputing labels at that point, using an iterative expectation-maximization algorithm. Three different implementations of the method exist. The binary classification method was first proposed by Warfield *et al.* [17], and Rohlfing *et al.* presented two multilabel generalizations of the method [18]. In the latter work, STAPLE was shown to outperform majority voting on confocal microscopy images of bee brains.

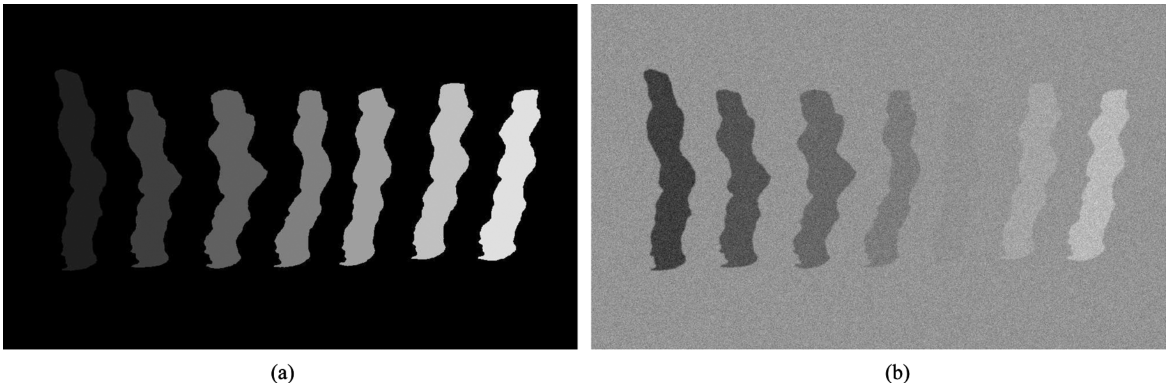


Fig. 2. Sample label mask and corresponding phantom with modified intensities and added noise. (a) Phantom mask. (b) Phantom.

III. Experiments on Digital Phantoms

A. Data

To study the performance of different atlas combination strategies, we built an artificial atlas of 18 images. A manually delineated elongated shape was copied 7 times on a 2-D image of size 1000×600. On this first image, random elastic deformations were applied using an online available ImageJ plugin[1]. Each transformation was calculated by deforming a *B*-spline grid on the image, with a random noise factor. The grid spacing was one pixel. We generated 18 different mask images, each of them containing seven regions. From each of these mask images, an atlas image was generated in three steps. First, image intensities were modified, so that there were regions that showed different contrast levels with the background. Second, a very slight Gaussian smoothing with $\sigma = 1$ was applied. Third, image intensities were modified to follow a Rician distribution, as MRI magnitude images do [26]. One of the phantom atlas images and the corresponding label mask can be seen in Fig. 2. The label mask assigns the label 0 to the background, and labels with increasing ordinal values to the regions from left to right, i.e., the leftmost region has label 1 and the right-most label 7. In the rest of the paper, we assimilate the label to the name of the region. One of the 18 images was used to set the parameters of the different segmentation algorithms, while the other 17 were used for algorithm evaluation.

B. Registration

The registrations required for the multi-atlas segmentation were done in two steps. First, an affine transform was used to account for translations, rotations, anisotropic scaling and shearing. Then, an elastic *B*-spline registration was performed, using an isotropic grid spacing of 8.0 pixels [27]. Mutual information was maximized in both cases [21]. The reason for using a much less dense grid than that employed for the deformation generation was to obtain nonideal registrations, as occurs in real complex 3-D images. Elastix, an open-source software for elastic image registration was used for all registrations [28]. To accelerate computation time, registration was done in a multiresolution fashion, with three resolution levels. A stochastic gradient descent optimizer was used, because it provides a good trade-off between precision and computation time [29].

C. Evaluation Measures

The similarity index (SI) is the most widely used measure to evaluate the performance of a segmentation algorithm. The SI between two segmentations, S_a and S_b, of the same object, is defined as

$$\mathrm{SI} = \frac{2[S_a \bigcap S_b]}{|S_a| + |S_b|} \tag{8}$$

where $\bigcap$ indicates the overlapping voxels between the two segmentations, and $|S_a|$ indicates the number of voxels of the corresponding segmentation [30]. SI has a value of 1 when there is a perfect match between labels and 0 when there is no overlap.

Apart from SI, the mean average surface distance (MASD) between the segmented structures was also evaluated [2], [31]. The MASD is the symmetric distance between the surfaces of the respective segmentations

$$\mathrm{MASD}(S_a, S_b) = \frac{1}{2}[d(S_a, S_b) + d(S_b, S_a)], \tag{9}$$

where $d(S_a, S_b)$ is the mean distance from the points of segmentation S_a to segmentation S_b.

For both measures, mask images were considered the ground truth.

D. Combination Algorithms

The combination algorithms were implemented based on the Insight Toolkit (ITK)[2], an online available toolkit for image processing. The gain (p) and region size (r) were set independently for each combination method, using a randomly chosen image from the set. For the global voting methods, we chose the p value (between 1 and 8) that maximized segmentation accuracy for the selected image. For the local methods, both p and r had to be set. The parameter tuning process consisted in setting r to an initial value of 10 and then varying p from 1 to 8, until obtaining the optimal value. Then, r was varied from 5 to 25 with a 5 pixel step, until finding the optimum value. Alternatively to choosing p, we also considered the option of selecting the image (in global voting) or pixel (in local voting) with the highest weight.

We tested the following combination algorithms:

- *Majority Voting*: this method does not require any parameter tuning.

[1] http://biocomp.cnb.uam.es/~iarganda/SplineDeformationGenerator/

[2] http://www.itk.org

TABLE I

Average SI for Different Phantom Regions With Different Combination Strategies. (*) Indicates $p < 0.05$ According to the Wilcoxon Matched-Pairs Signed-Ranks Test, When Compared One-to-One to all the Other Columns. (**) Indicates $p < 0.001$. If two Columns are Marked, it Means That the Difference Between Them is not Statistically Significant, but it is With the Rest of the Columns

Reg.	Combination Strategy							
	MV	STAPLE	GWV-MI	GWV-NCC	GWV-MSD	LWV-MI	LWV-NCC	LWV-MSD
1	0.97	0.95	0.98	0.97	0.96	0.98	0.97	**0.99 ****
2	0.97	0.96	0.98	0.97	0.97	0.98	0.97	**0.99 ****
3	0.97	0.95	0.97	0.97	0.96	0.97	0.97	**0.99 ****
4	0.92	0.88	0.93	0.92	0.87	0.92	0.92	**0.97 ****
5	0.83	0.79	**0.84 ***	0.83	0.78	0.83	0.83	**0.84 ***
6	0.91	0.85	0.91	0.91	0.85	0.91	0.91	**0.96 ****
7	0.95	0.91	0.95	0.95	0.89	0.95	0.93	**0.98 ****

TABLE II

Average MASD in Pixels for Different Phantom Regions With Different Combination Strategies. (*) Indicates $p < 0.05$ According to the Wilcoxon Matched-Pairs Signed-Ranks Test, When Compared One-to-One to all the Other Columns. (**) Indicates $p < 0.001$

Reg.	Combination Strategy							
	MV	STAPLE	GWV-MI	GWV-NCC	GWV-MSD	LWV-MI	LWV-NCC	LWV-MSD
1	1.60	2.37	1.18	1.61	2.06	1.12	1.55	**0.43 ****
2	1.35	2.12	0.99	1.37	1.62	1.02	1.35	**0.54 ****
3	1.58	2.35	1.36	1.58	2.26	1.44	1.58	**0.67 ****
4	3.91	5.95	3.52	3.89	5.85	3.87	3.89	**1.57 ****
5	8.00	11.29	7.59	7.97	11.42	7.97	7.96	**6.66 ****
6	4.61	7.58	4.31	4.65	9.46	4.61	4.63	**2.10 ****
7	2.73	4.44	2.35	2.76	8.53	2.46	4.41	**2.45 ***

- *Global Weighted Voting based on Normalized Cross-Correlation (GWV-NCC)*: Normalized cross-correlation (NCC) between two images is defined as

$$\mathrm{NCC} = \frac{\mathrm{Cov}(I_1, I_2)}{\sqrt{\mathrm{Var}(I_1)} \cdot \sqrt{\mathrm{Var}(I_2)}} \tag{10}$$

  where $\mathrm{Cov}(I_1, I_2)$ is the covariance of the images and $\mathrm{Var}(I_1)$ and $\mathrm{Var}(I_2)$ are the variances of each of the images. The gain factor p was set to 4.
- *Global Weighted Voting based on Mutual Information (GWV-MI)* with $p = 8$. The power of eight was required to amplify differences between the weights. Mutual information was computed as in [21], using the ITK implementation.
- *Global Weighted Voting based on Mean Square Distance (GWV-MSD)*: This is a particular case that does not require any gain factor, because the difference between the images is already very large. Therefore, the gain was simply set to -1 to account for the inverse relationship between mean square distance and image overlap after registration.
- *STAPLE*: We used an implementation by Rohlfing available online.[3] We did not limit the number of iterations and the termination threshold for the EM iterations was set to 10^{-5}.
- *Local Weighted Voting based on Normalized Cross-Correlation (LWV-NCC)*: The similarity measure employed in this local combination method was the same as in GWV-NCC, but computed in a square local neighborhood around each pixel of size $r = 10$. Best results were achieved with $p = 5$.
- *Local Weighted Voting based on Mutual Information (LWV-MI)*: For this local fusion method, we particularized the generalized local weighted voting method with normalized mutual information as similarity measure, defined as

$$\mathrm{NMI} = \frac{H(I_1) + H(I_2)}{H(I_1, I_2)} \tag{11}$$

  where $H(I_1)$ and $H(I_2)$ are the entropies of images 1, 2, and $H(I_1, I_2)$ is the joint enntropy of both images. The entropy of an image can be computed from its histogram $h(x)$ as

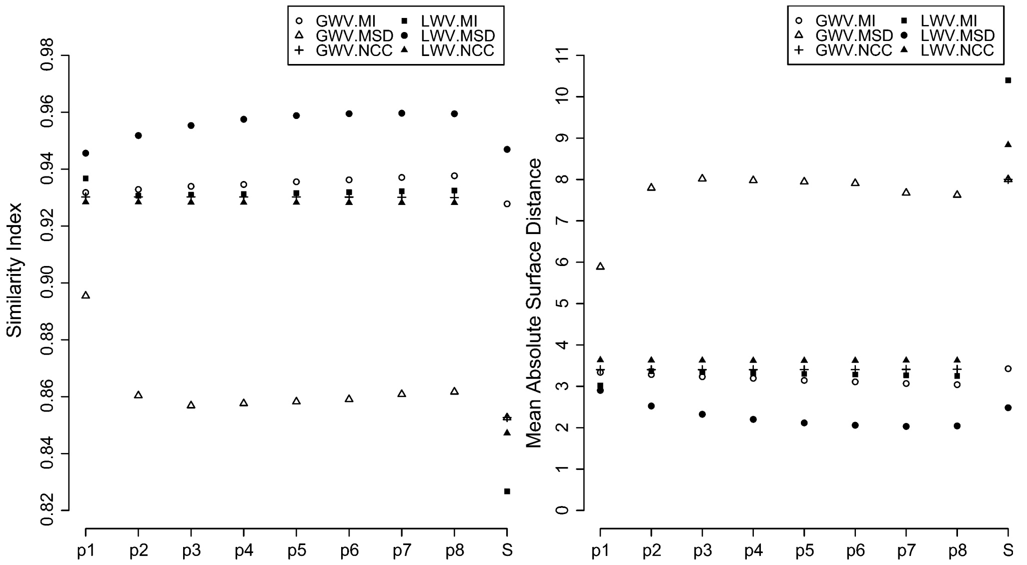
$$H(I_1) = -\sum_{i=1}^{N} h(x_i) \log_2 h(x_i) \tag{12}$$

  where N is the number of bins and x_i is the centroid of the ith histogram bin. The similarity measure was computed in a 2-D square neighborhood of isotropic size $r = 15$, and p was set to 8.
- *Local Weighted Voting based on Mean Square Distance (LWV-MSD)*: In this case, the chosen similarity measure was the mean squared distance computed in a square region around the voxel of interest of size 10 in each dimension. The power p was set to -6 to account for the inverse relationship between distance and similarity between images.

E. Results

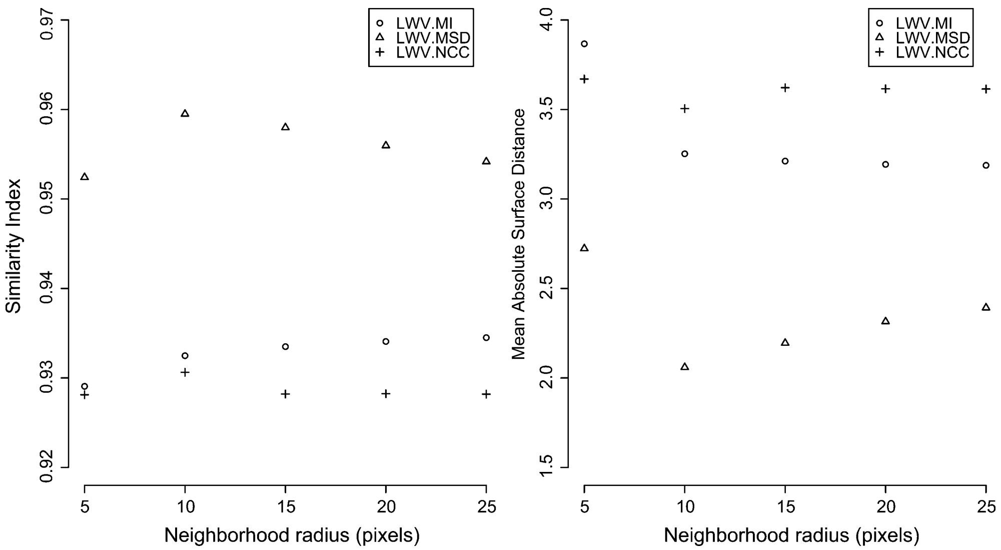
The evaluation measures obtained using different combination strategies on our phantom images are summarized in Tables I and II. The Wilcoxon matched-pairs signed-ranks test [32] was applied on the results on a region basis, to look for the best method for each region. If we look at SI (Table I), LWV-MSD is generally the best combination method, except

[3] http://www.stanford.edu/~rohlfing/software/index.html



(a) (b)

Fig. 3. Plots showing the effect of varying the gain factor p on mean SI and MASD over all regions for different combination strategies on phantom images. p is varied from 1 to 8. S indicates *selection*, that is, the voxel with the highest weight is selected, without any further voting process. $r = 10$ was set to for all LWV methods. (a) SI for varying gain factor p. (b) MASD for varying factor p.



(a) (b)

Fig. 4. Plots showing effect of varying neighborhood region radius r on mean SI and MASD over all regions for different combination strategies on phantom images. For all methods, p was the set to the value employed in the rest of experiments. (a) SI for varying neighborhood region radius r. (b) MASD for varying neighborhood region radius r.

for region 5. The particularity of this region is its extremely low contrast with the background [see Fig. 2(b)]. In terms of MASD (Table II), it is noteworthy that the difference in performance between LWV-MSD and the next best method is generally larger in regions with high contrast [regions 1, 2, and 3 are the ones with largest contrast, as shown in Fig. 2(b)].

It is also interesting that, in exceptional cases, a method can be significantly better than the other, even though the mean evaluation measure is worse. This is due to the fact that the employed statistical test works with the magnitude of the differences between the pairs of measures involved in the test. This is the case of region 7 in Table II, where a single region in one image degrades the mean MASD of LWV-MSD considerably, making it worse than the mean MASD of GWV-MI. However, the statistical test is able to detect that LWV-MSD performs consistently better (with $p < 0.05$) than the rest of the combination methods, including GWV-MI.

F. Influence of Tunable Parameter Selection

As mentioned in Section III-A, we used a single image to set the tunable parameters, both for global and local combining methods. This arises the question how much the final result can be influenced by a change in the selection of these parameters. To evaluate this, we looked at SI and MASD values as p varies for the different combination methods, as well as at the results of varying r. Results are shown in Figs. 3 and 4, respectively.

TABLE III
GAIN PARAMETER p FOR EACH COMBINATION ALGORITHM ON THE IBSR DATABASE

	GWV-MI	GWV-NCC	GWV-MSD	LWV-MI	LWV-NCC	LWV-MSD
Gain parameter p	4	6	-1	8	2	-1

TABLE IV
AVERAGE SI FOR DIFFERENT BRAIN STRUCTURES WITH DIFFERENT COMBINATION STRATEGIES. IN THE CASE OF PAIRED STRUCTURES (LEFT-RIGHT), THE FIRST NUMBER INDICATES MEAN SI OF THE LEFT STRUCTURE AND THE SECOND REFERS TO THE RIGHT STRUCTURE. THE HIGHEST VALUE FOR EACH STRUCTURE IS HIGHLIGHTED IN BOLD. (*) INDICATES $p < 0.05$ ACCORDING TO THE WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST, WHEN COMPARED ONE-TO-ONE TO ALL THE OTHER COLUMNS. (**) INDICATES $p < 0.001$. IF TWO COLUMNS ARE MARKED, IT MEANS THAT THE DIFFERENCE BETWEEN THEM IS NOT STATISTICALLY SIGNIFICANT, BUT IT IS WITH THE REST OF THE COLUMNS

Brain Structure	Combination Strategy							
	MV	STAPLE	GWV-MI	GWV-NCC	GWV-MSD	LWV-MI	LWV-NCC	LWV-MSD
Thalamus	0.86; 0.86	0.85; 0.85	0.87; 0.87	0.86; 0.86	0.87; 0.87	0.87; 0.87	0.86; 0.86	**0.87;0.88**
Caudate	0.78; 0.78	0.66; 0.69	0.79; 0.79	0.78; 0.78	0.79; 0.80	0.81; 0.80	0.79; 0.78	**0.83;0.83** *
Putamen	0.86; 0.85	0.81; 0.79	0.86; 0.86	0.86; 0.85	0.86; 0.86	**0.87;0.86**	0.86; 0.85	0.86; 0.86
Pallidum	0.80; 0.80	0.69; 0.70	0.80; 0.81	0.80; 0.80	0.80; 0.80	**0.81;0.81** *	0.80; 0.80	0.78; 0.79
Hippocampus	0.73; 0.75	0.52; 0.51	0.74; 0.75	0.73; 0.75	0.73; 0.75	**0.74;0.76** *	0.73; 0.75	**0.74;0.76** *
Amygdala	0.71; 0.71	0.65; 0.64	0.72; 0.72	0.71; 0.71	0.72; 0.72	0.72; 0.72	0.71; 0.71	0.72; 0.72
Accumbens area	0.66; 0.66	0.51; 0.47	0.67; 0.68	0.66; 0.66	0.66; 0.66	**0.68;0.68**	0.66; 0.66	0.67; 0.67
Ventral DC	0.81; 0.82	0.79; 0.77	0.82; 0.82	0.82; 0.82	0.82; 0.82	0.82; 0.82	0.82; 0.82	0.82; 0.82
Cerebral WM	0.75; 0.75	0.59; 0.54	0.75; 0.75	0.75; 0.75	0.75; 0.75	0.76; 0.76	0.75; 0.75	**0.78;0.78** **
Cerebral Cortex	0.78; 0.78	0.56; 0.47	0.79; 0.79	0.79; 0.78	0.79; 0.78	0.79; 0.79	0.78; 0.78	**0.81;0.81** **
Lateral Ventricle	0.77; 0.75	0.51; 0.50	0.78; 0.76	0.77; 0.75	0.77; 0.76	0.81; 0.79	0.78; 0.76	**0.83;0.82** **
Inferior Lat Vent	0.16; 0.15	0.15; 0.12	0.22; 0.21	0.18; 0.17	0.20; 0.18	0.22; 0.20	0.18; 0.17	**0.22;0.22**
Cerebellum Cortex	0.84; 0.84	0.73; 0.73	0.85; 0.86	0.84; 0.85	0.85; 0.85	0.85; 0.85	0.84; 0.84	**0.86;0.86**
Cerebellum WM	0.78; 0.78	0.70; 0.70	0.79; 0.79	0.78; 0.78	0.78; 0.78	**0.80;0.80** *	0.78; 0.78	0.79; 0.79
Third Ventricle	0.71	0.70	0.71	0.71	0.70	**0.73** *	0.71	**0.74** *
Fourth Ventricle	0.75	0.54	0.75	0.75	0.75	**0.77** **	0.75	**0.77** **
Brain Stem	0.89	0.85	0.90	0.90	0.90	0.90	0.90	**0.91**
CSF	0.57	0.55	0.60	0.58	0.55	**0.61**	0.58	**0.61**

The variation of p shows no clear peaks for any method, except for GWV-MSD. This could be expected because weight values are already very large with this metric, and amplifying them before the voting process is not required. The absence of peaks and the very slow variation between neighboring points on the rest of methods suggests that the selection of the gain factor can lead to slightly better or worse overall results, but it is not critical.

In the case of the neighborhood radius, it can be seen that if the smallest radius, i.e., $r = 5$, is neglected, the accuracies resulting from the different methods do not overlap with each other. That means that, on this dataset, the selection of r does not heavily influence the performance of the local combination methods. Very importantly, automatically selecting the parameters in a single image as described in Section III-D, we always obtained the p and r that provided the best segmentation results on the 17 image test set.

IV. EXPERIMENTS WITH BRAIN MR DATA

A. Data

To test the performance of the different combination strategies on real images, we employed 18 T1 MR brain images from the internet brain segmentation repository (IBSR) [33]. One of the images was used to tune the segmentation parameters, while the other 17 were employed for the evaluation of the algorithms. We studied the segmentation of 18 different structures, 14 of which were paired (left and right side). More structures were actually segmented on the images, but the segmentation failed almost completely on them due to their small size. All images had been normalized into the Talairach position and processed with biasfield correction routines. Image size is $256 \times 256 \times 128$. Resolutions are variable: In the $x - y$ dimensions they range from 0.837 to 1.0 mm and in the z dimension it is 1.5 mm in all cases.

B. Registration

The registration strategy used with the phantom images was extended to 3-D images. A similar approach has been previously used to register brain MR images [11], [34]. The voxel spacing for the *B*-spline grid was 8.0 voxels.

C. Combination Algorithms

We compared the same combination algorithms as we did on the phantoms, modified to work in 3-D images whenever required. The shape of the neighborhood used in the local combination methods was in this case cubical. The optimum gain parameters p, calculated on one of the images of the IBSR dataset, were generally different from the ones calculated for the phantoms, due to the different intensity distribution of the images. Table III shows the calculated p for each combination method. The process used to select the p and r values was the same as the one followed for the phantom images. Gain factors between 1 and 8 were considered. The tested r values were between 2 and 14 with a varying step of 3. The optimum value of r was 5 for all local combination strategies.

TABLE V

Average MASD for Different Brain Structures With Different Combination Strategies. In the Case of Paired Structures (Left-Right), The First Number Indicates Mean MASD of the Left Structure and the Second Refers to the Right Structure. The Highest Value for Each Structure is Highlighted in Bold. (*) Indicates $p < 0.05$ According to the Wilcoxon Matched-Pairs Signed-Ranks Test, When Compared One-to-One to all the Other Columns. (**) Indicates $p < 0.001$. If two Columns are Marked, it Means That the Difference Between Them is not Statistically Significant, but it is With the Rest of the Columns

Brain Structure	Combination Strategy							
	MV	STAPLE	GWV-MI	GWV-NCC	GWV-MSD	LWV-MI	LWV-NCC	LWV-MSD
Thalamus	0.85; 0.89	0.94; 0.98	0.81; 0.81	0.84; 0.88	0.84; 0.88	**0.79;0.80** *	0.81; 0.84	**0.76;0.74** *
Caudate	0.78; 0.81	1.74; 1.53	0.75; 0.75	0.78; 0.80	0.75; 0.75	0.69; 0.71	0.76; 0.79	**0.63;0.64** *
Putamen	0.70; 0.72	0.94; 1.14	0.67; 0.67	0.65; 0.72	0.68; 0.68	0.66; 0.68	0.70; 0.72	0.66; 0.67
Pallidum	0.70; 0.69	1.45; 1.27	0.70; 0.67	0.70; 0.75	0.70; 0.68	**0.68;0.66** *	0.70; 0.69	0.73; 0.70
Hippocampus	0.88; 0.79	2.98; 4.48	0.85; 0.78	0.87; 0.75	0.86; 0.79	**0.83;0.75** *	0.87; 0.79	**0.82;0.76** *
Amygdala	0.91; 0.92	1.30; 1.36	0.87; 0.89	0.91; 0.91	0.87; 0.89	0.88; 0.90	0.91; 0.92	**0.84;0.86**
Accumbens area	0.74; 0.71	3.14; 4.33	0.71; 0.0.68	0.74; 0.71	0.77; 0.74	0.70; 0.68	0.74; 0.71	**0.68;0.67**
Ventral DC	0.80; 0.78	0.96; 1.09	0.77; 0.74	0.80; 0.77	0.77; 0.75	0.77; 0.75	0.80; 0.78	**0.74;0.73**
Cerebral WM	1.31; 1.29	1.61; 1.78	1.24; 1.22	1.30; 1.29	1.25; 1.23	1.22; 1.21	1.30; 1.28	**1.09;1.07** **
Cerebral Cortex	1.34; 1.31	2.64; 3.85	1.24; 1.20	1.34; 1.31	1.26; 1.23	1.29; 1.25	1.35; 1.32	**1.13;1.10** **
Lateral Ventricle	0.88; 1.01	3.06; 3.10	0.89; 0.90	0.87; 0.96	0.89; 0.91	0.76; 0.80	0.85; 0.94	**0.67;0.70** *
Inferior Lat Vent	1.64; 1.55	2.62; 3.53	1.44; 1.43	1.55; 1.48	1.45; 1.44	**1.40;1.46**	1.55; 1.50	1.48; 1.45
Cerebellum Cortex	1.35; 1.35	1.76; 1.74	1.23; 1.19	1.34; 1.32	1.28; 1.25	1.26; 1.24	1.35; 1.33	**1.18;1.14** *
Cerebellum WM	1.19; 1.20	1.61; 1.65	1.09; 1.08	1.17; 1.19	1.12; 1.12	1.10; 1.08	1.18; 1.19	**1.04;1.01** *
Third Ventricle	0.68	0.72	0.65	0.68	0.68	**0.62** *	0.67	**0.59** *
Fourth Ventricle	0.68	2.54	0.68	0.68	0.67	**0.61** **	0.67	**0.62** **
Brain Stem	0.80	1.08	0.75	0.79	0.74	0.74	0.79	**0.69**
CSF	1.22	1.30	1.12	1.20	1.08	1.14	1.21	**1.05**

D. Evaluation Measures

As in the phantom experiments, SI and MASD were evaluated. In this case, the manual segmentation of the brain structures was considered to be the ground truth.

E. Results

Table IV shows the average SI for different brain regions using all the combination methods. Table V shows the corresponding MASD values. In general, local strategies rendered better accuracies than global methods, but performance varies from region to region. No combination method is better than the rest for all regions.

Out of the studied 18 structures, LWV-MSD was among the methods with highest average similarity index in 12 cases. This number was 8 for LWV-MI. The MASD results followed the same trend. LWV-MSD was the best method for 15 structures (including ties), and LWV-MI outperformed the others in six cases. It must be noted that ties occurred, as it can can be seen in Tables IV and V.

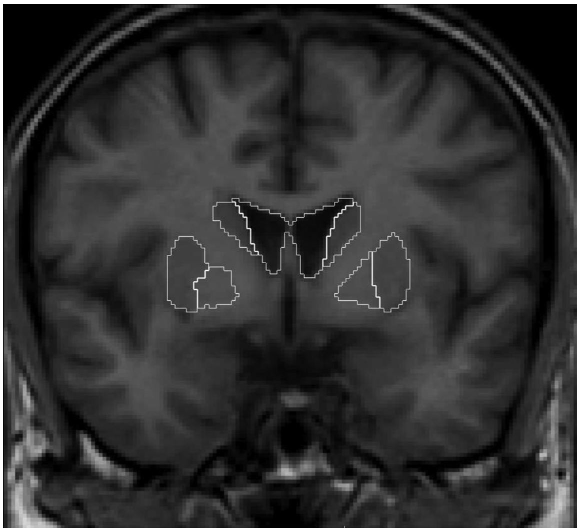


Fig. 5. T1 MR slice of human brain, with different anatomical structures delineated manually. Pallidum and putamen are the two paired structures below. The putamen is in the external part and the pallidum is in the interior. The lateral ventricles are the darkest structures delineated above, with the caudates next to them. The lack of gray level difference between pallidum, putamen and cerebral white matter voxels can be observed, together with the large contrast between the ventricles and the surrounding structures.

However, more interesting than the absolute number of regions in which one method was better than others, is the study of the particularities of those regions. Specifically, and for illustrative purposes, we will look at two particular structures: the lateral ventricles and the pallidum. The two lateral ventricles conform the largest part of the ventricular system of the brain and appear as a dark region on T1 MR images (see Fig. 5). They are mainly surrounded by cerebral white matter, which shows a much brighter gray level. The caudate, a nucleus within the basal ganglia, is also in contact with the lateral ventricles and, even though the intensity difference is smaller, it is still clearly distinguishable. It is in this kind of structures with large intensity contras where the advantage of using local weighting approaches is clearer. The average SI of the two-side lateral ventricles using LWV-MSD is 0.825, in contrast with 0.760 for the majority voting. This difference is very significant, according to the Wilcoxon matched-pairs signed-ranks test $(p \leq 10^{-7})$ [32]. An illustrative slice with the corresponding segmentation can be seen in Fig. 6. The cerebral and cerebellum cortex, the cerebral white matter and the caudate add to the group of structures with this kind of characteristics.

The pallidum or globus pallidus is part of the basal ganglia. It is surrounded by many regions, mainly cerebral white matter, and the putamen, but also the caudate, the amygdala and other close structures. The intensity contrast with most of them is

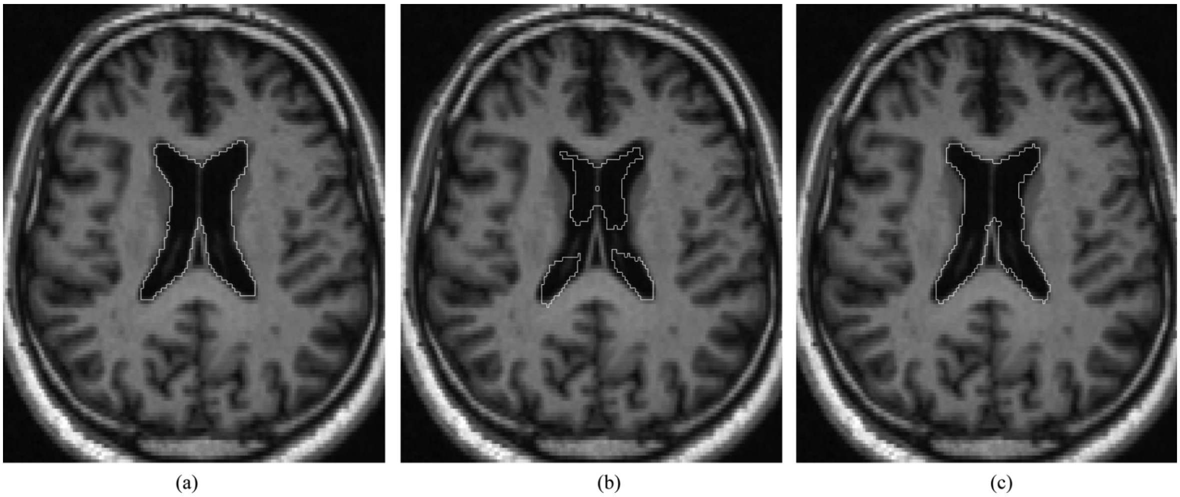


Fig. 6. Axial slice showing different segmentations of the lateral ventricles. The local weighted voting based on mean square distance approaches the manual segmentation better than the majority voting. (a) Manual segmentation. (b) Majority voting. (c) Local WV-MSD.

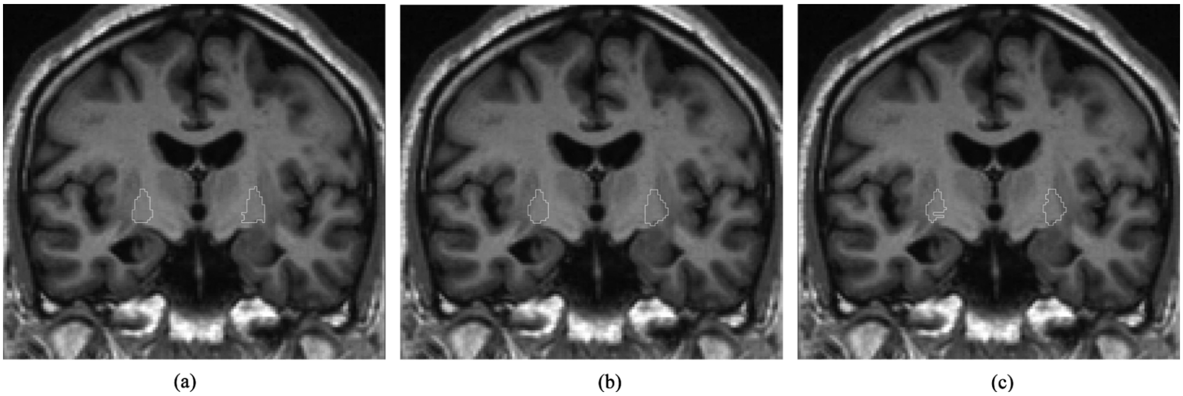


Fig. 7. Coronal slice showing different segmentations of the pallidum. The local weighted voting based on mean square distance results in a more noisy delineation than the global weighted voting based on mutual information. (a) Manual segmentation. (b) Global WV-MI. (c) Local WV-MSD.

TABLE VI
AVERAGE SI AND MASD FOR MAJORITY VOTING AND THE BEST SEGMENTATION COMBINATION METHOD FOR EACH STRUCTURE, SELECTED A POSTERIORI ACCORDING TO ITS SI VALUE

Evaluation Measure	Combination Strategy	
	Majority Voting	Best Combination Method
SI	0.733	**0.758**
MASD	0.967	**0.851**

very low (see Fig. 5). In this case, LWV-MSD performs worse than majority voting ($p \leq 0.03$ with Wilcoxon matched-pairs signed-ranks test), and GWV-MI achieves second best SI, after LWV-MI ($p \leq 0.02$). A slice comparing LWV-MSD and GWV-MI is shown in Fig. 7. A comparable results arises for other regions as the putamen or the cerebellum white matter, which also display low contrast with neighboring structures.

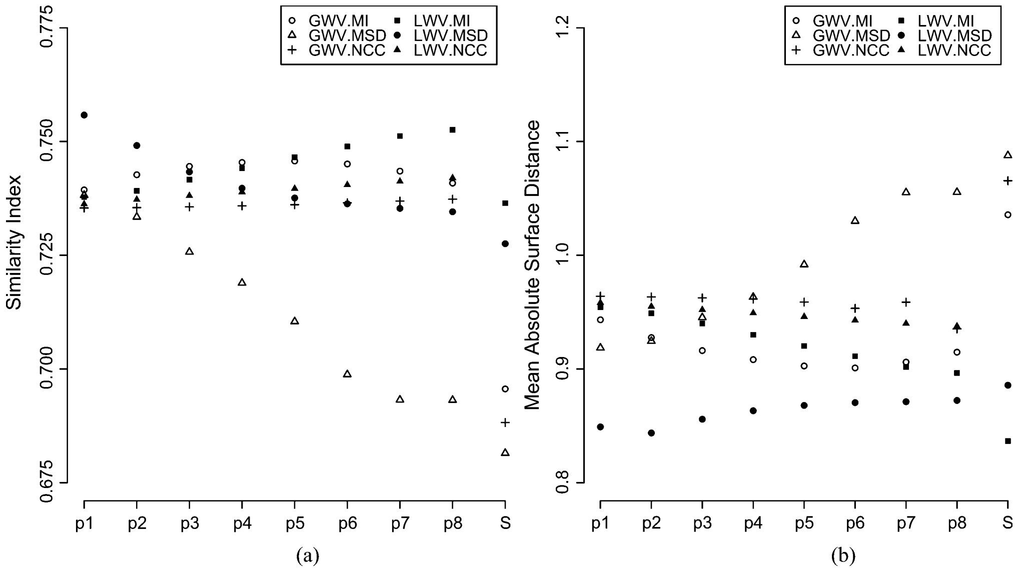
Table VI shows the mean values of SI and MASD obtained if we select the best fusion rule for each structure, based on the corresponding SI values. As a comparison, the SI and MASD values are shown for majority voting as well. In summary, there is a 3.41% gain in SI and 12.00% reduction in the average distance between surfaces with respect to majority voting.

With our current implementation and on an Intel Xeon 3.20-GHz processor, the approximate average computation times for segmentation combination in each brain scan were: 10 s for Majority Voting, 1 h for STAPLE, 2 min for GWV-MI, 4 min for GWV-NCC, 4 min for GWV-MSD, 5 min for LWV-MSD, 10 min for LWV-NCC, and 2 h for LWV-MI. The long time required by LWV-MI is due to the high computational burden of estimating the entropies from the histograms for each voxel where local weighted voting is required.

F. Influence of Tunable Parameter Selection

As it was done in the phantom images, we also studied the effect of varying both the gain (p) and neighborhood radius (r) parameters on the performance of the different combination methods. Results are summed up in Figs. 8 and 9.

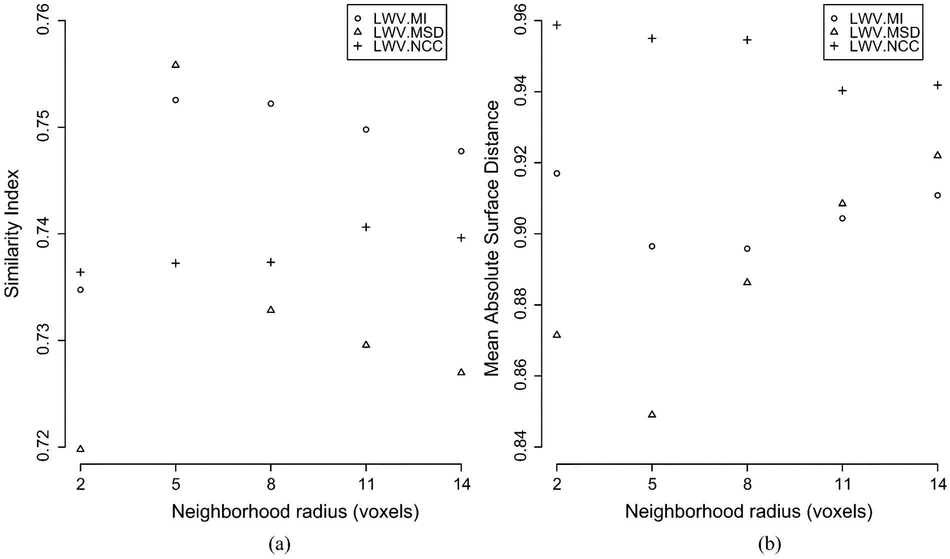
In the case of varying p, results are in principle not as clear as in the phantom dataset. As expected, GWV-MSD shows the greatest variability, $p = 1$ being the optimum solution. Both GWV-NCC and LWV-NCC vary very slightly through all values, implying that gain selection is not relevant. GWV-MI also varies smoothly and shows a maximum at $p = 5$. LWV-MI shows a continuous improvement from $p = 1$ to $p = 8$, in contrast with LWV-MSD which degrades its performance in



(a)

(b)

Fig. 8. Plots showing the effect of varying the gain factor p on mean SI and MASD over all regions for different combination strategies on IBSR images. p is varied from 1 to 8. S indicates *selection*, that is, the voxel with the highest weight is selected, without any further voting process. $r = 5$ was set for all LWV methods. (a) SI for varying gain factor p. (b) MASD for varying gain factor p.



(a)

(b)

Fig. 9. Plots showing effect of varying neighborhood region radius r on mean SI and MASD over all regions for different combination strategies on IBSR images. For all methods, p was the set to the value employed in the rest of experiments. (a) SI for varying neighborhood region radius r. (b) MASD for varying neighborhood region radius r.

that direction. In all cases the parameter selection method that we used was able to detect the optimum p value except for LWV-MI, where the second best was found.

For local combination strategies, varying the neighborhood radius r had different effects on the different combination techniques. For LWV-MSD, $r = 5$ proved the optimum global radius with a considerably sharp peak. In the case of LWV-MI and LWV-NCC the radius value does not greatly influence the overall performance, except for the $r = 2$ value for LWV-MI which clearly causes worse results.

V. Discussion

The results obtained from the experiments with phantom images and real brain MR images were slightly different. While for the phantom images LWV-MSD performs better or equal as all the other methods for all regions with a single exception, this is not the case with the brain images. On the phantom images, LWV-MSD is in general the best combination method, followed by LWV-MI and GWV-MI. The difference between them is smaller in regions that display low contrast with the background. We believe that this is due to the lack of contrasted edges. For each voxel, each segmentation is weighted according to a local similarity measure between the target image and the registered atlas image. If a high contrast between neighboring regions exists at that point, the similarity measure will be able to distinguish between accurate and inaccurate segmentations, because it will be sensitive to the overlap between the different regions. However, if no intensity difference exists, the similarity measure will not be able to assess the accuracy of the registration at that point. Moreover, the weights derived from the local mean square distance might be influenced by noise.

On the IBSR dataset, though, LWV-MSD is not the best choice for all regions. As pointed out in the results section, in regions that show low contrast with neighboring structures, other methods should be preferred. Even global voting strategies perform better than LWV-MSD in those cases. One of the reasons for global approaches to perform as good as or even better than local ones in low-contrast regions has been explained above. Another reason, which applies only when a manual segmentation is used as the ground truth, is related to the accuracy of manual segmentations in low-contrast regions. It is obvious that it is hard to manually delineate a structure such as the pallidum in an MR image, due to the lack of contrast. As a result, manual segmentations in those areas might be less accurate than those of other better defined regions. In those regions, features as shape or relative position play a much more important role than actual voxel intensity. Therefore, global voting approaches have an advantage over local methods, for they favour typical shapes and relative positions rather than voxel intensity distributions.

If LWV-MI and LWV-MSD are compared, the latter is better for more regions but the former proved less sensitive to noise. This means that in low-contrast regions, LWV-MSD can yield worse accuracy than majority voting, as seen in the case of the putamen or the pallidum. If differences in voxel gray values are only due to noise, as it is the case when there are no intrinsic intensity differences, using MSD as local similarity measure might lead to incorrect weights in the voting process. A possible solution would be to increase the neighborhood radius r in order to increase the intrinsic regularization. However, our experiments have shown that this can reduce the mean SI and MASD. It must be noted that LWV-MSD could be used on this image dataset because voxel gray levels had been previously normalized, and the basic assumption of intensity conservation among the images was fulfilled [25]. We believe this is the reason why LWV-MSD outperformed LWV-NCC.

The fact that STAPLE yields worse SI values than the rest of the methods is an unexpected outcome, for it departs from results reported in [18]. Reasons might be multiple. First, in the referred publication recognition rate is used for evaluation, instead of SI and MASD. Second, STAPLE was tested on a completely different image dataset. Third, the lack of any *a priori* probabilities may affect the algorithm negatively, and performance could improve significantly if performance estimates could be initialized within a reasonable range [17]. Moreover, in [20] majority voting and STAPLE showed mostly not statistically significant differences when combining segmentations of MR image of the prostate, and in some atlas sets majority voting performed better.

We propose using a representative image from the dataset to set the two variable parameters—gain factor and neighborhood radius—of the local combination algorithms, and then applying those parameters for all images of the same type. The simple assumption behind this approach is that the parameters that are optimum for a single image should not be far from the globally optimum parameters, given that all images have approximately the same intensity distribution. In our two test datasets, our parameter selection method proved robust enough and systematic comparison shows also that a wide range of parameters yield comparable results. However, higher robustness could be obtained by employing more images for the parameter selection.

Many publications on segmentation present a new method customized for a particular task, compare it with previous techniques and show the advantages of the novel contribution. Instead, our results suggest that, when combining multiple segmentations in atlas-based segmentation, the principle of the "no-panacea theorem" applies [35]: no method is better than others always, for all regions and images. One must select the best strategy among the existing (majority voting, global weighted voting, any kind of local weighted voting, etc.) according to the particular characteristics of the images and the regions. This can lead to substantial gain compared to majority voting, as results from Table VI show. We believe that a major contribution of our work is the study of the conditions in which local weighting methods perform consistently better than global methods. As it was explained, regions that have large intensity contrast with neighboring tissue, such as the ventricles in T1 MR images, benefit specially from local strategies. In contrast, there will be little or no benefit from using local approaches in regions that show similar intensities to the surrounding structures. In those cases global methods should be used. We believe that the same concept can be applied to segmenting images obtained using other imaging modalities, such as in computed tomography.

The main limitation of voting strategies is that they can not correct incorrect segmentations when all candidate segmentations have failed in a certain area. An extensive atlas which is representative of the whole population and a good registration algorithm can greatly limit this problem, by maximizing the probability that at least one of the candidate segmentations will be correct in all locations. Then, the local similarity measure would need to be sensitive enough to detect that correct registration and discard the incorrect ones.

The benefit that we have shown of using local approaches in high-contrast edges, suggests that it might be worth adding information on the edge strength when locally selecting a combination method. That is, a given structure might have different kind of borders: It might have a high-contrast border with one structure and a low-contrast border with another. Then, if local strategies were applied only on the high-contrast borders, overall accuracy might be better. This would be an interesting direction for future work.

VI. Conclusion

In this article, we have addressed the issue of combining segmentations to achieve the highest possible accuracy in multi-atlas medical image segmentation. We proposed a general local weighted voting method and showed how it can be applied with different similarity measures on different image datasets. We studied the performance of global and local weighted voting strategies for multi-atlas segmentation combination, and concluded that local methods should be preferred in regions that show high contrast with neighbor areas. To achieve optimum overall results, the best fusion method for each region must be found.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their insightful remarks, which have contributed greatly to improve the content of the paper. This work benefited from the use of Elastix, software developed by S. Klein and M. Staring, carried out at the Image Sciences Institute (ISI), on funding granted by the Netherlands Organization for Scientific Research (NWO). It is available at http://www.isi.uu.nl/Elastix/. The *B*-spline deformation generation code used for the artificial phantoms was developed by I. Arganda-Carreras, from the Spanish National Center of Biotechnology. The MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/.

REFERENCES

[1] H. Park, P. Bland, and C. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE Trans. Med. Imag.*, vol. 22, no. 4, pp. 483–492, Apr. 2003.

[2] I. Sluimer, M. Prokop, and B. van Ginneken, "Toward automated segmentation of the pathological lung in CT," *IEEE Trans. Med. Imag.*, vol. 24, no. 8, pp. 1025–1038, Aug. 2005.

[3] B. Dawant, S. Hartmann, J. Thirion, F. Maes, D. Vandermeulen, and P. Demaerel, "Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations. I. Methodology and validation on normal subjects," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 909–916, Oct. 1999.

[4] M. Lorenzo-Valdes, G. I. Sanchez-Ortiz, A. G. Elkington, R. H. Mohiaddin, and D. Rueckert, "Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm," *Med. Image Anal.*, vol. 8, no. 3, pp. 255–265, Sep. 2004.

[5] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, Jr., "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, 2004.

[6] J. Sethian*, Level Set Methods and Fast Marching Methods*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[7] S. Beucher and F. Meyer*, The Morphological Approach to Segmentation: The Watershed Transformation In: Mathematical Morphology in Image Processing*, E. Dougherty, Ed. New York: Marcel Dekker, 1992.

[8] A. Klein, B. Mensh, S. Ghosh, J. Tourville, and J. Hirsch, "Mindboggle: Automated brain labeling with multiple atlases," *BMC Med. Imag.*, vol. 5, no. 1, Oct. 2005.

[9] C. Svarer, K. Madsen, S. Hasselbach, L. Pinborg, S. Haugbol, V. Frojaer, S. Holm, O. B. Paulson, and G. Knudsen, "MR-based automatic delineation of interest in human brain PET images using probability maps," *NeuroImage*, vol. 24, no. 4, pp. 969–979, Feb. 2005.

[10] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer, Jr, "Quo vadis, atlas-based segmentation?," in *The Handbook of Medical Image Analysis—Volume III: Registration Models*. New York: Kluwer Academic/Plenum, 2005, ch. 11, pp. 435–486.

[11] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, Oct. 2006.

[12] T. Rohlfing, R. Brandt, C. R. Maurer, Jr, and R. Menzel, L. Staib, Ed., "Bee brains, B-splines and computational democracy: Generating an average shape atlas," in *IEEE Workshop Math. Methods Biomed. Image Anal.*, Los Alamitos, CA, 2001, pp. 187–194.

[13] P. Kochunov, J. L. Lancaster, R. W. P. Thompson, J. Mazziotta, J. Hardies, and P. Fox, "Regional spatial normalization: Toward an optimal target," *J. Comput. Assist. Tomogr.*, vol. 25, no. 5, pp. 805–816, 2001.

[14] T. Rohlfing and C. R. Maurer, Jr., "Multi-classifier framework for atlas-based image segmentation," *Pattern Recognit. Lett.*, vol. 26, no. 13, pp. 2070–2079, Oct. 2005.

[15] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[16] L. I. Kuncheva*, Combining Pattern Classifiers: Methods and Algorithms*. New York: Wiley, 2004.

[17] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[18] T. Rohlfing, D. Russakoff, and C. R. Maurer, Jr., "Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 983–994, Aug. 2004.

[19] T. Rohlfing and C. R. Maurer, Jr., "Shape-based averaging," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 153–161, Jan. 2007.

[20] S. Klein, U. van der Heide, I. Lips, M. van Vulpen, M. Staring, and J. Pluim, "Automatic segmentation of the prostate in 3-D MR images by atlas matching using localised mutual information," *Med. Phys.*, vol. 35, no. 4, pp. 1407–1417, 2008.

[21] D. Mattes, D. R. Haynor, H. Vesselle, T. K. Lewellyn, and W. Eubank, "Nonrigid multimodality image registration," in *Proc. SPIE Med. Imag. 2001: Image Process.*, M. Sonka and K. M. Hanson, Eds., 2001, vol. 4322, pp. 1609–1620.

[22] M. Wu, C. Rosano, P. Lopez-Garcia, C. S. Carter, and H. J. Aizenstein, "Optimum template selection for atlas-based segmentation," *NeuroImage*, vol. 34, no. 4, pp. 1612–1618, Feb. 2007.

[23] X. Artaechevarria, A. Munoz-Barrutia, and C. O. de Solórzano, "Efficient classifier generation and weighted voting for atlas-based segmentation: Two small steps faster and closer to the combination oracle," *SPIE Med. Imag.: Image Process.*, vol. 6914, no. 3, pp. 69141W-1–69141W-9, 2008.

[24] W. R. Crum, L. D. Griffin, D. G. Hill, and D. J. Hawkes, "Zen and the art of medical image registration: Correspondence, homology, and quality," *NeuroImage*, vol. 20, no. 3, pp. 1425–1437, Nov. 2003.

[25] A. Roche, G. Malandain, and N. Ayache, "Unifying maximum likelihood approaches in medical image registration," *Int. J. Imag. Syst. Technol.*, pp. 71–80, 2000.

[26] R. L. Gregg and R. D. Nowak, "Noise removal methods for high resolution MRI," in *IEEE 1997 Nucl. Sci. Symp.*, vol. 2, pp. 1117–1121.

[27] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hil, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imag.*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[28] S. Klein and M. Staring, Elastix [Online]. Available: http://www.isi.uu.nl/Elastix/

[29] S. Klein, M. Staring, and J. Pluim, "Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2879–2890, Dec. 2007.

[30] A. Zijdenbos, B. Dawant, R. Margolin, and A. Palmer, "Morphometric analysis of white matter lesions in MR images: Method and validation," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 716–724, Dec. 1994.

[31] G. Gerig, M. Jomier, and M. Chakos, "VALMET: A new validation tool for assesing and improving 3-D object segmentation," in *Lecture Notes in Computer Science*. New York: Springer-Verlag, 2001, vol. 2208, pp. 516–523.

[32] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics*, vol. 1, no. 6, pp. 80–83, Dec. 1945.

[33] Internet brain segmentation repository [Online]. Available: http://www.cma.mgh.harvard.edu/ibsr/

[34] E. van Rikxoort, Y. Arzhaeva, and B. van Ginneken, "A multi-atlas approach to automatic segmentation of the caudate nucleus in MR brain images," in *3-D Segmentation In The Clinic: A Grand Challenge*, T. Heimann, M. Styner, and B. van Ginneken, Eds., 2007, pp. 29–36.

[35] R. Hu and R. I. Damper, A "No Panacea Theorem" for classifier classification Aug. 2008, vol. 41, pp. 2665–2673.