

Multi-Atlas Segmentation with Joint Label Fusion

Hongzhi Wang, Member, IEEE, Jung W. Suh, Sandhitsu R. Das, John B. Pluta, Caryne Craige, and Paul A. Yushkevich, Member, IEEE

Introduction:atlas segmentation

- ★ Motivation: segmentation correlates with image appearance
- ★ Multi-Atlas Label fusion (MALF) Segmentation
- ★ Errors produced by atlas-based segmentation
- ★ Solution: optimally constructing a single atlas or multiple atlases:
 - 1-constructing one representative atlas for each mode obtained from clustering training images
 - 2-selecting the most relevant atlases for the unknown image on-the-fly

Label Fusion Techniques

★ **1. weighted voting:** Most existing label fusion method

$$\hat{S}_T^l(x) = \sum_{i=1}^n w_i(x) S_i^l(x),$$

- Spatially varying weights
- Common property: weights being computed independently for each atlas
- ❌ less effective when the label errors produced by the atlases are not independent (repeated atlases/bias)

★ **2. Majority voting(MV):** it is done among a small subset of atlases that globally or locally best match the target image, discarding poor matching atlases

- Removing independent noise
- ❌ lose the attractive property of voting
- ❌ Higher expected error
- ❌ Assign similar weights to different atlases with different registration quality


$$\hat{S}_T(x) = \operatorname{argmax}_{l \in \{1 \dots L\}} \sum_{i=1}^n S_i^l(x),$$

$$S_i^l(x) = \begin{cases} 1 & \text{if } S_i(x) = l; \\ 0 & \text{otherwise.} \end{cases}$$

Label Fusion Techniques

- ★ develop segmentation quality estimations based on local appearance similarity
- ★ images with similar appearance are more likely to have similar segmentation
- ★ Estimate weights by:
 - summed squared distance (SSD)
 - Gaussian weighting model

$$A_1 = (F_1, S_1), \dots, A_n = (F_n, S_n)$$

$$w_i(x) = \frac{1}{Z(x)} \left[\sum_{y \in \mathcal{N}(x)} (F_T(y) - F_i(y))^2 \right]^{-\beta}$$


$$w_i(x) = \frac{1}{Z(x)} e^{-\sum_{y \in \mathcal{N}(x)} [F_T(y) - F_i(y)]^2 / \sigma}$$

Beta and sigma: controlling the weight distribution

Novel alternative Label Fusion Technique?

★ Desired property:

reduce the bias due to the fact that atlases may produce correlated segmentation errors, without sacrificing the attractive properties of voting

Joint Label Fusion Approach

Goal: Minimize total labeling error while accounting for correlated errors between atlases.

- ★ assumption: binary segmentation
- ★ In multi-label segmentation problems: produce weight maps using weighted voting to compute a consensus segmentation for each label, and selecting at each voxel the label with the highest value of the consensus segmentation

$$S_T(x) = S_i(x) + \delta^i(x) \quad \begin{array}{l} \delta^i(x) \in \{-1, 0\} \text{ when } S_i(x) = 1 \\ \delta^i(x) \in \{0, 1\} \text{ when } S_i(x) = 0 \end{array}$$

$$\bar{S}(x) = \sum_{i=1}^n w_i(x) S_i(x)$$

$$\star E_{\delta^1(x), \dots, \delta^n(x)}[(S_T(x) - \bar{S}(x))^2 \mid F_T, F_1, \dots, F_n] = \dots = \mathbf{w}_x^t M_x \mathbf{w}_x,$$

$$M_x(i, j) = E_{\delta^i(x) \delta^j(x)}[\delta^i(x) \delta^j(x) \mid F_T, F_1, \dots, F_n]$$

$$= p(\delta^i(x) \delta^j(x) = 1 \mid F_T, F_1, \dots, F_n) .$$

Joint Label Fusion

- ★ combined label difference should be minimized

$$\mathbf{w}_x^* = \underset{\mathbf{w}_x}{\operatorname{argmin}} \mathbf{w}_x^t M_x \mathbf{w}_x \quad \text{subject to} \quad \sum_{i=1}^n \mathbf{w}_x(i) = 1.$$

- ★ Using lagrange multiplier: $\mathbf{w}_x = \frac{M_x^{-1} \mathbf{1}_n}{\mathbf{1}_n^t M_x^{-1} \mathbf{1}_n},$

- ★ If M_x is not full rank :quadratic programming

$$\mathbf{w}_x^t (M_x + \alpha I) \mathbf{w}_x = \mathbf{w}_x^t M_x \mathbf{w}_x + \alpha \|\mathbf{w}_x\|_2 \quad \text{subject to} \\ \sum_{i=1}^n \mathbf{w}_x(i) = 1.$$

Mx matrix Estimation Using Intensity similarity

Mx: expected pairwise joint label differences between the atlases and the target image

$$p(\delta^i(x)\delta^j(x) = 1 \mid F_T, F_i, F_j) =$$

$$p(\delta^i(x)\delta^j(x) = 1 \mid \{F_T(y), F_i(y), F_j(y) \mid y \in \mathcal{N}(x)\}).$$

adapt the inverse distance function:

$$Mx(i, j) = p(\delta^i(x)\delta^j(x) = 1 \mid \{F_T(y), F_i(y), F_j(y) \mid y \in \mathcal{N}(x)\})$$

$$\propto \left[\sum_{y \in \mathcal{N}(x)} |F_T(y) - F_i(y)| |F_T(y) - F_j(y)| \right]^\beta.$$

Refine label fusion by local patch search

- ★ Registration Errors: registration algorithm fails to accurately match corresponding structures between atlas and target images, leading to errors in multi-atlas label fusion (MALF).
 - constraints, along with other issues like failing to find the global optimum, prevent registration from achieving maximum local similarity between patches in the target and atlas images.
- ★ Local Patch Similarity: Sometimes, a nearby point in the atlas image patch provides a better match to the target patch than the originally aligned patch.
- ★ Local Patch Search Technique: By allowing for a search in the neighborhood for the most similar patch in the atlas image-→MALF performance can be moderately improved

$$\xi_i(x) = \arg \min_{x' \in \mathcal{N}'(x)} \|F_i(\mathcal{N}(x')) - F_T(\mathcal{N}(x))\|^2.$$

$$\bar{S}_T(x) = \sum_{i=1}^n w_i(\xi_i(x)) S_i(\xi_i(x)).$$

Experimental Setup

- ★ Two Segmentation Problems: The method is applied to segment the hippocampus using T1-weighted MRI and hippocampal subfields using T2-weighted MRI.
- ★ Free Parameters: optimized via exhaustive search using cross-validation
 - r : Radius of the local appearance window for similarity-based estimation.
 - r_s : Radius of the local search window for correcting registration errors.
 - σ : Parameter for transferring image similarities in joint label differences.
- ★ These parameters are optimized via exhaustive search using cross-validation.
- ★ Normalization and Smoothing:
 - The intensity vector from each image patch is normalized to have zero mean and constant norm.
 - Mean filter smoothing is applied to ensure spatial consistency in the voting weights across neighboring voxels in the atlas

Experimental Setup

- ★ Data: Hippocampal segmentation using MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI).
- ★ Dataset: 139 MRI images (57 control subjects, 82 MCI patients).
- ★ Methodology: Cross-validation with 20 test images, comparing Joint Label Fusion to majority voting, STAPLE, and local weighted voting methods (Gaussian, inverse distance)

Whole Hippocampus Segmentation

- ★ 1- Obtain manual segmentation: used a landmark-based semi-automatic hippocampal segmentation method to produce the initial segmentation for each image. Each fully labeled hippocampus was then manually edited using the paintbrush and polygon manual segmentation tools in ITK-SNAP by one of the authors
- ★ Outer cross validation: 20 atlas subset and 20 test -> inner cross validation for finding optimal parameters
- ★ 2- registration (all pairs of atlases and atlas-target)
 - Global
 - deformable

Label Fusion Strategy	Dice Similarity (Left Hippocampus)	Dice Similarity (Right Hippocampus)
Majority Voting	0.836 ± 0.084	0.829 ± 0.069
STAPLE [40]	0.846 ± 0.086	0.841 ± 0.086
LWGaussian	(0.885 ± 0.025) 0.886 ± 0.027	(0.873 ± 0.030) 0.875 ± 0.030
LWInverse	(0.884 ± 0.026) 0.885 ± 0.027	(0.872 ± 0.030) 0.873 ± 0.030
LWJoint	(0.893 ± 0.025) 0.897 ± 0.024	(0.884 ± 0.027) 0.888 ± 0.026

Average Dice similarity (\pm standard deviation) are computed across 10 outer cross-validation experiments, each having 20 test images. The results produced by each local weighted voting method without applying local search are shown in parentheses. The greatest similarity is obtained using the proposed LWJoint method.

	Left Hippocampus			Right Hippocampus		
Label Fusion Method	Volume (CTL)	Volume (MCI)	Cohen's d	Volume (CTL)	Volume (MCI)	Cohen's d
LWGaussian	2026 ± 277	1642 ± 334	1.7260	1947 ± 311	1553 ± 346	1.5576
LWInverse	2014 ± 274	1635 ± 326	1.7266	1930 ± 309	1544 ± 338	1.5504
LWJoint	2156 ± 285	1755 ± 353	1.7468	2083 ± 322	1668 ± 373	1.5700
Reference Seg.	2258 ± 325	1841 ± 368	1.5747	2201 ± 378	1785 ± 408	1.3643

The results are averaged over 10 cross-validation experiments, which together include test images from 94 control subjects and 106 MCI subjects. The last column shows the corresponding Cohen's d effect size, whose magnitude indicates the difference between the two populations. The hippocampus volume is normalized by intracranial volume for computing the Cohen's effect size.

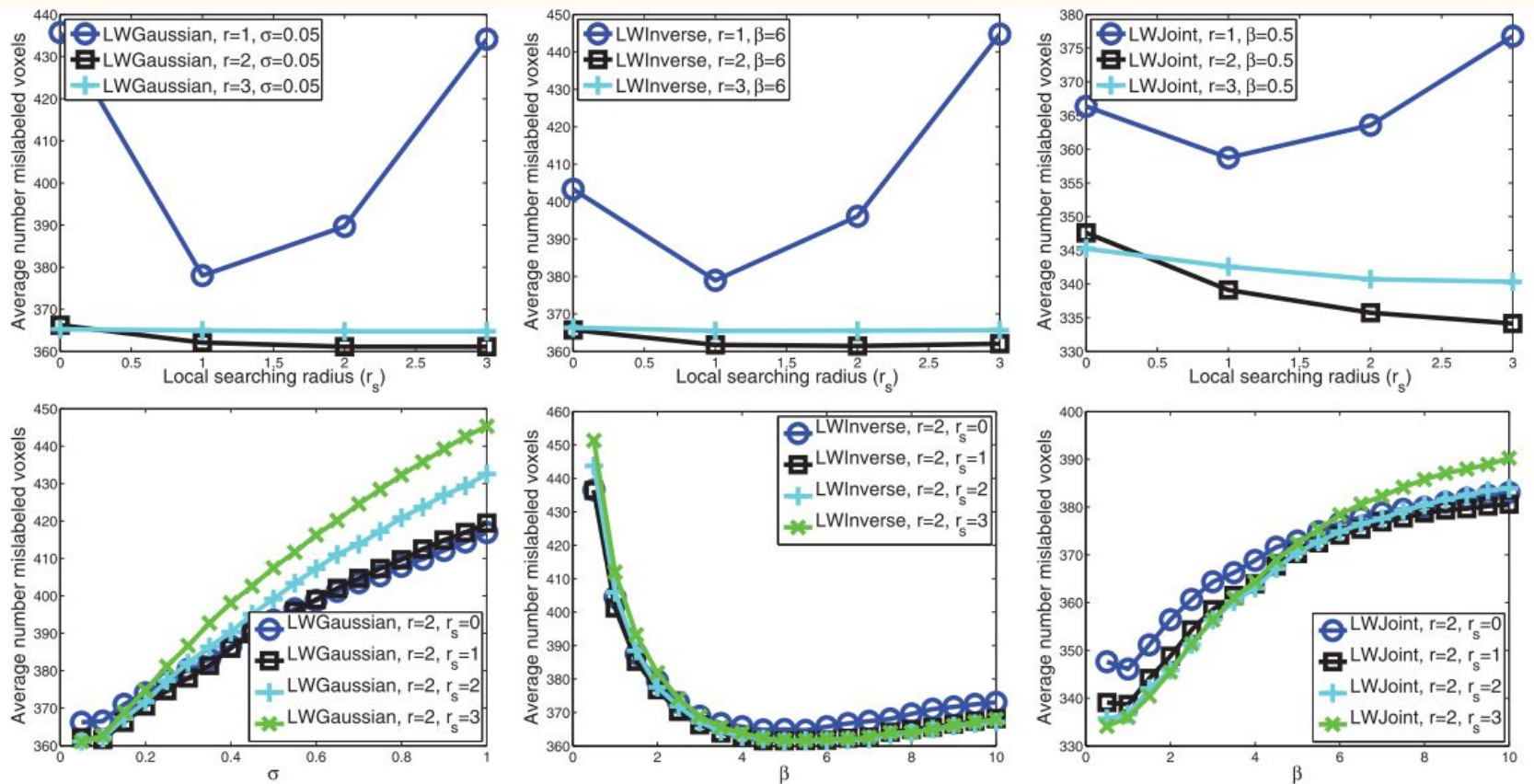


Fig. 1. Optimal label fusion parameter selection for LWGaussian (left), LWInverse (middle), and LWJoint (right) using leave-one-out cross validation. The upper figures plot the average number of mislabeled voxels against the local searching radius r_s and the appearance window radius r . The weighting function parameters σ , β are held fixed in these figures at its optimal value for the three methods, respectively. The lower figures plot the average number of mislabeled voxels against the local searching radius r_s and the weighting function parameter, σ and β , respectively. The appearance window radius r is held fixed in this figure at its optimal value.

Hippocampal Subfield Segmentation

- **Subfields:** CA1, CA3, dentate gyrus (DG), and others.
- Joint Label Fusion outperforms local weighted voting methods in almost all subfields.
- Local patch search and refinement further improve segmentation accuracy.

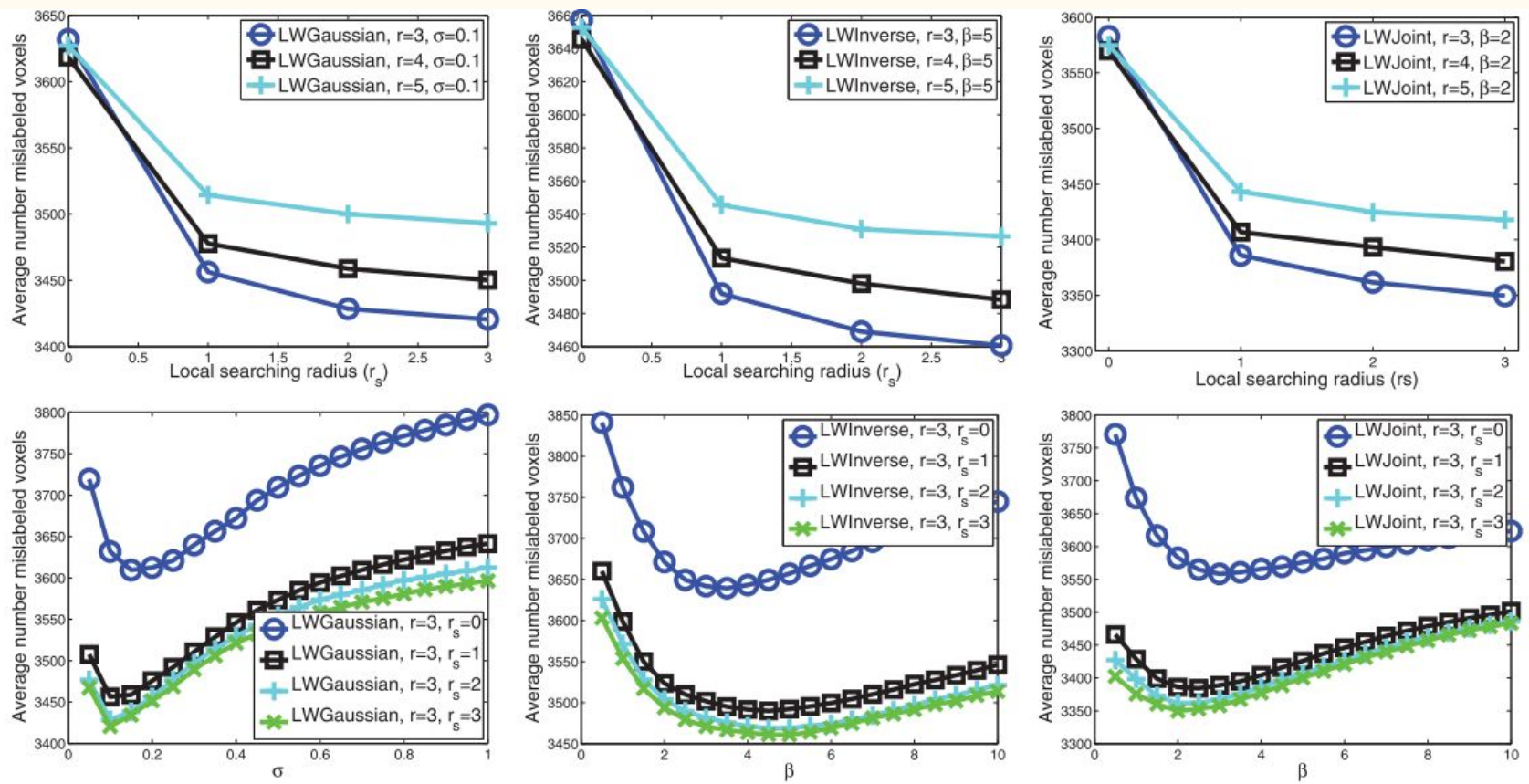


Fig. 3. Optimal label fusion parameter selection for LWGaussian (left), LWInverse (middle), and LWJoint (right) using leave-one-out cross validation. The upper figures plot the average number of mislabeled voxels against the local searching radius r_s and the appearance window radius r . The weighting function parameters σ , β are held fixed in these figures at its optimal value for the three methods, respectively. The lower figures plot the average number of mislabeled voxels against the local searching radius r_s and the weighting function parameter, σ and β , respectively. The appearance window radius r is held fixed in this figure at its optimal value.

TABLE 3
Average Performance of Different Label Fusion Strategies in Hippocampal Subfield Segmentation Experiments

Subfield	Maj. Voting	STAPLE [31]	[43] (Table 6)	LWInverse	LWGaussian	LWJoint
CA1	0.731±0.070	0.719±0.102	0.770±0.065	(0.766±0.064) 0.782±0.061	(0.769±0.063) 0.786±0.060	(0.773±0.059) 0.789±0.057**
CA2	0.322±0.180	0.357±0.191	0.422±0.175	(0.427±0.173) 0.427±0.177	(0.431±0.174) 0.431±0.178	(0.442±0.170) 0.455±0.176***
CA3	0.497±0.142	0.523±0.141	0.532±0.137	(0.537±0.124) 0.559±0.124	(0.541±0.125) 0.560±0.125	(0.549±0.121) 0.572±0.123***
DG	0.741±0.090	0.738±0.100	0.773±0.067	(0.768±0.070) 0.785±0.068	(0.767±0.070) 0.786±0.069	(0.767±0.070) 0.789±0.068*
head	0.864±0.028	0.860±0.033	0.874±0.025	(0.874±0.023) 0.882±0.024	(0.875±0.023) 0.883±0.024	(0.876±0.022) 0.885±0.023***
tail	0.739±0.123	0.720±0.144	0.744±0.119	(0.743±0.124) 0.752±0.128	(0.746±0.123) 0.755±0.129	(0.748±0.121) 0.759±0.125**
SUB	0.706±0.062	0.705±0.083	0.727±0.061	(0.726±0.067) 0.733±0.064	(0.729±0.066) 0.736±0.062	(0.734±0.061) 0.745±0.060***
ERC	0.606±0.138	0.603±0.131	0.627±0.123	(0.630±0.124) 0.647±0.128	(0.632±0.125) 0.648±0.129	(0.634±0.120) 0.652±0.126*
PHG	0.604±0.076	0.626±0.080	0.625±0.076	(0.629±0.075) 0.632±0.078	(0.629±0.076) 0.632±0.078	(0.630±0.074) 0.640±0.076***

* : $p < 0.05$ ** : $p < 0.01$ *** : $p < 0.0001$

Average Dice similarity (\pm standard deviation) are shown between the MALF result and the corresponding manual segmentation for each subfield, also averaging over left and right hemispheres and over 10 cross-validation experiments. The results produced by each local weighted voting methods without applying local search are shown in parentheses. Asterisks in the last column indicate results where the improvement of LWJoint over LWGaussian and LWInverse was statistically significant, as per paired Student's t -test.

TABLE 4
Average Performance of Different Label Fusion Strategies in Hippocampal Subfield Segmentation Experiments,
When Using Manual Head/Body/Tail Slice Partitioning (See Text)

Subfield	Maj. Voting	[43] (Table 1)	LWInverse	LWGaussian	LWJoint	Inter-Rater
CA1	0.804±0.059	0.851±0.040	(0.853±0.040) 0.869±0.038	(0.855±0.038) 0.870±0.037	(0.858±0.037) 0.874±0.035***	0.883±0.032
CA2	0.357±0.194	0.470±0.179	(0.474±0.183) 0.486±0.180	(0.477±0.182) 0.489±0.180	(0.487±0.177) 0.510±0.176***	0.522±0.160
CA3	0.530±0.145	0.583±0.133	(0.591±0.124) 0.619±0.123	(0.595±0.125) 0.620±0.123	(0.605±0.121) 0.634±0.123***	0.668±0.087
DG	0.813±0.087	0.859±0.045	(0.855±0.047) 0.870±0.046	(0.857±0.045) 0.871±0.045	(0.858±0.045) 0.875±0.044***	0.885±0.034
head	0.878±0.021	0.893±0.018	(0.892±0.016) 0.900±0.017	(0.893±0.015) 0.900±0.016	(0.894±0.016) 0.903±0.016***	0.900±0.016
tail	0.793±0.112	0.828±0.105	(0.804±0.105) 0.812±0.108	(0.806±0.104) 0.814±0.109	(0.809±0.103) 0.819±0.105***	0.901±0.059
SUB	0.715±0.062	0.742±0.063	(0.741±0.067) 0.747±0.065	(0.743±0.066) 0.749±0.064	(0.748±0.062) 0.758±0.062***	0.768±0.079
ERC	0.606±0.138	0.738±0.093	(0.745±0.096) 0.759±0.095	(0.745±0.096) 0.759±0.095	(0.750±0.094) 0.768±0.090***	0.786±0.123
PHG	0.627±0.074	0.658±0.073	(0.662±0.071) 0.666±0.075	(0.662±0.072) 0.666±0.075	(0.664±0.071) 0.675±0.072***	0.706±0.106

* : $p < 0.05$ ** : $p < 0.01$ *** : $p < 0.0001$

Average Dice similarity (± standard deviation) are shown between the MALF result and corresponding manual segmentation for each subfield, also averaging over left and right hemispheres and over 10 cross-validation experiments. The results produced by each local weighted voting method without applying local search are shown in parentheses. Asterisks in the LWJoint column indicate results where the improvement of LWJoint over LWGaussian and LWInverse was statistically significant, as per paired Student's t-test. The last column gives average Dice overlap between manual segmentations produced by two trained human raters.

Conclusion

Joint Label Fusion offers a significant improvement in multi-atlas segmentation by considering dependencies between atlases.

The method is computationally efficient and scales well to large datasets.