Available online at www.sciencedirect.com
SCIENCE DIRECT®
Medical Image Analysis 9 (2005) 69–86
MEDICAL IMAGE ANALYSIS
www.elsevier.com/locate/media

Detection and analysis of statistical differences in anatomical shape

Polina Golland [a,*], W. Eric L. Grimson [a], Martha E. Shenton [b], Ron Kikinis [c]

[a] *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[b] *Laboratory of Neuroscience, Clinical Neuroscience Division, Department of Psychiatry, VAMC-Brockton, Harvard Medical School, Brockton, MA 02401, USA*
[c] *Surgical Planning Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA*

Received 1 June 2002; received in revised form 21 January 2004; accepted 22 June 2004
Available online 29 September 2004

Abstract

We present a computational framework for image-based analysis and interpretation of statistical differences in anatomical shape between populations. Applications of such analysis include understanding developmental and anatomical aspects of disorders when comparing patients versus normal controls, studying morphological changes caused by aging, or even differences in normal anatomy, for example, differences between genders. Once a quantitative description of organ shape is extracted from input images, the problem of identifying differences between the two groups can be reduced to one of the classical questions in machine learning of constructing a classifier function for assigning new examples to one of the two groups while making as few misclassifications as possible. The resulting classifier must be interpreted in terms of shape differences between the two groups back in the image domain. We demonstrate a novel approach to such interpretation that allows us to argue about the identified shape differences in anatomically meaningful terms of organ deformation. Given a classifier function in the feature space, we derive a deformation that corresponds to the differences between the two classes while ignoring shape variability within each class. Based on this approach, we present a system for statistical shape analysis using distance transforms for shape representation and the support vector machines learning algorithm for the optimal classifier estimation and demonstrate it on artificially generated data sets, as well as real medical studies.

Keywords: Shape analysis; Discriminative analysis; Shape classification

1. Introduction

Anatomical shape, and its variation, remains an important topic of medical research. Understanding morphological changes caused by a particular disorder can help to identify the time of onset of a disease, quantify its development and potentially lead to better treatment. Other examples of morphological studies include investigating anatomical changes due to aging by comparing different age groups, and studies of anatomical differences between genders. Originally, image-based statistical studies of morphology were based on simple measurements of size, area and volume. Shape-based analysis promises to provide much more detailed descriptions of the anatomical changes due to the biological process of interest. In this paper, we present a computational framework for performing statistical comparison of populations based on complex shape descriptors. The analysis considers the entire set of shape features simultaneously and yields an assessment of how much the shape of the organ differs between the two populations, as well as a detailed description of the identified differences.

Image-based shape analysis typically consists of three main steps. First, quantitative measures of shape are extracted from each input image and are combined into a feature vector that describes the input shape. The set of

* Corresponding author.
E-mail address: polina@csail.mit.edu (P. Golland).

1361-8415/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2004.07.003

feature vectors is then used to construct either a generative model of shape variation within one population or a discriminative model of shape differences between two populations. This is followed by interpretation of the statistical model in terms of the original shape and image properties. Such interpretation is necessary for visualization and improved understanding of detected shape differences. In this section, we describe each of the three stages of the analysis, provide a review of related work and outline our approach.

1.1. Feature extraction

Shape analysis starts with extraction of shape features from input images. A great number of shape descriptors have been proposed for use in medical image analysis. They can be classified into several broad families, such as landmarks (Bookstein, 1997; Cootes et al., 1992; Dryden and Mardia, 1998), dense surface meshes (Brechbühler et al., 1995; Kelemen et al., 1998; Shenton et al., 2002; Staib and Duncan, 1992; Székely et al., 1996), skeleton-based representations (Bookstein, 1979; Fritsch et al., 1994; Golland et al., 1999; Golland and Grimson, 2000; Pizer et al., 1996), deformation fields that define a warping of a standard template to a particular input shape (Christensen et al., 1993; Davatzikos et al., 1996; Martin et al., 1994; Machado and Gee, 1998) and distance transforms that embed the outline of the object in a higher dimensional distance function over the image (Golland et al., 2000; Leventon et al., 2000). The choice of shape representation depends crucially on the application. For statistical modeling, the two most important properties of a shape descriptor are its sensitivity to noise in the input images and the ease of registration of the input examples into a common coordinate frame. [1] These determine the amount of noise in the training data, and therefore the quality of the resulting statistical model. In this work, we choose to use an existing approach based on distance transforms for feature extraction, mainly because of its simplicity and its smooth dependence on the noise in the object's boundary and its pose. The focus of this paper is on the later steps of the analysis that produce an interpretation of the statistical model, and not on the shape representation per se. Section 6 offers a discussion on employing other shape descriptors in conjunction with the statistical analysis tools presented in this paper.

1.2. Statistical analysis

Once the features have been extracted from the images, they are used to construct a statistical model of differences between the two groups of feature vectors. One approach is to treat the features as independent variables and to perform a simple hypothesis test on each feature separately (Bookstein, 1997; Machado and Gee, 1998; Yushkevich et al., 2001). If the features are local (e.g., voxels or boundary segments), the interpretation of the resulting model becomes straightforward, as we can create a mask in the image domain indicating the features that were deemed significantly different in the two groups. Unfortunately, it can be difficult to assess the significance of the entire pattern from the individual statistical tests. Moreover, if the features are global, such as Fourier coefficients of the outline curve, the interpretation of the detected differences becomes difficult. Alternatively, we can estimate a statistical model based on the entire vectors and possible dependencies among the features. In the generative case, this is typically done by applying principal component analysis (PCA) to estimate the mean and the covariance structure of the training set (Cootes et al., 1992, 1999; Kelemen et al., 1998). Earlier work on shape differences between populations employed PCA for dimensionality reduction, followed by training a simple (linear or quadratic) classifier in the reduced space (Csernansky et al., 1998; Martin et al., 1994). In this work, we use the support vector machines (SVMs) algorithm (Burges, 1998; Vapnik, 1995, 1998) to estimate the optimal classifier function directly in the original feature space while explicitly controlling its complexity. In addition to the theoretical reasons for its asymptotic optimality, Support Vector learning has been empirically demonstrated to be robust to overfitting and to generalize well even for small data sets. Furthermore, the algorithm provides a principled way to explore a hierarchy of increasingly complex classifier families, trading-off the training error and the complexity of the model.

1.3. Model interpretation in the image domain

To be useful in the clinical context, the resulting statistical model (eigenmodes or a classifier function) must be mapped back to the image domain, i.e., analyzed in terms of the input shape or image properties in order to generate a comprehensible description of the structure in the training data that was captured by the model. In the generative case, this is often done by sampling the implied Gaussian distribution with the mean and the covariance estimated from the data or by varying one principal component at a time. We previously used a similar approach for interpretation of a linear classifier by translating the original feature vector along the projection vector of the classifier function (Golland et al., 1999). More commonly, however, the resulting classifier is used only to establish statistical significance of morphological differences between the classes, and the generative models based on PCA are employed for

[1] This step might include explicitly establishing correspondences based on a discrete set of landmarks, or might produce implicit correspondences across different subjects by aligning their anatomy.

visualization of the shape variation within each group (Csernansky et al., 1998; Gerig et al., 2001; Martin et al., 1994), and approach that does not provide a direct comparison of the populations based on the estimated discriminative model.

1.4. Novel approach to model interpretation

In this paper, we demonstrate how to obtain a description of shape differences captured by the classifier function that was constructed in the analysis step. To understand the differences implicitly represented by the classification function, we study the function's sensitivity to changes in the input along different directions in the feature space. For every input example, we solve for the direction in the feature space that maximizes the change in the classifiers value while introducing as little irrelevant changes into the input vector as possible. We derive the sensitivity analysis for a large family of non-linear kernel-based classifiers. The results can be represented in the image domain as deformations of the original input shape, yielding both a quantitative description of the morphological differences between the classes and an intuitive visualization mechanism. Thus, in addition to the statistical descriptors, such as test error and confidence bounds, we also provide a mechanism for explicit visual interpretation of the detected shape variation between the two populations.

The remainder of this paper is organized as follows. In the next section, we explain our choice of the distance transforms for extracting shape features and introduce a local parameterization of the distance transform space which allows us to represent and visualize changes in the distance transform as deformations of the corresponding boundary surface. This is followed by a brief review of the Support Vector learning and a derivation of the discriminative direction as a description of differences between two classes captured by the classification function. We then combine shape description with statistical analysis and demonstrate the technique on a simple artificial example, as well as real medical studies. The paper concludes with a discussion of the lessons learned from the presented experiments and future research directions.

2. Shape representation

We chose to work with volumetric descriptors in order to avoid the implementation difficulties associated with establishing a common coordinate frame on the surface of such relatively smooth objects as subcortical structures. Our main interest lay in further developing the interpretation step of the analysis that provides a representation and visualization of the statistical model, leading to a choice of a very simple shape descriptor. And since the analysis assumes that the representation gives rise to an analytical function (i.e., one that changes smoothly as we slightly perturb the original shape), we augmented the binary segmentations to avoid discontinuities at the boundaries of the objects. Distance transform provided a simple linearization of the binary segmentation and therefore was chosen in this work. However, the analysis presented in the following sections is applicable to other representations of the shape, as we briefly discuss in Section 6.

2.1. Background: distance transforms

A distance transform, or distance map, is a function that for each point in the image (2D or 3D) is equal to the shortest distance from that point to the boundary of the object. The boundary is modeled implicitly as a zero level-set of the distance transform. We use a signed variant of the distance transform that negates the values of the distance transform outside the object, thus eliminating the singularity at the boundary. The distance transform is a piece-wise linear function, whose singularity ridges form the object's skeleton (Blum, 1967; Kimmel et al., 1995; Leymarie and Levine, 1992). [2] Distance transforms have been used in medical image analysis for extracting a medial axis representation of a shape (Fritsch et al., 1994; Golland et al., 1999; Pizer et al., 1996), and more recently, directly as a shape descriptor (Golland et al., 2000, 2001; Golland and Grimson, 2000; Leventon et al., 2000).

For shape studies, one would like to use a descriptor that is invariant under a family of rigid transformations, as the object's pose relative to the rest of the anatomy and the scanner coordinate frame should not affect its shape characteristics. Many volumetric representations are not invariant under rigid transformations, which is typically mitigated by aligning all shapes, or bringing them into a "canonical" pose, before constructing the descriptors. We align the shapes using moments of the distance transform inside the shape. In contrast to the moments of shape that weigh all points equally, the moments of the distance transform assign weights proportionally to the distance from the boundary, reflecting our belief that the interior points of the distance transform are estimated more reliably than the points close to the boundary of the object. This procedure is identical to the one used in (Leventon et al., 2000) when constructing a model of shape variation to guide a level-set segmentation. As we show later in this section, the errors in the resulting numerical descriptors are bounded by the errors in the outline introduced in

[2] To make the descriptor truly differentiable up to the order n, we would have to raise the distance values to the nth power. The distance transform itself, while continuous, is not differentiable at the skeleton points. However, we felt that it provided a reasonable first approximation that allowed us to carry on.

the segmentation and the alignment steps. This makes the distance transform an attractive choice for shape representation, as it guarantees stability in the presence of noise and allows rigorous analysis of errors in the training data.

The alignment also establishes implicit correspondences among locations in different input images. The problem of alignment, based on either the global functions of the shape such as moments, or distinct landmarks, has not been solved satisfactorily in the field of shape analysis. One of the main obstacles is that it is not clear what the "correct correspondence" between two examples of an anatomical shape is. Most methods use rigid alignment, operating under the premise that the non-rigid differences among the examples are exactly the shape differences we would like to model and study. Since we cannot assume perfect alignment, it is important to understand the descriptor's behavior under noise in the object's pose. Solving the alignment problem is outside the scope of this paper, but we will come back to this question in Section 6 when discussing the sensitivity of the analysis to shape representation. In our experience, the choice of representation influences the accuracy of the detection much more than the alignment procedure.

2.2. Local parameterization using surface meshes

The sensitivity analysis of the classifier function presented in the next section produces a direction of change in the input shapes that represents the differences between the two populations. In order to visualize or apply this change, or deformation, we must be able to move in the space of the distance transforms along a specified direction. However, the distance transforms do not form a linear vector space, but rather lie on a manifold in the higher dimensional space of real-valued images. This manifold is fully determined by the local constraints on the distance transform, but unfortunately, it does not have a *global* parametric description, which makes such local analysis difficult. We solve this problem by using a *local* parameterization of the manifold around any particular shape example based on the object's surface mesh. In a 3D image, a mesh contains nodes, edges and faces. In a 2D image, a mesh simplifies to a simple loop defined by its nodes and edges. We use the mesh node locations as features, similarly to the active shape models (Cootes et al., 1992, 1999). Ease of generative modeling based on fixed graph topology makes surface meshes an attractive choice for representing a family of possible deformations of the original shape. Here, we present the preliminary derivations that will allow us to investigate a small neighborhood of shapes around a particular input example; the usefulness of this analysis will become obvious in Section 3.

Formally, let $\mathbf{x}$ be a feature vector formed by concatenating the distance transform values at all voxels in the image. Let $\mathbf{s}$ be a vector of node displacements in the surface mesh of the same object: s_i is the displacement of node i along the normal to the boundary at that node. Since moving nodes along the surface of the shape does not change the shape, we will only consider changes in the node positions along the normal to the boundary, arbitrarily choosing the positive direction of the normal vector to point outwards. Zero vector $\mathbf{s}$ corresponds to the original shape and its distance map $\mathbf{x}$. We will say that a point on the boundary *influences* a particular location in the volume if it is the closest boundary point to that location. Obviously, the value of the distance transform at any voxel is equal to the distance from the voxel to the point(s) on the boundary that influence that voxel. Most voxels in the volume are influenced by one boundary point each, with the exception of the skeleton voxels. The distance transform feature vector $\mathbf{x}$ can be thought of as a union of two sub-vectors: vector $\hat{\mathbf{x}}$ that contains the distance transform values at the non-skeleton voxels, and vector $\check{\mathbf{x}}$ that contains the distance transform values at the skeleton voxels.

Given an infinitesimal displacement vector $\mathrm{d}\mathbf{s}$ that defines a deformation of the original shape, the value of the distance transform in any non-skeleton voxel changes by the amount equal to the displacement of the boundary point that influences it, which implies a local linear parameterization of the distance transform manifold around the original point $\hat{\mathbf{x}}$:

$$\mathrm{d}\hat{\mathbf{x}} = J_{\mathbf{x}}\,\mathrm{d}\mathbf{s}, \tag{1}$$

where $J_{\mathbf{x}}$ is the "influence matrix" that contains exactly one non-zero element in every row and is also the Jacobian of the parameterization:

$$J_{\mathbf{x}}(i,j) = \frac{\partial x_i}{\partial s_j} = \begin{cases} 1, & j \in S(i), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $S(i)$ is a set of mesh nodes that influence voxel i. The change in the value of the distance transform in any skeleton voxel is equal to the minimum of the (signed) displacements of all the nodes that influence that voxel:

$$\mathrm{d}\check{x}_i = \min_{j \in S(i)} \mathrm{d}s_j, \tag{3}$$

exhibiting a sub-linear behavior bounded by the linear model above.

Given an arbitrary infinitesimal displacement vector $\mathrm{d}\mathbf{x}$ of the original feature vector $\mathbf{x}$, we can find the deformation of the boundary that generates the closest on the distance transform manifold to the resulting feature vector:

$$\mathrm{d}\mathbf{s} \approx (J_{\mathbf{x}}^{\mathrm{T}} J_{\mathbf{x}})^{-1} J_{\mathbf{x}}^{\mathrm{T}} \mathrm{d}\hat{\mathbf{x}}, \tag{4}$$

where matrix $(J_{\mathbf{x}}^{\mathrm{T}} J_{\mathbf{x}})^{-1} J_{\mathbf{x}}^{\mathrm{T}}$ is the generalized inverse of the Jacobian $J_{\mathbf{x}}$. This projection operation is an approximation that ignores the value changes in the skeleton voxels

and assumes that the deformation d**s** does not change the set of skeleton voxels. We believe this is a reasonable model for several reasons. First, the number of skeleton voxels is usually small compared to the total number of voxels: in all our experiments, the number of skeleton voxels was 4–6% of the total number of voxels in the volume. In general, a skeleton of a shape is a surface of one dimension lower than the dimensionality of the image space (e.g., a skeleton of a 3D shape is a 2D surface) and typically contains a negligible number of voxels compared to the shape itself. Second, the value changes in these voxels are bounded by the linear model and therefore cannot have extreme contributions to the changes in the distance transform. And finally, we are interested in the change of the values of the distance transform voxels induced by the infinitesimal shape deformation, not their membership in the skeleton set. The approximation would become invalid if we were interested in large deformations of the original shape, as it would change not only the skeleton, but also the influence sets of the non-skeleton voxels.

The resulting linear change in the distance transform values at the non-skeleton voxels

$$d\hat{\mathbf{x}}^o = J_{\mathbf{x}}\mathrm{d}\mathbf{s} \approx J_{\mathbf{x}}(J_{\mathbf{x}}^{\mathrm{T}}J_{\mathbf{x}})^{-1}J_{\mathbf{x}}^{\mathrm{T}}\mathrm{d}\hat{\mathbf{x}} \tag{5}$$

can be extended to a full volume vector d$\mathbf{x}^o$ by computing the change at the skeleton voxels as a minimum of the displacements of the influencing nodes:

$$\mathrm{d}x_i^o = \begin{cases} \min\limits_{j\in S(i)} d_{s_j}, & \|S(i)\| > 1, \\ \mathrm{d}\hat{x}_i^o, & \|S(i)\| = 1. \end{cases} \tag{6}$$

This operation projects the infinitesimal displacement vector d**x** onto the distance transform manifold so that the resulting change in the distance transform d$\mathbf{x}^o$ defines a valid deformation of the shape while minimizing the distance between the two vectors. Fig. 1 illustrates this operation for a simple shape example. We will use this operation when analyzing changes in the original shapes suggested by the statistical analysis in the next section.

Before presenting the statistical analysis steps, let us summarize the feature extraction procedure:

- example images are segmented into relevant structures;
- these images are aligned by computing a distance map, then computing the transformation that aligns the moments of the maps;
- the aligned images are clipped to a common size;
- the aligned and clipped 3D distance transforms are used as feature vectors, with individual voxels being the vector components.

3. Statistical analysis

Once feature vectors have been extracted from the input images, they are used to construct a classifier for distinguishing between the two example groups. We use the SVMs learning algorithm to estimate the optimal classifier function. The classifier function constructed during the training phase implicitly encodes the differences in data between the two classes and we are interested in understanding the nature of the these differences. If expressed in terms of the original images or shapes, this information can provide an insight into the anatomical implications of detected shape differences. The analysis presented in this section addresses exactly this problem. We introduce and derive a *discriminative direction* at every point in the feature space (e.g., the space of distance transforms) with respect to a given classifier. Informally speaking, the discriminative direction tells us how to change any input example to make it look more like an example from another class without introducing any irrelevant changes that possibly make it more similar to other examples from the same class. It allows us to characterize shape differences captured by the classifier and to express them as deformations of the original shapes. And while we present the analysis for the classifier functions trained using the SVM algorithm, it can be extended to other classifiers that can be expressed as analytical functions.

3.1. Background: statistical modeling using support vector machines

Below, we state without proof the main facts on SVMs that are necessary for derivation of our results. The reader is referred to the original publications on

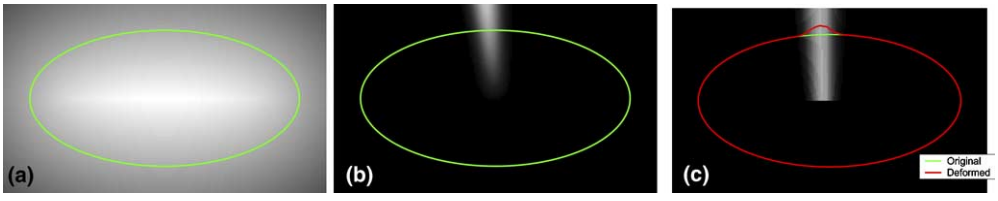

Fig. 1. Simple example of the distance transform parameterization. The images show an example outline and its signed distance transform **x** (a); an example of a displacement vector d**x** (b); the outline obtained by deforming the original boundary by the estimated deformation vector d**s** and the corresponding change in the distance transform d$\mathbf{x}^o$ (c).

SVMs (Burges, 1998; Vapnik, 1998) for more details on the method.

Given a training set of l pairs $\{(\mathbf{x}_k, y_k)\}_{k=1}^{l}$, where $\mathbf{x}_k \in \mathbb{R}^n$ are observations and $y_k \in \{-1,1\}$ are corresponding labels, the SVMs learning algorithm constructs a linear classifier

$$f(\mathbf{x}) = \langle \mathbf{x} \cdot \mathbf{w} \rangle + b \tag{7}$$

that maximizes the margin between the classes with respect to the separating hyperplane. It can be shown that the projection vector that maximizes the margin is equal to a linear combination of the training examples,

$$\mathbf{w}^* = \sum_{k=1}^{l} \alpha_k y_k \mathbf{x}_k, \tag{8}$$

where the non-negative coefficients α_k are determined by solving a constrained quadratic optimization problem. The vectors with non-zero coefficients are called *support vectors*, as they define, or "support", the boundary. The optimization problem of estimating the coefficients α_k and the resulting classifier

$$f(\mathbf{x}) = \sum_{k=1}^{l} \alpha_k y_k \langle \mathbf{x} \cdot \mathbf{x}_k \rangle + b \tag{9}$$

can be expressed entirely in terms of dot products of the training data, suggesting a natural extension of the technique to non-linear classifiers through use of kernel functions.

Function $K : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ is called a kernel if there exists a mapping function $\Psi_K : \mathbb{R}^n \mapsto \mathbb{F}$ from the original space to an arbitrary higher dimensional space $\mathbb{F}$ (for example, $\mathbb{R}^m$, $m \geqslant n$) such that the dot products of image vectors $\Psi_K(\mathbf{x})$ in the space $\mathbb{F}$ can be computed by applying K to the original vectors in $\mathbb{R}^n$:

$$K(\mathbf{u}, \mathbf{v}) = \langle \Psi_K(\mathbf{u}) \cdot \Psi_K(\mathbf{v}) \rangle \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n. \tag{10}$$

While mapping the points from the original space $\mathbb{R}^n$ to the higher dimensional space $\mathbb{F}$ and performing computations explicitly in $\mathbb{F}$ might be prohibitive because of the dimensionality of the space, we can compute certain quantities in $\mathbb{F}$ without ever computing the mapping if the answer depends only on dot products of the vectors in $\mathbb{F}$ (Fig. 2(a)). Thus, substituting the values of the kernel function instead of the dot products into the original optimization problem yields a non-linear classifier

$$f_K(\mathbf{x}) = \sum_{k=1}^{l} \alpha_k y_k K(\mathbf{x}, \mathbf{x}_k) + b \tag{11}$$

that defines a separating hyperplane in the higher dimensional space $\mathbb{F}$ whose normal is a linear combination of the support vector images $\Psi_K(\mathbf{x}_k)$'s:

$$\mathbf{w} = \sum_{k=1}^{l} \alpha_k y_k \Psi_K(\mathbf{x}_k). \tag{12}$$

However, the separating boundary in the original space can be an arbitrarily complex surface defined by the kernel function K and the support vector coefficients α_k.

For the linear kernel $K(\mathbf{u},\mathbf{v}) = \langle \mathbf{u} \cdot \mathbf{v} \rangle$, the mapping Ψ_K is the identity, and the classifier is a linear function in the original space. Several non-linear kernels have been also proposed. In this work, we use a family of Gaussian radial basis function (RBF) kernels

$$K(\mathbf{u}, \mathbf{v}) = \mathrm{e}^{-\|\mathbf{u}-\mathbf{v}\|^2/\gamma}, \tag{13}$$

where the parameter γ determines the width of the kernel. One of the important properties of this family of classifiers is its locality: moving a support vector slightly affects the separating boundary close to the vector, but does not change it in regions distant from the vector. Following the discussion in the previous section, this is a desirable property in the presence of noise in the training examples.

In order to select an optimal setting of the kernel parameters (e.g., the kernel width γ for the Gaussian RBF kernels), cross-validation techniques are often employed (Efron, 1982). In this work, we use leave-one-out cross-validation to estimate the expected accuracy of the resulting classifier and the traditional approach to esti-

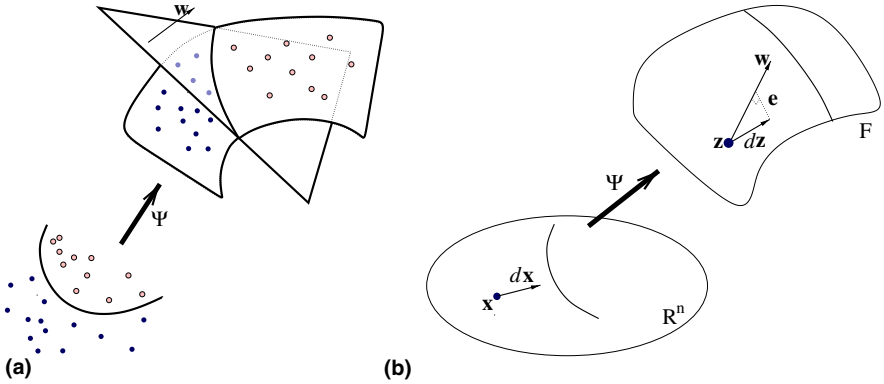

Fig. 2. Geometry of the kernel mapping and the discriminative direction. (a) Kernel mapping, (b) discriminative direction.

mating the confidence intervals based on the Law of Large Numbers and De Moivre–Laplace approximation: with probability at least $1-\eta$

$$|R-\hat{R}| \leqslant \Phi^{-1}\left(\frac{1-\eta}{2}\right)\sqrt{\frac{\hat{R}(1-\hat{R})}{l}}, \tag{14}$$

where R is the expected error of the trained classifier, $\hat{R}$ is the cross-validation error and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x \mathrm{e}^{-t^2/2}\,\mathrm{d}t \tag{15}$$

is the standard error function.

3.2. Discriminative direction

Eqs. (11) and (12) imply that the classifier function depends only on the projection of vector $\Psi_K(\mathbf{x})$ onto the vector $\mathbf{w}$ and it completely ignores the component of $\Psi_K(\mathbf{x})$ that is perpendicular to $\mathbf{w}$. This suggests that in order to create a displacement of $\Psi_K(\mathbf{x})$ that corresponds to the differences between the two classes, one should change the vector's projection onto $\mathbf{w}$ while keeping its perpendicular component the same. In the linear case, we can easily perform this operation, since we have access to the image vectors, $\Psi_K(\mathbf{x}) = \mathbf{x}$. This is similar to visualization techniques typically used in linear generative modeling, where the data variation is captured using PCA, and new samples are generated by changing a single principal component at a time. However, this approach is infeasible in the non-linear case, because we do not have access to the image vectors $\Psi_K(\mathbf{x})$'s. Our solution is to search for the direction around the feature vector $\mathbf{x}$ in the original space that minimizes the divergence of its image $\Psi_K(\mathbf{x})$ from the direction of the projection vector $\mathbf{w}$ (Fig. 2(b)). We call it a *discriminative direction*, as it represents the direction that affects the output of the classifier while introducing as little irrelevant change as possible into the input vector. Note that it might be impossible to move exactly along $\mathbf{w}$ as the image vectors do not populate the entire space $\mathbb{F}$, but rather form a manifold of lower dimensionality.

Formally, as we move from $\mathbf{x}$ to $\mathbf{x}+\mathrm{d}\mathbf{x}$ in $\mathbb{R}^n$, the image vector in space $\mathbb{F}$ changes by

$$\mathrm{d}\mathbf{z} = \Psi_K(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Psi_K(\mathbf{x}), \tag{16}$$

creating a deviation from $\mathbf{w}$:

$$\mathbf{e} = \mathrm{d}\mathbf{z} - \frac{\langle \mathrm{d}\mathbf{z}\cdot\mathbf{w}\rangle}{\|\mathbf{w}\|}\frac{\mathbf{w}}{\|\mathbf{w}\|} = \mathrm{d}\mathbf{z} - \frac{\langle \mathrm{d}\mathbf{z}\cdot\mathbf{w}\rangle}{\langle \mathbf{w}\cdot\mathbf{w}\rangle}\mathbf{w}. \tag{17}$$

Thus, the following constrained optimization problem defines the discriminative direction:

$$\text{minimize}\quad E(\mathrm{d}\mathbf{x}) = \|\mathbf{e}\|^2 = \langle \mathrm{d}\mathbf{z}\cdot\mathrm{d}\mathbf{z}\rangle - \frac{\langle \mathrm{d}\mathbf{z}\cdot\mathbf{w}\rangle^2}{\langle \mathbf{w}\cdot\mathbf{w}\rangle} \tag{18}$$

$$\text{s.t.}\quad \|\mathrm{d}\mathbf{x}\|^2 = \epsilon. \tag{19}$$

Since the cost function depends only on dot products of vectors in the space $\mathbb{F}$, it can be computed using the kernel function K. Substituting expressions for $\mathbf{w}$ and $\mathrm{d}\mathbf{z}$ from Eqs. (12) and (16), respectively, and using the kernel function to evaluate dot products, we obtain

$$\langle \mathbf{w}\cdot\mathbf{w}\rangle = \sum_{k,m}\alpha_k\alpha_m y_k y_m K(\mathbf{x}_k,\mathbf{x}_m), \tag{20}$$

$$\begin{aligned}\langle \mathrm{d}\mathbf{z}\cdot\mathbf{w}\rangle &= \sum\nolimits_k \alpha_k y_k (K(\mathbf{x}+\mathrm{d}\mathbf{x},\mathbf{x}_k) - K(\mathbf{x},\mathbf{x}_k))\\ &= \sum\nolimits_k \alpha_k y_k \sum_i \left.\frac{\partial K(\mathbf{u},\mathbf{v})}{\partial u_i}\right|_{\substack{\mathbf{u}=\mathbf{x}\\ \mathbf{v}=\mathbf{x}_k}} \mathrm{d}x_i \qquad (21)\\ &= \nabla f_K(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad (22)\end{aligned}$$

$$\begin{aligned}\langle \mathrm{d}\mathbf{z}\cdot\mathrm{d}\mathbf{z}\rangle &= K(\mathbf{x}+\mathrm{d}\mathbf{x},\mathbf{x}+\mathrm{d}\mathbf{x}) - 2K(\mathbf{x}+\mathrm{d}\mathbf{x},\mathbf{x}) + K(\mathbf{x},\mathbf{x})\\ &= \sum\nolimits_{i,j} \left.\frac{\partial^2 K(\mathbf{u},\mathbf{v})}{\partial u_i \partial v_j}\right|_{\substack{\mathbf{u}=\mathbf{x}\\ \mathbf{v}=\mathbf{x}}} \mathrm{d}x_i\,\mathrm{d}x_j \qquad (23)\\ &= \mathrm{d}\mathbf{x}^{\mathrm{T}} H_K(\mathbf{x})\mathrm{d}\mathbf{x}, \qquad (24)\end{aligned}$$

where row-vector $\nabla f_K(\mathbf{x})$ is the gradient of the classifier function f_K evaluated at $\mathbf{x}$, and matrix $H_K(\mathbf{x})$ is one of the (equivalent) off-diagonal quarters of the Hessian of the kernel function K, evaluated at $(\mathbf{x},\mathbf{x})$. The discriminative direction is therefore defined by the following optimization problem:

$$\text{minimize}\ E(\mathrm{d}\mathbf{x}) = \mathrm{d}\mathbf{x}^{\mathrm{T}}\left(H_K(\mathbf{x}) - \|\mathbf{w}\|^{-2}\nabla f_k^{\mathrm{T}}(\mathbf{x})\nabla f_k(\mathbf{x})\right)\mathrm{d}\mathbf{x} \tag{25}$$

$$\text{s.t.}\ \|\mathrm{d}\mathbf{x}\|^2 = \epsilon. \tag{26}$$

The solution is the eigenvector of matrix

$$Q_K(\mathbf{x}) = H_K(\mathbf{x}) - \|\mathbf{w}\|^{-2}\nabla f_k^{\mathrm{T}}(\mathbf{x})\nabla f_k(\mathbf{x}) \tag{27}$$

that corresponds to the smallest eigenvalue. Note that, with the exception of the linear case, the matrix $Q_K(\mathbf{x})$ and its smallest eigenvector are not the same for different points in the original space, and need to be estimated for every input vector $\mathbf{x}$. Furthermore, each solution defines two opposite directions in the input space, corresponding to the positive and the negative projections onto $\mathbf{w}$. We want to deform the input example towards the opposite class and therefore assign the direction of increasing function value to the examples with label -1 and the direction of the decreasing function to the examples with label 1.

3.2.1. Special cases: analytical solution

Obtaining a closed-form solution of this minimization problem could be desired, or even necessary, if the dimensionality of the feature space is high and computing the smallest eigenvector is computationally expensive. Here we show that an analytical solution exists for a large family of kernel functions, including the linear kernel and the Gaussian RBF kernels we use in this work.

We observe that the second component of the right hand side of Eq. (27) is a matrix of rank one whose only non-zero eigenvalue is equal to

$$\|\mathbf{w}\|^{-2}\|\nabla f_k(\mathbf{x})\|^2 \tag{28}$$

with the corresponding eigenvector $\nabla f_k^{\mathrm{T}}(\mathbf{x})$. The rest of the eigenvectors span the null-space of the matrix $\nabla f_k^{\mathrm{T}}(\mathbf{x})\nabla f_k(\mathbf{x})$. Therefore, if the matrix $H_K(\mathbf{x})$ is a multiple of the identity matrix, $H_K(\mathbf{x}) = cI$, the smallest eigenvector of matrix $Q_K(\mathbf{x})$ is equal to the largest eigenvector of the matrix $\nabla f_k^{\mathrm{T}}(\mathbf{x})\nabla f_k(\mathbf{x})$, and the discriminative direction is equal to the gradient of the classifier function:

$$\mathrm{d}\mathbf{x}^* = \nabla f_k^{\mathrm{T}}(\mathbf{x}), \tag{29}$$

$$E(\mathrm{d}\mathbf{x}^*) = c - \|\mathbf{w}\|^{-2}\|\nabla f_k^{\mathrm{T}}(\mathbf{x})\|^2. \tag{30}$$

It is well known that to achieve the fastest change in the value of a function, one should move along its gradient, but in this case, the gradient also corresponds to the direction that distinguishes between the two classes while minimizing inter-class variability.

We can easily verify that both for the linear and, more surprisingly, for the Gaussian RBF kernels we use in this work, the matrix $H_K(\mathbf{x})$ is of the right form to yield the solution outlined above. For the linear kernel,

$$\left.\frac{\partial^2 K(\mathbf{u},\mathbf{v})}{\partial u_i \partial v_j}\right|_{\substack{\mathbf{u}=\mathbf{x}\\ \mathbf{v}=\mathbf{x}}} = \delta_{ij}, \tag{31}$$

and therefore $H_K(\mathbf{x}) = I$ for all $\mathbf{x}$. The discriminative direction is defined as

$$\mathrm{d}\mathbf{x}^* = \nabla f_k^{\mathrm{T}}(\mathbf{x}) = \mathbf{w} = \sum \alpha_k y_k \mathbf{x}_k, \tag{32}$$

$$E(\mathrm{d}\mathbf{x}^*) = 0. \tag{33}$$

This is not entirely surprising, as the classifier is a linear function in the original space and we can move precisely along $\mathbf{w}$. For the Gaussian RBF kernels,

$$\left.\frac{\partial^2 K(\mathbf{u},\mathbf{v})}{\partial u_i \partial v_j}\right|_{\substack{\mathbf{u}=\mathbf{x}\\ \mathbf{v}=\mathbf{x}}} = -\frac{2}{\gamma}\delta_{ij}, \tag{34}$$

which yields a closed form solution for the discriminative direction:

$$\begin{aligned}\mathrm{d}\mathbf{x}^* &= -\frac{2}{\gamma}\sum_k \alpha_k y_k \mathrm{e}^{-\frac{\|\mathbf{x}-\mathbf{x}_k\|^2}{\gamma}}(\mathbf{x}-\mathbf{x}_k)\\ &= \sum_k \alpha_k y_k K(\mathbf{x},\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k),\end{aligned} \tag{35}$$

$$E(\mathrm{d}\mathbf{x}^*) = \frac{2}{\gamma} - \|\mathbf{w}\|^{-2}\|\nabla f_k^{\mathrm{T}}(\mathbf{x})\|^2. \tag{36}$$

Unlike the linear case, we cannot achieve zero error, and the discriminative direction is only an approximation. The exact solution is unattainable in this case, as it has no corresponding direction in the original space.

To gain a better intuition about the solution presented above, let us consider the geometry of the problem. We perform the search in the original space by considering all points on an infinitesimally small sphere centered at the original input vector $\mathbf{x}$. In the range space of the mapping function Ψ_K, the images of points $\mathbf{x} + \mathrm{d}\mathbf{x}$ form an ellipsoid defined by the quadratic form

$$\mathrm{d}\mathbf{z}^{\mathrm{T}}\,\mathrm{d}\mathbf{z} = \mathrm{d}\mathbf{x}^{\mathrm{T}} H_K(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{37}$$

For $H_K(\mathbf{x}) \sim I$, the ellipsoid becomes a sphere, all $\mathrm{d}\mathbf{z}$'s are of the same length, and the minimum of error in the displacement vector $\mathrm{d}\mathbf{z}$ corresponds to the maximum of the projection of $\mathrm{d}\mathbf{z}$ onto $\mathbf{w}$. Therefore, the discriminative direction is parallel to the gradient of the classifier function. If $H_K(\mathbf{x})$ is of any other form, the length of the displacement vector $\mathrm{d}\mathbf{z}$ changes as we vary $\mathrm{d}\mathbf{x}$, and the minimum of the error in the displacement is not necessarily aligned with the direction that maximizes the projection.

Ideally, if we could access vectors in $\mathbb{F}$ directly, we would move the input vector $\Psi_K(\mathbf{x})$ along the projection vector $\mathbf{w}$ and study the change in $\mathbf{x}$ introduced by this process. However, we cannot explicitly manipulate elements of the space $\mathbb{F}$, but must explore it through search in the original space, effectively constraining the search to the manifold of the image vectors $\Psi_K(\mathbf{x})$. This corresponds to projecting $\mathbf{w}$ onto the manifold and moving along the projection vector. There are various ways to perform such projection, and in this work we chose to minimize the error between the approximation vector and the exact solution. We also note that different measures (e.g., the length of the projection of the resulting displacement vector onto $\mathbf{w}$) might be more appropriate for other applications.

3.3. Selecting inputs

Given any input example, we can compute the discriminative direction that represents the differences between the two classes captured by the classifier in the neighborhood of the example. But how should we choose the input examples for which to compute the discriminative direction? We argue that in order to study the differences between the classes, one has to examine the input vectors that are close to the separating boundary, namely, the support vectors. Note that this approach is significantly different from generative modeling, where a "typical" representative, often constructed by computing the mean of the training data, is used for analysis and visualization (e.g., to compare two different classes, one would compare their typical representatives (Csernansky et al., 1998; Machado and

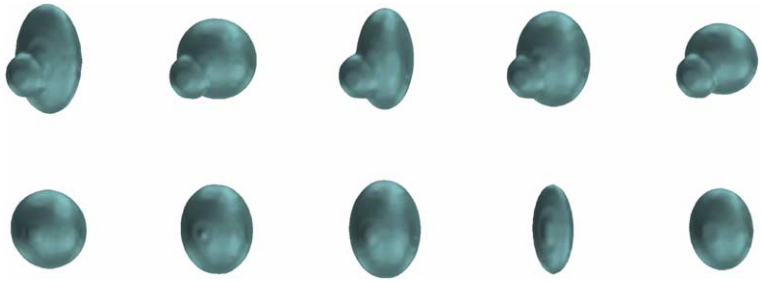

Fig. 3. Simple artificial study. Six example shapes from both classes are shown.

Gee, 1998)). In the discriminative framework, we are more interested in the examples that lie close to the opposite class, as they define the differences between the two classes and the optimal separating boundary.

Support vectors define a margin corridor whose shape is determined by the kernel type used for training. We can estimate the distance from any support vector to the separating boundary by examining the gradient of the classification function for that vector. A large gradient indicates that the support vector is close to the separating boundary and therefore can provide more information on the spatial structure of the boundary. This provides a natural heuristic for assigning importance weighting to different support vectors in the analysis of the discriminative direction.

To summarize, we presented an approach to quantifying the classifier's behavior with respect to small changes in the input vectors, trying to answer the following question: what changes would make the original input look more like an example from the other class without introducing irrelevant changes? We introduced the notion of the discriminative direction, which corresponds to the maximum changes in the classifier's response while minimizing irrelevant changes in the input. In our application, this can be used to interpret the differences between the two classes as deformations of the original input shapes, as explained in the next section.

4. System overview: a simple example

Before presenting the experimental results for the real medical studies, we explain how the components of the analysis described in the previous sections are combined into a system. We will illustrate the steps of the algorithm on a simple simulated shape study that contains 30 volumetric images of ellipsoidal shapes of varying sizes. The width, height and thickness of the shapes were sampled uniformly out of a ±10 voxel range centered at 20, 30 and 40 voxels, respectively. We randomly divided the data set into two classes of 10 and 20 examples, respectively, and added a spherical bump to the shapes in the first class. The bump location was sampled out of a ±3 voxel range centered on the side of the main ellipsoid. [3] Fig. 3 illustrates both types of shapes. Knowledge of the shape differences between the groups in this simulated experiment makes it relatively easy to assess the effectiveness of the analysis. Evaluation of the results is more difficult in the real medical studies, where the true shape differences, if they exist, are unknown.

4.1. Feature extraction

For every input scan, we compute the distance transform and use its moments to establish a new coordinate system in the volume, placing the origin at the center of mass and aligning the coordinate frame with the principal axes of inertia. The values of the distance transform are then sampled along the new axes at uniform intervals and are concatenated into a feature vector. This step creates the training set $\{(\mathbf{x}_k, y_k)\}_{k=1}^{l}$, where $\mathbf{x}_k$ are the feature vectors obtained by sampling the distance transforms, and $y_k \in \{-1,1\}$ are the corresponding labels defining the membership in one of the two populations.

4.2. Training a classifier

The training feature vectors and their labels are used by the Support Vector Machines learning algorithm to construct a classifier function $f(\mathbf{x})$ for labeling new examples. In each experiment, we trained a linear and a Gaussian RBF classifier. To determine the best setting of the kernel width γ for the non-linear classifier, we systematically explore the range of possible values, and for each setting of the parameter γ, train a classifier and perform leave-one-out cross-validation to estimate its generalization performance. We then select the classifier with the highest cross-validation accuracy.

[3] We employ an artificial example in this section for illustration purposes only, to demonstrate how the shape differences detected in the training phase are expressed and visualized using the concept of the discriminative direction. We chose such a simple shape to make the visualization easier for the readers to interpret before we proceed to the real medical studies.

The shapes in our simple example were easily separable using a linear classification function, but we also trained a non-linear classifier for comparison. Both classifiers separate the data perfectly. The cross-validation accuracy is 100% as well.

4.3. From the classifier to shape differences

Once the classifier has been constructed, we proceed to estimate the discriminative direction, which is equal to the gradient of the classification function for both the linear and the Gaussian kernels. We compute the discriminative direction $d\mathbf{x}^*$ at the support vectors identified in the training phase and use the magnitude of the classifier gradient to rank the vectors' importance. The next step is to compute the deformation of the support vectors $d\mathbf{s}$ that corresponds to the discriminative direction $d\mathbf{x}^*$ by projecting the vector $d\mathbf{x}^*$ onto the space of infinitesimal changes of the distance transform. Note that while the gradient of the classifier function has an analytical form, the corresponding deformation must be computed individually for every shape using Eq. (4).

Fig. 4 shows the estimated deformation $d\mathbf{s}$ for the six support vectors of the linear classifier. The color coding is used to indicate the direction and the magnitude of the deformation, changing from blue (inwards) to green (no deformation) to red (outwards). The same colormap will be used in all other figures depicting the discriminative direction. The shape differences defined by the deformation $d\mathbf{s}$ are localized to the area of the bump. The support vectors are redundant, representing almost identical deformations. We encountered similar redundancy in all our experiments, even in the non-linear case. We also note that the deformation of the boundary is fairly smooth, even though we did not augment the learning algorithm with any explicit information on the type of dependencies or spatial variation in the features. The resulting hypothesis of differences is consistent with the smooth nature of the shape representation we used in this work.

As discussed in Section 3.3, our approach of seeking an explicit interpretation of the classifier function in the original domain of the anatomical shape differs from the more commonly used technique of identifying the mean of the population as a representative shape and visualizing the differences between the two means, possibly adjusted for the covariance structure, as an illustration of the differences between the two populations. In the case of linear classification, the discriminative direction is the same everywhere in the space of input shapes, and our solution is essentially equivalent to the traditional approach, as the discriminative direction can be displayed on the surface of the representative example from each population, such as the mean shape. But once we accept the possibility that the separating boundary between the two populations can be an arbitrarily complex surface in the feature space, it behooves us to analyze the boundary and understand what it represents in terms of the original shapes. The resulting deformation might be dramatically different in different parts of the space, which could lead us to conclude that at least one of the populations is composed of several distinct groups, or to attempt to find a representation that yields a relatively flat separating surface between the populations. However, this information on the nature of shape differences cannot be conveyed by the traditional approach that reduces such differences to a single deformation between the population means.

The surface-based representation of the shape differences between the two classes is significantly easier to interpret than the volumetric results (not shown here). This advantage becomes even more apparent when we work with real anatomical shapes and use non-linear classifiers that yield a different volumetric discriminative direction for every support vector. In order to infer a top-level description of shape differences from the changes in the distance transform, one has to argue about shape differences in a way that effectively reduces the volumetric representation to a surface-based defor-

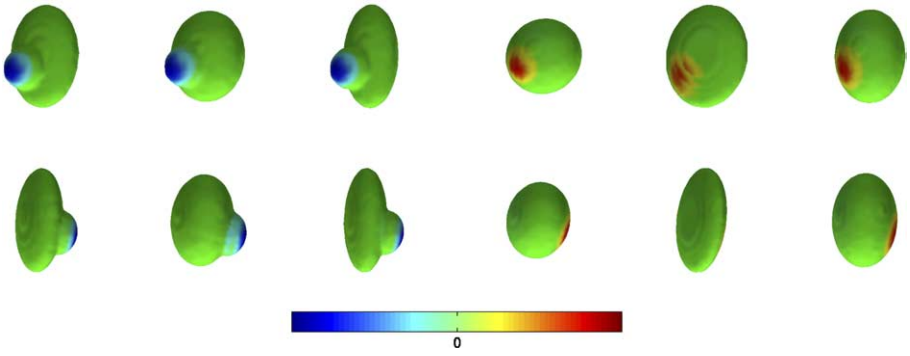

Fig. 4. Discriminative direction expressed as deformation $d\mathbf{s}$ for the three support vectors from each group in the simple shape study. Two views of each shape are shown. The color coding is used to indicate the direction and the magnitude of the deformation, changing from blue (inwards) to green (no deformation) to red (outwards).

Table 1
Training and cross-validation accuracy (with 95% confidence intervals) for the hippocampus in schizophrenia study

Accuracy (%)	Volume		Shape, linear		Shape, RBF	
	Right	Left	Right	Left	Right	Left
Training	60.0	63.3	100	100	100	100
Cross-validation	60.0 ± 17.5	63.3 ± 17.2	53.3 ± 17.8	56.7 ± 17.6	76.7 ± 15.1	70.0 ± 16.3

mation description. Our analysis presented in Section 2.2 effectively formalizes this line of argument.

4.4. Scaling

Volume and area measurements are extensively used in statistical studies of anatomical organs. In morphological studies, volume differences might be indicative of shape differences. However, using volume measurements by themselves has been questioned repeatedly in the medical research community because of wide variation in size among subjects. This problem is mitigated by normalizing the volume measurements with respect to some measure of the global scale (for example, the size of the brain can be used to normalize the hippocampus volume). Whether the scale is part of the object's shape is a controversy of its own. One could argue that scaling the object uniformly does not change its shape. In this work, we scale the shapes to the same volume. Such scaling can be easily incorporated into the feature extraction step. In the next section, unless specified otherwise, the training shapes are normalized with respect to their volume.

5. Experimental results

In this section, we consider studies of two different anatomical organs: the hippocampus–amygdala complex in schizophrenia and corpus callosum in the first episode affective disorder. The hippocampus study includes two separate studies, as we compare the right and the left hippocampus separately.

5.1. Hippocampus in schizophrenia

In this study, we compared the shape of hippocampus–amygdala complex using MRI scans of 15 schizophrenia patients and 15 matched controls. In each scan, the hippocampus–amygdala complex was manually segmented. Details on the subject selection and data acquisition can be found in (Shenton et al., 1992). The same paper reports statistically significant reduction in the volume of the left hippocampus, which suggests that shape differences might also be present in this study.

Table 1 contains the summary of performance estimates for this study. In order to present and compare the results of different experiments in a uniform fashion, we first trained a classifier based on the volume measurements of the structure. The standard t-test employed in the volumetric studies can be used only if the feature space is one-dimensional and is therefore not applicable to the case of multi-dimensional shape descriptors. Treating the one-dimensional volume descriptor similarly to the shape descriptors allows us to compare them directly. The confidence interval for the cross-validation results were estimated for 95% confidence level ($\eta = 0.05$) for all experiments in this work. Note that statistical significance does not necessarily mean perfect separation: the volume-based leave-one-out cross-validation accuracy for the left hippocampus is 63.3% (±17.2%). [4] By visually inspecting the shape of the hippocampus, we could see no distinct features that would guarantee 100% classification accuracy. This is common in the medical studies, where the global anatomical shape is similar in both groups, and the small deformations due to a particular disorder, if such exist, are of interest.

While the volume-based differences are more prominent in the left hippocampus in this study, the shape-based performance estimates for the left hippocampus are lower than those for its right counterpart. But since we scaled the structures to an identical volume, thus separating shape from size, we would not expect the volume-based results and the shape-based finding to be perfectly correlated. Here, we present the discriminative direction for the best RBF classifier both for the right and the left hippocampus, followed by the discussion on both structures, on the significance of the findings and the issues this experiment highlighted for the future extensions of the current analysis framework.

5.1.1. Shape differences

Fig. 5 shows three support vectors from the normal control group and the schizophrenia group. Four views (front, center-out, back, outside-in) are shown for each shape. These shapes were chosen from the list of the support vectors sorted in the descending order of the mag-

[4] Note that the cross-validation accuracy for volume-based descriptors is close to the training accuracy. Removing a single training example can only affect the training result if the example is close to the threshold, i.e., it is a support vector, and there could be only few such vectors in the low-dimensional space. However, it is not surprising that the training accuracy in shape-based experiments is 100%, as the space is so high-dimensional that we would expect the data to be easily separable.

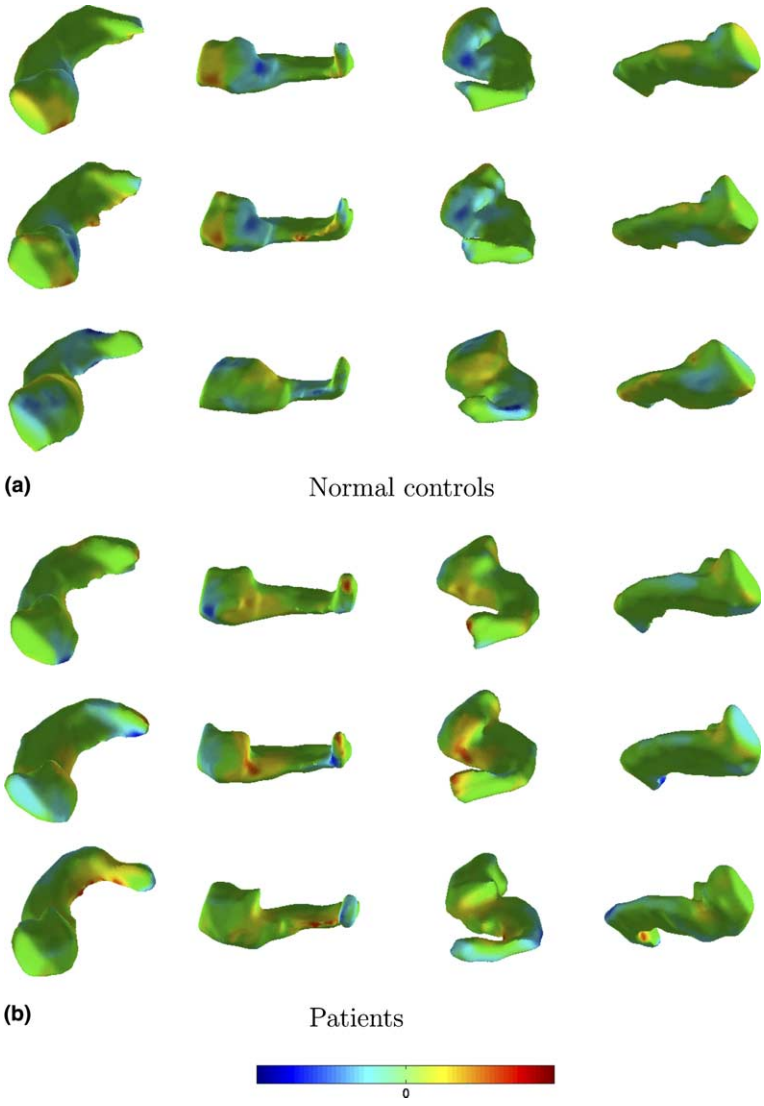

Fig. 5. Discriminative direction for the right hippocampus shown as deformations of three support vectors from the patient group. Four views of each shape are shown (one row per subject). (a) Normal controls; (b) patients. The color coding is used to indicate the direction and the magnitude of the deformation, changing from blue (inwards) to green (no deformation) to red (outwards).

nitude of the classifier gradient. Similarly to the artificial example in the previous section, the algorithm produces several support vectors for the same type of deformation. We omit support vectors with very similar deformations to the ones shown in the figures. As a result, the shapes displayed in Fig. 5 are support vectors 1, 3 and 6 from the normal control group and support vectors 1, 3 and 5 from the schizophrenia group.

We can see that the deformations identified by the analysis are smooth and localized. Furthermore, the protrusions are separated from indentations by areas where no deformation is required. [5] We also note that the support vectors from different classes define deformations of very similar nature, but of opposite signs. We believe that such pairs of support vectors "oppose each other" across the separating boundary, but a more precise definition and analysis of this notion has to be developed before we can characterize it quantitatively.

A significant amount of deformation is localized in the anterior region of the structure, which indicates that the bulbous "head" of the amygdala is curved-in, or tucked-in, relative to the main body in normal controls more than in schizophrenia patients. This deformation is prominent in the first two support vectors from each group. In normal controls, there is a significant deformation inward in the inner part of the amygdala and a corresponding expansion on the outside part of it. The second interesting shape difference is located in the posterior part of the hippocampus and is represented by the

[5] A series of small changes of opposite sign close to each other would raise a concern that the structure captured by the classifier does not correspond to the smooth way the anatomical shapes deform and is induced by noise and errors in the boundary.

third support vector in the figures. It seems that the "tail" is thinner and possibly shorter in schizophrenics in the region of the shape space close to this support vector.

Figs. 6 and 7 show the discriminative direction for the left hippocampus as a deformation of the top support vectors from the normal control group and the schizophrenia group, respectively. The first two support vectors in each group indicate that the posterior "tail" of the structure is folded-in, or curved, in normal controls more than in schizophrenics. In addition, the last three support vectors contain a deformation in the anterior part of the structure. The support vectors in the normal control group contain a slight deformation inward and a protrusion of a higher magnitude in the anterior part. This deformation is of a similar nature for the three support vectors, but it is localized in different parts of the bulbous head. Besides the obvious explanation that the location of this deformation is not fixed in the population, this could also correspond to a general enlargement of the anterior part relative to the whole structure in schizophrenics. Slight misalignments of the structures in the feature extraction step can cause such size differences to be detected in different areas of the affected surface. Since statistically significant volume reduction was detected in the left hippocampus, this could mean that the posterior part of the structure is affected by the volume reduction in a more significant way than the anterior part.

5.1.2. Notes

This section demonstrates our technique on the real medical data. We obtained a detailed description of the shape differences between the schizophrenia group and the normal control group expressed as deformations of the example shapes in the two groups. While the medical significance of these findings has to be further investigated, such visualizations can clearly facilitate exploration of the shape differences in organs affected by the disease.

Note that the shape differences found in the two hippocampi are asymmetric. We considered the two structures independently of each other, but studying the relationship between the two sides of the brain could help us understand the effects of a disorder better. An even higher cross-validation accuracy (87%) was recently reported on the same data set in (Gerig et al., 2001) based on the average distance between the aligned surfaces of the left hippocampus and a mirror image of the right hippocampus as an asymmetry measure. Unfortunately, such global asymmetry measures are too specific to be generally applicable, and furthermore,

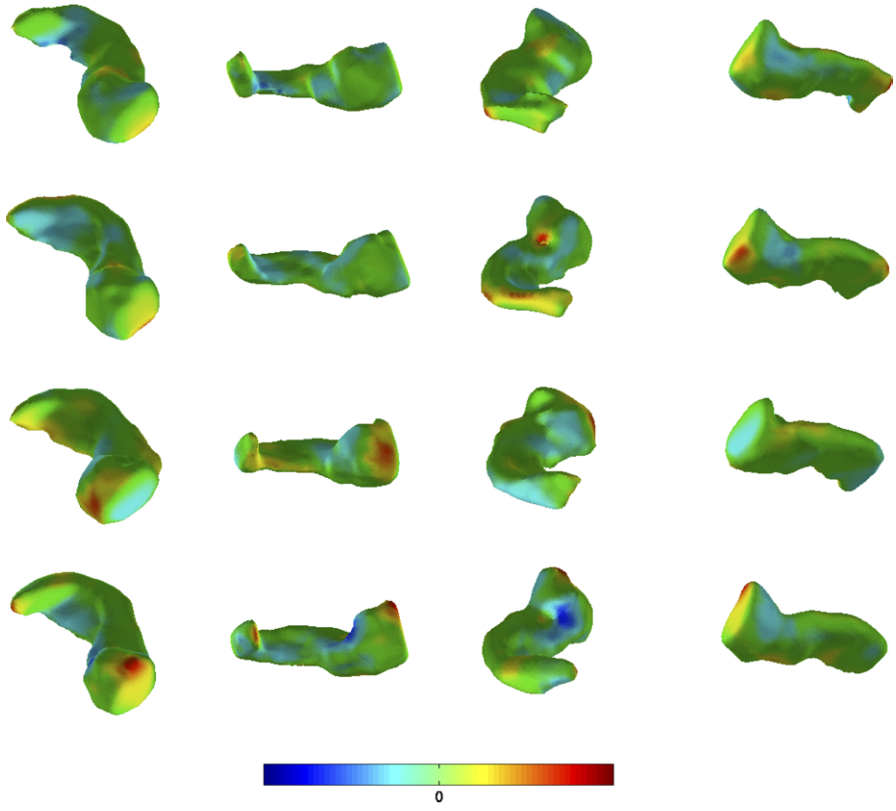

Fig. 6. Discriminative direction for the left hippocampus shown as deformations of four support vectors from the normal control group. Four views of each shape are shown (one row per subject). The color coding is used to indicate the direction and the magnitude of the deformation, changing from blue (inwards) to green (no deformation) to red (outwards).

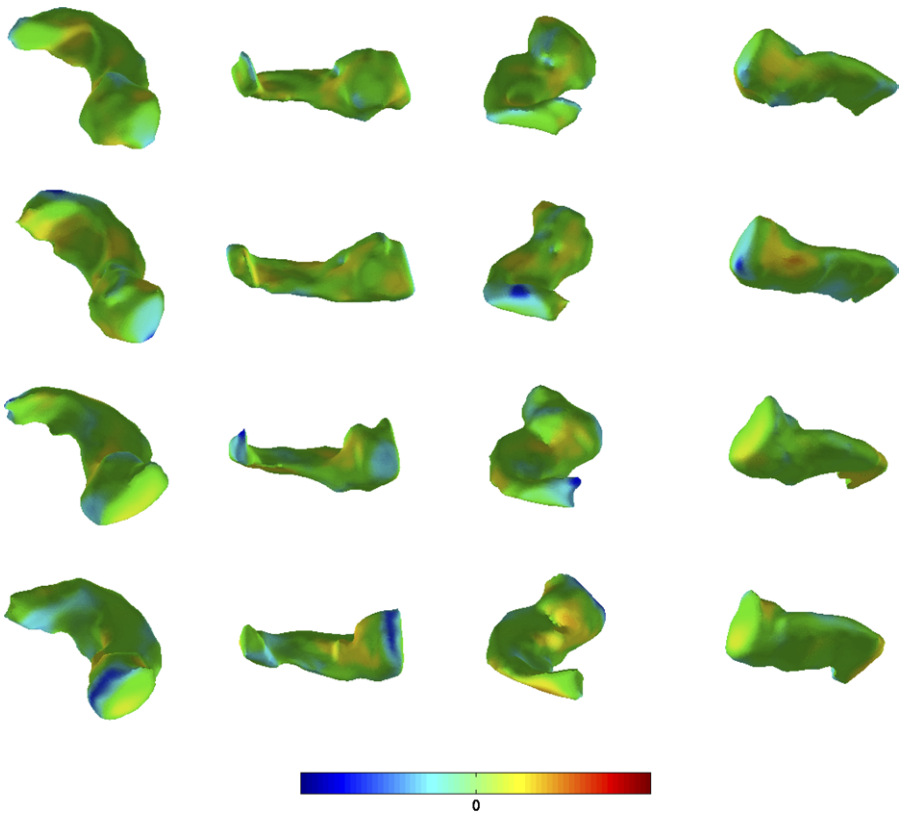

Fig. 7. Discriminative direction for the left hippocampus shown as deformations of four support vectors from the patient group. Four views of each shape are shown (one row per subject). The color coding is used to indicate the direction and the magnitude of the deformation, changing from blue (inwards) to green (no deformation) to red (outwards).

they do not provide information on the details of the shape differences. In this work, we concentrated on a single structure at a time and provided the analysis and the detailed interpretation of shape differences based on general descriptors. The separation between the classes could be improved by combining information from different structures, up to building a description of the differences for the entire sub-cortical region.

The results reported in this section indicate that the training data set is too small to provide accurate estimates of the generalization performance of the classifier. Wide confidence intervals indicate that more data is needed to guarantee that the resulting classifier and the detected shape variation reflect the true differences in the population. Unfortunately, the ground truth, i.e., the true differences between the classes, or even whether such differences exist, is not known for the medical studies we are working with. Anatomical shape analysis is a relatively new field, and not much is known about the deformations caused by the disorders of interest. We therefore believe that developing principled algorithms for investigating morphology of the organs, along with thorough data collection and analysis, can have a significant impact on the field.

In the next section, we demonstrate our method on the study of the corpus callosum shape in first episode affective disorder patients. Our findings in the study indicate differences between the groups, with a similar concern about the size of the training data set as for the hippocampus study.

5.2. Corpus callosum in affective disorder

Corpus callosum is a bundle of white matter fibers connecting the two hemispheres of the brain. The two-dimensional cross-section of the bundle is actually what is studied in the medical research. To ensure consistency, all the scans in the study have been aligned manually by trained physicians so that the cross-section is indeed perpendicular to the bundle. For every scan in the study, the physician drew three lines on the coronal view that, based on pre-specified anatomical landmarks, defined the best mid-sagittal slice, followed by the drawing of three lines defining the mid-sagittal slice on the axial view of the same scan. The scan was then rotated and resampled using the least-squares fit to the specified lines as a mid-sagittal slice and the AC-PC line as the normal to the coronal direction. Further details on the data collection and scan alignment can be found in (Frumin et al., 2002). In this study, we compared 18 affective disorder patients with 20 normal controls. Table 2 summarizes the performance estimates for the linear and the

Gaussian RBF classification on this data set. The gap between performance estimates for the linear and the non-linear classification is not as wide as in the hippocampus study, but it is still substantial. The 95% confidence interval for the linear classification touches the 50% baseline, while the confidence interval for the best RBF classifier is well above it.

Fig. 8 shows the detected shape differences as deformations of the first six support vectors from each group. Since the shape of interest in this study is two-dimensional, one view is sufficient to visualize the shape. The colormap used in Fig. 8 is identical to the previous figures, but the color now indicates motion along the normal to the 2D outline in the mid-sagittal plane. Thus blue corresponds to a local "contraction" in the white matter and red corresponds to the local "swelling". Similarly to the hippocampus study, there is a lot of redundancy in the deformation represented by the support vectors. The most prominent difference captured by the classifier is the deformation in the anterior part of the structure (the left end of corpus callosum in the images). We observe a significant amount of horizontal expansion and vertical contraction in the anterior part in the patient group and the deformation of the opposite sign in the normal control group. The amount of deformation varies across the shapes, but it is present in almost all of them. This deformation corresponds to the anterior of the corpus callosum being "squashed" horizontally and elongated vertically in the affective disorder patients compared to the normal control group.

Vectors #2, #3 in the normal control group and vectors #5, #6 in the patient group indicate that for some of the cases, the width of the middle part of the corpus callosum is wider in the affective disorder patient: there

Table 2
Training and cross-validation accuracy (with 95% confidence intervals) for the corpus callosum in the affective disorder study

Accuracy (%)	Linear	RBF
Train	100	100
Cross-validate	65.8 ± 15.0	73.7 ± 13.9

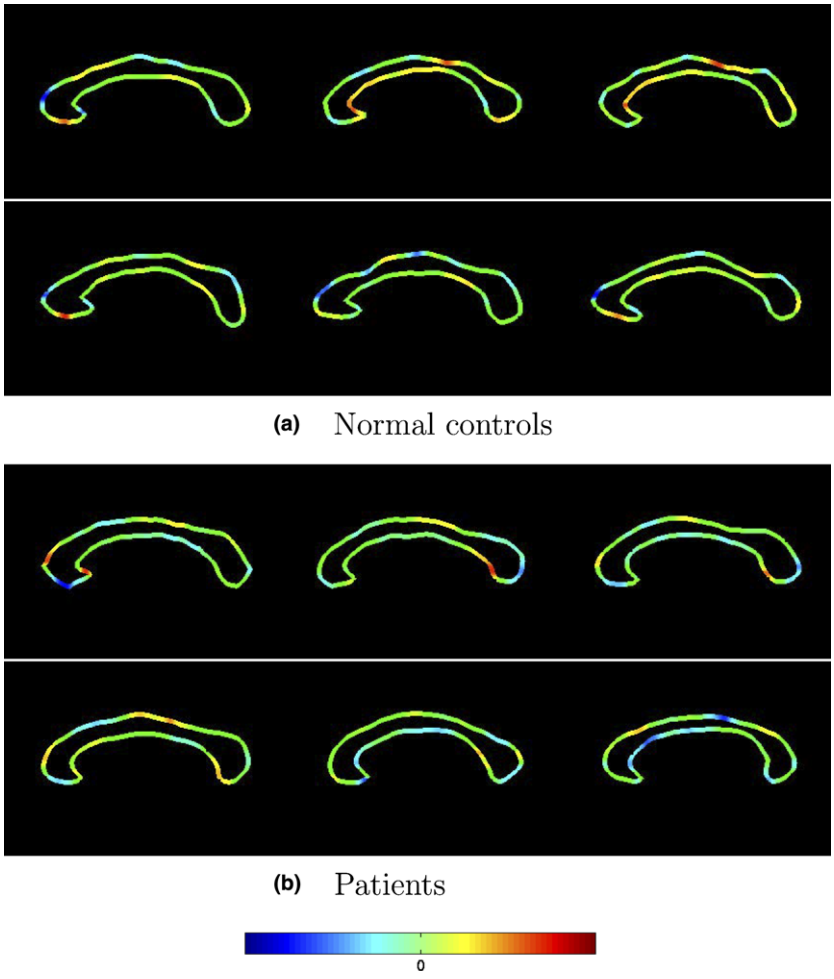

Fig. 8. Discriminative direction for corpus callosum in the affective disorder study shown as deformations of six support vectors from each group. (a) Normal controls; (b) patients. The color coding is used to indicate the direction and the magnitude of the deformation, changing from blue (inwards) to green (no deformation) to red (outwards).

is a consistent contraction associated with the patient examples and expansion associated with the normal controls.

Vectors #2 and #3 in the patient group indicate some amount of deformation in the posterior part of the structure. Interestingly, this deformation is not represented at all in the normal control group. We noted earlier that many support vectors have a matching counterpart from the other class that represents a deformation of opposite sign, but of very similar nature. This is an example when there seems to be a "gap" in the boundary support on the normal control side. This is an interesting phenomenon that needs further investigation. Potentially, one might be able to construct "virtual" support vectors by artificially reflecting the existing support vectors across the boundary.

To summarize, there is a consistent deformation of the anterior part of the structure that corresponds to horizontal narrowing and vertical extension of the anterior part of the corpus callosum in the affective disorder patients. In addition to that, the middle part is widened in some of the patients compared to normal controls. Similarly to the hippocampus study, more data will have to be collected for validation of these results.

6. Discussion

In this section, we reflect on our experience with the technique, unexpected problems that arose in the experiments and the insights they provided into the nature of the statistical shape analysis.

6.1. Morphological differences as deformations

The visualized shape differences contain detailed morphological information which can be correlated with functional information on the organ of interest, hopefully leading to a better understanding of the disease and its development. The representation of morphological differences as deformations of the original input shapes is significantly more detailed, localized and informative than the global measurements traditionally used in the medical research, such as volume changes. Note that the deformation maps can also be viewed as maps of local volume changes. Deformation of the object boundary causes local changes in the volume, and the total change of the object volume is an integral of the local deformation. Thus shape analysis represents a natural extension of the commonly used volumetric studies.

6.2. Detection and significance

The experiments highlighted the importance of statistical testing of the resulting hypothesis. While the training algorithm will produce a classifier which can be analyzed for discriminative direction for any two sets of examples, the shape differences found in the training set are useful for understanding the phenomenon in question only if they accurately reflect the morphological differences in the entire population. Therefore, accurately estimating the expected performance of the resulting classifier and the statistical significance of the morphological variability it represents is a crucial component of the analysis which requires further development. The cross-validation procedure provides an unbiased estimate of the expected accuracy of the resulting classifier function, but for a small number of examples, the variance of this estimator might lead to a substantial difference in the estimated accuracy and the true one. Furthermore, since we used the cross-validation to select the kernel parameters, a more principled approach would be to test the resulting settings on a separate large hold-out set. As the number of available examples in out studies increases, such thorough testing will allow a more accurate estimation of the expected accuracy. We are currently investigating an open question of how to convert the testing accuracy estimates to the statistical significance indicators that are traditionally used in the clinical studies. The problem arises since we are using a non-parametric approach of fitting the classifier function to the training data without assuming a strong model that would allow defining the corresponding probability distribution under the null hypothesis.

6.3. Shape representation

We also note that the training algorithm and the discriminative direction can be used as a very effective tool for investigating the power of different shape descriptors for representing morphological variability. Our approach to classification and further interpretation of the results allows us to compare the descriptors empirically based on their performance in shape-based statistical tests. More often than not, the shape analysis methodology is presented in the literature as a monolithic structure where the shape description and the statistical analysis are inseparable. In reality, these two components can be improved independently of each other. For example, we are currently applying the same statistical analysis methodology to a different shape descriptor on the same data sets (Timoner et al., 2002; Timoner, 2003). In that work, we use the deformation field produced as part of matching a template to each of the input shapes as a shape descriptor. And while the detected differences are qualitatively very similar to the ones demonstrated in this study, the cross-validation accuracy on the same data sets is improved substantially by changing the representation. Moreover, we tested different alignment techniques, from moments of the shape

to elastic registration (Timoner et al., 2002). Our empirical experience suggests that the alignment procedure has a smaller effect on the statistical indicators (≈2−3% improvement in the cross-validation accuracy) than the choice of the representation (≈15% improvement in the cross-validation accuracy). We believe that shape representation does and should depend on the organ of interest (we would not expect the same descriptor work equally well for the hippocampus and for the cortical folds), while the statistical analysis can be easily adapted to work with a large family of descriptors. It can therefore be used as a test-bed for various shape representations.

6.4. Studies with no shape differences found

While our method identified significant differences between the groups in the experiments reported in the previous sections, it failed to do so in several other studies. However, the experiments that failed to produce statistically significant differences between the two classes can still help us to better understand the problem and potentially improve the technique. In such experiments, different statistical indicators, such as the cross-validation accuracy, are very sensitive to the settings of the kernel parameters, often predicting close to 50% baseline classification accuracy on new examples. Alternatively, the predicted performance could be higher than the baseline, but the confidence intervals might be too wide to allow us to make any claims on how well the differences detected in the training set represent the situation in the whole population. We have observed some of these warning signs in a study of corpus callosum in the first episode schizophrenia that we performed in parallel with the reported corpus callosum study in affective disorder (Frumin et al., 2002). We compared the shape of the corpus callosum between 16 patients and 20 normal controls, with the best cross-validation accuracy of $69.4 \pm 15.0\%$. Moreover, the cross-validation results were significantly more sensitive to the settings of the kernel parameters than in the studies reported above. And while we can visualize the differences between the classes represented by the classifier, our concern is that such low indicators of the generalized performance cast doubts in the robustness of the detected differences.

Such experiments bring up an interesting general question of when one should stop the search for shape differences. Any morphological study starts with a hypothesis of shape differences which is to be confirmed by the empirical evidence from the collected images. We could get statistically unsatisfactory results described above either because our representation and analysis cannot capture the shape differences present in the population, or because such differences do not exist. In the former case, we can improve the technique for shape representation and statistical analysis to include more complex models of morphology and its variability and collect more data to reduce the confidence intervals. But the fundamental question remains, when should we abandon the search for better analysis techniques and more training data and declare that there are no differences between the two populations? This problem is common in many fields of research, as the current theoretical framework provides us only with tools for establishing the fact of existence of a particular phenomenon. It is nearly impossible, at least with our current system of reasoning, to prove the absence of the hypothesized effect.

7. Conclusions

The focus of this paper is the interpretation of the classifier function constructed to distinguish between two populations. We present a novel technique for classifier analysis in terms of the input features in the general context of the statistical learning theory and instantiate the technique for shape analysis by establishing a locally linear parameterization of the distance transform space using deformations of the corresponding boundary surface. Such parameterization yields a representation of the shape differences captured by the classifier as deformations of the input shapes relative to the examples from the opposite class. We demonstrate the method on both artificial examples that illustrate the approach and the real medical studies. Experimental studies suggested several directions of future work from refining the analysis technique for interpretation of shape differences to collecting more data for strengthening the statistical confidence indicators. The analysis generates a detailed description of shape changes due to diseases and can facilitate the studies of the disorders through understanding of the induced anatomical changes.

Acknowledgements

Quadratic optimization was performed using PR_LOQO optimizer written by Alex Smola.

This research was supported in part by NSF IIS 9610249 grant. The authors would like to acknowledge Dr. Shenton's Grants NIMH K02, MH 01110 and R01 MH 50747 Grants, Dr. Kikinis's Grants NIH PO1 CA67165, R01RR11747, P41RR13218 and NSF ERC 9731748 Grants.

The authors thank the anonymous reviewers, whose careful proofreading and thoughtful comments and suggestions helped us to substantially improve the clarity of presentation throughout the paper.

References

Blum, H., 1967. A transformation for extracting new descriptors of shape. In: Models of the Perception of Speech and Visual Form. MIT Press, Cambridge, MA.

Bookstein, F.L., 1979. The line skeleton. CGIP: Computer Graphics and Image Processing 11, 123–137.

Bookstein, F.L., 1997. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. Medical Image Analysis 1 (3), 225–243.

Brechbühler, C., Gerig, G., Kübler, O., 1995. Parameterization of closed surfaces for 3-D shape description. CVGIP: Image Understanding 61, 154–170.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Christensen, G., Rabbitt, R.D., Miller, M.I., 1993. A deformable neuroanatomy textbook based on viscous fluid mechanics. In: Prince, J., Runolfsson, T. (Eds.), Proceedings of the 27th Conference on Information Sciences and Systems, pp. 211–216.

Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J., 1992. Training models of shape from sets of examples. In: Proceedings of the British Machine Vision Conference. Springer, Berlin, pp. 9–18.

Cootes, T.F., Beeston, C., Edwards, G.J., Taylor, C.J., 1999. A unified framework for atlas matching using active appearance models. In: Proc. IPMI'99LNCS, 1613, pp. 322–333.

Csernansky, J.G., Joshi, S., Wang, L., Haller, J.M., Gado, M., Miller, J.P., Grenander, U., Miller, M.I., 1998. Hippocampal morphometry in schizophrenia by high dimensional brain mapping. Proceedings of the National Academy of Science 95 (19), 11406–11411.

Davatzikos, C., Vaillant, M., Resnick, S., Prince, J.L., Letovsky, S., Bryan, R.N., 1996. A computerized method for morphological analysis of the corpus callosum. Journal of Computer Assisted Tomography 20, 88–97.

Dryden, I.L., Mardia, K.V., 1998. Statistical Shape Analysis. Wiley, New York.

Efron, B., 1982. The Jacknife, The Bootstrap, and Other Resampling Plans. SIAM, Philadelphia, PA.

Fritsch, D.S., Pizer, S.M., Morse, B., Eberly, D.H., Liu, A., 1994. The multiscale medial axis and its applications in image registration. Patter Recognition Letters 15, 445–452.

Frumin, M., Golland, P., Kikinis, R., Hirayasu, Y., Salisbury, D.F., Hennen, J., Dickey, C.C., Anderson, M., Fischer, I.A., Yurgelun-Todd, D., Jolesz, F.A., Grimson, W.E.L., McCarley, R.W., Shenton, M.E., 2002. Shape differences in the corpus callosum in first psychotic episode schizophrenia and first psychotic episode affective disorder. American Journal of Psychiatry 159, 866–868.

Gerig, G., Styner, M., Shenton, M.E., Lieberman, J.A., 2001. Shape versus size: improved understanding of the morphology of brain structures. In: Proc. MICCAI'2001LNCS, 2208, pp. 24–32.

Golland, P., Grimson, W.E.L., 2000. Fixed topology skeletons. In: Proc. CVPR'2000, pp. 10–17.

Golland, P., Grimson, W.E.L., Kikinis, R., 1999. Statistical shape analysis using fixed topology skeletons: corpus callosum study. In: Proc. IPMI'99LNCS, 1613, pp. 382–387.

Golland, P., Grimson, W.E.L., Shenton, M.E., Kikinis, R., 2000. Small sample size learning for shape analysis of anatomical structures. In: Proc. MICCAI'2000LNCS, 1935, pp. 72–82.

Golland, P., Grimson, W.E.L., Shenton, M.E., Kikinis, R., 2001. Deformation analysis for shaped based classification. In: Proc. IPMI'01LNCS, 2082, pp. 517–530.

Kelemen, A., Székely, G., Gerig, G., 1998. Three-dimensional model-based segmentation. In: Proceedings of the IEEE International Workshop on Model Based 3D Image Analysis, pp. 87–96.

Kimmel, R., Shaked, D., Kiryati, N., Bruckstein, A.M., 1995. Skeletonization via distance maps and level sets. CVIU: Computer Vision and Image Understanding 62 (3), 382–391.

Leventon, M.E., Grimson, W.E.L., Faugeras, O., 2000. Statistical shape influence in geodesic active contours. In: Proc. CVPR'2000, pp. 316–323.

Leymarie, F., Levine, M.D., 1992. Simulating the grassfire transform using an active contour model. IEEE Transactions PAMI 14 (1), 56–75.

Machado, A.M.C., Gee, J.C., 1998. Atlas warping for brain morphometry. In: Proc. SPIE Medical Imaging 1998: Image Processing, SPIE, 3338, pp. 642–651.

Martin, J., Pentland, A., Kikinis, R., 1994. Shape analysis of brain structures using physical and experimental models. In: Proc. CVPR'94, pp. 752–755.

Pizer, S.M., Fritsch, D.S., Yushkevich, P., Johnson, V., Chaney, E., 1996. Segmentation, registration, and measurement of shape variation via image object shape. IEEE Transactions on Medical Imaging 18 (10), 851–865.

Shenton, M.E., Kikinis, R., Jolesz, F.A., Pollak, S.D., Lemay, M., Wible, C.G., Hokama, H., Martin, J., Metcalf, D., Coleman, M., McCarley, R.W., 1992. Abnormalities in the left temporal lobe and thought disorder in schizophrenia: a quantitative magnetic resonance imaging study. New England Journal of Medicine 327, 604–612.

Shenton, M.E., Gerig, G., McCarley, R.W., Szekely, G., Kikinis, R., 2002. Amygdala–hippocampus shape differences in schizophrenia: the application of 3D shape models to volumetric MR data. Psychiatry Research Neuroimaging 115, 15–35.

Staib, L., Duncan, J., 1992. Boundary finding with parametrically deformable models. IEEE PAMI 14 (11), 1061–1075.

Székely, G., Kelemen, A., Brechbüler, C., Gerig, G., 1996. 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.

Timoner, S.J., Golland, P., Kikinis, R., Shenton, M.E., Grimson, W.E.L., Wells III, W.M., 2002. Performance issues in shape classification. In: Proc. MICCAI 2002LNCS, 2488, pp. 508–515.

Timoner, S.J., 2003. Compact representations for fast nonrigid registration of medical images. Ph.D. Thesis, MIT.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer, Berlin.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley, New York.

Yushkevich, P., Pizer, S.M., Joshi, S.C., Marron, J.S., 2001. Intuitive, localized analysis of shape variability. In: Proc. IPMI'2001LNCS, 2082, pp. 402–408.