Shape-Based Averaging for Combination of Multiple Segmentations

T. Rohlfing[1] and C.R. Maurer, Jr.[2]

[1] Neuroscience Program, SRI International, Menlo Park, CA, USA
torsten@synapse.sri.com
[2] Department of Neurosurgery, Stanford University, Stanford CA, USA
crmaurer@stanford.edu

Abstract. Combination of multiple segmentations has recently been introduced as an effective method to obtain segmentations that are more accurate than any of the individual input segmentations. This paper introduces a new way to combine multiple segmentations using a novel shape-based averaging method. Individual segmentations are combined based on the signed Euclidean distance maps of the labels in each input segmentation. Compared to label voting, the new combination method produces smoother, more regular output segmentations and avoids fragmentation of contiguous structures. Using publicly available segmented human brain MR images (IBSR database), we perform a quantitative comparison between shape-based averaging and label voting by combining random segmentations with controlled error magnitudes and known ground truth. Shape-based averaging generated combined segmentations that were closer to the ground truth than combinations from label voting for all numbers of input segmentations (up to ten). The relative advantage of shape-based averaging over voting was larger for fewer input segmentations, and larger for greater deviations of the input segmentations from the ground truth. We conclude that shape-based averaging improves the accuracy of combined segmentations, in particular when only a few input segmentations are available and when the quality of the input segmentations is low.

1 Introduction

Combination of multiple segmentations has recently been introduced as an effective method to obtain segmentations that are more accurate than any of the individual input segmentations [1,2,3,4]. Typically, such algorithms are based on local (i.e., voxel-wise) decision fusion schemes, such as voting, or on probability-theoretical combination of Bayesian classifiers that assign likelihoods to the possible output classes [5,6].

For classification of voxels in multi-dimensional images, this paper introduces a new way to combine multiple segmentations using a novel shape-based averaging method. Unlike many other classification problems, there is a natural distance relationship between the voxels of an n-dimensional image. We exploit this relationship to combine segmentations based on the signed Euclidean distance maps of the labels in each input segmentation. Compared to label voting, the new combination method produces smoother, more regular output segmentations, and it also produces segmentations that are closer to the ground truth as measured by the recognition rate.

J. Duncan and G. Gerig (Eds.): MICCAI 2005, LNCS 3750, pp. 838–845, 2005.

Our method is related to shape-based interpolation, which was introduced by Raya & Udupa [7] as a method for the interpolation of binary images. Grevera & Udupa [8] later extended the method to gray-level images by embedding an n-dimensional gray-level image into an $(n{+}1)$-dimensional binary image space. Our approach is similar in that we consider images with multiple classes of a segmentation. However, our approach is different insofar as it combines multiple such images on a common grid into one image, rather than resamples one image onto a new grid. In this sense, our method is a shape-based averaging method.

2 Methods

Let L be the number of classes in the segmentation. For simplicity, each class is identified with a number in the set $\Lambda = \{0, \dots, L{-}1\}$, where class 0 without loss of generality represents the image background. For K different (input) segmentations of the same image, let $s_k(\mathbf{x}) \in \Lambda$ for $k = 1, \dots, K$ be the class assigned to voxel $\mathbf{x}$ in segmentation k. We are particularly interested in atlas-based segmentations that are generated by mapping the coordinates of an image onto those of a segmented atlas image. For s_k, let the atlas image be A_k and the transformation $\mathbf{T}_k$, so that

$$s_k : \mathbf{x} \mapsto A_k(\mathbf{T}_k(\mathbf{x})) \in \Lambda. \tag{1}$$

2.1 Shape-Based Averaging of Segmentations

Let $d_{k,l}(\mathbf{x})$ be the signed Euclidean distance of the voxel at $\mathbf{x}$ from the nearest surface voxel with label l in segmentation k. The value of $d_{k,l}(\mathbf{x})$ is negative if $\mathbf{x}$ is inside structure l, positive if $\mathbf{x}$ is outside, zero if and only if $\mathbf{x}$ is a voxel on the surface of structure l in segmentation k. Note that in effect, we derive from each of the abstract-level classifications s_k a measurement-level classification $d_{k,*}$.

```
for all $\mathbf{x}$ do                          ▷ Loop over all voxels to initialize data structures
    $S(\mathbf{x}) \leftarrow L$                                  ▷ Set label to “undecided”
    $D_{\min}(\mathbf{x}) \leftarrow \infty$                      ▷ Initialize distance map as “far outside”
end for
for $l = 0, \dots, L{-}1$ do                                      ▷ Loop over all labels
    for all $\mathbf{x}$ do                                       ▷ Loop over all voxels
        $D \leftarrow \sum_k d_{k,l}(\mathbf{x})$     ▷ Total signed distances for this voxel and label
        if $D < D_{\min}(\mathbf{x})$ then            ▷ Is new distance smaller than current minimum?
            $S(\mathbf{x}) \leftarrow l$                          ▷ Update combined label map
            $D_{\min}(\mathbf{x}) \leftarrow D$                   ▷ Update minimum total distance
        end if
    end for
end for
```

Fig. 1. Shape-based averaging algorithm. See text for details

Based on the distance maps of all structures in all input segmentations, we define the total distance of voxel $\mathbf{x}$ from label l as

$$D_l(\mathbf{x}) = \sum_{k=1}^{K} d_{k,l}(\mathbf{x}). \tag{2}$$

Note that since the number of segmentations K is constant, the total distance is directly proportional to the average distance. The combined segmentation $S(\mathbf{x})$ for voxel $\mathbf{x}$ is now determined by minimizing the total distance from the combined label (and, equivalently, the average distance) as

$$S(\mathbf{x}) = \arg\min_{l \in \Lambda} D_l(\mathbf{x}). \tag{3}$$

This can be iteratively computed using the algorithm in Fig. 1. Note that at any given time, due to the incremental nature of our algorithm, it requires space for three distance maps independent of the number of classes L: 1) the individual distance map $d_{k,l}$ for the current input segmentation k and class l, 2) the total distance map D_l over all segmentations for class l, and 3) the minimum total distance map $D_{\min}$ over all classes so far.

The main computational burden of our method stems from repeatedly computing the Euclidean distance transformation. We use an efficient algorithm by Maurer *et al.* [9] that computes the exact Euclidean distance in linear time $O(N)$, where N is the number of voxels in the image.

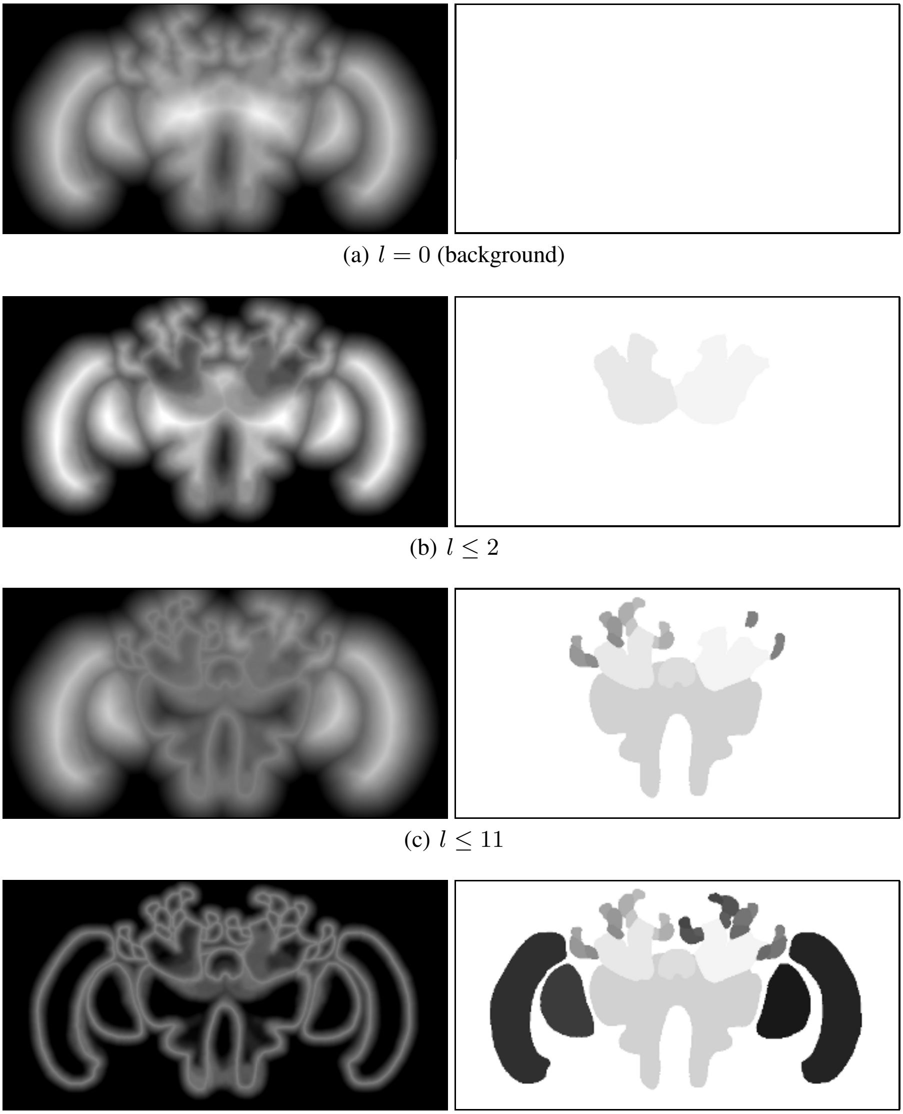
Examples of intermediate closest distance maps and corresponding label maps are illustrated in Fig. 2. For better graphical presentation, the images shown are from a simpler segmentation problem with a smaller number of less complex structures than there are in the human brain. As the algorithm iterates over all labels, areas that have been assigned to a label turn negative in the minimum total distance map (Fig. 2(b) and 2(c)), representing their location "inside" a structure. When all labels have been processed, the boundaries between structures are identified by the zero-level set in the final total distance map (Fig. 2(d)).

2.2 Label Voting

For comparison with our new method, we have implemented a standard segmentation combination scheme based on label voting. In atlas-based segmentation, labels need to be computed for non-grid locations in the atlas by interpolation. One interpolation technique that is applicable to label data and produces better results then nearest neighbor interpolation (NN) is partial volume (PV) interpolation, introduced by Maes *et al.* [10] for histogram generation in entropy-based image registration. Using PV interpolation, a vector of weights is returned as the classifier output, where each weight represents the relative share of one label. From these weights, we compute the combined segmentation for each voxel by sum fusion, i.e., by adding all weight vectors from the individual segmentations and selecting the class with the highest weight in the sum.

Note that by applying PV interpolation and sum fusion we effectively take advantage of the inherent sub-pixel resolution[1] of atlas-based segmentations. Other segmen-

[1] NB: sub-pixel *resolution* does not imply sub-pixel *accuracy*.



(a) $l = 0$ (background)

(b) $l \leq 2$

(c) $l \leq 11$

(d) $l \leq 22$ (final)

Fig. 2. Example of evolving minimum total distance maps $D_{\min}(\mathbf{x})$ (left image in each pair) and label maps $S(\mathbf{x})$ (right image in each pair). Brighter values in the distance maps correspond to positive values (i.e., outside of structures), darker values correspond to negative values (i.e., inside of structures). (a) Image background is canonically treated as an ordinary label. (b) First two non-background structures. (c) First 11 non-background structures. (d) Final combined segmentation (22 structures). For illustrative purposes, the images shown in this figure are from a different, simpler segmentation problem with a smaller number of less complex structures than there are in the human brain images used for quantitative evaluation in this paper.

tation methods, most notably manual segmentation, may not require label interpolation, in which case combination by sum fusion reduces to combination by vote fusion, i.e., counting of discrete votes for each label. The same is true for atlas-based segmentation when nearest (NN) is used instead of PV interpolation.

2.3 Evaluation

For quantitative evaluation of the combination method independent of the performance of a particular segmentation algorithm, we apply a strategy introduced by Rohlfing *et al.* [3]. This method evaluated classifier combination methods in atlas-based segmentation using segmentations with controlled error magnitudes and known ground truth. Based on the ground truth segmentation of an image, a random segmentation is generated by applying a nonrigid coordinate transformation of random magnitude. In particular, we apply B-spline free-form deformations (FFD) [11] with control point positions perturbed from the identity transformation by adding Gaussian-distributed random numbers. The choice of the FFD transformation model is motivated by its compact representation, as well as by the fact that it is also used in a popular nonrigid registration algorithm by Rueckert *et al.* [12].

A set of publicly available expert-segmented human brain MR images (T_1-weighted anatomical images) from ten subjects was obtained from the Internet Brain Segmentation Repository (IBSR; `http://www.cma.mgh.harvard.edu/ibsr/`). The corresponding segmentations with 43 anatomical structures provide the ground truths for the random segmentation evaluation outlined above (since we do not perform an actual segmentation, the anatomical MR images were not actually used in this study). All images had the same size, 256×256×128 voxels, with coronal slice orientation. The in-plane pixel size was either 0.9 mm or 1.0 mm. The slice spacing of all images was 1.5 mm.

For each image in our test set, twenty random segmentations were generated: ten with a standard deviation of the random perturbation of the FFD control points of $\sigma = 10\,\mathrm{mm}$, and another ten with $\sigma = 20\,\mathrm{mm}$. Note that larger values of σ correspond to larger magnitudes of the random FFDs, and thus to larger deviations of the random segmentations from the (undeformed) ground truth.

For each value of σ, we then computed combinations of two through ten of the respective random segmentations at that error level, once using shape-based averaging and once using label voting. The accuracy of each combined segmentation was then quantified by computing the recognition rate, i.e., the fraction of correctly labeled voxels as compared to the ground truth segmentation.

3 Results

Examples of combined segmentations using shape-based averaging and label voting are shown as 3-D renderings in Fig. 3. Shape-based averaging produced visually superior results. In particular, the inherent spatial continuity of the Euclidean distance maps avoided fragmentation of contiguous structures.

The recognition rates of combined human brain MR segmentations (simulated segmentations) using shape-based averaging and label voting are plotted in Fig. 4. Results

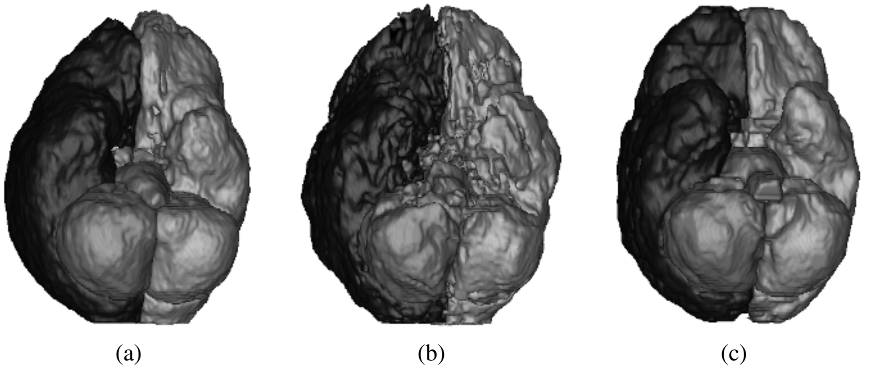
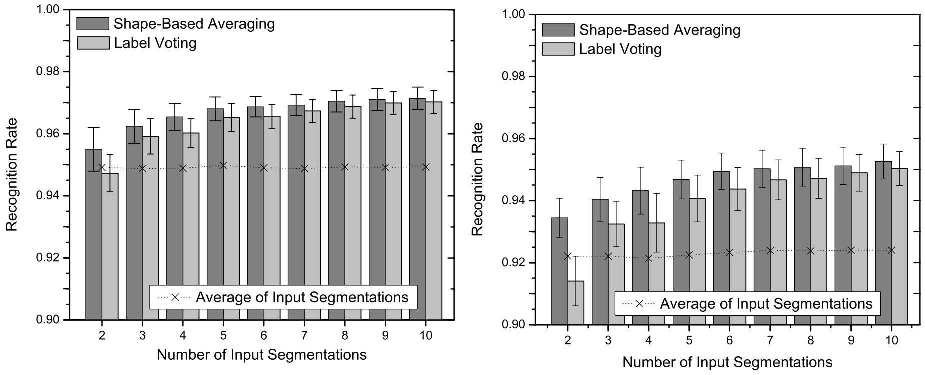


Fig. 3. Three-dimensional rendering of the inferior brain surface of the human brain MR data used for evaluation. (a) Shape-based averaging. (b) Label voting. (c) Ground truth. The renderings (a) and (b) are the result of averaging the same five simulated segmentations generated with FFD perturbation $\sigma = 20\,\mathrm{mm}$.



(a) σ = 10 mm (b) σ = 20 mm

Fig. 4. Recognition rates of combined segmentations for shape-based averaging and label voting. (a) Input segmentations generated using random FFD with $\sigma = 10\,\mathrm{mm}$. (b) Input segmentations generated using random FFD with $\sigma = 20\,\mathrm{mm}$. In both graphs, the columns represent the mean recognition rates over ten subjects. The error bars represent the respective standard deviations. The dashed lines show the averaged recognition rates of the individual segmentations used as inputs for the combination methods.

using random deformations with $\sigma = 10\,\mathrm{mm}$ are shown in Fig. 4(a), results using $\sigma = 20\,\mathrm{mm}$ in Fig. 4(b).

Both segmentation combination methods generated outputs closer to the ground truth the more input segmentations were provided to them. Between the two combination methods, shape-based averaging clearly outperformed label voting in all cases. The relative advantage of shape-based averaging was larger for smaller numbers of input segmentations, and it was larger for greater deviations of the input segmentations from the ground truth ($\sigma = 20\,\mathrm{mm}$). The recognition rate of the combined segmentation using shape-based averaging improved consistently with added input segmentations, while label voting benefited less from even numbers of inputs than it did from odd numbers.

Note that combination of two segmentations by voting is not entirely reasonable as there is no way to decide the winning label in cases of disagreement. Therefore, the combined classification will fail wherever the two input segmentations disagree. As a result, the combination of only two segmentations by label voting has a worse recognition rate than the individual segmentations. Similarly, even numbers of input segmentations in general increase the likelihood of equal numbers of votes for more than one label in label voting. Neither is the case for shape-based averaging, which clearly improves recognition rates even for only two input segmentations, because each segmentation assigns a weight to every voxel based on its distance from the nearest structure boundary.

4 Discussion

This paper has introduced a method for shape-based averaging that can be applied to combine multiple segmentations of the same image. In a quantitative evaluation study using simulated segmentations, which makes it independent of the performance of any particular segmentation algorithm, we have demonstrated the superiority of our method to label voting. Applied to identical input segmentations, shape-based averaging generated combined segmentations that were substantially closer to the ground truth than those generated by label voting. The improvement achieved by shape-based averaging was larger for smaller number of input segmentations, and larger for input segmentations that deviate more from the ground truth.

While evaluation using randomly deformed ground truth segmentations borrows from concepts of atlas-based segmentations, our method is straight forward to apply to segmentations generated by arbitrary labeling methods. It works on as few as two segmentations, whereas label voting requires at least three and is prone to undecided voxels for small numbers of segmentations. These properties make our method potentially interesting for combination of multiple manual segmentations, where the number of available segmentations is typically small.

A potentially useful extension of our method is to use robust averaging rather than the arithmetic means of the individual distance functions as in the present paper (Eq. 3). This may improve the combination results in the presence of outliers, and ultimately also provide an effective way to address the problem of diversity [13] among the input segmentations.

Acknowledgments

The authors thank R. Menzel (Institute für Neurobiologie, Freie Universität Berlin, Berlin, Germany) and for R. Brandt (Mercury Computer Systems GmbH, Berlin, Germany) providing segmented bee brain microscopy images used for illustration in Fig. 2.

The normal 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. Rohlfing, T., Brandt, R., Menzel, R., Maurer, Jr., C.R.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage **21** (2004) 1428–1442
2. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Trans. Med. Imag. **23** (2004) 903–921
3. Rohlfing, T., Russakoff, D.B., Maurer, Jr., C.R.: Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. IEEE Trans. Med. Imag. **23** (2004) 983–994
4. Rohlfing, T., Maurer, Jr., C.R.: Multi-classifier framework for atlas-based image segmentation. Pattern Recogn. Lett. (2005, in press)
5. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans. Syst. Man Cybern. **22** (1992) 418–435
6. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. Machine Intell. **20** (1998) 226–239
7. Raya, S.P., Udupa, J.K.: Shape-based interpolation of multidimensional objects. IEEE Trans. Med. Imag. **9** (1990) 32–42
8. Grevera, G.J., Udupa, J.K.: Shape-based interpolation of multidimensional grey-level images. IEEE Trans. Med. Imag. **15** (1996) 881–892
9. Maurer, Jr., C.R., Qi, R., Raghavan, V.: A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. IEEE Trans. Pattern Anal. Machine Intell. **25** (2003) 265–270
10. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximisation of mutual information. IEEE Trans. Med. Imag. **16** (1997) 187–198
11. Sederberg, T.W., Parry, S.R.: Free-form deformation and solid geometric models. Comput. Graph. (ACM) **20** (1986) 151–160
12. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast MR images. IEEE Trans. Med. Imag. **18** (1999) 712–721
13. Kuncheva, L.I.: Diversity in multiple classifier systems. Inform. Fusion **6** (2005) 3–4