

# Regularized Discriminative Direction for Anatomical Shape Difference Analysis

Luping Zhou, Richard Hartley, Lei Wang, Paulette Lieby, Nick Barnes

### Abstract

Identifying the shape difference between two groups of anatomical objects is important for medical image analysis and computer-aided diagnosis. A method called “discriminative direction” in the literature has been proposed to solve this problem. In that method, the shape difference between groups is identified by deforming a shape along the discriminative direction. This paper conducts a thorough study about inferring this discriminative direction in an efficient and accurate way. First, finding the discriminative direction is reformulated as a pre-image problem in kernel-based learning. This provides a complementary but conceptually simpler solution than the previous method. More importantly, we find that a shape deforming along the original discriminative direction cannot faithfully maintain its anatomical correctness. This unnecessarily introduces spurious shape differences and leads to inaccurate analysis. To overcome this problem, this paper further proposes a *regularized* discriminative direction by requiring a shape to conform to its underlying distribution when it deforms. Two different approaches are developed to impose the regularization, one from the perspective of probability distributions and the other from a geometric point of view, and their relationship is discussed. After verifying their superior performance through controlled experiments, we apply the proposed methods to detecting and localizing the hippocampal shape difference between sexes. We get results consistent with other independent research, providing a more compact representation of the shape difference compared with the established discriminative direction method.

### Keywords

Statistical shape analysis, Discriminative direction, Pre-image problem, Shape distribution, Hippocampal shapes.

## I. Introduction

Identifying the morphological differences between anatomical shapes related to disorders or aspects of normal anatomy such as ageing and sex is an important area of medical image analysis. Once identified, this difference may be used to facilitate the early stage diagnosis of diseases. Detecting the difference is typically formulated as a classification problem which aims optimally to separate two groups of anatomical shapes from each other. Groups of anatomical shapes are seldom easy to differentiate even for a medical expert, for example difference in hippocampal shape between sexes [1]. In recent years, thanks to their capacity for differentiating linearly nonseparable classes, kernel classifiers such as Support Vector Machines (SVMs) [2] and Kernel Fisher Discriminant Analysis (KFDA)[3] have been widely used. They have also been employed to discriminate anatomical shapes [4], [5]. This works as follows. Via a kernel function, the training shapes from two groups are mapped nonlinearly from a shape descriptor space  $\mathbb{R}^d$  to a higher dimensional space  $\mathcal{F}$  (known as the *feature space*), where they will most likely become linearly separable. In this way, a linear classifier  $f(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}) + b$  can be sought in  $\mathcal{F}$ , where  $\Phi(\cdot)$  denotes the mapping function. Its normal  $\mathbf{w}$  indicates the direction that best discriminates the two groups. Golland et al. in [6], [7], [4] put forward that the shape difference between groups can be manifested by deforming a shape along  $\mathbf{w}$ . Considering that the shape difference identified in  $\mathcal{F}$  is only *mathematically* meaningful, they project such difference back to  $\mathbb{R}^d$  and explain it in an *anatomically* meaningful way. For this purpose, Golland et al. defined the “*discriminative direction*” for each point (here denoting a shape) in  $\mathbb{R}^d$ . It is a direction such that when a shape moves along it, the image of this shape in  $\mathcal{F}$  will move in a way that best agrees with the normal  $\mathbf{w}$ . According to [6], [7], [4], travelling along this “direction” will form a curve in the shape descriptor space  $\mathbb{R}^d$ , having the property that when a shape deforms along its course, only the differences related to group discrimination appear and the within-group variability is excluded. Therefore the shape difference between groups can be captured by observing how a particular shape in one group gradually changes to look like a shape in another group. The “discriminative direction” method has been used to localize the distinctive difference in hippocampal shapes between normal controls and the schizophrenia patients [4].

This paper conducts a thorough study on inferring the optimal direction along which the shape difference between groups can best be manifested. It has two main contributions. Firstly, we revisit the original discriminative direction and solve it in a better way along lines suggested by [4]. As pointed out in [4], it is better to move a shape’s image in  $\mathcal{F}$  along the normal  $\mathbf{w}$  and observe the change of this shape in  $\mathbb{R}^d$ . Clearly, this way has the advantage of conceptual simplicity and agrees

well with intuition. We find that it can be conveniently realized by formulating the estimation of the discriminative direction as a pre-image problem that has recently arisen in kernel-based learning methods. Its connection to and difference from Golland’s method are discussed. Experimental study shows that the two methods produce comparable results in shape difference analysis. Secondly, and this is the more significant contribution of this paper, we propose a new direction which is more accurate than the original discriminative direction in revealing the shape difference between groups. Following the terminology in [6], [7], [4], we name the new direction the “regularized discriminative direction”.

Our work is driven by our observation that when deforming a shape along the original discriminative direction proposed in [4], spurious shape difference may appear. In this paper, we analyze the problem and discover the possible causes: the original “discriminative direction” ignores the fact that the high dimensional shape descriptors often reside on a lower dimensional manifold. Deforming along this direction will deviate from the manifold and the underlying shape distribution. This will generate shapes which do not really exist for a given anatomical object. This drawback has to be removed before the discriminative direction approach can be applied to any practical study. We overcome this drawback by making the newly deformed shapes comply with the underlying distribution of the shape population. For this purpose, we develop two approaches that impose regularization from two different perspectives, that of probability distribution and that of geometry. The relationship between the two approaches and their advantages are discussed respectively. Moreover, for one of the approaches, an analytical solution to the regularized discriminative direction is derived, which avoids iterative optimization. Our experimental study demonstrates the advantages of the regularized discriminative direction over the original one by carrying out controlled experiments. After this, the proposed regularized discriminative direction is employed to detect and localize the group difference of the hippocampal shapes between sexes.

It is worth noting that the direction proposed in this paper and that proposed in [4] have a significance not limited to shape difference analysis. Essentially, inferring these directions is an inverse problem for the kernel mapping. In the field of machine learning, such a problem has been studied for a *generative* model in the context of Kernel Principal Component Analysis (KPCA) with the applications of image denoising and compression [3], [8], [9]. However the inverse kernel mapping problem for *discriminative* models such as classifiers has not been given enough attention. This may be due to the fact that people often care more about obtaining a good classification performance than characterizing the nature of the difference between classes. The latter however has significance in medical image analysis. The decoding of the shape difference in  $\mathbb{R}^d$  not only helps in understanding the morphological difference between anatomical organs, which may have an association with the organ functions [10], [11], [12], but also helps to improve the classification performance by selecting the discriminative features.

An earlier and briefer description of these results has appeared in our previous paper [13].

## II. Related work

We first give a brief introduction to the established “discriminative direction” method, which is the foundation of this paper. After that, two approaches to solving the preimage problem for KPCA are reviewed. The basic ideas behind these two approaches inspire our reformulation of the “discriminative direction” problem described in Section III. There they are extended to solve a different preimage problem for kernel classifiers such as KFDA or SVMs.

### A. Discriminative direction

Let  $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  ( $\mathbf{x}_i \in \mathbb{R}^d$ ) denote a set of  $n$  training shapes labelled in two groups. Training a kernel classifier implicitly performs a mapping  $\Phi(\cdot)$  from an input space (for example, a shape descriptor space)  $\mathbb{R}^d$  to a high dimensional Hilbert space (known as the feature space)  $\mathcal{F}$ . By training an SVM classifier or conducting KFDA, an optimal separating hyperplane with unit normal

$\mathbf{w}$  and bias  $b$  can be obtained in  $\mathcal{F}$  as

$$f(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}) + b, \quad (1)$$

where  $\mathbf{w}$  and  $b$  are estimated from the training data in  $\mathcal{D}$ . Vector  $\mathbf{w}$  indicates the direction that best discriminates the two groups. Given an unseen data point  $\mathbf{x}$ , ideally  $\Phi(\mathbf{x})$  should move along direction  $\mathbf{w}$  strictly to reflect only the shape difference between two groups. However there is a problem. Given a kernel function, the associated mapping  $\Phi(\cdot)$  often maps the training data onto a lower dimensional manifold in  $\mathcal{F}$ <sup>1</sup>. For example, when a Gaussian Radial Basis Function (RBF) kernel is used, all the training data (indeed the whole  $\mathbb{R}^d$ ) will be mapped onto a unit hypersphere in  $\mathcal{F}$ . The normal  $\mathbf{w}$  does not necessarily reside on this manifold. As a result, if  $\Phi(\mathbf{x})$  moves strictly along  $\mathbf{w}$ , the resulting images may not have a pre-image in  $\mathbb{R}^d$  any more. In other words, by forcing the pre-image of  $\Phi(\mathbf{x})$  to exist,  $\Phi(\mathbf{x})$  cannot move strictly along  $\mathbf{w}$ .

**Golland’s Method.** In [4], Golland et al. used the following strategy to handle this problem (Fig. 1). They searched for a direction  $d\mathbf{x}$  in  $\mathbb{R}^d$ . When  $\mathbf{x}$  moves along  $d\mathbf{x}$ , a displacement  $d\phi = \Phi(\mathbf{x} + d\mathbf{x}) - \Phi(\mathbf{x})$  will be induced in  $\mathcal{F}$  accordingly. The divergence from  $d\phi$  to  $\mathbf{w}$ , i.e., the displacement component of  $d\phi$  which is perpendicular to  $\mathbf{w}$ , is computed as  $d\phi - \langle d\phi, \mathbf{w} \rangle \mathbf{w}$ . Minimizing the divergence from  $\mathbf{w}$  makes the movement of  $\Phi(\mathbf{x})$  agree with  $\mathbf{w}$  as much as possible. The optimization problem is given as:

$$\begin{aligned} \text{Find } d\mathbf{x} = \arg \min_{d\mathbf{x} \in \mathbb{R}^d} \|d\phi - \langle d\phi, \mathbf{w} \rangle \mathbf{w}\|^2 &= \arg \min_{d\mathbf{x} \in \mathbb{R}^d} \langle d\phi, d\phi \rangle - \langle d\phi, \mathbf{w} \rangle^2 \\ \text{such that } \|d\mathbf{x}\|^2 &= \epsilon \\ \text{and } d\phi &= \Phi(\mathbf{x} + d\mathbf{x}) - \Phi(\mathbf{x}), \end{aligned} \quad (2)$$

where  $\epsilon$  is a preset small positive real number. Note that the constraint of  $\|d\mathbf{x}\|^2 = \epsilon$  is used, allowing  $d\mathbf{x}$  to be searched identically along *all* directions in  $\mathbb{R}^d$ . This method implicitly assumes that the distribution of the shapes occupies the whole of the space  $\mathbb{R}^d$ , which brings up the problem of spurious shape differences addressed in this paper. Also, it is pointed out in [4] that a conceptually simpler way to infer the discriminative direction is to move  $\Phi(\mathbf{x})$  strictly along  $\mathbf{w}$  and observe the change of  $\mathbf{x}$  in  $\mathbb{R}^d$ . We achieve this by using the pre-image techniques discussed in Section II-B.

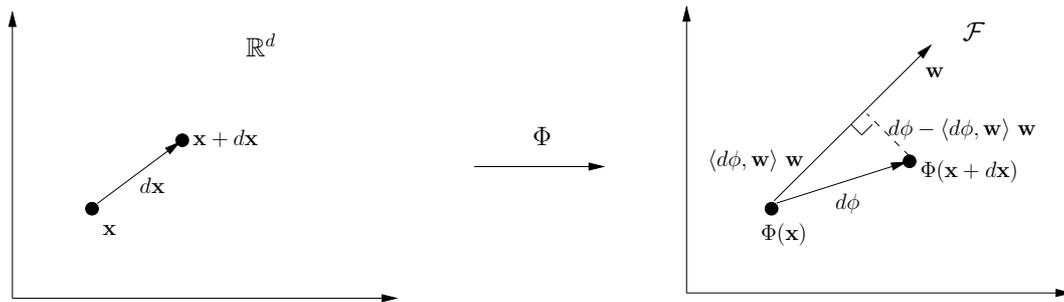


Fig. 1. *Explanation of how to infer the “discriminative direction” in Golland’s method. A point  $\mathbf{x}$  in the input space  $\mathbb{R}^d$  is mapped to a point  $\Phi(\mathbf{x})$  in the feature space  $\mathcal{F}$ . When  $\mathbf{x}$  moves a small step  $d\mathbf{x}$  in  $\mathbb{R}^d$ , a displacement  $d\phi$  will be induced in  $\mathcal{F}$  accordingly. The divergence of  $d\phi$  from  $\mathbf{w}$  is  $d\phi - \langle d\phi, \mathbf{w} \rangle \mathbf{w}$ , which is perpendicular to  $\mathbf{w}$ . Minimizing this divergence with respect to  $d\mathbf{x}$  can make  $\Phi(\mathbf{x})$  move along  $\mathbf{w}$  as much as possible in  $\mathcal{F}$ . Such a direction  $d\mathbf{x}$  is the “discriminative direction” at point  $\mathbf{x}$  in  $\mathbb{R}^d$ .*

The “discriminative direction” method is different from the traditional way of analyzing the shape difference for populations. The traditional method takes the mean of a population as a representative

<sup>1</sup>Please note that this manifold is induced by the kernel mapping  $\Phi(\cdot)$ . It is in the feature space  $\mathcal{F}$ . It should not be confused with the manifold in the shape descriptor space on which the shapes reside. The latter is used in this paper to develop the regularization discriminative direction.

shape and reveals the population difference by comparing the mean shapes of two populations. This approach works when the population distribution follows a simple model, for instance, a single mode Gaussian distribution. However, when the population distribution is complex, such as a mixture of Gaussian distributions, the shape difference varies among different parts within one population. Simply comparing two means will fail to capture such change in shape differences. Golland’s “discriminative direction” provides a remedy to this problem. In the case of linear classification, this method is comparable to the traditional method since the “discriminative direction” is consistent everywhere within a population. In the case of non-linear classification, the “discriminative direction” varies among different parts within a population. Hence it conveys a more complex shape difference between two populations. In [4], it is also pointed out that instead of deforming the mean of the training data, the input vectors close to the opposite class, i.e., the support vectors, should be selected for analysis. This is because the support vectors define the optimal separating boundary between two classes, and thus determine the shape class difference.

### B. Preimage techniques in kernel methods

Traditionally, the preimage problem for kernel methods is studied in the context of kernel PCA (KPCA) which models the variation of one class ([14], [15], [8], [9]). This problem is different from that of the kernel methods for two-class classification which we want to address in this paper. However, the basic ideas behind the existing approaches inspire the development of the solution for our problem.

In the same manner as kernel classifiers (Section I), KPCA implicitly maps the data from the input space  $\mathbb{R}^d$  to a high dimensional feature space  $\mathcal{F}$  via a non-linear mapping  $\Phi$ . Then a linear PCA is performed in  $\mathcal{F}$ , giving a set of orthogonal eigenvectors. Data points in  $\mathcal{F}$  are projected onto the space spanned by these eigenvectors and reconstructed there using only the  $n$  leading eigenvectors ranked by their eigenvalues. Finding the preimages of the reconstructed data is important in applications such as image denoising and statistical shape analysis [14], [15], [8], [9]. Fig. 2 illustrates the preimage problem for KPCA. Two non-iterative approaches for solving this problem are reviewed below.

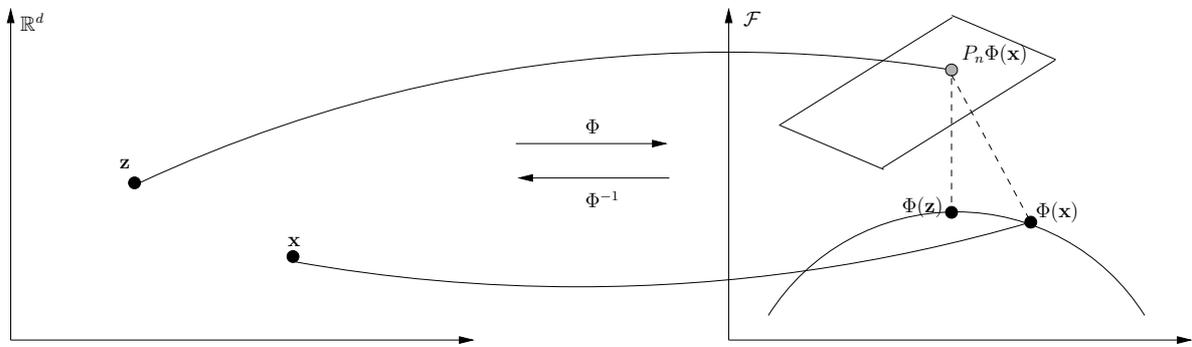


Fig. 2. Preimage problem in KPCA. The point  $\mathbf{x}$  in the input space  $\mathbb{R}^d$  is mapped to the point  $\Phi(\mathbf{x})$  in the feature space  $\mathcal{F}$ . Projecting  $\Phi(\mathbf{x})$  onto the first  $n$  significant eigenvectors found by KPCA in  $\mathcal{F}$  gives the point  $P_n\Phi(\mathbf{x})$ . The preimage of  $P_n\Phi(\mathbf{x})$  in  $\mathbb{R}^d$  is approximated as  $\mathbf{z}$ , provided that  $\Phi(\mathbf{z})$  is the point nearest to  $P_n\Phi(\mathbf{x})$  in  $\mathcal{F}$ .

**Rathi’s Approach.** In [9], Rathi et al. solve this preimage problem by minimizing the squared residual error  $\rho(\mathbf{z}) = \|P_n\Phi(\mathbf{x}) - \Phi(\mathbf{z})\|^2$ , where  $P_n\Phi(\mathbf{x})$  represents the reconstructed point of  $\Phi(\mathbf{x})$  from its projections on the  $n$  leading eigenvectors, and  $\mathbf{z}$  is the estimate of the preimage of  $P_n\Phi(\mathbf{x})$ . Using an RBF kernel  $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$  and setting  $\nabla_{\mathbf{z}}\rho = 0$  at extrema gives

$$\mathbf{z} = \frac{\sum_{i=1}^n \tilde{\gamma}_i \exp(-\|\mathbf{z} - \mathbf{x}_i\|^2/(2\sigma^2)) \mathbf{x}_i}{\sum_{i=1}^n \tilde{\gamma}_i \exp(-\|\mathbf{z} - \mathbf{x}_i\|^2/(2\sigma^2))}, \quad (3)$$

where  $\mathbf{x}_i$  is training data, and  $n$  is the total number of  $\mathbf{x}_i$ . The coefficients  $\tilde{\gamma}_i$  are learned by training a Kernel PCA. The novelty of Rathi's method is that the authors do not use the fixed point iteration scheme as proposed in [15], which is an iterative method and causes numerical instability. Instead, they replace the term  $\|\mathbf{z} - \mathbf{x}_i\|^2$  in the kernel function by a distance  $d^2(\mathbf{z}, \mathbf{x}_i) = \|\mathbf{z} - \mathbf{x}_i\|^2$  in  $\mathbb{R}^d$ , which is further replaced by a corresponding distance  $d_{\mathcal{F}}^2(\Phi(\mathbf{z}), \Phi(\mathbf{x}_i)) = \|\Phi(\mathbf{z}) - \Phi(\mathbf{x}_i)\|^2$  in  $\mathcal{F}$  via a simple relationship. Though  $\Phi$  is unknown,  $d_{\mathcal{F}}^2(P_n\Phi(\mathbf{x}), \Phi(\mathbf{x}_i))$  is formulated in terms of inner products, and hence may be calculated by kernel functions. Assuming that  $\Phi(\mathbf{z}) = P_n\Phi(\mathbf{x})$  ( $\rho(\mathbf{z}) = 0$ ), the distance  $d_{\mathcal{F}}^2(\Phi(\mathbf{z}), \Phi(\mathbf{x}_i))$  becomes computable and thus a closed-form solution of  $\mathbf{z}$  is achieved.

**Kwok's Approach.** Kwok et al. proposed a different approach in [8], which does not minimize the squared residual error  $\rho(\mathbf{z})$  as in Rathi's approach. The basic idea assumes that the local structure, i.e., the distance relationship between an input vector  $\mathbf{x}$  and its neighbouring training points  $\mathbf{x}_i$  ( $i = 1, \dots, m$ ), is preserved when mapping from  $\mathcal{F}$  to  $\mathbb{R}^d$ . That is, the  $m$ -nearest neighbours of  $P_n\Phi(\mathbf{x})$  in  $\mathcal{F}$  remain the  $m$ -nearest neighbours of its preimage  $\mathbf{z}$  in  $\mathbb{R}^d$ . Moreover, Kwok assumes that distances  $\|\mathbf{x}_i - \mathbf{x}_j\|$  between points  $\mathbf{x}_i$  and  $\mathbf{x}_j$  in the input space  $\mathbb{R}^d$  can be computed explicitly from the distances  $\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)\|$  between their corresponding points  $\Phi(\mathbf{x}_i)$  and  $\Phi(\mathbf{x}_j)$  in the feature-space  $\mathcal{F}$ . This is true for a Gaussian RBF kernel, or more particularly for isotropic kernels, for which  $\|\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_j)\|^2 = \kappa(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$  for some function  $\kappa$ . Given a point  $\mathbf{z}_{\mathcal{F}}$  in feature space, and "neighbouring" points  $\Phi(\mathbf{x}_i)$ ;  $i = 1, \dots, m$ , the back-projected point  $\mathbf{z}$  obtained by Kwok's method is the point that minimizes the function

$$\sum_{i=1}^m (\|\mathbf{x}_i - \mathbf{z}\|^2 - d_i^2)^2 \quad \text{where } d_i^2 = \kappa^{-1}(\|\Phi(\mathbf{x}_i) - \mathbf{z}_{\mathcal{F}}\|^2), \quad (4)$$

subject to the constraint that  $\mathbf{z}$  lies in the subspace spanned by points  $\mathbf{x}_i$ .

There is a closed-form solution to this problem as follows. Define the matrix  $\mathbf{X} = [\mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_m - \bar{\mathbf{x}}]$  where  $\bar{\mathbf{x}}$  is the centroid  $\bar{\mathbf{x}} = (1/m) \sum_{i=1}^m \mathbf{x}_i$ . Let  $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$  be the Singular Value Decomposition. Then the solution is given by

$$\mathbf{z} = -\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{V}^T(\mathbf{d}^2 - \mathbf{d}_0^2)/2 + \bar{\mathbf{x}}, \quad (5)$$

where  $\mathbf{d}^2$  is the vector  $(d_1^2, \dots, d_m^2)^T$  and  $\mathbf{d}_0^2$  is the vector with  $i$ -th entry equal to  $\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$ . For details, see [8].

### III. Reformulation of the original discriminative direction

In this section, we reformulate Golland's "discriminative direction" from a different perspective, which is the first contribution of this paper. As argued in [6], [4], for a particular shape descriptor  $\hat{\mathbf{x}}$ , its image  $\Phi(\hat{\mathbf{x}})$  in  $\mathcal{F}$  should be moved along the direction  $\mathbf{w}$  in order to study the changes of  $\hat{\mathbf{x}}$  introduced by this process. However the mapping  $\Phi(\cdot)$  is generally too complex to work out, preventing  $\Phi(\hat{\mathbf{x}})$  from being accessed directly. Hence in [4], the discriminative direction is explored via a search in the shape descriptor space  $\mathbb{R}^d$ . In this paper, we circumvent the explicit manipulation of  $\Phi(\hat{\mathbf{x}})$  by formulating the discriminative direction problem solely in terms of kernel functions. Then we approximate the preimages of  $\Phi(\hat{\mathbf{x}})$  in  $\mathbb{R}^d$  while  $\Phi(\hat{\mathbf{x}})$  is moving strictly along the direction  $\mathbf{w}$  in  $\mathcal{F}$ . In this way, we provide a complementary and conceptually simpler means of inferring the discriminative direction. The basic idea is shown in Fig. 3. To solve our reformulation, the preimage methods in [9], [8] are modified for use by considering two key issues: (i) we need to formulate our cost function in terms of inner products only, so that the unsolvable  $\Phi$  can be eliminated; and (ii) we need to derive the distance  $d_{\mathcal{F}}^2(\Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j))$  in the feature space  $\mathcal{F}$ , and formulate it in terms of inner products, so that it is computable via the kernel function. Our work is elaborated below.

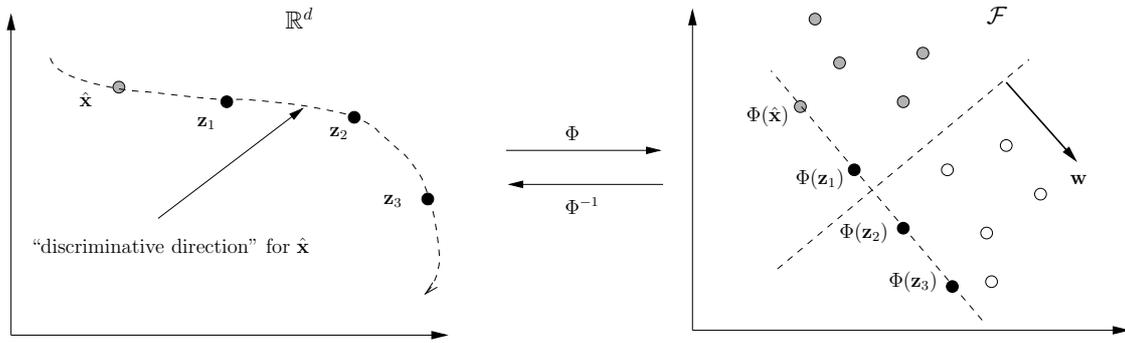


Fig. 3. *Discriminative direction for a point,  $\hat{\mathbf{x}}$ . A nonlinear mapping  $\Phi$  maps the shape descriptor space  $\mathbb{R}^d$  onto a feature space  $\mathcal{F}$ , where the two classes (gray and white dots) become linearly separable. The  $\mathbf{w}$  is the normal of a hyperplane found in  $\mathcal{F}$  to best discriminate the two classes. Move  $\Phi(\hat{\mathbf{x}})$  along  $\mathbf{w}$  to a new position  $\Phi(\mathbf{z}_1)$ , and project  $\Phi(\hat{\mathbf{x}})$  back to  $\mathbb{R}^d$  as  $\mathbf{z}_1$ . The vector  $\mathbf{z}_1 - \hat{\mathbf{x}}$  is the “discriminative direction” of  $\hat{\mathbf{x}}$ . Similar results can be obtained for  $\Phi(\mathbf{z}_2)$  and  $\Phi(\mathbf{z}_3)$ .*

Without loss of generality, assume that  $\mathbf{w}$  has been normalized to a unit vector. Moving  $\Phi(\hat{\mathbf{x}})$  along  $\mathbf{w}$  in  $\mathcal{F}$  for a step  $s$  leads to a new position  $\Phi(\hat{\mathbf{x}}) + s\mathbf{w}$ . Let  $\mathbf{z}$  be the best estimate of the pre-image of  $\Phi(\hat{\mathbf{x}}) + s\mathbf{w}$ , representing the new shape of  $\hat{\mathbf{x}}$  after deformation. The vector  $\mathbf{z} - \hat{\mathbf{x}}$  is just the discriminative direction. To estimate the preimage, the residual error  $\rho(\mathbf{z})$  should be minimized with respect to  $\mathbf{z}$ :

$$\mathbf{z}^* = \arg \min_{\mathbf{z} \in \mathbb{R}} \rho(\mathbf{z}) = \arg \min_{\mathbf{z} \in \mathbb{R}} \|\Phi(\hat{\mathbf{x}}) + s\mathbf{w} - \Phi(\mathbf{z})\|^2. \quad (6)$$

Notice that

$$\rho(\mathbf{z}) = \langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\hat{\mathbf{x}}) + s\mathbf{w} \rangle + \langle \Phi(\mathbf{z}), \Phi(\mathbf{z}) \rangle - 2\langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\mathbf{z}) \rangle,$$

and that  $\langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\hat{\mathbf{x}}) + s\mathbf{w} \rangle$  is constant with respect to the variable  $\mathbf{z}$ .

We now assume a Gaussian RBF kernel  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$ . In this case,  $\langle \Phi(\mathbf{z}), \Phi(\mathbf{z}) \rangle$  is constant, and so

$$\rho(\mathbf{z}) = -2\langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\mathbf{z}) \rangle + C,$$

where  $C$  is a constant value. Therefore, the minimization problem in (6) is equivalent to

$$\mathbf{z}^* = \arg \max_{\mathbf{z} \in \mathbb{R}} \langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\mathbf{z}) \rangle. \quad (7)$$

**The Rathi-inspired method.** Noting that  $\mathbf{w}$  lies in a space spanned by the training points, we may write  $\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$ . Now, to find the minimum of  $\rho(\mathbf{z})$  we set the derivative with respect to  $\mathbf{z}$  to zero and rearrange terms. For a Gaussian RBF kernel, this results in the equation

$$\mathbf{z} = \frac{k(\hat{\mathbf{x}}, \mathbf{z})\hat{\mathbf{x}} + s \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{z})\mathbf{x}_i}{k(\hat{\mathbf{x}}, \mathbf{z}) + s \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{z})}$$

Observe that  $\mathbf{z}$  appears on both sides of this equation. However, using a trick of Rathi, we can eliminate  $\mathbf{z}$  from the right hand side of the equation. Assuming that  $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle = 1$  for any  $\mathbf{x}$ , which is true for normalized kernels and some isotropic kernels such as the RBF kernel, we may write

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = 1 - \frac{1}{2} \|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|^2,$$

so

$$\mathbf{z} = \frac{(2 - \|\Phi(\hat{\mathbf{x}}) - \Phi(\mathbf{z})\|^2)\hat{\mathbf{x}} + s \sum_{i=1}^n \alpha_i (2 - \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{z})\|^2)\mathbf{x}_i}{2 - \|\Phi(\hat{\mathbf{x}}) - \Phi(\mathbf{z})\|^2 + s \sum_{i=1}^n \alpha_i (2 - \|\Phi(\mathbf{x}_i) - \Phi(\mathbf{z})\|^2)}.$$

We now assume that  $\Phi(\mathbf{z}) = \Phi(\hat{\mathbf{x}}) + s\mathbf{w}$ , and recall that  $\|\mathbf{w}\| = 1$ . This allows us to eliminate  $\mathbf{z}$  from the right hand side of this formula, resulting in a closed-form solution for  $\mathbf{z}$ , namely

$$\mathbf{z}^* = \frac{(2 - s^2)\hat{\mathbf{x}} + s \sum_{i=1}^n \alpha_i (2 - \|\Phi(\mathbf{x}_i) - (\Phi(\hat{\mathbf{x}}) + s\mathbf{w})\|^2) \mathbf{x}_i}{2 - s^2 + s \sum_{i=1}^n \alpha_i (2 - \|\Phi(\mathbf{x}_i) - (\Phi(\hat{\mathbf{x}}) + s\mathbf{w})\|^2)} \quad (8)$$

Finally,  $\|\Phi(\mathbf{x}_i) - (\Phi(\hat{\mathbf{x}}) + s\mathbf{w})\|^2$  can be computed by as follows.

$$\begin{aligned} \|\Phi(\mathbf{x}_i) - (\Phi(\hat{\mathbf{x}}) + s\mathbf{w})\|^2 &= \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_i) \rangle + \langle \Phi(\hat{\mathbf{x}}), \Phi(\hat{\mathbf{x}}) \rangle + s^2 \langle \mathbf{w}, \mathbf{w} \rangle \\ &\quad + 2s \langle \Phi(\hat{\mathbf{x}}), \mathbf{w} \rangle - 2 \langle \Phi(\mathbf{x}_i), \Phi(\hat{\mathbf{x}}) \rangle - 2s \langle \Phi(\mathbf{x}_i), \mathbf{w} \rangle \\ &= 1 + 1 + s^2 + 2s \sum_{j=1}^n \alpha_j k(\mathbf{x}_j, \hat{\mathbf{x}}) \\ &\quad - 2k(\mathbf{x}_i, \hat{\mathbf{x}}) - 2s \sum_{j=1}^n \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \\ &= 2 + s^2 + 2s \boldsymbol{\alpha}^\top \mathbf{k}_{\hat{\mathbf{x}}} - 2k(\mathbf{x}_i, \hat{\mathbf{x}}) - 2s \boldsymbol{\alpha}^\top \mathbf{k}_{\mathbf{x}_i}. \end{aligned} \quad (9)$$

where, given  $n$  training data,  $\boldsymbol{\alpha}$  denotes  $(\alpha_1, \dots, \alpha_n)^\top$ ,  $\mathbf{k}_{\hat{\mathbf{x}}}$  denotes  $(k(\hat{\mathbf{x}}, \mathbf{x}_1), \dots, k(\hat{\mathbf{x}}, \mathbf{x}_n))^\top$ , and  $\mathbf{k}_{\mathbf{x}_i}$  denotes  $(k(\mathbf{x}_i, \mathbf{x}_1), \dots, k(\mathbf{x}_i, \mathbf{x}_n))^\top$ .

**The Kwok-inspired method.** Although Kwok described his method specifically for the KPCA problem, he remarked that the method can be generally used for other inverse kernel problems. We apply it here for our discriminative direction problem.

Considering a point  $\mathbf{z}_{\mathcal{F}} = \Phi(\hat{\mathbf{x}}) + s\mathbf{w}$  in  $\mathcal{F}$ , we find a set of training points  $\mathbf{x}_i$  such that the  $\Phi(\mathbf{x}_i)$  are the  $k$  nearest neighbours to  $\mathbf{z}_{\mathcal{F}}$ . Next, we compute the distances  $\|\Phi(\mathbf{x}_i) - (\Phi(\hat{\mathbf{x}}) + s\mathbf{w})\|$  in  $\mathcal{F}$  using (9). We need to relate these distances  $d_{\mathcal{F}}^2$  to distances  $d^2$  in the shape-descriptor space  $\mathbb{R}^d$ . For a Gaussian RBF kernel, we can write

$$d_{\mathcal{F}}^2(\mathbf{x}, \mathbf{y}) = \|\Phi(\mathbf{x}) - \Phi(\mathbf{y})\|^2 = 2 - 2k(\mathbf{x}, \mathbf{y}) = 2 \left( 1 - \exp \left( \frac{-d^2(\mathbf{x}, \mathbf{y})}{2\sigma^2} \right) \right).$$

Therefore,

$$d^2 = -2\sigma^2 (\log(1 - d_{\mathcal{F}}^2/2)). \quad (10)$$

The back-projected point  $\mathbf{z}$  corresponding to  $\mathbf{z}_{\mathcal{F}} = \Phi(\hat{\mathbf{x}}) + s\mathbf{w}$  is then computed using (5).

Both Rathi-inspired and Kwok-inspired methods are non-iterative and only involve linear algebra. Given  $\hat{\mathbf{x}}$  and  $s$ ,  $\mathbf{z}$  can be obtained immediately. Experimental study in Section VI-A shows that the discriminative direction estimated using the pre-image techniques produces results similar to those of Golland's method. Therefore, in the rest of this paper, for the purpose of conceptual simplicity we keep using the formulation in (6) to model the discriminative direction. The main contribution of the present section is how we reformulate the discriminative direction problem compared with Golland's work. In the rest of this paper, we point out that strictly following such a discriminative direction is problematic either in Golland's method or in the Rathi-inspired or Kwok-inspired methods. In the following sections, variants of the discriminative direction are proposed to address this problem. These are the most important contributions of this paper.

#### IV. Problems of the original discriminative direction

The original discriminative direction approach projects the shape difference between groups captured by the kernel classifiers back to the shape descriptor space where the deformed shape can be interpreted (for example, visualized) as an anatomically meaningful object. By comparing the changes

of a shape before and after its deformation along the discriminative direction, the class difference is therefore localized or isolated.

However, it is observed that simply deforming along the optimal discriminative direction provided by a classifier cannot maintain a shape's intrinsic characteristic which makes it belong to a particular shape group (called "anatomical correctness" in this paper). This is because the deformed shape deviates from the underlying distribution of the original shapes. In such cases, when the shapes before and after the deformation are compared, artefact differences will be introduced.

Take the two classes of right-angled triangles in Fig. 4 for example. They share the same hypotenuse  $AB$  with different vertices  $C$  corresponding to the right angle. All these right angle vertices are located on a semi-circle which takes  $AB$  as its diameter. Label triangles whose vertex  $C$  belongs to the first quadrant as a positive class and triangles whose vertex  $C$  belongs to the second quadrant as a negative class. The triangle  $ACB$  is fully determined by  $C$  once  $AB$  has been fixed. Thus when representing  $\triangle ACB$  by the position of  $C$ , the optimal discriminative direction  $\mathbf{w}$  given by a classifier to best separate the two classes is perpendicular to the  $y$  axis. Moving the vertex  $C$  strictly following the direction  $\mathbf{w}$  causes it to deviate from the semi-circle where it lies. This causes  $\triangle ACB$  to lose the right angle during the deformation. When comparing the triangles before and after the deformation, two kinds of differences - the position changes of the vertex  $C$  and the angle changes of  $\angle ACB$  - will be observed. However, the angle change is a spurious change because all the triangles in the two groups are right-angled. Evidently this is because the fact that the point  $C$  lies on a semi-circle is unreasonably neglected.

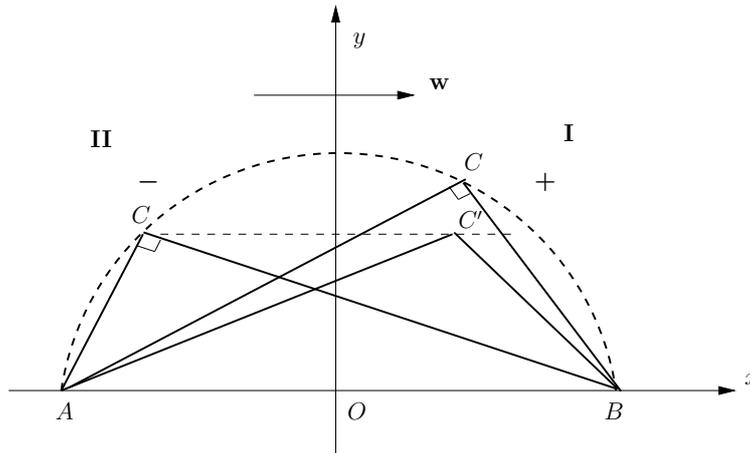


Fig. 4. The problem caused by strictly following the original discriminative direction provided by the classifier. Two groups of right-angled triangles which share the same hypotenuse are differentiated by the position of the vertex  $C$ , which corresponds to the right angle. The positive class has  $C$  in quadrant I, while the negative class has  $C$  in quadrant II. Take the position of  $C$  as the shape descriptor. Moving  $C$  along the optimal discriminative direction  $\mathbf{w}$  given by the classifier will cause the point  $C$  to deviate from the semi-circle where it resides. This causes the deformed triangle  $\triangle AC'B$  to lose its property of right-angledness. As a result, an angle difference will be falsely observed when comparing the triangles  $\triangle ACB$  and  $\triangle AC'B$ .

To remedy such problems, we argue that the underlying distribution of the shapes should not be neglected because (i) the shapes often reside on a lower dimensional manifold though the shape descriptor space has high dimensionality, and (ii) the deformation of an anatomical object may be spatially restricted by its surroundings; for example we may say that a deformation of an organ such as the liver will be in part constrained by the organs surrounding it. This consideration underpins the regularized discriminative direction approach described below.

## V. Regularized discriminative direction

The key idea of our approach is that when seeking the pre-image of a point in the feature space  $\mathcal{F}$ , the possible solutions should be confined to the underlying distribution of the shapes rather than the whole shape descriptor space  $\mathbb{R}^d$  as before. That is, the preimage should be computed as follows.

$$\mathbf{z}^* = \arg \min_{\mathbf{z} \in \Omega} \rho(\mathbf{z}), \quad (11)$$

where  $\rho(\mathbf{z})$  is defined in (6), and  $\Omega$  may be a submanifold in  $\mathbb{R}^d$ . The key change is the use of  $\mathbf{z} \in \Omega$  rather than  $\mathbf{z} \in \mathbb{R}^d$  as before. Now the problem becomes finding a suitable definition of  $\Omega$  and efficiently solving the resulting constrained optimization problem. In this paper, we propose two approaches to regularize the solution of  $\mathbf{z}$ , one from a local distribution perspective, and the other from a geometric point of view.

### A. Regularization via local distribution

Let  $\hat{\mathbf{x}}$  denote a particular shape to be deformed. Recall that moving  $\Phi(\hat{\mathbf{x}})$  along  $\mathbf{w}$  in  $\mathcal{F}$  for a step  $s$  arrives at a new position  $\Phi(\hat{\mathbf{x}}) + s\mathbf{w}$ . Let  $\mathbf{z}$  approximate the pre-image of  $\Phi(\hat{\mathbf{x}}) + s\mathbf{w}$ . We argue that to ensure ‘‘anatomical correctness’’, the pre-image  $\mathbf{z}$  should comply with the probability distribution of the shape  $\hat{\mathbf{x}}$ . For example, when  $\hat{\mathbf{x}}$  resides on a lower dimensional manifold,  $\mathbf{z}$  should reside on it too. Let  $N_\epsilon(\hat{\mathbf{x}}) = \{\mathbf{x} \mid \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon\}$  be a neighbourhood of  $\hat{\mathbf{x}}$  and  $p(\mathbf{x} \mid \mathbf{x} \in N_\epsilon(\hat{\mathbf{x}}))$  be an empirical probability density function of  $\mathbf{x}$  in  $N_\epsilon(\hat{\mathbf{x}})$  estimated from  $n$  training shapes. Here we model  $p(\mathbf{x})$  as a normal distribution<sup>2</sup> with mean  $\boldsymbol{\mu} = \hat{\mathbf{x}}$  and covariance matrix  $\boldsymbol{\Sigma} = (1/n) \sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{x}})(\mathbf{x}_i - \hat{\mathbf{x}})^\top$ . Note that for an RBF kernel, a small distance  $d_{\mathcal{F}}$  in  $\mathcal{F}$  corresponds to a small distance  $d$  in  $\mathbb{R}^d$ . This can be clearly seen by the distance relationship in (10), i.e.,  $d^2 = -2\sigma^2 \log(1 - d_{\mathcal{F}}^2/2)$ . Hence moving  $\Phi(\hat{\mathbf{x}})$  with a sufficiently small step  $s$  in  $\mathcal{F}$  will always ensure that  $\mathbf{z}$  stays in  $N_\epsilon(\hat{\mathbf{x}})$ . Following the argument that  $\mathbf{z}$  should comply with the probability distribution of  $\mathbf{x}$ , we require that  $p(\mathbf{z})$  should be large enough, or equally that  $(\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})$  be adequately small, provided that  $\boldsymbol{\Sigma}$  has full rank. In this way, the optimal  $\mathbf{z}$  is defined as

$$\mathbf{z}^* = \arg \min_{\mathbf{z} \in \mathbb{R}^d} \rho(\mathbf{z}) + 2\eta (\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}) \quad (12)$$

where  $\eta \geq 0$  is a regularization parameter. When  $\eta$  is 0, this problem reduces to the minimization of  $\rho(\mathbf{z})$  in (6).

Consider the case when the shapes reside on a lower dimensional manifold in  $\mathbb{R}^d$ , causing  $\boldsymbol{\Sigma}$  to be rank-deficient. Decompose  $\boldsymbol{\Sigma}$  as  $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top$ , where each column of  $\boldsymbol{\Gamma}$  is an eigenvector and  $\boldsymbol{\Lambda}$  is  $\text{diag}\{\lambda_1, \dots, \lambda_r, 0, \dots, 0\}$ . The  $\lambda_i$  is the  $i$ -th positive eigenvalue and  $r$  is the rank of  $\boldsymbol{\Sigma}$ . An optimal solution  $\mathbf{z}^*$  should satisfy:

$$\mathbf{z}^* \in \{\mathbf{z} \mid (\boldsymbol{\Gamma}^\top(\mathbf{z} - \boldsymbol{\mu}))_i = 0 \text{ for } i = r + 1, \dots, d\} = \{\mathbf{z} \mid \mathbf{z} = \boldsymbol{\mu} + \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Lambda}}^{\frac{1}{2}}\mathbf{u}\}, \quad (13)$$

where  $\mathbf{u}$  is a vector in  $\mathbb{R}^r$ . The  $r \times r$  matrix  $\hat{\boldsymbol{\Lambda}}$  is  $\text{diag}\{\lambda_1, \dots, \lambda_r\}$ . The  $d \times r$  matrix  $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_r)$  contains the eigenvectors corresponding to  $\hat{\boldsymbol{\Lambda}}$ .

The result in (13) is explained as follows. Let  $\mathcal{M}$  be the manifold where the shapes reside, and  $T_{\boldsymbol{\mu}}(\mathcal{M})$  be a tangent plane of  $\mathcal{M}$  at  $\boldsymbol{\mu}$ . This tangent plane is spanned by the eigenvectors in  $\hat{\boldsymbol{\Gamma}}$ . Since a manifold can be locally approximated by its tangent plane, the result in (13) can be thought of as confining the solution  $\mathbf{z}$  to the manifold  $\mathcal{M}$ . Moreover note that the shapes do not necessarily uniformly distribute in  $T_{\boldsymbol{\mu}}(\mathcal{M})$ . Our regularized method naturally incorporates the variance along

<sup>2</sup>Using a more complicated model may be dangerous in the sense that its parameters may not be reliably estimated because the number of training samples in  $N_\epsilon(\hat{\mathbf{x}})$  is quite limited in practice.

different directions of the orthogonal basis via  $\hat{\Lambda}^{\frac{1}{2}}$  in (13). This makes it achieve a better performance than merely projecting  $\mathbf{z}^*$  which minimizes (6) onto the tangent plane<sup>3</sup>, as shown later in Section VI-B.

Finally, combining (13) and (12), the problem in (12) can be simplified by optimizing  $\mathbf{u}$  as:

$$\mathbf{u}^* = \arg \min_{\mathbf{u} \in \mathbb{R}^r} \rho(\boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{\frac{1}{2}} \mathbf{u}) + 2\eta \langle \mathbf{u}, \mathbf{u} \rangle, \quad (14)$$

and  $\mathbf{z}^*$  is then computed by (13). In practice, the number of the training shapes is often much less than the dimensionality of the shape descriptors. This results in a rank-deficient  $\boldsymbol{\Sigma}$ , whose rank  $r \ll d$ . Hence, optimizing over  $\mathbf{u}$  greatly reduces the number of parameters to estimate in comparison to directly optimizing over  $\mathbf{z}$ . Iterative optimization methods can be used to estimate  $\mathbf{u}$ . However, when  $r$  is large, optimizing  $\mathbf{u}$  is still cumbersome. Below we propose a new differential equation based solution so that for a given step  $s$ , first  $\mathbf{u}^*$  and then  $\mathbf{z}^*$  can be directly worked out as an analytical solution.

### A.1 An analytic solution to the pre-image $\mathbf{z}^*$

According to (7), the problem in (14) is equivalent to maximizing  $\langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\boldsymbol{\mu} + \hat{\Gamma}^{\top} \hat{\Lambda}^{\frac{1}{2}} \mathbf{u}) \rangle - \eta \langle \mathbf{u}, \mathbf{u} \rangle$  provided  $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle$  is a constant, which is the case for an RBF kernel. As before,  $\mathbf{w}$  lies in a space spanned by the training samples:  $\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$  with  $\alpha_i \in \mathcal{R}$ . We maximize the following expression:

$$\begin{aligned} f(s, \mathbf{u}) &= \langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{\frac{1}{2}} \mathbf{u}) \rangle - \eta \langle \mathbf{u}, \mathbf{u} \rangle \\ &= k(\hat{\mathbf{x}}, \boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{\frac{1}{2}} \mathbf{u}) + s \sum_i \alpha_i k(\mathbf{x}_i, \boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{\frac{1}{2}} \mathbf{u}) - \eta \langle \mathbf{u}, \mathbf{u} \rangle \\ &\triangleq g(\mathbf{u}) + s \cdot h(\mathbf{u}) - \eta \cdot l(\mathbf{u}). \end{aligned}$$

For each given  $s$ , there will be a  $\mathbf{u}^*$  which maximizes  $f(s, \mathbf{u})$ . This optimization problem is not convex and has multiple isolated local maxima. Here we propose an approach which does not directly solve the static optimization over  $\mathbf{u}$  for a given  $s$ . Instead it makes use of the fact that  $(0, \mathbf{0})$  is a global maximum of  $f(s, \mathbf{u})$  and traces the change of the global maximum with respect to  $s$ . As long as  $\mathbf{u}^*(s)$  is continuous and differentiable, our solution remains the global maximum or at least locally maximum. The change of  $\mathbf{u}^*$  with respect to  $s$  can be considered as a curve  $\mathbf{u}^*(s)$  in  $\mathbb{R}^r$  parameterized by  $s$ , such that  $\mathbf{u}^*(0) = \mathbf{0}$ . The curve can be traced out by computing its tangent  $d\mathbf{u}^*/ds$ . We approximate  $f(s, \mathbf{u})$  by a second order Taylor expansion

$$\begin{aligned} f(s, \mathbf{u}) &\approx g(\mathbf{u}_0) + s h(\mathbf{u}_0) - \eta l(\mathbf{u}_0) \\ &\quad + (\mathbf{J}_g + s\mathbf{J}_h - \eta\mathbf{J}_l)(\mathbf{u} - \mathbf{u}_0) \\ &\quad + \frac{1}{2}(\mathbf{u} - \mathbf{u}_0)^{\top} (\mathbf{H}_g + s\mathbf{H}_h - \eta\mathbf{H}_l)(\mathbf{u} - \mathbf{u}_0), \end{aligned} \quad (15)$$

where  $\mathbf{J}$  and  $\mathbf{H}$  are the Jacobian and Hessian of the functions  $g$ ,  $h$  and  $l$  with respect to  $\mathbf{u}$ , evaluated at  $\mathbf{u}_0$ . Here  $\mathbf{u}_0$  maximizes  $f(s, \mathbf{u})$  when  $s = s_0$ . The first order derivative of  $f$  with respect to  $\mathbf{u}$  vanishes at  $\mathbf{u}_0$  and any other extremum  $\mathbf{u}^*$ . From  $\left. \frac{\partial f}{\partial \mathbf{u}} \right|_{(\mathbf{u}_0, s_0)} = 0$ , we get  $\mathbf{J}_g - \eta\mathbf{J}_l = -s_0\mathbf{J}_h$ . Combining it in

$\left. \frac{\partial f}{\partial \mathbf{u}} \right|_{(\mathbf{u}^*, s)} = 0$ , we get

$$s\mathbf{J}_h + (\mathbf{J}_g - \eta\mathbf{J}_l) + (\mathbf{H}_g + s\mathbf{H}_h - \eta\mathbf{H}_l)(\mathbf{u}^* - \mathbf{u}_0) = (s - s_0)\mathbf{J}_h + (\mathbf{H}_g + s\mathbf{H}_h - \eta\mathbf{H}_l)(\mathbf{u}^* - \mathbf{u}_0) = 0.$$

<sup>3</sup>This approach is called ‘‘tangent plane projection’’ later in our experiments.

This gives

$$\left. \frac{d\mathbf{u}^*}{ds} \right|_{s=s_0} = -(\mathbf{H}_g + s_0\mathbf{H}_h - \eta\mathbf{H}_l)^{-1}\mathbf{J}_h. \quad (16)$$

This curve passes through an initial point  $(0, \mathbf{0})$  with the tangent direction

$$\left. \frac{d\mathbf{u}^*}{ds} \right|_{s=0} = -(\mathbf{H}_g - \eta\mathbf{H}_l)^{-1}\mathbf{J}_h. \quad (17)$$

The curve of  $\mathbf{u}^*(s)$  can be therefore traced out and the regularized discriminative direction is formed as a sequence of points  $\hat{\mathbf{x}}^{(0)}, \hat{\mathbf{x}}^{(1)}, \dots, \hat{\mathbf{x}}^{(t)}, \dots$  according to

$$\begin{aligned} \mathbf{u}^{*(t)} &= \mathbf{u}^{*(t-1)} + \left. \frac{d\mathbf{u}^{*(t-1)}}{ds} \right|_{s=0} (s_t - s_{t-1}) \\ \hat{\mathbf{x}}^{(t)} &= \mathbf{z}^* = \hat{\mathbf{x}}^{(t-1)} + \hat{\Gamma}^{(t-1)}[\hat{\Lambda}^{(t-1)}]^{1/2}\mathbf{u}^{*(t)} \end{aligned} \quad (18)$$

where  $\mathbf{u}^{*(0)} = \mathbf{0}$ ,  $\hat{\Gamma}^{(t-1)}$  and  $\hat{\Lambda}^{(t-1)}$  are estimated from  $\hat{\mathbf{x}}^{(t-1)}$ , and  $\hat{\mathbf{x}}^{(0)} = \hat{\mathbf{x}}$ . The mean  $\boldsymbol{\mu}$  and covariance  $\Sigma^{(t-1)}$  are computed at each step from the training points close to  $\hat{\mathbf{x}}^{(t-1)}$ .

**Runge-Kutta Method.** A four-stage Runge Kutta method is integrated to suppress the lower-order error terms of this ordinary differential equation. Runge Kutta methods are iterative methods to approximate the solution of ordinary differential equations. They use trial steps at the middle points of an interval to remove the lower-order error terms. Our algorithm is summarized in Table I.

TABLE I  
Algorithm: An Analytical Solution

- 
1. Let  $t = 0$ ,  $\hat{\mathbf{x}}^{(0)} = \hat{\mathbf{x}}$
  2. Let  $\mathbf{u}^{(t)} = \mathbf{0}$ ,  $s^{(t)} = 0$
  3. Estimate  $\hat{\Gamma}$  and  $\hat{\Lambda}$  at  $\hat{\mathbf{x}}^{(t)}$
  4. Evaluate  $\mathbf{J}_h$ ,  $\mathbf{H}_g$ ,  $\mathbf{H}_h$  at  $\mathbf{u}^{(t)}$  with an RBF kernel.
 
$$\begin{aligned} \mathbf{J}_h &= -\frac{1}{\sigma^2}(\boldsymbol{\alpha} * \mathbf{K}_x^\top)^\top (\mathbf{D}_x \hat{\Gamma} \hat{\Lambda}^{1/2}) \\ \mathbf{H}_g &= -\frac{1}{\sigma^2} \hat{\Lambda} \\ \mathbf{H}_h &= \frac{1}{\sigma^4}(\boldsymbol{\alpha} * \mathbf{K}_x^\top)^\top (\text{diag}(\mathbf{D}_x \mathbf{D}_x^\top)) \mathbf{I}_d - \frac{1}{\sigma^2}(\mathbf{K}_x \boldsymbol{\alpha}) \mathbf{I}_d \\ \mathbf{H}_l &= 2\mathbf{I}_d \end{aligned}$$
  5. Compute  $\mathbf{u}^{(t+1)}$  and the new position  $\hat{\mathbf{x}}^{(t+1)}$  using a four-stage Runge Kutta method.
    - (1)  $\mathbf{u}_1 \leftarrow \mathbf{u}^{(t)}$ ,  $s_{u1} \leftarrow s^{(t)}$ , compute  $\mathbf{t}_1 = \left. \frac{d\mathbf{u}^*}{ds} \right|_{s=s_{u1}, \mathbf{u}=\mathbf{u}^{(t)}}$ .
    - (2)  $\mathbf{u}_2 \leftarrow \mathbf{u}_1 + \mathbf{t}_1 \Delta s / 2$ ,  $s_{u2} \leftarrow s^{(t)} + \Delta s / 2$ , compute  $\mathbf{t}_2 = \left. \frac{d\mathbf{u}^*}{ds} \right|_{s=s_{u2}, \mathbf{u}=\mathbf{u}^{(t)}}$
    - (3)  $\mathbf{u}_3 \leftarrow \mathbf{u}_1 + \mathbf{t}_2 \Delta s / 2$ ,  $s_{u3} \leftarrow s^{(t)} + \Delta s / 2$ , compute  $\mathbf{t}_3 = \left. \frac{d\mathbf{u}^*}{ds} \right|_{s=s_{u3}, \mathbf{u}=\mathbf{u}^{(t)}}$
    - (4)  $\mathbf{u}_4 \leftarrow \mathbf{u}_1 + \mathbf{t}_3 \Delta s / 2$ ,  $s_{u4} \leftarrow s^{(t)} + \Delta s$ , compute  $\mathbf{t}_4 = \left. \frac{d\mathbf{u}^*}{ds} \right|_{s=s_{u4}, \mathbf{u}=\mathbf{u}^{(t)}}$
    - (5) Compute  $\mathbf{u}^{(t+1)} = \mathbf{u}^{(t)} + \Delta s \times (1/6\mathbf{t}_1 + 1/3\mathbf{t}_2 + 1/3\mathbf{t}_3 + 1/6\mathbf{t}_4)$
    - (6) Compute the new position  $\hat{\mathbf{x}}^{(t+1)}$ :  $\hat{\mathbf{x}}^{(t+1)} = \hat{\mathbf{x}}^{(t)} + \boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{1/2} \mathbf{u}$
  6.  $\mathbf{z}^* = \hat{\mathbf{x}}^{(t+1)}$ ,  $t = t + 1$
  7. Repeat step 2 ~ 6 to get the pre-images of the movement along  $\mathbf{w}$  in  $\mathcal{F}$ .
- 

Operator  $\text{diag}(\cdot)$  is the diagonal of a matrix; operator  $*$  the component-wise multiplication of two matrices;  $n$  the number of training data;  $d$  the dimensionality of the input space;  $\mathbf{I}$  an identity matrix;  $\mathbf{K}_x = (k(\mathbf{x}_1, \boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{1/2} \mathbf{u}^{(t)}), k(\mathbf{x}_2, \boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{1/2} \mathbf{u}^{(t)}), \dots, k(\mathbf{x}_n, \boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{1/2} \mathbf{u}^{(t)}))$ ; and  $\mathbf{D}_x = ((\boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{1/2} \mathbf{u}^{(t)} - \mathbf{x}_1), (\boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{1/2} \mathbf{u}^{(t)} - \mathbf{x}_2), \dots, (\boldsymbol{\mu} + \hat{\Gamma} \hat{\Lambda}^{1/2} \mathbf{u}^{(t)} - \mathbf{x}_n))^\top$ .

## A.2 Discussion

The covariance matrix  $\Sigma$  has to be estimated from the neighbourhood of  $\hat{\mathbf{x}}$ . When the number  $n_{nb}$  of training samples in the neighbourhood is sufficiently larger than the dimensionality  $d$  of the data, we can safely assume the sample-based estimate to be its true value. However, the common case may be that  $n_{nb}$  is much smaller than  $d$ . It is known that in such cases, the sample-based estimate tends to bias the larger eigenvalues towards values that are too high and the smaller ones towards values that are too low (artificially zero) [16]. Therefore, we may over-estimate the variances of the points in  $N_\epsilon(\hat{\mathbf{x}})$  along the eigenvectors corresponding to large eigenvalues, and underestimate the variances along the eigenvectors corresponding to small eigenvalues. When  $n_{nb}$  is not large enough, we employ a ‘‘shrinking’’ [16] scheme which shrinks  $\Sigma$  towards a multiple of the identity matrix by using  $\hat{\Sigma} = (1 - \gamma)\Sigma + \gamma\bar{\lambda}\mathbf{I}$ , where  $\gamma$  ( $0 \leq \gamma < 1$ ) controls the shrinkage, and  $\bar{\lambda}$  is the average value of the eigenvalues  $\lambda_i$  obtained from the sample-based estimate  $\Sigma$ .

The parameter  $\eta$  in (12) and (14) balances the fitting of the data and the fitting of the local distribution model. In our experiment, it is empirically set as  $\eta = \frac{\sqrt{\bar{\lambda}}}{\sigma^2}$ , where  $\bar{\lambda}$  is defined above, and  $\sigma$  is the parameter of the RBF kernel used by the classifier. The parameter  $\sigma$  is automatically selected by a grid search with a 5-fold cross-validation. The experiment in Section VI-B.2 shows that our algorithm is not sensitive to the selection of  $\eta$  in a reasonably wide range (0.1 to 10 times  $\frac{\sqrt{\bar{\lambda}}}{\sigma^2}$ ).

There is an interesting observation about the analytic solution when  $\eta = 0$ . In such case, we maximize  $\hat{f}(s, \mathbf{z}) = \langle \Phi(\hat{\mathbf{x}}) + s\mathbf{w}, \Phi(\mathbf{z}) \rangle$  with respect to  $\mathbf{z}$  directly. Let  $\hat{g}(\mathbf{z}) = k(\hat{\mathbf{x}}, \mathbf{z})$  and  $\hat{h}(\mathbf{z}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{z})$ , we have  $\hat{f}(s, \mathbf{z}) = \hat{g}(\mathbf{z}) + s \cdot \hat{h}(\mathbf{z})$ . Using the idea mentioned above, the optimal  $\mathbf{z}^*$  which maximizes  $\hat{f}(s, \mathbf{z})$  for a given  $s$  is obtained by tracing out the curve  $\mathbf{z}^*(s)$ . It is not difficult to derive that  $\left. \frac{d\mathbf{z}^*}{ds} \right|_{s=0} = -(\hat{\mathbf{H}}_g + s\hat{\mathbf{H}}_h)^{-1}\hat{\mathbf{J}}_h = -(\hat{\mathbf{H}}_g)^{-1}\hat{\mathbf{J}}_h$ , where  $\hat{\mathbf{J}}$  and  $\hat{\mathbf{H}}$  are the Jacobian and Hessian of the functions  $\hat{g}$  and  $\hat{h}$  with respect to  $\mathbf{z}$ , evaluated at  $\mathbf{z} = \hat{\mathbf{x}}$ . Using an RBF kernel, we have  $\hat{\mathbf{H}}_g = -(1/\sigma^2)\mathbf{I}_d$ , which is a multiple of the identity matrix. Therefore the regularized discriminative direction is parallel to the direction of  $\hat{\mathbf{J}}_h$  when  $\eta = 0$ . Note that  $\hat{\mathbf{J}}_h$  is same as the gradient of the classifier function used in [4]. This result coincides with that of Golland’s method in [4]. It indicates that Golland’s method is a special case of our regularized discriminative direction when  $\eta = 0$ .

## B. Regularization via a convex combination of the neighbours

In this section, we provide an alternative method to confine the solution of the preimage  $\mathbf{z}$  to the shape manifold using local geometric restrictions. As mentioned previously in Section V-A, for an RBF kernel, moving  $\Phi(\mathbf{x})$  for a small step always ensures that  $\mathbf{z}$  lies in  $N_\epsilon(\hat{\mathbf{x}})$ , the neighbourhood of  $\hat{\mathbf{x}}$ . Hence here we constrain  $\mathbf{z}$  to be a convex combination of the  $\mathbf{x}_i$  ( $\mathbf{x}_i \in N_\epsilon(\hat{\mathbf{x}})$ ), that is,  $\mathbf{z} = \sum_{i=1}^n \omega_i \mathbf{x}_i$ , where  $\omega_i \geq 0$ ,  $\sum_{i=1}^n \omega_i = 1$ , and  $n$  is the number of neighbours. In other words,  $\mathbf{z}$  lies in the convex hull of  $N_\epsilon(\hat{\mathbf{x}})$ . Then finding the regularized discriminative direction comes down to solving the following optimization problem over the weights  $\omega_i$ , which can be solved by iterative methods:

$$\begin{aligned} \min_{\omega_i} & \|\Phi(\hat{\mathbf{x}}) + s\mathbf{w} - \Phi\left(\sum_{i=1}^n \omega_i \mathbf{x}_i\right)\|^2 \\ & \text{such that } \omega_i \geq 0 \\ & \text{and } \sum_{i=1}^n \omega_i = 1 \quad . \end{aligned} \quad (19)$$

Consider the relationship between the convex combination method in (19) and the local distribution regularized method in Section V-A. Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$  be a sequence of random variables sampled independently and identically from a distribution  $p(\mathbf{X})$  defined over  $N_\epsilon(\hat{\mathbf{x}})$ , taking the values

$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ . Assume  $\mathbf{X}$  has a mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ . Let  $\mathbf{Z}$  denote a random variable  $\sum_i \omega_i \mathbf{X}_i$ , representing the possible positions of the preimage  $\mathbf{z}$  with respect to  $\mathbf{X}_i$ . It can be shown that  $\mathbf{Z}$  has the same mean as the neighbours  $\mathbf{X}_1, \dots, \mathbf{X}_n$ :

$$E[\mathbf{Z}] = E\left[\sum_{i=1}^n \omega_i \mathbf{X}_i\right] = \sum_i \omega_i E[\mathbf{X}_i] = \sum_i \omega_i \boldsymbol{\mu} = \boldsymbol{\mu}, \quad \text{since } \sum_i \omega_i = 1.$$

Therefore,

$$\begin{aligned} Cov(\mathbf{Z}) &= E[(\mathbf{Z} - E[\mathbf{Z}])(\mathbf{Z} - E[\mathbf{Z}])^\top] \\ &= E\left[\left(\sum_i \omega_i (\mathbf{X}_i - \boldsymbol{\mu})\right)\left(\sum_i \omega_i (\mathbf{X}_i - \boldsymbol{\mu})\right)^\top\right] \\ &= \sum_i \sum_j \omega_i \omega_j E[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})^\top]. \end{aligned} \quad (20)$$

Since  $\mathbf{X}_i$  and  $\mathbf{X}_j$  are independent of each other,  $E[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})^\top] = \mathbf{0}$  when  $i \neq j$ . Hence

$$Cov(\mathbf{Z}) = \sum_i \omega_i^2 E[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^\top] = \left(\sum_i \omega_i^2\right) \boldsymbol{\Sigma}.$$

That is,  $Cov(\mathbf{Z})$  has the same eigenvectors as  $\boldsymbol{\Sigma}$  and the eigenvalues which are  $\sum_i \omega_i^2$  times the eigenvalues of  $\boldsymbol{\Sigma}$ . Since  $\omega_i \in [0, 1]$ , according to Cauchy's inequality, we have

$$1/n = \left(\sum_{i=1}^n \omega_i\right)^2/n \leq \sum_{i=1}^n \omega_i^2 \leq 1,$$

showing that the eigenvalues of  $Cov(\mathbf{Z})$  are not larger than the eigenvalues of  $\boldsymbol{\Sigma}$ . Combining the results above, we see that compared with the local distribution regularized method, the distribution of  $\mathbf{Z}$  given by the convex combination method has the same mean, and a covariance matrix with the same eigenvectors but smaller eigenvalues. Hence the convex combination method implicitly makes its solution  $\mathbf{Z}$  comply to the local distribution of its neighbourhood but in a more strict manner.

Compared with the local distribution regularized method, the most salient advantage of the convex combination method is that there is no need to empirically set the parameter  $\eta$ . However its connection to the requirement of complying to the underlying shape distribution is not as obvious as the local distribution regularized method. Further it does not have an analytic solution which the local distribution regularized method has. Its constrained optimization problem has to be solved in an iterative way, which is more computationally expensive. Moreover, to reconstruct the curve of the regularized discriminative direction from a sequence of small steps, the convex combination method may encounter problems when the points in the neighbourhood are sparse. In the worst case, for a small step the neighbours remain unchanged, and so does their convex combination. The curve of the regularized discriminative direction hence cannot be further extended. Increasing the neighbourhood size may mitigate this problem.

Note that in the Kwok-inspired method in Section III, the local geometric relationship is also considered for estimating the deformed shape  $\mathbf{z}$ . There the new shape  $\mathbf{z}$  is confined by its neighbours in the shape descriptor space  $\mathbb{R}^d$  through a distance relationship which is carried from the feature space  $\mathcal{F}$ . However, the constraint based on maintaining the distance relationship between  $\mathcal{F}$  and  $\mathbb{R}^d$  does not confine the deformation tightly enough as our convex combination method does.

Our experiment in Section VI-B.1 demonstrates the advantages of the convex combination method over the Kwok-inspired method.

## VI. Experiments

Different types of controlled experiments are conducted to test the performance of the variants of the discriminative direction proposed in this paper, such as the comparability between our reformulation of the discriminative direction and Golland’s original method, the fidelity of the deformed shapes obtained by our regularized discriminative direction methods, and the sensitivity of our methods to the regularization parameter  $\eta$ . But more importantly, and also being our ultimate target, the proposed regularization method is applied to characterize shape differences in sex for human hippocampi: key class differences are localized.

### A. Comparability of the methods for the original discriminative direction

In this section, we demonstrate that reformulating the original discriminative direction problem as shown in Section III gives comparable solutions to Golland’s method. The residual error in (11) is minimized using the Rathi-inspired method and tested on two classes of points (dark or light in Fig. 5) lying on two concentric annuli respectively. The two classes of points are classified by a support vector machine with an RBF kernel. Sixteen randomly selected support vectors are moved for 10 steps along the direction  $\mathbf{w}$  in  $\mathcal{F}$  toward the opposite class. The result is shown in Fig. 5. Note that in  $\mathcal{F}$  all points that have the same projection value on  $\mathbf{w}$  are located in the same hyperplane perpendicular to  $\mathbf{w}$ . Such a hyperplane is visualized as a level contour in  $\mathbb{R}^2$ . As shown in Fig. 5 (a), the “discriminative directions” obtained by the Rathi-inspired method are consistently radial at different points and best differentiate the two annuli, which agrees well with our intuition and that of Golland’s method (Fig. 5 (b)).

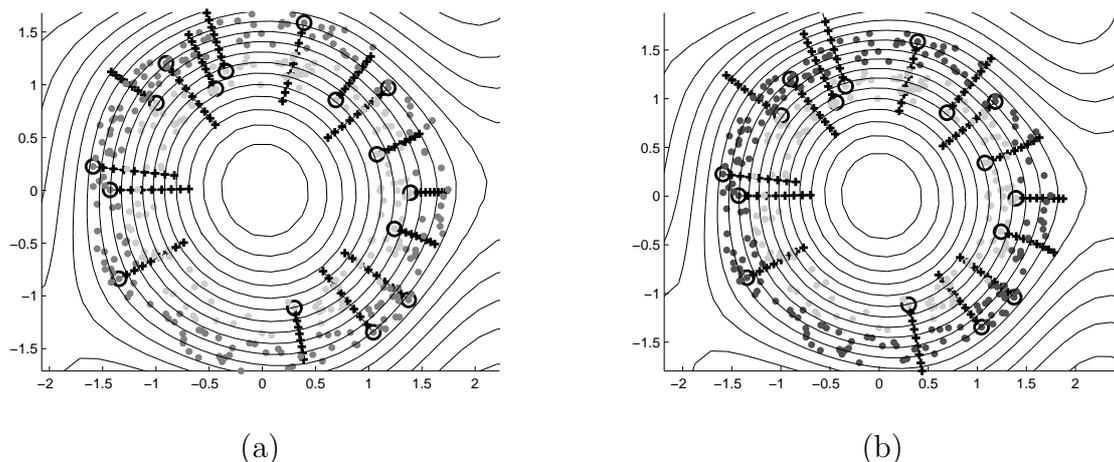


Fig. 5. *Discriminative direction recovered by (a) the Rathi-inspired method and (b) the Golland’s method. The discriminative directions at different points consistently point to the direction that best differentiates the two classes.*

### B. Verification for the regularized discriminative direction

Our main purpose is to use the regularized discriminative direction to localize the class difference for human hippocampal shapes between sexes. This remains an open problem and lacks ground truth. Hence, first we have to verify our proposed methods with data for which we know what kind of deformations to expect, and compare it with Golland’s method.

The verification is performed on the USPS handwritten digit image database and a subset of the UMIST facial image database [17]. The USPS database contains  $16 \times 16$  thumbnail images of ten handwritten digits (0-9). For the UMIST database, we manually labelled 55 images belonging to 8 persons as left-side view and right-side view. Each image is represented by a high-dimensional feature vector comprising all pixels, similar to the landmark representation of shapes. The images

are known to only reside in a low-dimensional manifold [18]. In our local distribution regularized method, a neighbourhood size of 20 is used for the USPS data, and 5 for the UMIST data. We aim to discriminate (i) the shapes of two groups of digits, and (ii) two classes of human faces: left-side view and right-side view. In the experiments, a particular feature vector is moved from one class towards the other along the discriminative direction. Note that the newly generated images in this course do not exist in the database.

### B.1 Fidelity of the deformed images

Fig. 6 is the result on USPS. As shown, Golland’s method introduces much more noise (spurious difference), while our regularized method localizes the discrimination well, adding the minimum necessary shape changes. The advantages of the regularized method are more obvious on UMIST data as shown in Fig. 7. During deformation, it only introduces the class difference (the change of view), leaving the individual variability (the owner of the face) unchanged (Fig. 7 (c)). Most importantly, the newly generated images remain faces. However Golland’s method cannot guarantee this (Fig. 7 (a)), and the authentic difference is overwhelmed by noise. Fig. 7 (b) shows the result obtained by the “tangent plane projection” (see Footnote 3, Page 110). It is better than Golland’s method, but still worse than the local distribution regularized method (see the ghost around the glasses). This demonstrates the benefit of using  $\hat{\Lambda}$  in the local distribution regularized method, as shown in (14).

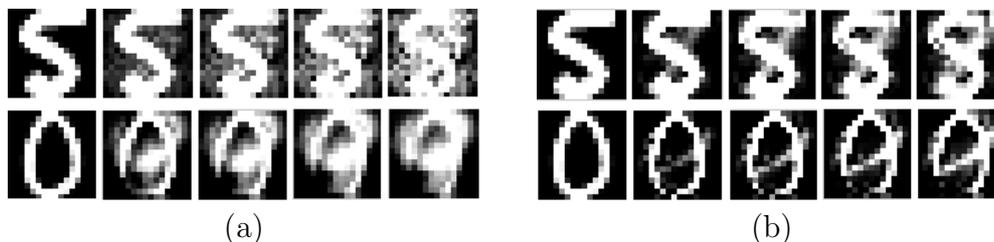


Fig. 6. Verification on the USPS data by (a) Golland’s method and (b) our regularized method. The top row shows the deformation from digit 5 to digit 8. The bottom row shows the deformation from digit 0 to digit 9.



Fig. 7. Verification on the UMIST data. (a) Golland’s method, (b) the tangent plane projection (see Footnote 3, Page 110), (c) the regularized method. During the deformation, a right-side view face (the leftmost image) turns towards left gradually (keep adding class difference) while remaining to be a face image of the same person (filter individual variability) in (b) and (c). However Golland’s method in (a) does not guarantee this.

In addition to the visual comparison above, quantitative analysis has also been performed on the UMIST data to determine how likely a newly generated image belongs to the distribution of the training images in the regularized method, Golland’s method and the tangent plane projection method

respectively. In this paper, a one-class SVM [19] which estimates the probability density is used for this purpose. Given a set of  $n$  training data, one-class SVMs infer a function  $f$  which has positive values in the region where most of the training data is distributed (a percentage is preset by the user) and has negative values elsewhere. The function  $f$  can be expressed as  $f(\mathbf{z}) = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{z}) - \rho$ . Its value decreases from the densely distributed areas to the sparsely distributed areas. The values of  $\alpha_i$  and  $\rho$  are learned by one-class SVMs and  $k(\cdot, \cdot)$  is the employed kernel function. In our experiment, we learn the distribution using an RBF kernel from the training images in both classes, and test on a left-side view face deformed in 45 steps. The corresponding values of the  $f$  function are calculated for each newly generated facial image during the deformation. Fig. 8 clearly demonstrates that the regularized method has larger  $f$  value (mostly positive) than Golland’s method (mostly negative) for the image deformed at every step, which indicates that a new image generated by the regularized method more likely conforms to the distribution of the original images. The performance of the tangent plane projection method is in the middle.

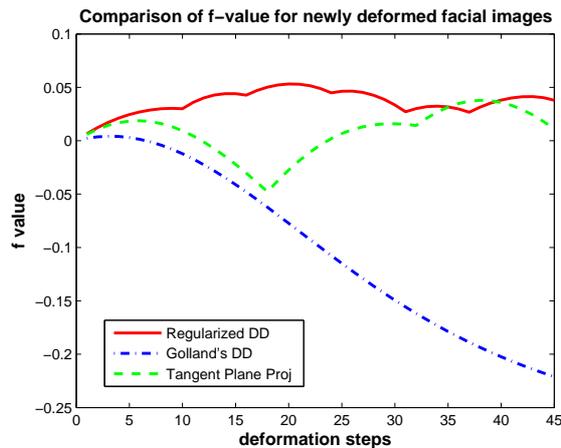


Fig. 8. Comparing the values of  $f$  function in a one-class SVM on the newly deformed facial images obtained by the regularized method, Golland’s method, and the tangent plane projection respectively. As shown, the deformed images obtained by the regularized method (red solid line) have highest  $f$  values, indicating that these images better conform to the distribution of the original facial images compared with the deformed images obtained by the other two methods.

Similarly, we compare the performance of the convex combination method with the Kwok-inspired method because both methods use the neighbouring information as constraints. This is shown in Fig. 9 and Fig. 10. It can be seen that the Kwok-inspired method generates ghost faces which are hard combinations of different views of faces at the intermediate steps, though the final result is comparable to that of the convex combination method. In contrast, the convex combination always generates clear faces, giving smooth transitions at the intermediate steps. The change of the  $f$  value in Fig. 10 affirms this result. The convex combination has an obvious advantage over the Kwok-inspired method in the middle part of the curves.

## B.2 Sensitivity to the regularization parameter

To measure how the regularization parameter  $\eta$  affects the performance of the local distribution regularized method proposed in Section V-A, we conduct experiments as described below. Firstly, we compute the Euclidean distance between each face image in the UMIST database and its nearest neighbour and average such a distance over all the faces belonging to the same person. This distance indicates the average distance between two neighbouring face image descriptors belonging to the same person and is called “average minimum neighbouring distance” in this paper. A left-side view face is deformed for 45 steps using three different values of  $\eta$  respectively, i.e.  $\eta = 0.1 \times \eta_0$ ,  $\eta = \eta_0$ , and  $\eta = 10 \times \eta_0$ , where  $\eta_0 = \frac{\sqrt{\lambda}}{\sigma^2}$  (see Section V-A.2). At each of the 45 steps, we obtain three deformed



Fig. 9. Comparison of the convex combination method (b) and the Kwok-inspired method (a) on the UMIST data.

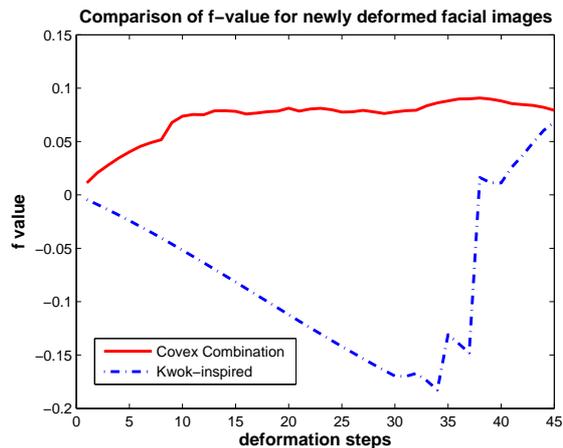


Fig. 10. Comparing the values of  $f$  function in a one-class SVM on the newly deformed facial images obtained by the convex combination and the Kwok-inspired method. As shown, the deformed images obtained by the convex combination method (red solid line) have higher  $f$  values, indicating that these images better conform to the distribution of the original facial images than the deformed images obtained by the Kwok-inspired methods.

shapes by using each of the three different values of  $\eta$ . The pairwise Euclidean distances among these three deformed faces are computed. The results for all the 45 steps are plotted in Fig 11 as three curves, i.e., the distances between the corresponding deformed shapes obtained using (a)  $\eta = \eta_0$  and  $\eta = 10 \times \eta_0$  (indicated by the magenta dashed line), (b)  $\eta = 10 \times \eta_0$  and  $\eta = 0.1 \times \eta_0$  (indicated by the blue dashed line), and (c)  $\eta = \eta_0$ , and  $\eta = 0.1 \times \eta_0$  (indicated by the green dashed line). Notice that even when changing the value of  $\eta$  from  $0.1 \times \eta_0$  to  $10 \times \eta_0$ , which is 100 times larger, the pairwise distance between two deformed images (the blue dashed line) at the corresponding steps is no more than 2.1332. It is still lower than the average minimum neighbouring distance 3.2985 (the red solid line) for the same person. Hence the change in performance of the local distribution regularized method is not significant with respect to different values of  $\eta$  within a reasonably large range.

### C. Hippocampal shape analysis using the regularized discriminative direction

The hippocampus serves as a biomarker for ageing disease and is involved in neurodevelopmental processes. It has aroused great interest in neuroscience. Though the commonly used volumetric measures can provide some indication of normal variation and anomaly, they lack sensitivity and specificity. Therefore the analysis on the hippocampal anatomy is of great significance. Using our proposed regularized discriminative direction methods, we now investigate the shape difference of hippocampi between sex. This is part of a longitudinal study, PATH through Life, from the Centre for Mental Health Research, the Australian National University.

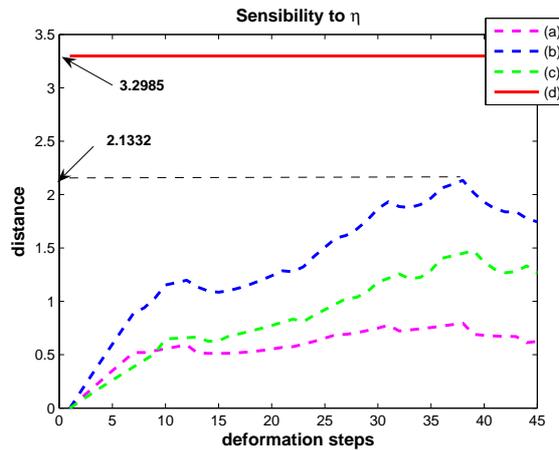


Fig. 11. The distances between the corresponding pairs of deformed facial images obtained at each deformation step using different  $\eta$ : (a)  $\eta = \eta_0$  and  $\eta = 10 \times \eta_0$ , indicated by the magenta dashed line; (b)  $\eta = 10 \times \eta_0$  and  $\eta = 0.1 \times \eta_0$ , indicated by the blue dashed line; and (c)  $\eta = \eta_0$  and  $\eta = 0.1 \times \eta_0$ , indicated by the green dashed line. The average minimum neighbouring distance is indicated by (d), the red solid line. It can be seen that when changing the value of  $\eta$  from  $0.1 \times \eta_0$  to  $10 \times \eta_0$ , which is 100 times larger, the pairwise distance between two deformed images (the blue dashed line) at the corresponding steps is no more than 2.1332. It is still lower than the average minimum neighbouring distance 3.2985 (the red solid line).

## C.1 Data

A database of left hippocampi of healthy individuals is used, which comprises 219 females and 181 males in an age span of 40-44. The hippocampal surfaces have been hand-traced using the Watson *et al.* protocol [20]. They contain the head and tail, but not the posterior section of the tail. Each shape is represented by landmarks reconstructed from the spherical harmonics (SPHARM) representation [21], [22], [23], [24] of different degrees. They are normalized with respect to volume. Using high degree SPHARM, more details appear in the shape model, but the chances of including noise also increase. As shown in our previous work [25], using SPHARM degree 5, we are already able to detect statistically significant ( $\alpha > 0.005$ ) difference between sex by a fast permutation/bootstrap test [25], [26]. In our experiment below, each shape is represented by 642 landmarks obtained from SPHARM degree 5 with established landmark correspondence. To remove the translation and rotation, hippocampal shapes are represented by ellipsoids with degree one SPHARM expansion, and then aligned by the three axes provided by the ellipsoids.

An SVM classifier with the RBF kernel is employed for classification. Following [4], support vectors are selected from the input shape descriptors to study the discriminative direction. A neighbourhood size of 20 is used in the local distribution regularized method.

## C.2 Result

The localized difference between shapes is shown in Fig. 12. These hippocampi belong to six individuals (a column for each one), three females and three males. The colour code indicates the nature of deformation that an actual hippocampal shape undergoes to become a shape akin to the opposite class. Take the leftmost hippocampus in Fig. 12 for example. To make this female hippocampus to be male-like, the blue areas should shrink. As observed, the shape changes are not uniform over the whole hippocampus: small changes (either compression or expansion, in green colour) occur on most of the shape, while sharp changes are localized on the head and the tail. Comparing the deformations in both ways (female to male and vice versa), the regularized method consistently captures the compression in the lateral parts at the head and the tail for male hippocampi. Compared with Golland's method which shows a different pattern (a compression next to an expansion in the head), our results are also

more compact, with changes concentrated in fewer regions but at greater magnitude. Interestingly, the work in [27] has reported findings similar to that of our method. In [27], the hippocampal shapes are represented by medial models [28], [29], which are totally different from our SPHARM-based shape descriptors. Shape difference for sex is observed and estimated to correspond to volume loss in males in the lateral areas of the hippocampus head and tail. This change happens in young male adults, a phenomenon not observed in females. This finding supports that of our regularized method.

Quantitative analysis similar to that of the verification on UMIST face data is performed on the hippocampal shapes deformed from 18 support vectors. The result is summarized in Fig. 13. Each shape descriptor is moved for 35 small steps. The results at the final step are compared here. To achieve an accurate estimation of the probability distribution for the hippocampal shapes, a relatively small  $\sigma$  is used in the RBF kernel for training the one-class SVM. This produces a complex decision boundary which tightly encloses these shapes by the one-class SVM. In this case, all the shapes including the deformed ones are close to the decision boundary where the values of  $f$  function is zero. This is why the  $f$ -values in Fig. 13 are small. However, comparing the relative relationship between the  $f$ -values, Fig. 13 still demonstrates that our regularized discriminative direction gives higher values for most of the cases, indicating that the deformed shapes better conform to the underlying distribution than those obtained by Golland's method.

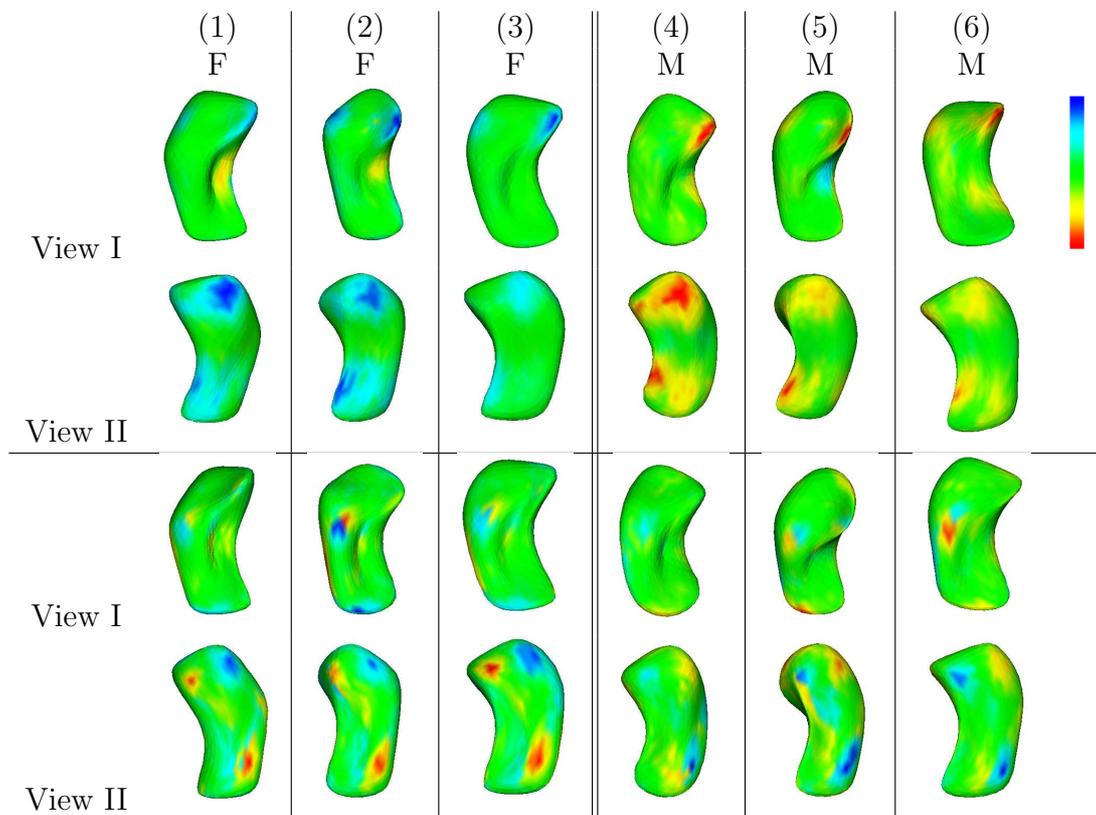


Fig. 12. Localized discrimination for sex on hippocampi of six individuals (three females on the left, three males on the right) from two perspective of views. The top two rows are generated by our regularized method, while the bottom two rows are generated by Golland's method. The colour code indicates the deformation that a female/male hippocampus undergoes to become a male-like/female-like one. Green indicates small shape change. From green to red, the amount of protrusion increases. From green to blue, the amount of shrinkage increases.

## VII. Conclusion

In this paper, we conducted a thorough study of the discriminative direction method which is used for localizing and visualizing differences between groups of anatomical shapes. We approached the established method from a different perspective, pointed out the limitations of this method, proposed

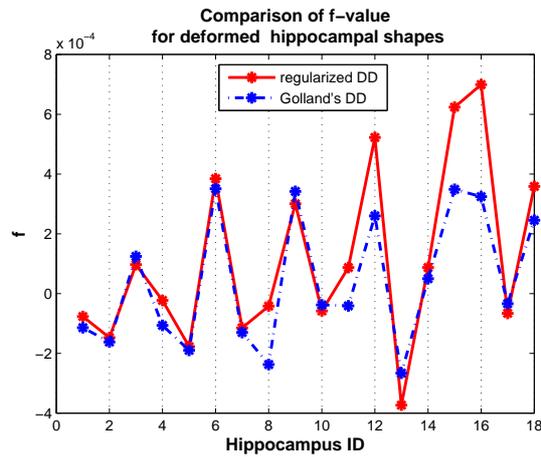


Fig. 13. The values of  $f$  function in a one-class SVM on 18 support vectors in the hippocampus database are compared between the local distribution regularized method and Golland's method. The hippocampal shapes are represented by SPHARM degree 5 and normalized with respect to volume. It can be seen that for most of the cases the deformed hippocampal shapes obtained by our local distribution regularized method have higher  $f$ -values (the red line) than those deformed shapes obtained by Golland's method (the blue line).

and demonstrated two new regularized variants of the established method that respect the underlying distribution of the shapes as well as capturing the essential class difference. The result on controlled experiments shows a significant improvement in the fidelity of the generated shapes of our approaches. Finally, our proposed regularized discriminative direction approach is applied to studying the sex difference of hippocampal shapes, localizing the key difference at the lateral parts of the head and the tail. More applications are expected in our future work.

#### ACKNOWLEDGMENTS

National ICT Australia is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australia Research Council. The authors thank the PATH research team at the center for Mental Health Research, ANU, Canberra, and the Neuroimaging Group (Neuropsychiatric Institute), Prince of Wales Hospital, Sydney, for providing the original MR and segmented data sets.

#### REFERENCES

- [1] Maller J. J., Réglade-Meslin C., Anstey K. J., and Sachdev P., "Sex and symmetry differences in hippocampal volumetrics: Before and beyond the opening of the crus of the fornix," *Hippocampus*, vol. 16, pp. 80–90, 2006.
- [2] N. Cristianini and J. Shawe-Taylor, "An Introduction to Support Vector Machines," *Cambridge University Press*, 2000.
- [3] S. Mika, G. Raetsch, J. Weston, B. Schoelkopf, and K-R. Mueller, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, 1999, pp. 41–48.
- [4] P. Golland, W.E. Grimson, M.E. Shenton, and R. Kikinis, "Detection and analysis of statistical differences in anatomical shape," *Med. Image Analysis*, vol. 9, no. 1, pp. 69–85, 2005.
- [5] S. Kodipaka, B.C. Vemuri, A. Rangarajan, C.M. Leonard, I. Schmallfuss, and S. Eisenschenk, "Kernel fisher discriminant for shape-based classification in epilepsy," in *MICCAI 2007*, 2007, pp. 79–90.
- [6] Polina Golland, "Discriminative direction for kernel classifiers," in *Advances in Neural Information Processing Systems (NIPS)*, 2001, pp. 745–752.
- [7] P. Golland, B. Fischl, M. Spiridon, N. Kanwisher, R. L. Buckner, M. E. Shenton, R. Kikinis, A. Dale, and W. E. Grimson, "Discriminative analysis for image-based studies," in *MICCAI 2002*, 2002, pp. 508–515.
- [8] James T. Kwok and Ivor W. Tsang, "The pre-image problem in kernel methods," *IEEE Transactions on Neural Networks*, vol. 15, no. 6, pp. 1517–1525, 2004.
- [9] Rathi Y., Dambreville S., and Tannenbaum A., "Statistical shape analysis using kernel PCA," in *Proceedings of SPIE Electronic Imaging 2006*, 2006, pp. 425–432.
- [10] de Leon M. J., George A. E., Golomb J., Tarshish C., Convit A., Kluger A., DeSanti S., McRae T., Ferris S. H., Reisberg B., Ince C., Rusinek H., Bobinski M., Quinn C. B., Miller B. L., and Wisniewski H., "Frequency of hippocampal formation atrophy in normal aging and Alzheimer's disease," *Neurobiology of Aging*, vol. 18, pp. 1–11, 1997.
- [11] Kabani N. J., Sled J. G., Shuper A., and Chertkow H., "Frequency of hippocampal formation atrophy in normal aging and Alzheimer's Disease," *Magnetic Resonance in Medicine*, vol. 47, pp. 143–148, 2002.

- [12] Mega M. S., Small G. W., Xu M. L., Felix J., Manese M., Tran N. P., Dailey J. I., Ercoli L. M., Bookheimer S. Y., and Toga A. W., "Frequency of hippocampal formation atrophy in normal aging and Alzheimer's Disease," *Psychosomatic Medicine*, vol. 64, pp. 487–492, 2002.
- [13] L. Zhou, R. Hartley, L. Wang, P. Lieby, and N. Barnes, "Regularized discriminative direction for shape difference analysis," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2008 (accepted)*, 2008.
- [14] B. Schoelkopf, S. Mika, A. Smola, G. Raetsch, and K-R. Mueller, "Kernel PCA pattern reconstruction via approximate pre-images," in *Proceedings Eighth International Conference Artificial Neural Networks*, 1998, pp. 147–152.
- [15] S. Mika, B. Schoelkopf, A. J. Smola, K-R. Mueller, M. Scholz, and G. Raetsch, "Kernel PCA and de-noising in feature spaces," in *Proceedings of Advances in Neural Information Processing Systems*, 1999, pp. 536–542.
- [16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. Wiley-Interscience, November 2000.
- [17] D. B. Graham and N. M. Allinson, "Face recognition: From theory to applications," *NATO ASI Series F, Computer and Systems Sciences*, vol. 163, pp. 446–456, 1998.
- [18] Tenenbaum J. B., de Silva V., and Langford J. C., "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.
- [19] Bernhard Scholkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1471, 2001.
- [20] Watson C., Jack CR Jr., and Cendes F., "Volumetric magnetic resonance imaging. clinical applications and contributions to the understanding of temporal lobe epilepsy," *neuroblock Archives of Neurology*, vol. 54, pp. 1521–1531, 1997.
- [21] A. Kelemen, G. Szekely, and G. Gerig, "Elastic model-based segmentation of 3-D neuroradiological data sets," *IEEE Trans. on Medical Imaging*, vol. 18, no. 10, pp. 828–839, 1999.
- [22] G. Gerig, M. Styner, D. Jones, D. Weinberger, and J. Lieberman, "Shape analysis of brain ventricles using spharm," in *Proceedings of Workshop on Mathematical Methods in Biomedical Image Analysis MMBIA*, 2001, pp. 171–178.
- [23] G. Gerig, M. Styner, M.E. Shenton, and J. Lieberman, "Shape versus size: Improved understanding of the morphology of brain structures," in *Proceedings of Fourth International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI'2001)*, 2001, pp. 24–32.
- [24] L. Shen, J. Ford, F. Makedon, and A. Saykin, "Hippocampal shape analysis surface-based representation and classification," in *Proceedings of SPIE-Medical Imaging*, 2003, pp. 253–264.
- [25] L. Zhou, R. Hartley, P. Lieby, N. Barnes, K. Anstey, N. Cherbuin, and P. Sachdev, "A study of hippocampal shape difference between genders by efficient hypothesis test and discriminative deformation," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2007, pp. 375–383.
- [26] P. M. Gill, "Efficient calculation of  $p$ -values in linear-statistic permutation significance tests," *Journal of Statistical Computation and Simulation*, vol. 77, no. 1, pp. 55–61, 2007.
- [27] Bouix S., Pruessner J. C., Collins D. L., and Siddiqi K., "Hippocampal shape analysis using medial surfaces," *NeuroImage*, vol. 25, pp. 1077–1089, 2005.
- [28] M. Styner, J. Lieberman, and G. Gerig, "Boundary and medial shape analysis of the hippocampus in schizophrenia," in *MICCAI, vol. 2*, 2003, pp. 464 – 471.
- [29] M. Styner, J. Lieberman, D. Pantazis, and G. Gerig, "Boundary and medial shape analysis of the hippocampus in schizophrenia," *Medical Image Analysis*, vol. 8, no. 3, pp. 197 – 2003, 2004.