Multi-Classifier Framework for Atlas-Based Image Segmentation

Torsten Rohlfing and Calvin R. Maurer, Jr.

Image Guidance Laboratories, Department of Neurosurgery, Stanford University
Stanford, CA 94305-5327
rohlfing@stanford.edu, calvin.maurer@igl.stanford.edu

Abstract

We develop and evaluate in this paper a multi-classifier framework for atlas-based segmentation, a popular segmentation method in biomedical image analysis. An atlas is a spatial map of classes (e.g., anatomical structures), which is usually derived from a reference individual by manual segmentation. An atlas-based classification is generated by registering an image to an atlas, that is, by computing a semantically correct coordinate mapping between the two. In the present paper, the registration algorithm is an intensity-based non-rigid method that computes a free-form deformation (FFD) defined on a uniform grid of control points. The transformation is regularized by a weighted smoothness constraint term. Different atlases, as well as different parameterizations of the registration algorithm, lead to different and somewhat independent atlas-based classifiers. The outputs of these classifiers can be combined in order to improve overall classification accuracy. In an evaluation study, biomedical images from seven subjects are segmented 1) using three individual atlases; 2) using one atlas and three different resolutions of the FFD control point grid; 3) using one atlas and three different regularization constraint weights. In each case, the three individual segmentations are combined by Sum Rule fusion. For each individual and for each combined segmentation, its recognition rate (relative number of correctly labeled image voxels) is computed against a manual gold-standard segmentation. In all cases, classifier combination consistently improved classification accuracy. The biggest improvement was achieved using multiple atlases, a smaller gain resulted from multiple regularization constraint weights, and a marginal gain resulted from multiple control point spacings. We conclude that multi-classifier methods have a natural application to atlas-based segmentation and the potential to increase classification accuracy in real-world segmentation problems.

1. Introduction

Combinations of multiple independent classifiers can be substantially more accurate than each of the individual classifiers alone. Numerous applications of this principle have been reported in the pattern recognition literature over the past years. Xu *et al.* [1] evaluated different combination schemes of classifiers for the recognition of unconstrained handwritten numerals. They observed an increase of the recognition rates from 93 percent (best out of four individual classifiers) to 98.6 percent for several of their combination methods. Similarly, Kittler *et al.* [2] combined four classifiers for optical character recognition (OCR) of uppercase letters and digits, and found the classification rate to improve from 95 percent (best individual classifier) to over 98 percent for a combined classifier.

A common problem, however, is the construction of independent classifiers. A general solution for this problem has been proposed by Breiman [3] under the name "bootstrap aggregation", or simply "bagging". If the training of a classifier is unstable, that is, sensitive to the training set, multiple classifiers can be generated by using different training sets. In this paper, we propose and evaluate a classifier approach to atlas-based image registration. This approach allows for the systematic generation of a virtually unlimited number of classifiers in a natural way, similar to the bagging method.

Atlas-based segmentation is particularly successful in current applications to segment three-dimensional (3-D) biomedical images [4, 5, 6]. The segmentation is performed by computing an anatomically correct coordinate transformation between an image and an already segmented image, the so-called atlas. The process of computing the transformation between the two images is called registration, and is itself subject of active research [7, 8].

When comparing atlas-based segmentation to classifier techniques, the atlas assumes the role of the test set, and training of the classifier is achieved by computing the registration between image and atlas. We will take a more detailed look at the parallels between atlas-based segmentations and classifiers in Section 2.

Analogously to the bagging method, multiple independent classifiers arise in a natural way by using multiple training sets, that is multiple different atlases. Most pub-

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)
1063-6919/04 $20.00 © 2004 **IEEE**

Table 1: Analogies between neural networks and atlas-based classifiers.

	Neural Network	Atlas-Based Classifier
Input	Pattern	Coordinate in Image Domain
Output	Class	Segmentation Label
Internal Structure	Network Topology	Transformation Model
Parameters	Connection Weights	Transformation Parameters
Parameter Adjustment	Learning	Registration
Learning Input	Training Set	Atlas
Typical Learning Algorithm	Backpropagation	Intensity-Based Non-Rigid Registration

lished work on atlas-based segmentation uses atlases that were generated by (manually) segmenting one individual. If two atlases are generated from different individuals, then they are guaranteed to be independent. Likewise, a different training method, e.g., a different registration algorithm, can be used to generate multiple different classifiers using the same atlas. In the absence of multiple working registration algorithms, where applicable multiple different parameterisations of the same algorithm may be used.

In Section 3 we evaluate three ways to generate multiple atlas-based classifiers. Biomedical images from seven subjects are segmented using multiple individual atlases, or using one atlas and multiple different values for two parameterizations of the registration algorithm. In each case, the individual segmentations are combined into a final segmentation. For each individual and for each combined segmentation, its recognition rate (relative number of correctly labeled image voxels) is computed against a manual gold-standard segmentation. We conclude in Section 4 with a discussion of our results, their relevance, and their practical implications.

2. Atlas-Based Classifiers

This section introduces a classifier view of atlas-based segmentation. Based on this foundation we then demonstrate the application of multi-classifier principles to atlas-based segmentation.

2.1. Analogies

Let us first take a look at the analogies between a generic classifier, say a neural network, and an atlas-based classifier. Table 1 gives an overview of the related concepts. From an operational perspective, an atlas-based classifier takes as its input a coordinate inside the domain of the image and produces as its output the label assigned to that coordinate. Internally, this label is determined by lookup from a discrete label field. The unsegmented image and the atlas are related to each other by a coordinate transformation. The parameters of the transformation are the internal parameters of the classifier, so that training the classifier is equivalent to performing a registration between the image and the atlas.

The coordinate transformation of an atlas-based classifier is continuous, typically even smooth, and does not in general map grid points of the image to grid points in the atlas. Label lookup therefore requires an interpolation of labels of some sort.

The simplest interpolation that can be applied to non-numerical data such as labels is nearest neighbor (NN) interpolation, which returns the unique label of the nearest atlas grid point as the classifier output. A more sophisticated technique that is applicable to label data is partial volume (PV) interpolation, originally introduced by Maes *et al.* [9] 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. Compared to NN interpolation, using PV interpolation of the label map avoids jagged edges when generating oblique slices. From the classifier perspective, the fractional label weights resulting from PV interpolation virtually eliminate the possibility of a tie in the classifier voting, thus greatly reducing the number of rejected patterns.

2.2. Image-to-Atlas Registration

We use an intensity-based registration method based on the normalized mutual information (NMI) image similarity measure [10]. The coordinate mapping between image and atlas is computed by a non-rigid registration algorithm introduced by Rueckert *et al.* [11]. The transformation model is a free-form deformation [12] defined on a uniformly spaced grid of discrete control points. Between the control points, a smooth deformation field is interpolated using approximating third-order B-splines.

To improve segmentation accuracy and robustness, and to be able to equally model small and large deformations, we employ a multi-resolution deformation strategy. A coarse initial control point grid is repeatedly refined [13] until a final resolution is reached. This final resolution, that is the final spacing between the FFD control points, is the

Table 2: Recognition rates for three different training sets (i.e., atlases).

	Recognition Rates			
Subject	Min	Max	Mean	Combined
#1	0.9611	0.9668	0.9645	0.9696
#2	0.9580	0.9723	0.9647	0.9691
#3	0.9703	0.9762	0.9736	0.9774
#4	0.9647	0.9691	0.9676	0.9731
#5	0.9536	0.9574	0.9553	0.9605
#6	0.9599	0.9688	0.9651	0.9716
#7	0.9686	0.9794	0.9755	0.9802

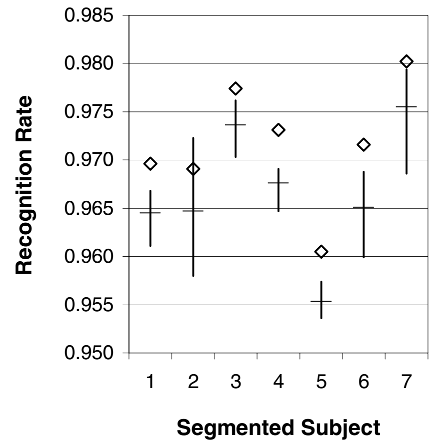

Figure 1: Recognition rates of individual vs. combined classifiers using three different atlases. The diamonds represent the recognition rates of the combined segmentation for each segmented subject. The vertical lines represent the range of recognition rates achieved by the individual segmentations. The horizontal lines represent the numerical averages of the individual recognition rates.

most important parameter of the algorithm.

In order to prevent unrealistic transformations due to incomplete or noisy data, the optimization function of the non-rigid registration algorithm is regularized with a constraint term $E_{\text{constraint}}$ that enforces smoothness of the deformation field. The overall optimization function E_{total} is a weighted combination of the NMI image similarity measure and the smoothness constraint:

$$E_{\text{total}} = (1-w)E_{\text{NMI}} + wE_{\text{constraint}}. \tag{1}$$

It is important to note that other non-rigid registration algorithms can also be used for atlas-based segmentation. The method used for the present paper is merely one example of a technique that we have empirically found to be accurate and computationally efficient.

2.3 Bagging of Atlas-Based Classifiers

The principle idea of bagging [3] is to generate multiple independent classifiers by exploiting instability of classifier learning under changes to the learning set. In addition, one can exploit instability under different internal parameterizations of the classifier.

In the context of atlas-based segmentation, using a different learning set means using a different atlas to register the unsegmented image to. Among other, less important settings, the non-rigid registration algorithm used in this paper has two major parameters: the spacing of the control point grid (or rather, the final spacing in a multi-resolution strategy), and the relative weight of the regularization term in the overall cost function (Eq. (1)). Changing either of these two parameters changes the registration outcome, and therefore results in a different classifier, even when applied to the same atlas.

3. Evaluation Study

We evaluate bagging of multiple atlas-based classifiers by segmenting three-dimensional biomedical images [14, 15] from a small animal brain mapping study. Each image contained 84–114 slices with thickness $8\,\mu\text{m}$ and each slice had 610–749 pixels in x direction and 379–496 pixels in y direction with pixel size $3.8\,\mu\text{m}$. This is relevant insofar as the image resolution determines the ranges of meaningful values for various parameters of the non-rigid registration, most notably the FFD control point spacing.

For each image, a trained expert performed a complete manual segmentation. This segmentation can either serve as an atlas to segment other images, or as a gold standard to quantify the accuracy of an automatic segmentation. From a total of 10 individuals, a random subset of three individuals is designated for use as atlases. The images from the remaining seven subjects serve as the images to be segmented. The manual segmentations for these subjects serve as the gold standard of the atlas-based segmentations.

The three subsections below describe the setups and results of three different trials. In the first trial, each of the seven test images is registered independently to each of the three atlases. In the second trial, one atlas is randomly chosen, and each test image is registered to this atlas three times, using three different control point spacings of the FFD. Finally, in the third trial each test image is registered to the same atlas using the same control point spacing, but using three different weights of the smoothness constraint.

All classifier combinations are computed by sum rule fusion of the label weights that result from PV interpolation.

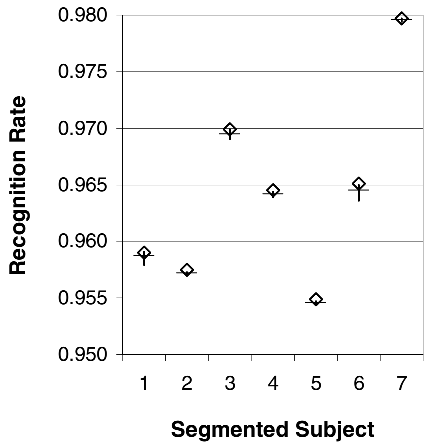

Figure 2: Recognition rates of individual vs. combined classifiers using three different control point spacings. For a description of the graphical presentation see Fig. 1.

3.1. Multiple Training Sets (Atlases)

Closely following the original idea of bagging, we generate multiple classifiers by using different independent training sets. In our context, this means using multiple atlases derived from different reference subjects. Each of the seven images to be segmented is registered independently to each of three atlases.

The recognition rates achieved using each of three different atlases are visualized and compared to the recognition rate of the combined segmentation in Fig. 1. The actual numerical values are given in Table 2.

The minimum, maximum, and mean recognition rates were computed over the individual classifiers, each resulting from a different atlas. Note that for all but one segmented subjects (#2), the combined classifier had a higher recognition rate than the *best* individual classifier. For all subjects, the combined classifier had a better recognition rate than the average of the three individual classifiers. On average, the recognition rate improved by 0.5 percent.

3.2. Multiple Control Point Spacings

The recognition rates achieved when segmenting each of the seven test subjects using a single atlases, but with three different control point spacings of the non-rigid registration algorithm (50, 60, and 70 μm) are visualized and compared to the recognition rate of the combined segmentation in Fig. 2. The actual numerical values are given in Table 3.

The minimum, maximum, and mean recognition rates were computed over the individual classifiers, each resulting from a different spacing. For six out of seven subjects, the combined classifier had a better or equal recognition rate compared the best individual classifier. For all subjects, the combined classifier had a better recognition rate than the average of the three individual classifiers. On average, the recognition rate improved by 0.03 percent.

Table 3: Recognition rates for three different control point spacings of the FFD registration.

	Recognition Rates			
Subject	Min	Max	Mean	Combined
#1	0.9579	0.9591	0.9587	0.9590
#2	0.9570	0.9573	0.9572	0.9575
#3	0.9690	0.9699	0.9695	0.9699
#4	0.9639	0.9644	0.9642	0.9645
#5	0.9544	0.9547	0.9546	0.9549
#6	0.9636	0.9650	0.9645	0.9651
#7	0.9794	0.9797	0.9796	0.9797

Note that for all individuals the range of the individual classifiers' recognition rates in this trial is much smaller than for example using multiple atlases in the previous subsection (compare Fig. 2 vs. Fig. 1). This suggests that there is relatively little change of the segmentation due to changes in the control point spacings, resulting in relatively little benefit of the classifier combination.

3.3. Multiple Smoothness Constraint Weights

The recognition rates achieved when segmenting each of the seven test subjects using a single atlases, but with three different smoothness constraint weights of the non-rigid registration algorithm ($w = 0.05$, 0.1, and 0.3) are visualized and compared to the recognition rate of the combined segmentation in Fig. 3. The actual numerical values are given in Table 4.

The minimum, maximum, and mean recognition rates were computed over the individual classifiers, each resulting from a different constraint weight. For all seven subjects, the combined classifier had a better or equal recognition rate compared the best individual classifier. On average, the recognition rate improved by 0.19 percent.

4. Discussion

The main advantage of a classifier view of atlas-based segmentation is that it opens the field to the application of multi-classifier decision fusion techniques. We have shown in previous work [14, 15] that segmentation accuracy can be significantly improved when more than a single atlas is used. However, multiple atlases are not always available, since their generation is time consuming and tedious. This makes the message of the present paper ever more important – it has demonstrated that improvements of segmentation

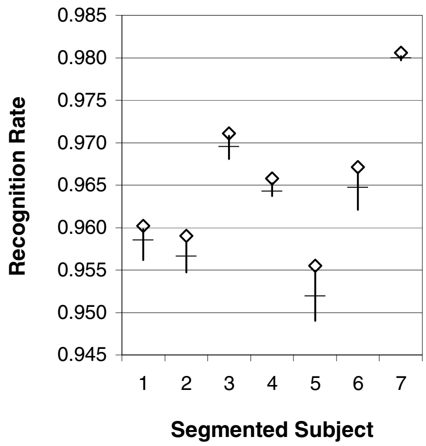

Figure 3: Recognition rates of individual vs. combined classifiers using three different smoothness constraint weights. For a description of the graphical presentation see Fig. 1.

Table 4: Recognition rates for three smoothness constraint weights of the FFD registration.

	Recognition Rates			
Subject	Min	Max	Mean	Combined
#1	0.9562	0.9598	0.9585	0.9602
#2	0.9547	0.9582	0.9566	0.9590
#3	0.9681	0.9708	0.9695	0.9711
#4	0.9637	0.9652	0.9643	0.9658
#5	0.9490	0.9549	0.9519	0.9555
#6	0.9621	0.9665	0.9647	0.9671
#7	0.9797	0.9802	0.9800	0.9806

accuracy are possible with only a single atlas, but different parameterizations of the non-rigid registration technique.

This paper has evaluated three different ways of generating multiple atlas-based classifiers: use of multiple atlases, deformation control point spacings, and multiple regularization constraint weights. Our results suggest that there is a quantitatively bigger accuracy gain from using multiple atlases than there is from using different regularization weights, which itself produced bigger gains than multiple control point spacings. We believe that the reason for this is that atlases obtained from multiple reference individuals are truly independent and also reflect, to some extent, the variability of a population. Whenever possible, multiple atlases should therefore be used. Of course, in addition to using multiple atlases, further improvements may be possible by registering to each atlas several times with different registration parameters.

Between the two registration parameters we considered, control point spacing and smoothness constraint weight, the latter appeared to have a bigger impact on the segmentation result. The fact that decision fusion effectively benefits from multiple classifiers resulting from variations of this parameter is particularly important as there is no *a priori* correct value for it. Since the constraint weight relates fundamentally unrelated quantities, image similarity and deformation energy, classifier fusion allows us to cover a range of possible values without having to pick one. For control point spacing, on the other hand, we can basically say that "more is better", that is, the finer the control point grid, the more successfully the transformation can model inter-individual shape differences, and the more accurate the registration and thus the segmentation will be (unpublished results). The effectiveness of variations of the constraint weight is also relevant, because, unlike the control point spacing, this type of parameter is also present in any other registration algorithms based on computing an equilibrium between internal forces (i.e., image similarity) and external forces (i.e., a physics-based constraint such as elasticity or volume preservation) [16].

The obvious disadvantage of the suggested procedure is the repeated application of a computationally expensive non-rigid registration step. We have therefore limited our consideration in this paper to three segmentations per subject, the minimal number of classifiers that can be combined in a meaningful way. Even so, we observed noticeable and consistent accuracy improvements. In the future, with ever increasing computer speeds, we consider registration times to be not too serious of an issue. Furthermore, non-rigid registration can be parallelized with very low overhead [17], and multiple independent registrations can easily be performed on a cluster of inexpensive computation nodes.

Moving past the combination of multiple atlas-based classifiers, our framework provides a theoretical foundation for combining atlas-based segmentations with segmentations obtained from fundamentally different methods. They could be combined, for example, with results generated by level set techniques [18], and with such originating from active contour methods [19]. After all, the more dissimilar the methods combined, the more likely their errors are independent.

Finally, more advanced methods for classifier combination become applicable. We are currently working on techniques to estimate the individual classifier performance in the absence of ground truth [20]. Given these estimates, we hope to be able to identify more accurate individual classifiers and disregard less accurate ones in the decision fusion. These algorithms have a similar aim as the Behavior Knowledge Space method introduced by Huang and Suen [21], but suffer substantially less from scaling problems when the number of classes and classifiers are increased.

5. Acknowledgment

The authors thank Robert Brandt and Randolf Menzel for providing segmented bee brain 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.

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] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[4] D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans, "Automatic 3-D model-based neuroanatomical segmentation," *Hum. Brain Map.*, vol. 3, no. 3, pp. 190–208, 1995.

[5] D. L. Collins and A. C. Evans, "Animal: Validation and applications of nonlinear registration-based segmentation," *Intern. J. Pattern Recognit. Artif. Intell.*, vol. 11, no. 8, pp. 1271–1294, 1997.

[6] B. M. Dawant, S. L. Hartmann, J. P. 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: Part I, methodology and validation on normal subjects," *IEEE Trans. Med. Imag.*, vol. 18, no. 10, pp. 909–916, 1999.

[7] J. M. Fitzpatrick, D. L. G. Hill, and C. R. Maurer, Jr., "Image registration," in *Handbook of Medical Imaging, Volume 2: Medical Image Processing and Analysis*, M. Sonka and J. M. Fitzpatrick, Eds., chapter 8, pp. 447–513. SPIE Press, Bellingham, WA, 2000.

[8] J. V. Hajnal, D. L. G. Hill, and D. J. Hawkes, Eds., *Medical Image Registration*, CRC Press, Boca Raton, FL, 2001.

[9] 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.

[10] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.

[11] 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.

[12] T. W. Sederberg and S. R. Parry, "Free-form deformation and solid geometric models," *Comput. Graph. (ACM)*, vol. 20, no. 4, pp. 151–160, 1986.

[13] D. R. Forsey and R. H. Bartels, "Hierarchical B-spline refinement," *Comput. Graph. (ACM)*, vol. 22, no. 4, pp. 205–212, 1988.

[14] 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., Feb. 2003, vol. 5032 of *Proceedings of the SPIE*, pp. 363–374.

[15] 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.

[16] B. Fischer and J. Modersitzki, "FLIRT: A flexible image registration toolbox," in *Biomedical Image Registration: Second International Workshop, WBIR 2003, Philadelphia, PA, USA, June 23-24, 2003, Revised Papers*, J. C. Gee, J. B. A. Maintz, and M. W. Vannier, Eds., Berlin Heidelberg, June 2003, vol. 2717 of *Lecture Notes in Computer Science*, pp. 261–270, Springer-Verlag.

[17] 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.

[18] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modelling with front propagation: A level set approach," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 2, pp. 158–175, 1995.

[19] A. Yezzi, Jr., S. Kichenassamy, A. Kumar, P. Olver, and A. Tannenbaum, "A geometric snake model for segmentation of medical imagery," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 199–209, 1997.

[20] 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*, C. Taylor and J. A. Noble, Eds., Berlin Heidelberg, July 2003, vol. 2732 of *Lecture Notes in Computer Science*, pp. 210–221, Springer-Verlag, 18th International Conference, IPMI 2003, Ambleside, UK, July 2003.

[21] Y. S. Huang and C. Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 1, pp. 90–94, 1995.