

Small Sample Size Learning for Shape Analysis of Anatomical Structures

Polina Golland¹, W. Eric L. Grimson¹, Martha E. Shenton², and Ron Kikinis³

¹ Artificial Intelligence Laboratory,
Massachusetts Institute of Technology, Cambridge, MA.

`{polina,welg}@ai.mit.edu`,

² Laboratory of Neuroscience, Clinical Neuroscience Division, Department of
Psychiatry, VAMC-Brockton, Harvard Medical School, Brockton, MA.
`mshenton@warren.med.harvard.edu`

³ Surgical Planning Laboratory Brigham and Women's Hospital,
Harvard Medical School, Boston, MA.
`kikinis@bwh.harvard.edu`

Abstract. We present a novel approach to statistical shape analysis of anatomical structures based on small sample size learning techniques. The high complexity of shape models used in medical image analysis, combined with a typically small number of training examples, places the problem outside the realm of classical statistics. This difficulty is traditionally overcome by first reducing dimensionality of the shape representation (e.g., using PCA) and then performing training and classification in the reduced space defined by a few principal components. We propose to learn the shape differences between the classes in the original high dimensional parameter space, while controlling the capacity (generalization error) of the classifier. This approach makes significantly fewer assumptions on the properties and the distribution of the underlying data, which can be advantageous in anatomical shape analysis where little is known about the true nature of the input data. Support Vector Machines with Radial Basis Function kernels are used as a training method and the VC dimension is used for the theoretical analysis of the classifier capacity. We demonstrate the method by applying it to shape classification of the hippocampus-amygdala complex in a data set of 15 schizophrenia patients and 15 normal controls. Using our technique, the separation between the classes and the confidence intervals are improved over a volume based analysis (63% to 73%). Thus exploiting techniques from small sample size learning theory provides us with a principled way of utilizing shape information in statistical analysis of the disorder effects on the brain.

1 Introduction

Statistical shape analysis, or shape based classification, attempts to identify statistical differences between two groups of images of the same organ based on 3D images of the organ. It can be used to study a disease (patients vs. normal

controls) or changes caused by aging (different age groups). Size and volume measurements have been widely used for this purpose, but they capture only a small subset of organ shape differences. If utilized properly, shape information can significantly improve our understanding of the anatomical changes due to a particular disorder.

The first step in shape based statistical analysis is extraction of the shape parameters, or a shape descriptor. Most work in constructing models of shape and its variation has been motivated by and used for segmentation tasks, where a generative model of shape variation guides the search over the space of possible deformations of the “representative” shape template, with the goal of matching it to a new input image. Several analytical models for shape variation have been used in conjunction with different shape models [2,14], followed by introduction of a statistical approach by Cootes and Taylor [5]. They proposed to learn the variation of the shape within a class of examples from the input data by extracting a shape descriptor from every training example and using Principal Component Analysis (PCA) to build a linear model of variation and compress it to a few principal modes of deformation. This method was explored in the domain of image segmentation by several authors using different shape models [9,11,15].

When applied to shape analysis, this approach employs PCA to reduce the dimensionality of the shape parameter space and then performs the training in the reduced, low dimensional space [6,8,10]. The main reason for dimensionality reduction in shape analysis is the concern that robust parameter estimation is infeasible in the original high dimensional space when only a small number of examples is available. Underlying this concern is the assumption that in order to reliably train a classifier, one has to estimate the density distribution of the data. Using PCA for dimensionality reduction is essentially equivalent to assuming that the examples are generated by a Gaussian distribution, and estimating its parameters – mean and covariance – from the input data. An interesting question then arises: what if the data are not generated by a Gaussian distribution? In the generative case, the Law of Large Numbers justifies using this method for estimating mean and covariance of the class. The estimates are unbiased and as the number of examples grows, they become increasingly accurate. However, the situation is different in the discriminative case. As the number of examples grows, the class mean and covariance estimates converge to their true values, but the Gaussian model based on the estimates does not describe the data more accurately. As a result, a classifier derived from this model does not approach the optimal solution. Since we have no evidence that the shape parameters for any group of subjects (either patients or normal controls) follow a high dimensional Gaussian distribution, it is desirable to avoid making this assumption in the analysis of the data.

We examine an approach to shape classification that makes much weaker assumptions than the ones used by traditional methods. This work is based on small sample size learning theory developed in the Machine Learning community. All approaches to learning from examples share the following very important no-

tion: the trade-off between generalization ability and training error determines the optimal classifier function. The training error, which is equal to the number of misclassified examples in the training set, can be reduced by using increasingly complex classifier functions that can capture the differences between the classes more easily. But the goal of learning is to optimize the classifier performance on new, unseen examples. The expected error rate of the classifier on future examples is called test error, or generalization error. Unfortunately, as the complexity of the classifier increases, its estimation from the training data becomes less robust, and as a result, the test error increases. The optimal classifier function is the one that can capture the class differences in the training data set (low training error) and generalize well the concept of the differences to the examples that had not been seen in the training phase (low expectation of test error). Traditionally, the classifier complexity has been associated with the number of parameters defining it. But the relationship between the complexity, or capacity, of the classifier and the number of parameters is not necessarily a monotonic one (see [3,17] for examples). Vapnik [16] introduced a better measure of capacity, called VC dimension, and derived theoretical bounds on the generalization error of a classifier based on it. One of the most important observations made within the theoretical analysis is that it is not necessary to solve the density estimation problem as an intermediate step in the learning process. Instead, the optimal classifier can and should be estimated directly from the input data. Based on this theory, a new learning method, called Support Vector Machines (SVMs), was proposed [16,17]. The main principle of the method is to train the classifier using the input data, while controlling its VC dimension, rather than the number of parameters. In this paper, we demonstrate a framework for statistical shape analysis based on SVMs that for small data sets provides results comparable to other techniques, and is guaranteed to converge to the optimal solution as the number of examples increases. It makes no assumptions on the distribution of the data, and can therefore be applied even when we do not have enough knowledge to estimate the distribution that produced the input data.

SVMs have recently been used in computer vision for classification tasks on grayscale images, such as character recognition [17], face detection [12] and others. While there are similarities between these applications and the problem of anatomical shape analysis, there are also significant differences. Among serious challenges in the general pattern recognition domain are non-linear imaging effects, such as occlusions, or changes in pose and illumination, and creating negative examples for object detections (e.g., a class of “non-faces” in face detection). Moreover, these applications are characterized by a wealth (tens of thousands) of training examples and therefore exploit asymptotic behavior of the learning method, rather than its convergence properties.

The rest of the paper is organized as follows. In the next section, we review Support Vector Machines and the capacity analysis based on VC dimension. It is followed by discussion of shape representation using a distance transform and the description of our algorithm for shape analysis based on SVMs. Section 4 reports the results of the method’s application to a data set of 30 examples (15

schizophrenia patient and 15 matched controls) of the hippocampus-amygdala complex and the comparison with a volume based analysis for the same data set. We conclude with the discussion of the results and future research directions.

2 Small Sample Size Learning and Support Vector Machines

In this section, we present a brief overview of Support Vector Machines (SVMs) and the capacity analysis based on the VC dimension. The reader is referred to the tutorial [3] for an extensive introduction and to the book [17] for more details on the theoretical foundation and the proofs.

VC bound. Given a training set of l pairs $\{(\mathbf{x}_i, y_i)\}_1^l$, where $\mathbf{x}_i \in R^n$ are observations and $y_i \in \{-1, 1\}$ are corresponding labels, and a family of classifier functions $\{f_\omega(\mathbf{x})\}$ parametrized by ω , the learning task is to select a member of the family that assigns labels to new, unseen examples while making as few errors as possible. Ideally, we would like to minimize the expectation of the test error, also called *expected risk*,

$$R(\omega) = \int \frac{1}{2} |y - f_\omega(\mathbf{x})| dP(\mathbf{x}, y), \quad (1)$$

where $P(\mathbf{x}, y)$ is the probability distribution that generates the observations. In practice, however, it is difficult to implement, as $P(\mathbf{x}, y)$ is unknown. Instead, the training error, also called *empirical risk*, can be computed:

$$R_{\text{emp}}(\omega) = \frac{1}{2l} \sum_{i=1}^l |y_i - f_\omega(\mathbf{x}_i)|. \quad (2)$$

One can show that for most families of classifier functions used in learning, the minimum of empirical risk converges in probability to the minimum of expected risk as the number of training examples increases. Furthermore, for any η ($0 \leq \eta \leq 1$), with probability at least $1 - \eta$, the classifier $f_{\omega^*}(\mathbf{x})$ that minimizes empirical risk $R_{\text{emp}}(\omega)$ on the given training set satisfies

$$R(\omega^*) \leq R_{\text{emp}}(\omega^*) + \sqrt{\frac{h}{l} \left(\log \frac{2l}{h} + 1 \right) - \frac{1}{l} \log \frac{\eta}{4}}, \quad (3)$$

where h is a quantity called *VC dimension*. The VC dimension of a family of classifiers is a measure of the complexity of the space of all hypotheses they can generate on any given training set. The right hand side of (3) is often referred to as the *VC bound* and its second term is called *VC confidence*. Note that it is a distribution-free bound, i.e., one does not need to know the distribution of the input data to estimate the convergence rate of the learning algorithm. This also implies that the bound is usually fairly loose for any particular distribution, and tighter bounds could be derived if the data distribution function were known.

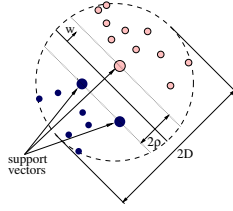


Fig. 1. Linearly separable classes with margin ρ and bounding sphere of radius D .

Support Vector Machines. Let's first consider a situation when the two classes are linearly separable (Fig. 1), i.e., there exists a vector \mathbf{w} and a constant b such that

$$\forall i: y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1, \quad (4)$$

where (\cdot) denotes a dot product of two vectors. The Support Vector learning machine searches for a \mathbf{w} that maximizes the margin between the two classes

$$\rho(\mathbf{w}) = \frac{1}{2\|\mathbf{w}\|} (\min_{y_i=1} (\mathbf{w} \cdot \mathbf{x}_i) - \max_{y_i=-1} (\mathbf{w} \cdot \mathbf{x}_i)), \quad (5)$$

which can be shown to be equivalent to minimizing $J(\mathbf{w}) = \|\mathbf{w}\|^2$ subject to constraints (4). Using Lagrange multipliers, the solution to this constrained quadratic optimization problem is a linear combination of the training examples $\mathbf{w}^* = \sum_{\mathbf{x}_i} \alpha_i y_i \mathbf{x}_i$ ($\alpha_i \geq 0$) with only a few non-zero coefficients α_i . Moreover, the non-zero α_i 's correspond to the training vectors that satisfy the inequality constraint (4) with equality. These are called *support vectors*, as they “support” the separating boundary between the classes (Fig. 1). The resulting classifier $f(\mathbf{x}) = \text{sign}((\mathbf{w}^* \cdot \mathbf{x}) + b) = \text{sign}(\sum \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b)$ is a linear function of dot products between the input vector \mathbf{x} and the support vectors and defines a hyperplane in R^n .

To extend this to a non-separable case, we introduce non-negative slack variables ξ_i that measure by how much each training example violates constraint (4). The optimization problem is transformed to minimizing $J(\mathbf{w}) = \|\mathbf{w}\|^2 + C \sum \xi_i$ subject to constraints $\forall i: y_i((\mathbf{w} \cdot \mathbf{x}_i) + b) \geq 1 - \xi_i$. The constant C determines a trade-off between maximizing the margin and minimizing the number of errors. The same techniques that were used in the separable case can be applied to this optimization problem as well, and the optimal vector \mathbf{w}^* is still a linear combination of a few support vectors.

This technique can also be extended to non-linear classification by observing that we only use dot products of data vectors to perform training and classification. Consider a *kernel function* $K: R^n \times R^n \mapsto R$, such that for some function $\Psi: R^n \mapsto R^m$ that maps the data into a higher dimensional space, the value of the dot product in R^m can be computed by applying K to the vectors in R^n : $\forall \mathbf{u}, \mathbf{v} \in R^n K(\mathbf{u}, \mathbf{v}) = (\Psi(\mathbf{u}) \cdot \Psi(\mathbf{v}))$. We can effectively train a linear classifier in the higher dimensional space R^m without explicitly evaluating Ψ , but rather using K to compute the dot products in R^m . This clas-

sifier produces a non-linear decision boundary back in the original space R^n : $f(\mathbf{x}) = \text{sign}(\sum \alpha_i y_i (\Psi(\mathbf{x}_i) \cdot \Psi(\mathbf{x})) + b) = \text{sign}(\sum \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b)$.

Several different kernel functions have been proposed in the machine learning community. In this work, we use a family of Gaussian kernel functions $K(\mathbf{u}, \mathbf{v}) = \exp\{-\|\mathbf{u} - \mathbf{v}\|^2/\gamma\}$, where the parameter γ determines the width of the kernel.

Classifier selection. It can be shown that the VC dimension of an optimal hyperplane is bounded by

$$h \leq \min(D^2/\rho^2, n) + 1, \quad (6)$$

where ρ is the margin of the classifier, D is the radius of the smallest sphere that contains all the training examples, and n is the dimensionality of the space (Fig. 1). This bound can also be computed in the non-linear case, as the radius of the bounding sphere in the target space can be estimated using the kernel function. This suggests a method for classifier selection: among optimal classifiers obtained for a hierarchy of function families (e.g., of different kernel width γ), choose the classifier that minimizes the VC bound (3), using the margin and the radius of the bounding sphere to estimate the VC dimension of the classifier.

For problems with a small number of training examples, the VC bound might be too loose to be helpful for classifier selection, and other methods, such as cross-validation, are employed. The relationship between VC dimension and cross-validation is discussed in detail in [17]. The traditional approach to estimating the expected test error from the cross-validation results uses the Law of Large Numbers and De Moivre - Laplace approximation to arrive at the following bound: with probability at least $1 - \eta$

$$|R(\omega^*) - \hat{R}| \leq \Phi^{-1}\left(\frac{1-\eta}{2}\right) \sqrt{\frac{\hat{R}(1-\hat{R})}{l}}, \quad (7)$$

where \hat{R} is the error rate of the cross-validation and $\Phi(x) = \int_0^x \frac{1}{\sqrt{2\pi}} \exp^{-t^2/2} dt$ is the standard error function.

3 Distance Transforms as Shape Descriptors

Numerous models have been proposed for shape description, as discussed in the introduction. In this section, we describe how we extract the shape information from segmented (binary) images using a distance transform and use it for classification.

A distance transform, or distance map, is a function that for any point inside an object is equal to the distance from the point to the closest point on an outline. Since the values of the distance transform at neighboring voxels are highly correlated, using it as a shape descriptor provides the learning algorithm with a lot of information on the structure of the feature space.

Another important property of the distance transform is what we call a *continuity of mapping*. The shape information extraction is an inherently noisy



Fig. 2. Data example: an example sagittal slice with the hippocampus-amygdala complex segmented (left), a 3D surface model of the hippocampus-amygdala complex (middle), and a slice of the 3D distance transform (right).

process, consisting of imaging, segmentation and feature extraction, with every step introducing errors. Small changes in the input image (rigid transformation, errors in segmentation, etc.) cause the corresponding feature vector to change as well. The vectors in the training set are therefore computed with uncertainty, and it is important that small errors in any of the steps do not cause significant displacements of the resulting feature vectors. Since the gradient magnitude of the distance map is bounded by 1 everywhere in the image, and changes in the outline have only local effect on the distance map, the distance between the new and the old feature vectors in the feature space is bounded by the “magnitude of transformation” in the image space.

Finally, there is a question of establishing correspondence between the components of different feature vectors, or alignment. We use moments of the distance transform to align the images into a canonical representation, and other techniques, e.g., rigid registration, can further improve the initial alignment of the shapes. This step must use rigid transformations, as the non-rigid deformation of the shape is exactly what should be captured and generalized by the learning algorithm. Moments of the distance map place higher weights on interior points than on points that are closer to the boundary, which reflects our belief that the interior points of the distance transform are estimated more reliably than the outline points. The images are also scaled to be of the same volume, as we consider shape properties independently of the volume.

To summarize, our algorithm takes segmented images, computes a distance transform and its moments, aligns and scales the images so that the object volume and the center of gravity of the distance transform is the same for all example shapes, and the principal axes of the distance transform coincide with the volume sampling directions. Then the distance transform can be either resampled using the same transformation or recomputed using the aligned images. To minimize resampling errors, we perform all of the operations on a sub-voxel level. The resulting distance maps are sampled uniformly in 3D and stored as vectors to be used for learning. Each component of the feature vector corresponds to a particular location in space and is equal to the value of the distance map in that location. Each anatomical structure is processed separately, and the resulting feature vector is a concatenation of the vectors for each structure of interest.

Structure	Right hippocampus		Left hippocampus	
Descriptor	volume	shape	volume	shape
Training accuracy(%)	60.0	83.3	63.3	83.3
Cross-validation(%)	60.0 \pm 17.5	66.7 \pm 16.9	63.3 \pm 17.2	60.0 \pm 17.5
Structure	Both hippocampi			
Descriptor	volume	shape	shape & volume	shape & rel. vol.
Training accuracy(%)	66.7	86.7	83.3	83.3
Cross-validation(%)	63.3 \pm 17.2	70.0 \pm 16.4	73.3 \pm 15.8	70.0 \pm 16.4

Table 1. Training and cross-validation accuracy for volume and shape. The results for cross-validation consist of estimated expected error, as well as 95% confidence interval, computed based on Eq. (7). The training accuracy is reported for the parameter setting that yielded the best cross-validation results.

4 Experimental Results

In this section, we report the results of the method applied to a data set that contains MRI scans of 15 schizophrenia patients and 15 matched controls. In each scan, the hippocampus-amygdala complex was manually segmented (Fig. 2a,b). More details on the subject selection and data acquisition can be found in [13]. The same paper reports statistically significant differences in left anterior hippocampus based on relative volume measurements (the volume of the structure normalized by the total volume of intracranial cavity or ICC). For each of the experiments reported in this paper, we systematically explore the space of parameters (the kernel width γ and the soft margin constant C) by sampling it on a logarithmic scale, training a Gaussian kernel classifier and estimate its VC dimension as described in Section 2, and also performing leave-one-out cross-validation. Table 1 contains the summary of training and the cross-validation accuracy for all our experiments.

In order to compare the shape based analysis to the volume based analysis, we first trained a classifier and performed leave-one-out cross-validation based on the relative volume measurements only¹. We then repeat the experiment with the shape descriptors (distance maps). Fig. 3a,b shows the accuracy sampling for different parameter settings for shape classification, and Table 1 contains the training and the cross-validation accuracy for the best classifier selected based on the cross-validation results. If used on a single side of hippocampus, the shape based results are not significantly different from the volume based result, but there is a significant improvement (63% to 70%) when the two sides are considered together. This suggests that the disease might affect the shape of both hippocampi, and it is the combined shape that is the best indicator of those differences.

¹ Note that for volume based classification, the cross-validation accuracy does not differ significantly from the training accuracy, because removing a single training example can only affect the training result if the example is close to the threshold (i.e., it's a support vector), and there could only be very few of those in the low-dimensional space.

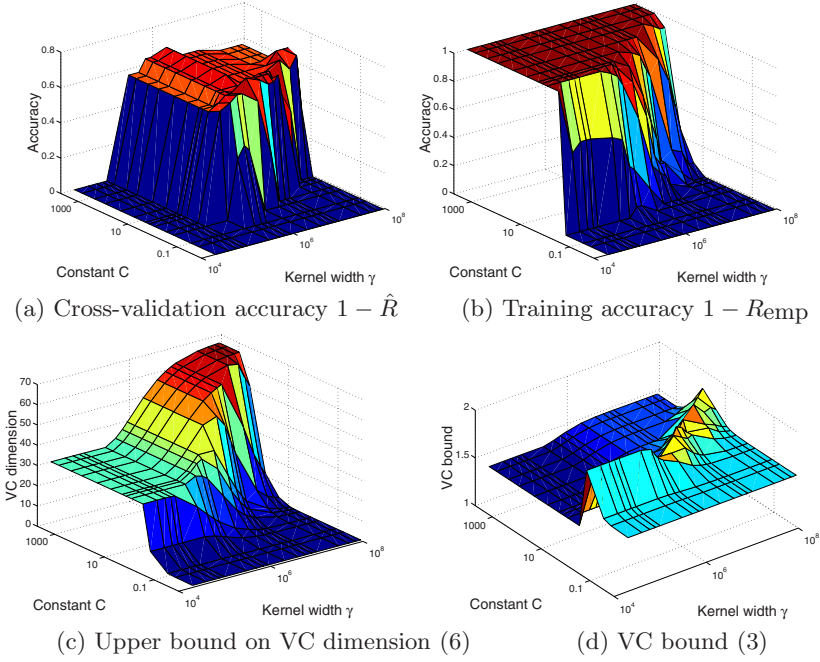


Fig. 3. Learning results for different parameter settings using both hippocampi.

Given the results reported so far, the natural question to ask is what happens if we combine shape information with volume measurements? If they are decorrelated, better estimates could potentially be obtained by using both. In order to investigate this question, we re-ran the experiment with a slightly modified feature extraction procedure. In the first experiment we kept the volume of the structure unchanged, and then ran the experiment again while scaling the images so that the resulting volume for every image is proportional to the relative volume as reported in the original paper [13]. The results of these experiments are also shown in Table 1. We can see that combining volume and shape information yields additional improvement (cross-validation accuracy 73%).

In order to perform the capacity analysis of the classifier, we estimated VC dimension for each one of the parameter settings and computed the VC bound (3) for the shape based experiment (see Fig. 3). Unfortunately, the bound is too loose to be helpful for classifier selection (it is greater than 1), and we have to resort to cross-validation. Examining the graphs in Fig. 3, we observe that empirical risk R_{emp} and the estimated VC dimension behave in opposite ways: one increases when the other goes down. But when they are combined in (3), the VC confidence dominates the empirical risk term and the sum does not have a distinct minimum. As the number of input examples increases, the estimates of the VC dimension will become more accurate, and the bound can be useful in investigating the functional behavior of the generalization error for different setting of parameters γ and C .

5 Discussion and Conclusion

We have proposed and demonstrated a novel approach to statistical shape analysis of anatomical structures. Our technique is based on the body of knowledge developed in machine learning for situations when the distribution of the data is unknown and the number of examples is limited. It makes no assumptions on the distribution of the data and moreover, is guaranteed to converge to the optimal solution (minimizing the expected error on future examples) as the number of training examples increases. It provides a principled way to attack a problem of modeling statistical shape differences in the medical domain. When applied to the hippocampus study in schizophrenia, shape information allowed us to improve over a purely volume based approach. The results are comparable to the traditional approach (e.g., Csernansky *et al* [6] report 80% accuracy in cross-validation using deformation fields as a shape model and PCA for dimensionality reduction), and if more examples are available, can potentially tighten the bound on the test error.

We used distance maps as shape descriptors in this work, but the learning techniques presented here are applicable to other shape models as well. The main contribution of this paper is in the way we perform learning on the shape features. Comparison between different shape representations and their performance when combined with our analysis is an interesting question that needs to be investigated in the future.

There are several other topics that we would like to explore next, and the most important one is the interpretation of the learning results in terms of shape deformation. Linear models provide a very intuitive interpretation of the parameters as weights assigned to the corresponding vectors (e.g., modes of deformation), but it cannot be extended in a straightforward manner to any family of non-linear models. Finding methods for mapping a non-linear classifier function back into the image domain and providing an interpretation of the results for the medical researches is our current interest.

Another important direction of research in this framework is incorporating invariants into the learning process. To enable efficient learning, information on the structure of the feature space has to be provided to the training algorithm. Some of this is achieved by selecting a good representation. Invariants are important constraints on the feature space that often cannot be explicitly modeled by the shape representation. In some cases, a family of classifiers can be changed to guarantee that the resulting function satisfies the invariance constraint (see [4], for example), or artificial examples can be generated using the invariants and added to the original training set [1]. We plan to explore this direction as a way of further restricting the classifier family and thus improving the accuracy.

To summarize, recent advances in statistical learning theory enabled a new approach to statistical analysis of high dimensional data. As the results presented in this paper demonstrate, the field of statistical shape analysis, and medical research in general, can benefit from these techniques. While there are still many open questions on interpretation of the results and incorporation of prior knowl-

edge into the method, it is a promising direction of research that can help medical researches to get a better insight into various anatomical phenomena.

Acknowledgments. Quadratic optimization was performed using PR-LOQO optimizer written by Alex Smola.

This research was supported in part by NSF IIS 9610249 grant. M. E. Shenton was supported by NIMH K02, MH 01110 and R01 MH 50747 grants, R. Kikinis was supported by NIH PO1 CA67165, R01RR11747, P41RR13218 and NSF ERC 9731748 grants.

References

1. Y.S. Abu-Mostafa. Learning from hints. *J. Complexity*, 10(1):165-178, 1994.
2. F. L. Bookstein. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis*, 1(3):225-243, 1996.
3. C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121-167, 1998.
4. C. J. C. Burges. Geometry and Invariance in Kernel Based Methods. In *Advances in Kernel Methods: Support Vector Learning*, Eds. B. Schölkopf, C.J.C. Burges, A.J. Smola, MIT Press, 89-116, 1998.
5. T. F. Cootes *et al.* The Use of Active Shape Models for Locating Structures in Medical Images. *Image and Vision Computing*, 12(6):355-366, 1994.
6. J. G. Csernansky *et al.* Hippocampal morphometry in schizophrenia by high dimensional brain mapping. In *Proc. Nat. Acad. of Science*, 95(19):11406-11411, 1998.
7. D. S. Fritsch *et al.* The multiscale medial axis and its applications in image registration. *Pattern Recognition Letters*, 15:445-452 1994.
8. J. C. Gee and R. Bajcsy, Personal communication.
9. A. Kelemen, G. Székely, and G. Gerig. Three-dimensional Model-Based Segmentation. In *Proc. IEEE International Workshop on Model Based 3D Image Analysis*, Bombay, India, 87-96, 1998.
10. J. Martin, A. Pentland, and R. Kikinis. Shape Analysis of Brain Structures Using Physical and Experimental Models. In *Proc. CVPR'94*, 752-755, 1994.
11. M. E. Leventon, W. E. L. Grimson and O. Faugeras. Statistical Shape Influence in Geodesic Active Countours. In *Proc. CVPR'2000*, 316-323, 2000.
12. E. E. Osuna, R. Freund, F. Girosi. Training Support Vector Machines: An Application to Face Detection. In *Proc. CVPR'97*, 130-136, 1997.
13. M. E. Shenton, *et al.* Abnormalities in the left temporal lobe and thought disorder in schizophrenia: A quantitative magnetic resonance imaging study. *New England J. Medicine*, 327:604-612, 1992.
14. L. Staib and J. Duncan. Boundary finding with parametrically deformable models. *IEEE PAMI*, 14(11):1061-1075, 1992.
15. G. Székely *et al.* Segmentation of 2D and 3D objects from MRI volume data using constrained elastic deformations of flexible Fourier contour and surface models. *Medical Image Analysis*, 1(1):19-34, 1996.
16. V. N. Vapnik. The Nature of Statistical Learning Theory. *Springer*, 1995.
17. V. N. Vapnik. Statistical Learning Theory. *John Wiley & Sons*, 1998.