Mixture of Segmenters with Discriminative Spatial Regularization and Sparse Weight Selection [*]

Ting Chen[1], Baba C. Vemuri[1][**],
Anand Rangarajan[1], and Stephan J. Eisenschenk[2]

[1] Department of CISE, University of Florida, Gainesville, FL, USA
{tichen,vemuri,anand}@cise.ufl.edu
[2] Department of Neurology, University of Florida, Gainesville, FL, USA
stephan.eisenschenk@neurology.ufl.edu

Abstract. This paper presents a novel segmentation algorithm which automatically learns the combination of weak segmenters and builds a strong one based on the assumption that the locally weighted combination varies w.r.t. *both* the weak segmenters and the training images. We learn the weighted combination during the training stage using a discriminative spatial regularization which depends on training set labels. A closed form solution to the cost function is derived for this approach. In the testing stage, a sparse regularization scheme is imposed to avoid overfitting. To the best of our knowledge, such a segmentation technique has never been reported in literature and we empirically show that it significantly improves on the performances of the weak segmenters. After showcasing the performance of the algorithm in the context of atlas-based segmentation, we present comparisons to the existing weak segmenter combination strategies on a hippocampal data set.

1 Introduction

Brain MR image analysis and its associated application in the diagnosis and treatment of brain-based diseases has attracted immense attention in the past two decades. The segmentation of brain neuroanatomy is one of the key steps in medical image analysis. For example, researchers are interested in the study of hippocampal structures due to the critical role they play in many neuro-disorders including dementia, epilepsy and schizophrenia. In order to avoid the tedium involved in manual segmentation, a technique that is able to automatically segment the hippocampi from 3D brain MR scans is of great clinical interest.

Several techniques have been proposed in the literature to segment the hippocampus. One direct approach in [1] seeks to build an atlas from the training images with manual labels and deform it to the test image using a deformable

registration. However, it has limitations in accuracy for segmenting small structures like the hippocampus with low contrast intensity boundaries. Alternatively, there exists a class of methods that achieve more robustness and full automation by extracting features from the image at each voxel and resorting to machine learning techniques to label the voxel. In [2], Golland *et al.* use the support vector machine (SVM) to classify the features chosen by PCA from a large feature set. In [3], Morra *et al.* adopt Adaboost to select the features and resort to the SVM for classification. Both frameworks obtain relatively good approximation to the boundary of the hippocampus and the feature selection method simplifies the experts' effort in choosing informative features from a large feature set. However, one still needs to develop a very large feature pool containing potentially useful features,an non-trivial task in general.

Recently, it was shown that combining multiple atlas based segmentations improves the segmentation accuracy [4, 5]. As one of the most popular combination strategies, majority voting was shown to improve the accuracy and robustness of weak hypotheses. In [4], Artaechevarria *et al.* propose an image segmentation algorithm that combines multiple atlas-based segmenters based on weighted voting with the weights estimated from the local similarity between each atlas and the test image. Due to this specific design, the weak segmenters for this algorithm are limited to atlas-based segmentation. Besides, in order for the voting based technique to work well, a sufficiently large number of atlases are needed and a robust registration algorithm is required so as to have a relatively accurate segmentation for each atlas. A different combination strategy called SuperDyn was proposed by Khan *et al.* [5], wherein supervised learning was used for computing the weighted combination. Subsequently, dynamic information based on registration accuracy between the atlas and the test image was employed for the weight selection. SuperDyn *independently* estimates the weights for each weak segmenter at each voxel. However, this is clearly inadequate since it has been known for long that strong spatial dependencies exist in most real images [6]. Moreover, the aforementioned dynamic selection also restricts this technique to atlas-based weak segmenters.

In this paper, we propose a novel segmentation algorithm dubbed SegMix which is different from *all* of the aforementioned frameworks. Spatial smoothness and boundary discontinuities in the anatomical structures are explicitly incorporated into a discriminative regularizer in the training stage, resulting in a general technique capable of utilizing a vast variety of weak segmenters. SegMix assumes that the combination weights depend not only on the weak learners but also on the training data. *This is analogous to a medical consultation carried out by a group of doctors on a number of patients. It is well-justified to assume that each patient's personal condition has a different effect on the experts' final decision.* We carefully treat the problem of overfitting (which can occur from having too many weights) by utilizing the previously mentioned spatial weight regularization and via a novel non-parametric testing sieve. This makes SegMix substantially and thematically different from the other combination strategies, e.g. SuperDyn and Voting, which basically assume that the weights *only vary*

w.r.t. the weak segmenters. Due to this novel aspect of our framework, combining a very small number of weak segmenters can lead to dramatically significant improvements which is validated in our experiments. Furthermore, a novel scheme is imposed in the testing phase, wherein a strong segmenter is constructed by approximating the test image with a *sparse* combination of the training data and only choosing the learned weights corresponding to those sparsely selected training images. Intuitively speaking, in the training stage, we use a *cooperation* mechanism on the weak segmenters so as to achieve the best segmentation for each training image, while in the testing phase, we use a *competition* mechanism to select only the relevant images from the training set for a particular test image. The result is an algorithm driven by "*co-opetition*" which uses the previously learned cooperation skills of the competitively selected training data to let the weak segmenters collaborate and obtain a strong segmentation. *Note that as more expert driven manual delineations become available, they can be used as weak segmenters in our framework.*

2 Methodology
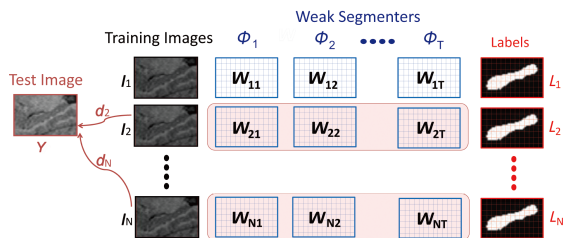
2.1 The Segmentation Mixture Setup

Fig. 1. Framework of the proposed algorithm.

In this section, we present the methodology of our algorithm. We begin with illustrating the basic framework of the technique in Fig.1. In the training stage, the optimal locally weighted combination of the weak segmentations are estimated to best approximate the ground truth label for each training image. As shown in the figure, the weight matrix W_{nt} is associated with the nth training image and t^{th} weak segmenter. Since we compute local weights, the spatial interactions among voxels are modeled closely following the Discriminative Random Field (DRF) [6]. In the testing stage, we compute the sparse combination of the training images to approximate the test image and only the learned weights associated with those selected training images are used to construct the strong segmenter for that particular test image. The figure shows an example of picking I_2 and I_N as the sparse representation of the test image and only the weights associated with them, $W_{21}, \ldots W_{2T}$ and $W_{N1}, \ldots W_{NT}$, are used to construct the final strong segmenter for test image Y.

Before getting into the details of the algorithm, we first introduce two key terms that will be used in the rest of this paper.

Signed Distance Functions (SDF) are used to represent the shape of the structure in the segmentation. Each weak segmentation output is represented by a SDF and the locally weighted combination of them corresponds to a strong segmenter. Several techniques exist in literature for linear combination of SDFs. In [7], Pohl *et al.* embed SDFs into the linear space of LogOdds, where addition and scalar multiplication are in closed form, while in [8] an SDF is mapped to the square-root density space via a Schrödinger wave function, where a variety of Riemannian operations are easily computed. However, we will resort to a simpler idea pointed out by Leventon *et al.* in [9]. They claim that using the signed distance transform shape representation is tolerant to slight misalignment and hence a rough alignment of the data during pre-processing will avoid solving for the general correspondence problem in SDF combination.

A **Neighborhood Graph G** is computed for each training data to store the spatial interactions and dependencies between voxels. Given a training sample, let $L(x)$ be the signed distance transform of the label image and M be the number of voxels. The $M \times M$-dimensional neighborhood graph matrix G is computed from the following DRF formulation: $G(i,j) = \exp(-||(L(x_i) - L(x_j))||_2^2)$, where $j \in \mathcal{N}(i)$. In 3D, the neighborhood region of the ith voxel $\mathcal{N}(i)$ is computed through $6, 18$ or 26-connectivity. This neighborhood graph will be used to gate the distance between the weights corresponding to the ith and j^{th} voxels.

2.2 Training Stage: Discriminative Spatial Weight Regularization

Our segmentation algorithm takes as input a set of weak segmentation results and combines them via a regression model. Assume there are T weak segmenters and the outputs of them are binary images $b_t(x), t = 1, 2, \ldots, T$ distinguishing the structures from the background. *For a given training image $I(x)$, $\Phi_t(x)$ is the signed distance transform computed from the tth weak segmentation output $b_t(x)$ and $L(x)$ is the signed distance transform of the ground truth label image, representing the true segmentation of $I(x)$. The algorithm assumes that $L(x)$ is the locally weighted combination of $\Phi_t(x)$.*

Since the parameters to be optimized depend on both the voxel locations and the weak segmenters, we solve this minimization problem voxel-wise by re-arranging each $\Phi_t(x)$ into a column vector and stacking them together column by column in the $M \times T$ matrix $\bar{\Phi}$, where M is the number of voxels in the image. Let the column vector $\boldsymbol{l}$ be the re-arrangement of $L(x)$ and denote column vector $\boldsymbol{\phi}_i$ as taken from the ith row of $\bar{\Phi}$, $\boldsymbol{l}_i$ as the ith entry of $\boldsymbol{l}$ and column vector $\boldsymbol{w}_i$ as the weights associated with $\boldsymbol{\phi}_i$.

It is well-justified to assume that the weights $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are expected to be similar if the jth voxel is in the neighborhood of the ith voxel. We therefore adopt a regularization term based on the pre-computed matrix G, which captures the similarity of the labels within the neighborhood. We eventually formalize our cost function in the following:

$$\boldsymbol{w}^* = arg \min_{\boldsymbol{w}} \sum_{i=1}^{M} ||\boldsymbol{w}_i \cdot \boldsymbol{\phi}_i - \boldsymbol{l}_i||_2^2 + \lambda \sum_{i,j=1}^{M} G(i,j)||\boldsymbol{w}_i - \boldsymbol{w}_j||_2^2.$$

A closed form solution can be derived for this objective function. We start by expanding the cost function and get

$$E(\{\boldsymbol{w}_i\}) = \sum_{i=1}^{M} \boldsymbol{w}_i^t \boldsymbol{\phi}_i \boldsymbol{\phi}_i^t \boldsymbol{w}_i + \sum_{i=1}^{M} \boldsymbol{l}_i^2 - 2 \sum_{i=1}^{M} \boldsymbol{l}_i \boldsymbol{w}_i^t \boldsymbol{\phi}_i$$

$$+ \lambda \sum_{i,j=1}^{M} G(i,j)(\boldsymbol{w}_i^t \boldsymbol{w}_i + \boldsymbol{w}_j^t \boldsymbol{w}_j - 2\boldsymbol{w}_i^t \boldsymbol{w}_j).$$

With the following notations: (1) $H_i = \boldsymbol{\phi}_i \boldsymbol{\phi}_i^t$, (2) $W^t = [\boldsymbol{w}_1^t, \dots \boldsymbol{w}_M^t]$, (3) $B_k = H_k + 2\lambda(\sum_{i=k,j\neq1} G(i,j) + \sum_{i\neq1,j=k} G(i,j))I_{T \times T}$, (4) $p^t = [\boldsymbol{l}_1 \boldsymbol{\phi}_1^t, \dots \boldsymbol{l}_M \boldsymbol{\phi}_M^t]$ and after some algebra, the cost function is re-arranged into a matrix form:

$$E = W^t \begin{pmatrix} B_1 & -2\lambda G(1,2)I_{T \times T} & \dots & -2\lambda G(1,N)I_{T \times T} \\ -2\lambda G(2,1)I_{T \times T} & B_2 & \dots & -2\lambda G(2,N)I_{T \times T} \\ \dots & \dots & \dots & \dots \\ -2\lambda G(N,1)I_{T \times T} & -2\lambda G(N,2)I_{T \times T} & \dots & B_N \end{pmatrix} W$$

$$- 2p^t W + \sum_{n=1}^{N} \boldsymbol{l}_n^2.$$

We take the derivative of E w.r.t. W and set the result to 0 in order to solve for the weights. We then have $\frac{\partial E}{\partial W} = (D^t + D)W - 2p^t = 0$, with D being the matrix in the equation above that contains B_k as diagonal. The problem is finally reduced to solving the following linear system $(D^t + D)W = 2p^t$. Note that since $D^t + D$ is a sparse matrix, we finally solve a *sparse least-squares problem* which can be efficiently performed.

2.3 Testing Stage: Sparse Linear Combination

Assume there are N training images and for each image, we solve for the local weights W to combine the weak segmenters. We denote by W_{nt}, the weight matrix for the nth training image and the tth weak segmenter, which is basically a single matrix involving the re-arrangement of W that we solved for in the training stage. To avoid overfitting, in the testing stage, not all the training results are used. *This is similar to the situation when a new patient comes in for medical consultation, we expect that a good strategy for the experts involves searching for useful relevant case studies from the old patients in order to arrive at a consensus diagnosis.* Therefore, only a subset of the trained parameters are helpful in testing. Several techniques can be used to achieve this goal, for instance the K-Nearest Neighbor (kNN) and the Sparse Representation methods. The kNN based search for the most similar cases to represent the testing data will potentially fail when all the training images differ from the test sample. Besides, one has to resort to a relatively complicated data structure for fast kNN implementation when the feature dimension is high. Recently, sparsity has been

investigated for feature selection [10]. Following this philosophy, we compute the optimal sparse combination of the training images to approximate the test data.

Formally, let A be a matrix with N columns where the ith column contains the ith training image and let Y be the given test image. To obtain a sparse combination of training set images, we use an L_1 norm regularizer. The problem is formalized as follows: $\gamma^* = arg\min_\gamma ||A\gamma - Y||_2^2 + \alpha||\gamma||_1$, which can be solved using existing techniques, such as LASSO. The final strong segmenter is then given by $S(x) = \sum_{n=1}^N \sum_{t=1}^T \gamma_n W_{nt}(x)\Phi_t(x)$.

3   Experiments

In this section, we empirically validate our SegMix algorithm and compare it with the widely used voting based methods. Our algorithm significantly improves the weak segmentation results given a small number of low-accuracy weak segmenters.

Hippocampus Data Set: This experiment is performed on a hippocampal data set containing 60 brain MRI (T1) images, with the right hippocampi manually segmented by an expert neurologist. We divide the data set into 2 groups. The first group contains 20 images used in building the multiple atlases. 10-fold cross validation is applied to the remaining 40 images. The original brain images are first corrected for intensity inhomogeneity and normalized, then registered to the same coordinate system using a similarity transformation. Since the hippocampi are within a certain region of the brain, we therefore define a bounding box that approximately encloses each hippocampus and only take these ROIs as the input to our segmentation algorithm. The size of ROI is $56 \times 39 \times 30$. We extract the ROI for each test image by first deforming the brain MRI scan to a labeled brain template and finding the ROI based on the template information.

Weak Segmenters: *Note that any segmentation method is applicable as a weak segmenter within our framework.* However, in order to demonstrate the robustness and performance of our algorithm and compare it to the existing multi-atlas segmentation methods, we use atlas-based segmentation [1] as the weak segmenter. We first cluster the 20 images into a set of groups based on the hippocampal shape information from the labels. The signed distance transform representation of each shape is mapped to the square-root density space via a Schrödinger wave function [8]. Hence, each shape corresponds to a single point on the high dimensional sphere and the similarity between the shapes is computed intrinsically using the geodesic distance on the unit sphere. Armed with this intrinsic similarity measure, any clustering method may be used here. We employ affinity propagation [11] since it does not require the number of clusters to be specified. We get 5 clusters and the atlases/centers for each group are used for the weak segmentations.

Performance Measure: The evaluation metrics for measuring the performance of the algorithm used in this paper include the *similarity index* (known as Dice coefficient) SIM= $\frac{2\mathcal{V}(A\cap B)}{\mathcal{V}(A)+\mathcal{V}(B)}$, which computes the overlap of two volumes and the *difference index* DIF= $\frac{2|\mathcal{V}(A)-\mathcal{V}(B)|}{\mathcal{V}(A)+\mathcal{V}(B)}$, which measures the size difference

of the two volumes. Here $\mathcal{V}(\cdot)$ computes the volume. A good segmentation has larger overlap with the ground truth hence higher similarity index, but lower difference index.

Experiment Setting: The free parameters involved in our SegMix algorithm include λ and α for the regularization in the training and testing stages respectively. We (empirically) discovered that the algorithm is not very sensitive to the particular choice of λ and α, so we set $\lambda = 0.5$ and $\alpha = 0.5$ throughout the comparison experiments. An 18-connected neighborhood is used for computing the graph G. The nonrigid registration algorithm used for atlas-based segmentation is Demons, where we use all the *default* parameter settings, i.e. all the registrations are performed with the same parameters. *Due to the use of standard defaults, atlas-based segmentation leads to low-accuracy for each weak segmenter.* However, the experimental results indicate that SegMix significantly improves on those weak segmentations.

Experimental Results: To validate our proposed algorithm, the experiments are performed on the 3D hippocampal MRI images using (1) SegMix, (2) Global Weighted Voting (GWV) and (3) Local Weighted Voting (LWV)[4] with 5 weak segmenters. In the following table, we list the average performance evaluation for the 10-fold cross validation of the 40 images.

Table 1. The average SIM and DIF indices for SegMix, GWV and LWV.

	weak SIM	final SIM	Increased	weak DIF	final DIF	Decreased
SegMix	0.68 ± 0.04	0.80 ± 0.03	**17.65%**	0.42 ± 0.10	0.09 ± 0.06	**78.57%**
GWV	0.68 ± 0.04	0.73 ± 0.04	7.35%	0.42 ± 0.10	0.37 ± 0.11	11.90%
LWV	0.68 ± 0.04	0.74 ± 0.04	8.82%	0.42 ± 0.10	0.39 ± 0.10	7.14%

Since better segmentation corresponds to larger SIM but smaller DIF, we show the increased SIM value and decreased DIF value w.r.t. the weak segmentations. Due to the low accuracy and limited number of the weak segmenters, the performance of the voting-based methods are poor as expected. We also present the SIM and the DIF values for both the weak segmentations and the final strong segmentation and show the improvements of our techniques w.r.t. the weak segmenters for each test image in Fig.2.

4   Conclusion

In this paper, we introduced a novel weak segmentation combination strategy based on the assumption that the locally weighted combination varies w.r.t. *both* the weak segmenters and the training images. We learned the weighted combination during the training stage using a discriminative spatial regularization which depends on training set labels. In the testing stage, a sparse regularization scheme was imposed to avoid overfitting. The experimental results indicated
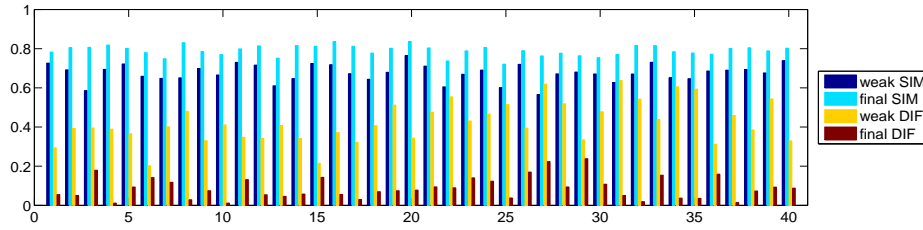
8

Fig. 2. The figure shows the SIM and DIF for each image.

that our algorithm not only outperforms the voting based methods but also significantly improves the performances of the weak segmenters.

References

1. Carmichael, O.T., Aizenstein, H.A., Davis, S.W., Becker, J.T., Thompson, P.M., Meltzer, C.C., Liu, Y.: Atlas-based Hippocampus Segmentation in Alzheimer's Disease and Mild Cognitive Impairment. NeuroImage 27(4), 979-990 (2005)
2. Golland, P., Grimson, W., Shenton, M., Kikinis, R.: Detection and Analysis of Statistical Differences in Anatomical Shape. Med. Image Anal. 9(1), 69-86 (2005)
3. Morra, J.H., Tu, Z., Apostolova, L.G., Green, A., Toga, A.W., Thompson,P.M.: Comparison of AdaBoost and Support Vector Machines for Detecting Alzheimer's Disease through Automated Hippocampal Segmentation. IEEE Trans. Med. Image 29(1), 30-42 (2010)
4. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de-Solorzano, C.: Combination Strategies in Multi-atlas Image Segmentation: Application to Brain MR Data. IEEE Trans. Med. Image 28(8), 1266-1277 (2009)
5. Khan, A., Cherbuin, N., Wen, W., Anstey, K.J., Sachdev, P., Beg, M.F.: Optimal Weights for Local Multi-atlas Fusion using Supervised Learning and Dynamic Information (SuperDyn): Validation on Hippocampus Segmentation. NeuroImage 56(1), 126-139 (2011)
6. Kumar, S., Hebert, M.: Discriminative Random Fields. Int. J. Comput. Vision 68(2), 179-201 (2006)
7. Pohl, K.M., Fisher, J., Shenton, M., McCarley, R.W., Grimson, W.E., Kikinis, R., Wells, W.M.: Logarithm Odds Maps for Shape Representation. In: Larsen, R., Nielsen M., Sporring, J. (eds.) MICCAI 2006, Part II. LNCS, vol.4191, pp. 955-963. Springer, Heidelberg (2006)
8. Chen, T., Rangarajan, A., Eisenschenk, S.J., Vemuri, B.C.: Construction of Neuroanatomical Shape Complex Atlas from 3D Brain MRI. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part III. LNCS, vol.6363, pp. 65-72. Springer, Heidelberg (2010)
9. Leventon, M.E., Grimson, W.E.L., Faugeras,O.: Statistical Shape Influence in Geodesic Active Contours. In: IEEE Conf. CVPR, pp. 316-323 (2000)
10. Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., Metaxas, D.: Automatic Image Annotation Using Group Sparsity. In: IEEE Conf. CVPR, pp. 3312-3319 (2010)
11. Frey, B.J., Dueck, D.: Clustering by Passing Messages Between Data Points. Science 315, 972-976 (2007)