PERFORMANCE-BASED MULTI-CLASSIFIER DECISION FUSION FOR ATLAS-BASED SEGMENTATION OF BIOMEDICAL IMAGES

T. Rohlfing[1], *D. B. Russakoff*[1,2], *R. Brandt*[3,4], *R. Menzel*[3], *and C. R. Maurer, Jr.*[1]

[1] Image Guidance Laboratories, Department of Neurosurgery, Stanford University, Stanford, CA

[2]Department of Computer Science, Stanford University, Stanford, CA

[3]Department of Neurobiology, Freie Universität Berlin, Berlin, Germany

[4]Indeed – Visual Concepts GmbH, Berlin, Germany

ABSTRACT

Combinations of multiple classifiers have been found to be consistently more accurate than a single classifier. The construction of multiple independent classifiers, however, is typically a non-trivial problem. In atlas-based segmentation, multiple classifiers arise naturally, for example, from using multiple atlases. This paper evaluates the application of performance-based decision fusion methods to multi-classifier atlas-based segmentation. In a leave-one-out study, each of 20 subjects is segmented using each of the remaining 19 as the atlas. The resulting 19 segmentations per subject are combined into a final segmentation using three different methods: 1) simple decision fusion using the sum rule; 2) using a binary classifier performance model; 3) using a multi-label classifier performance model. The accuracy of each combined segmentation is computed by comparing it to the manual ground truth segmentation. The two methods that incorporate classifier performance outperform sum rule fusion, with the multi-label model performing better than the binary model.

1. INTRODUCTION

Combinations of classifiers are consistently more accurate than single classifiers [1, 2]. For atlas-based segmentation, multiple independent classifiers arise naturally from the use of multiple atlases derived from different individuals [3]. We evaluate in this work methods to estimate the performance parameters of multiple atlas-based classifiers. The performance estimates can be used in the classifier combination to assign higher confidence to more reliable individual classifiers. The goal of this strategy is to improve the overall classification accuracy of the combined segmentation by focusing on the more accurate individual classifiers.

TR was supported by the National Science Foundation under Grant No. EIA-0104114. DBR was supported by the Interdisciplinary Initiatives Program, which is part of the Bio-X Program at Stanford University, under the grant "Image-Guided Radiosurgery for the Spine and Lungs."

2. METHODS

2.1. Image Data

We demonstrate and evaluate the methods described in this paper by applying them to three-dimensional (3-D) confocal microscopy images of the brains of honeybees [4]. The brains of 20 adult worker bees were imaged with a resolution of $3.8\,\mu\mathrm{m}$ in x and y, and $8\,\mu\mathrm{m}$ in z direction. The resulting 3-D images had 84–114 slices (sections), each slice having 610–749 pixels in x direction and 379–496 pixels in y direction. An example of a slice from one of the microscopy images and the corresponding label image is shown in Fig. 1. For use as atlases during atlas-based segmentation, as well as for providing a gold standard, all images were manually segmented, distinguishing 22 major compartments of the brain.

2.2. Image Registration and Atlas-Based Segmentation

An atlas is registered to a given image by first computing an affine [5] transformation, followed by a free-form deformation based on B-splines [6] to account for inter-individual shape differences. Our implementation of both methods is highly efficient and takes advantage of SMP multiprocessing [7]. This facilitates repeating in particular the (computationally expensive) non-rigid registration with different atlases.

2.3. Performance-Based Decision Fusion

Independent segmentations are combined into a final segmentation using several different decision fusion methods. The simplest method is sum rule fusion [2], where each

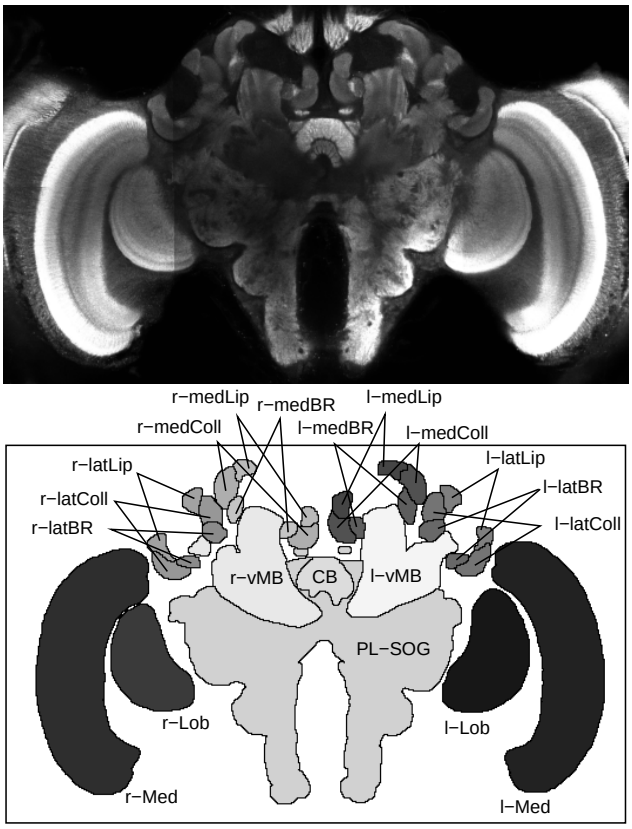

Fig. 1. Example of a central axial slice from a 3-D confocal microscopy image of a bee brain (*top*) with corresponding label image (*bottom*). Each gray level in the label image corresponds to a different anatomical structure. See Ref. [4] for a detailed list of structures and abbreviations.

voxel in the final segmentation is assigned the label that receives the most combined weight from the individual segmentations. The individually assigned weights can be fractional, due to the use of partial volume interpolation [8] within the atlas, which avoids staircase artifacts typically resulting from nearest neighbor interpolation.

Each atlas-based segmentations is more or less accurate, and knowledge of each segmentation's accuracy can be used to weight its contribution in the decision fusion. Warfield *et al*. [9] recently described an EM algorithm for estimating the binary segmentation performance of multiple experts in the absence of a ground truth, acronymed STAPLE for Simultaneous Truth and Performance Level Evaluation. The performance of each segmentation is different for each label in the segmentation, so we apply two multi-label extensions of the Warfield method to estimate the per-label performance of the individual atlas-based segmentations.

The first method independently applies the Warfield algorithm to each label in the segmentation. Let $e_k(x)$ be the decision of classifier k for voxel x and C_j the class that the voxel belongs to, i.e., the correct label. The method estimates the common binary performance parameters sensitivity $p_k^{(j)}$ and specificity $q_k^{(j)}$ for each classifier k and each label j, defined as

$$p_k^{(j)} = P(e_k(x) = j | x \in C_j), \quad (1)$$

$$q_k^{(j)} = P(e_k(x) \neq j | x \notin C_j). \quad (2)$$

The second method is based on a multi-label performance parameter model (row-normalized confusion matrix of a Bayesian classifier) that takes into account cross-label misclassifications [10]. The entries of this matrix are the following conditional probabilities:

$$\lambda_k^{(i,j)} = P(e_k(x) = j | x \in C_i). \quad (3)$$

Note that the binary model is a special case of the multi-label model with two classes, 0 and 1 for background and foreground, and $p_k^{(1)} \equiv \lambda_k^{(1,1)}$ and $q_k^{(1)} \equiv \lambda_k^{(0,0)}$.

Analogously to the binary performance model, the parameters of the multi-label model are estimated by an EM algorithm (see Ref. [10] for details). Given the classifier decisions and the estimated performance parameters of either model, the label probabilities can be computed for each voxel using Bayes' theorem, e.g.,

$$P(x \in C_i | \mathbf{e_k}(x), \boldsymbol{\lambda}) = \frac{P(x \in C_i) \prod_k P(e_k(x) | x \in C_i, \boldsymbol{\lambda})}{\sum_j P(x \in C_j) \prod_k P(e_k(x) | x \in C_j, \boldsymbol{\lambda})} \quad (4)$$

for the multi-label model. The label with the highest probability is then taken as the outcome of the decision fusion and assigned to the voxel in the combined segmentation:

$$E(x) = \arg\max_i P(x \in C_i | \mathbf{e_k}(x), \boldsymbol{\lambda}). \quad (5)$$

2.4. Evaluation Study

Using the manual segmentation as a gold standard, we evaluate the segmentation accuracies of both EM methods relative to each other and to classifier combination by sum rule decision fusion [2]. For each of 20 individuals we compute 19 atlas-based segmentations, using each of the remaining 19 individuals as the atlas. These segmentations are combined into a final segmentation using each of the following classifier decision fusion methods:

1. sum rule fusion,
2. the binary performance parameters estimated by the Warfield algorithm, and
3. the multi-label performance parameters estimated using our algorithm.

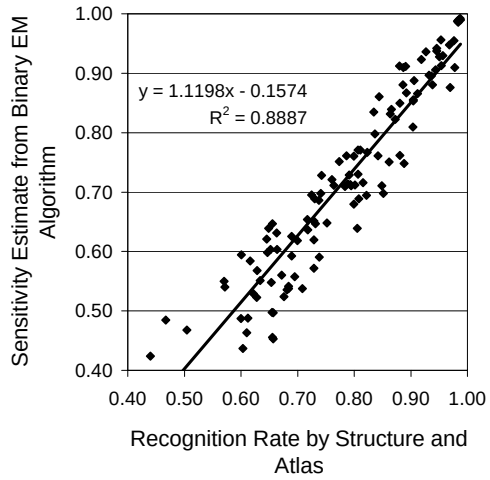

Fig. 2. Estimated vs. actual performance parameters based on binary performance model.

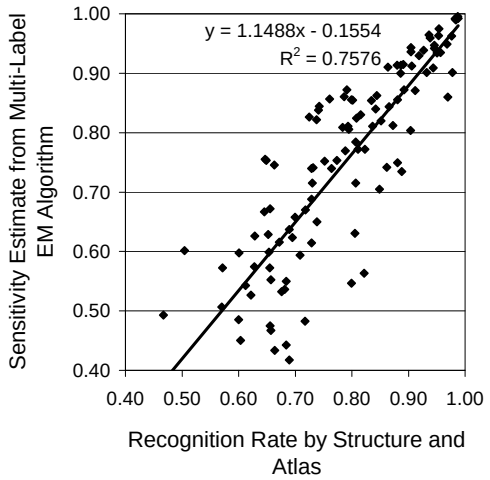

Fig. 3. Estimated vs. actual performance parameters based on multi-label performance model.

The accuracies of all segmentations are computed by comparing them to the manual segmentation, which provide the gold standard for this study.

The recognition rate $r_k^{(j)}$ for label j in the segmentation by classifier k is the relative number of correctly labeled voxels, i.e.,

$$r_k^{(j)} = \frac{\#\{e_k(x) = j\}}{\#\{x \in C_j\}}. \quad (6)$$

It is easy to see these values are the *a posteriori* sensitivities of the classifiers, i.e., $p_k^{(j)}$ in the binary performance model and $\lambda_k^{(j,j)}$ in the multi-label performance model. The accuracy of the performance parameter estimation methods can therefore be computed by comparing the estimated parameters with the actual recognition rates.

3. RESULTS

3.1. Performance Parameter Estimation Accuracy

The sensitivity performance parameters estimated by the EM methods are plotted against the actual *a posteriori* recognition rates in Fig. 2 and Fig. 3 for the binary performance model and the multi-label performance model, respectively. Each plot shows the parameters computed for all 22 classes (structures). In order to improve the visual presentation, only five segmented images were included in the plots. The five images were selected randomly and were identical for both plots.

Both EM methods computed reasonably accurate estimates of the true performance parameters (linear regression, $R^2 = 0.89$ and $R^2 = 0.76$ for the binary and the multi-label model, respectively).

3.2. Segmentation Accuracy

The recognition rates achieved using the three decision fusion methods are compared to each other in Fig. 4. For reference, this figure also shows the recognition rates using a single individual atlas with no decision fusion. The single atlas was chosen based on the *a posteriori* recognition rates, that is, it was the one image out of the 20 subjects in our study, which, when used as the atlas, gave the best recognition rates for segmentations of the remaining 19 images.

It is easy to see that each of the three decision fusion methods produced more accurate segmentations (i.e., higher recognition rates) than atlas-based segmentation using a single individual atlas. Note again that we used the best possible individual atlas, so no other subject, when chosen as the atlas, produced a higher recognition rate.

Between the decision fusion methods, the EM algorithm based on the binary performance model outperforms sum rule fusion. Both methods are outperformed by the EM algorithm based on the multi-label performance model. The mean recognition rates were: 93% for Warfield's algorithm, 95% for multi-label estimation, 91% for Sum Rule fusion with no performance estimates.

4. DISCUSSION

Performance-based combinations of multiple segmentations can substantially improve segmentation accuracy. The methods applied in this paper are easily applicable beyond the application described here. Firstly, they can combine segmentations generated by fundamentally different algorithms, e.g., by incorporating level set-based methods, or segmentations using active contours. Next, they are applicable to more general biomedical classification problems, such as

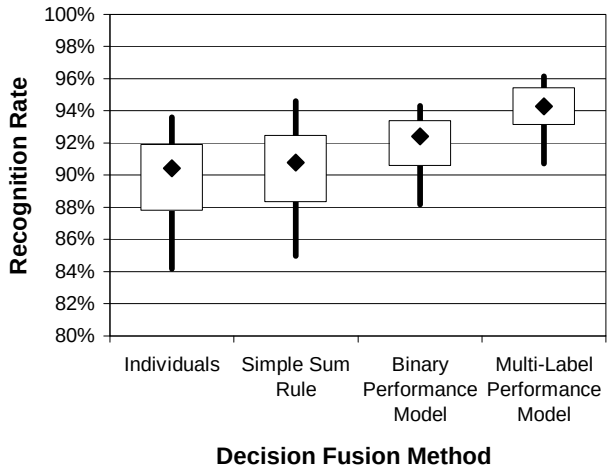

Fig. 4. Comparison of segmentation accuracy using a single atlas vs. different decision fusion methods. The diamonds show the median recognition rate for each method over 20 segmented subjects. The ends of the solid lines represent the minimum and maximum recognition rates, while the upper and lower edges of the boxes represent the 75^{th} and 25^{th} percentiles, respectively.

the classification of breast lesions as benign or malignant. Finally, performance-based decision fusion with EM performance parameter estimation represents a general framework of combining classifiers in applications outside image processing, for example including handwriting recognition or speaker identification.

5. ACKNOWLEDGMENT

The authors thank Andreas Steege and Charlotte Kaps for manually tracing the microscopy images. All non-rigid registrations were performed on an SGI Origin 3800 supercomputer in the Stanford University Bio-X core facility for Biomedical Computation. All classifier combinations were computed on a workstation cluster running Condor. The Condor Software Program (Condor) was developed by the Condor Team at the Computer Sciences Department of the University of Wisconsin-Madison. All rights, title, and interest in Condor are owned by the Condor Team.

6. REFERENCES

[1] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Trans. Syst. Man Cybern.*, vol. 22, no. 3, pp. 418–435, 1992.

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

[3] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, Jr., "Segmentation of three-dimensional images using non-rigid registration: Methods and validation with application to confocal microscopy images of bee brains," in *Medical Imaging: Image Processing*, M. Sonka and J. M. Fitzpatrick, Eds., 2003, vol. 5032 of *Proceedings of the SPIE*, pp. 363–374.

[4] R. Brandt, T. Rohlfing, A. Steege, M. Westerhoff, and R. Menzel, "An average three-dimensional atlas of the honeybee brain based on confocal images of 20 subjects," Submitted to Journal of Comparative Neurology.

[5] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures," *Med. Phys.*, vol. 24, no. 1, pp. 25–35, 1997.

[6] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, 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, 1999.

[7] T. Rohlfing and C. R. Maurer, Jr., "Non-rigid image registration in shared-memory multiprocessor environments with application to brains, breasts, and bees," *IEEE Trans. Inform. Technol. Biomed.*, vol. 7, no. 1, pp. 16–25, 2003.

[8] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximisation of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, 1997.

[9] S. K. Warfield, K. H. Zou, and W. M. Wells, "Validation of image segmentation and expert quality with an expectation-maximization algorithm," in *Proceedings of Fifth International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Part I*, Berlin Heidelberg, 2002, vol. 2488 of *Lecture Notes in Computer Science*, pp. 298–306, Springer-Verlag.

[10] T. Rohlfing, D. B. Russakoff, and C. R. Maurer, Jr., "Expectation maximization strategies for multi-atlas multi-label segmentation," in *Information Processing in Medical Imaging*, Berlin Heidelberg, 2003, vol. 2732 of *Lecture Notes in Computer Science*, pp. 210–221, Springer-Verlag.