# Is Pinocchio's Nose Long or His Head Small? Learning Shape Distances for Classification

Daniel Gill[1], Ya'acov Ritov[1], and Gideon Dror[2]

[1] Department of Statistics, The Hebrew University, Jerusalem 91905, Israel
`gill@mta.ac.il, yaacov@mscc.huji.ac.il`
[2] Department of Computer Science, The Academic College of Tel-Aviv Yaffo,
Tel-Aviv 64044, Israel
`gideon@mta.ac.il`

**Abstract.** This work presents a new approach to analysis of shapes represented by finite set of landmarks, that generalizes the notion of Procrustes distance - an invariant metric under translation, scaling, and rotation. In many shape classification tasks there is a large variability in certain landmarks due to intra-class and/or inter-class variations. Such variations cause poor shape alignment needed for Procrustes distance computation, and lead to poor classification performance. We apply a general framework to the task of supervised classification of shapes that naturally deals with landmark distributions exhibiting large intra class or inter-class variabilty. The incorporation of Procrustes metric and of a learnt general quadratic distance inspired by Fisher linear discriminant objective function, produces a generalized Procrustes distance. The learnt distance retains the invariance properties and emphasizes the discriminative shape features. In addition, we show how the learnt metric can be useful for kernel machines design and demonstrate a performance enhancement accomplished by the learnt distances on a variety of classification tasks of organismal forms datasets.

## 1 Introduction

The mathematical notion of shape is an equivalence class under certain type of group of transformations. The most common transformations are: translation, scaling, and rotation. This definition refers only to the question whether two shapes are identical, but in many cases we want to measure shape similarity or shape distance. Shape definitions in statistics were given by Bookstein [1] and Kendall [2], whose attitudes assume that correspondences between the two shapes are known. These latter approaches make sense while assuming that the two shapes are similar and have homologues features. A common and useful representation of planar shape is by landmark points. This approach can be easily extended to 3D shapes.

Parsimonious representation by landmarks has its advantages from the computational point of view and is very useful in many computer vision applications. Indeed, geometric features can represent the shape and location of facial components and are used in face analysis and synthesis [3]. Landmark analysis is

also being used in medical imaging, robotics, dental medicine, anthropology, and many more applications.

In a supervised learning setting a desirable metric for shape classification should not only satisfy certain invariance properties but also capture the discriminative properties of the inputs. In this paper we present a learning algorithm which produces a metric, that satisfies these demands. Moreover, we show how this metric can be used for the design of kernel machines for shape classification.

## 2    Shape Space and Distances

A natural choice of landmarks is a finite set of particularly meaningful and salient points which can be identified by computer and humans. Several types of landmarks were suggested in previous works (see [1]). In the general case, there is a considerable loss of information by extracting only landmarks, and the transformed shape cannot be restored exactly from the landmarks. Yet, many essential characteristics may remain in such representation. A set of $k$ ordered landmark points in 2D plane can be represented as a $2k$-dimensional vector. Comparing two shapes is usually based on corresponding landmarks which are termed *homologies*.

The general notion of distance (or similarity) between two shapes is quite vague. This term can be easily defined when using $2k$-vectors by taking only their coordinates as attributes. It is obvious that the order of the landmarks matters. Another convenient representation is called planar-by-complex and uses complex values to represent each 2-dimensional landmark point, so the whole shape is represented as an $n \times 1$ complex vector. The configuration matrix is a $k \times m$ matrix of real Cartesian coordinates of $k$ landmarks in an $m$-dimensional Euclidian space. In a planar-by-complex representation the configuration is a $k$ dimensional column vector of complex entries. From now on we will assume that all the shapes we deal with are two-dimensional and are given in the planar-by-complex representation.

### 2.1    Shape Metric

A desired distance measure between two planar landmark based shapes should be insensitive to translation, scaling and rotation. Consider a configuration $\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k) \in \mathcal{C}^k$, a centered configuration $\mathbf{x}$ satisfies $\mathbf{x}^* \mathbf{1}_k = 0$, which is accomplished by: $\mathbf{x} \to \mathbf{x} - \mathbf{1}_k^T \mathbf{x} \mathbf{1}_k$, where $\mathbf{x}^*$ denotes the complex conjugate of $\mathbf{x}$.

The *full Procrustes fit* of $\mathbf{x}$ onto $\mathbf{y}$ is: $\mathbf{x}^P = (\widehat{a} + i\widehat{b})\mathbf{1}_k + \widehat{\beta}e^{i\widehat{\vartheta}}\mathbf{x}$ where the parameters values $(\widehat{a}, \widehat{b}, \widehat{\beta}, \widehat{\vartheta})$ are chosen to minimize the Euclidean distance between $\mathbf{y}$ and the transformed configuration of $\mathbf{x}$, and their values are (see [4]):

$$\widehat{a} + i\widehat{b} = 0, \ \widehat{\vartheta} = \arg(\mathbf{x}^* \mathbf{y}), \ \widehat{\beta} = \frac{(\mathbf{x}^* \mathbf{y} \mathbf{y}^* \mathbf{x})^{\frac{1}{2}}}{\mathbf{x}^* \mathbf{x}} \tag{1}$$

Removing the similarity operations from a k-landmark planar configuration space leaves a $2k - 4$ dimensional shape space manifold (2 dimensions for translation, one dimension for scaling, and one dimension for rotation).

The *full Procrustes distance* between two configurations $\mathbf{x}$ and $\mathbf{y}$ is given by:

$$d_F(\mathbf{x}, \mathbf{y}) = \inf_{\beta, \vartheta, a, b} \left\| \frac{\mathbf{y}}{\|\mathbf{y}\|} - \frac{\mathbf{x}}{\|\mathbf{x}\|} \beta e^{i\vartheta} - a - bi \right\| = \left( 1 - \frac{\mathbf{y}^* \mathbf{x} \mathbf{x}^* \mathbf{y}}{\mathbf{x}^* \mathbf{x} \mathbf{y}^* \mathbf{y}} \right)^{\frac{1}{2}}. \qquad (2)$$

The *full Procrustes mean shape* $\widehat{\boldsymbol{\mu}}$ of a set of configurations $\{\mathbf{w}_{i=1}^n\}$ is the one that minimizes the sum of square full Procrustes distances to each configuration in the set, i.e.

$$\widehat{\boldsymbol{\mu}} = \arg\inf_{\boldsymbol{\mu}} \sum_{i=1}^n d_F^2(\mathbf{w}_i, \boldsymbol{\mu}). \qquad (3)$$

It can be shown that the full Procrustes mean shape, $\widehat{\boldsymbol{\mu}}$, is the eigenvector corresponding to the largest eigenvalue of the following matrix:

$$S = \sum_{i=1}^n \frac{\mathbf{w}_i \mathbf{w}_i^*}{\mathbf{w}_i^* \mathbf{w}_i}. \qquad (4)$$

(see [5]). The eigenvector is unique (up to rotations - all rotations of $\widehat{\boldsymbol{\mu}}$ are also solutions, but these all correspond to the same shape) provided there is a single largest eigenvalue of S. In many morphometric studies several configurations are handled and pairwise fitted to a single common consensus in an iterative procedure [6]. This process is called *generalized Procrustes analysis*. Scatter analysis, using generalized Procrustes analysis handle the superimposed configurations in an Euclidean manner and provide good linear approximation of the shape space manifold in cases where the configurations variability is small.

Though Procrustes distances and least-squares superimpositions are very common, they can sometimes give a misleading explanation of the differences between a pair of configurations [6,7], especially when the difference is limited to a small subset of landmarks. The Procrustes superimposition tends to obtain less extreme magnitudes of landmark shifts. The fact that in least-squares superimposition landmarks are treated uniformly irrespective of their variance results in poor estimation, and reaches its extreme when all of the shape variation occurs at a single landmark, which is known as the *Pinocchio effect* [8]. This effect is demonstrated in Fig. 1. Due to proportions conservation, trying to minimize the sum-of-squares differences affects all landmarks and thus tilts the head and diminishes its size. Moreover, such variations affect the configuration's center of mass and thus affect translation as well. Actually, the Procrustes fit does not do what would have been expected from pre-classification alignment to do. A desirable fit would be an alignment that brings together non-discriminative landmarks and separates discriminative landmarks, and, in addition, gives appropriate weights for the features according to their discriminative significance.

## 2.2   General Quadratic Shape Metric

A general quadratic distance metric, can be represented by a symmetric positive semi-definite $k \times k$ matrix $Q$ (we use the $Q = A^* A$ decomposition and estimate
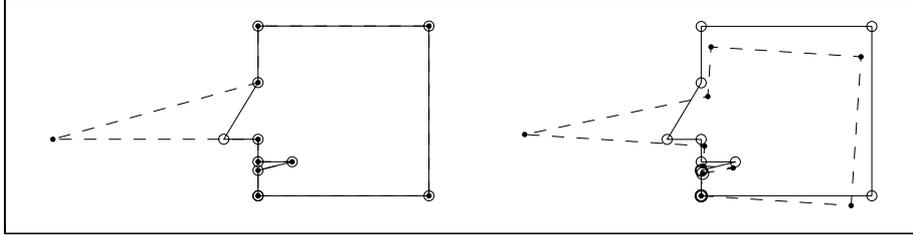
**Fig. 1.** Pinocchio effect. Two faces which differ only by the tip of their nose are super-imposed by their similar features (*left*). Minimization of the sum-of-squares differences affects all landmarks of the longed-nose face: diminishes the head size and tilts it (*right*).

A). Centering a configuration $\mathbf{x}$ according to the metric induced by $Q$ means that $\mathbf{x}^*Q\mathbf{1}_k = 0$, and this is done by: $\mathbf{x} \rightarrow \mathbf{x} - \mathbf{1}_k^T Q\mathbf{x}\mathbf{1}_k$. For the rest of this section, we assume that all configurations are centered according to the metric induced by $Q$.

The *general quadratic full Procrustes fit* of $\mathbf{x}$ onto $\mathbf{y}$ is:

$$\mathbf{x}^P = (\widehat{a}^Q + i\widehat{b}^Q)1_k + \widehat{\beta}^Q e^{i\widehat{\vartheta}^Q}\mathbf{x} \tag{5}$$

where the parameters values $(\widehat{a}^Q, \widehat{b}^Q, \widehat{\beta}^Q, \widehat{\vartheta}^Q)$ are chosen to minimize:

$$D_Q^2(\mathbf{x}, \mathbf{y}) = \left\| A\mathbf{y} - A\mathbf{x}\beta e^{i\vartheta} - A(a + bi)\mathbf{1}_k \right\|^2. \tag{6}$$

**Claim 1.** *The minimizing parameters* $(\widehat{a}^Q, \widehat{b}^Q, \widehat{\beta}^Q, \widehat{\vartheta}^Q)$ *values are:*

$$\widehat{a}^Q + i\widehat{b}^Q = 0, \ \widehat{\vartheta}^Q = \arg(\mathbf{x}^*Q\mathbf{y}), \ \widehat{\beta}^Q = \frac{(\mathbf{x}^*Q\mathbf{y}\mathbf{y}^*Q\mathbf{x})^{\frac{1}{2}}}{\mathbf{x}^*Q\mathbf{x}}, \tag{7}$$

(the proof is similar to the Euclidean case).

The *general quadratic full Procrustes distance*, according to matrix $Q = A^*A$, between two configurations $\mathbf{x}$ and $\mathbf{y}$ is given by:

$$d_Q^2(\mathbf{x}, \mathbf{y}) = \inf_{\beta,\vartheta,a,b} \left\| A\frac{\mathbf{y}}{\|\mathbf{y}\|_Q} - A\frac{\mathbf{x}}{\|\mathbf{x}\|_Q}\beta e^{i\vartheta} - a - bi \right\| \tag{8}$$

$$= \left(1 - \frac{\mathbf{y}^*Q\mathbf{x}\mathbf{x}^*Q\mathbf{y}}{\mathbf{x}^*Q\mathbf{x}\mathbf{y}^*Q\mathbf{y}}\right)^{\frac{1}{2}},$$

where $\|\mathbf{x}\|_Q^2 = \mathbf{x}^*Q\mathbf{x}$ is the square of the generalized norm.

The general quadratic Procrustes mean shape $\widehat{\mu}^Q$, with a matrix $Q = A^*A$, of a set of configurations $\{\mathbf{w}_i\}_{i=1}^n$ is the one that minimizes the sum of square generalized distances to each configuration in the set, i.e.

$$\widehat{\boldsymbol{\mu}}^Q = \arg\inf_{\boldsymbol{\mu}} \sum_{i=1}^n d_Q^2(\mathbf{w}_i, \boldsymbol{\mu}). \tag{9}$$

**Claim 2.** *The general quadratic Procrustes mean shape is the eigenvector corresponding to the largest eigenvalue of the following matrix:*

$$S^Q = \sum_{i=1}^{n} \frac{A\mathbf{w}_i\mathbf{w}_i^* A^*}{\mathbf{w}_i^* A^* A\mathbf{w}_i}, \tag{10}$$

(the proof is similar to the Euclidean case).

## 3   Metric Learning

Many pattern recognition algorithms use a distance or similarity measures over the input space. The right metric should fit the task at hand, and understanding the input features and their importance for the task may lead to an appropriate metric. In many cases there is no such prior understanding, but estimating the metric from the data might result in a better performance than that achieved by off the shelf metrics such as the Euclidean [9,10,11]. Fisher Linear Discriminant (FLD) is a classical method for linear projection of the data in a way that maximizes the ratio of the between-class scatter and the within-class scatter of the transformed data (see [12]).

Given a labeled data set consisting of 2D input configurations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ where $\mathbf{x}_i \in \mathcal{C}^k$ and corresponding class labels $c_1, c_2, \ldots, c_n$, we define between-class scatter and within-class scatter both induced by the metric $Q$. In a similar way to FLD the desired metric $Q$ is the one that maximizes the ratio of the generalized between-class and within-class scatters.

We denote the general quadratic Procrustes mean shape of the members of class $j$ by $\widehat{\boldsymbol{\mu}}_j^Q$, and the full general quadratic Procrustes mean shape of all configurations by $\widehat{\boldsymbol{\mu}}^Q$. Denote

$$\Delta_{k,l}^Q = \left( \frac{\mathbf{x}_l}{\|\mathbf{x}_l\|_Q} \widehat{\beta}_{k,l}^Q e^{i\widehat{\vartheta}_{kl}^Q} \right) - \widehat{\boldsymbol{\mu}}_k^Q \tag{11}$$

and

$$\Delta_k^Q = \widehat{\boldsymbol{\mu}}_k^Q \widehat{\beta}_k^Q e^{i\widehat{\vartheta}_k^Q} - \widehat{\boldsymbol{\mu}}^Q \tag{12}$$

where $\widehat{\beta}_{k,l}^Q, \widehat{\beta}_k^Q$ are the scaling solutions of eq. 8 for the $l$-th configuration towards the mean of class $k$, and scaling of the $k$-th mean configuration towards the global mean respectively. The angles $\widehat{\vartheta}_{kl}^Q, \widehat{\vartheta}_k^Q$ are those which satisfy eq. 8 for rotation the $l$-th configuration towards the mean of class $k$, and rotation of the $k$-th mean configuration towards the global mean correspondently (the translations equal to zero if all configurations are previously centered).

The within class scatter according to a matrix $Q$ is:

$$s_W^Q = \sum_{j=1}^{m} \sum_{i=1}^{n} r_{ij} d_Q^2 \left( \mathbf{w}_i, \widehat{\boldsymbol{\mu}}^Q \right) = \sum_{j=1}^{m} \sum_{i=1}^{n} r_{ij} \left( \Delta_{j,i}^Q \right)^* Q \Delta_{j,i}^Q \tag{13}$$

where

$$r_{kl} = \begin{cases} 1 & \mathbf{x}_l \in \text{Class } k \\ 0 & \text{Otherwise} \end{cases} \tag{14}$$

and $m$ is the number of classes.

The between class scatter according to a matrix $Q$ is:

$$s_B^Q = \sum_{k=1}^{m} n_k \left( \Delta_k^Q \right)^* Q \Delta_k^Q, \tag{15}$$

where $n_k$ is the number of samples belong to class $k$.

The desired metric $Q_{opt}$ is the one that maximizes the ratio of the between-class scatter and within-class scatter:

$$Q_{opt} = \arg\max_Q \frac{s_B^Q}{s_W^Q}. \tag{16}$$

The rank of $Q_{opt}$ is at most $m-1$. Contrary to the standard FLD, the suggested objective function $f$ may have many local maxima. Thus, maximizing the objective function should be carried out carefully, and only a local maximum is guaranteed.

## 4  Procrustes Distance Based Classifiers

One of the goals of distance learning is the enhancement of the performance of classifiers. In recent years, many studies have dealt with the design and analysis of kernel machines [13]. Kernel machines use inner-products functions where the decision function is not a linear function of the data. Replacing the predefined kernels with ones that are designed for the task at hand and are derived from the data itself, is likely to improve the performance of the classifier considerably, especially when training examples are scarce [14]. In this section we introduce new kernels based on the general quadratic full procrustes distance where the learnt metric can be plugged in to produce new kernels with improved capabilities of shape classification.

### 4.1  General Quadratic Procrustes Kernels

Certain condition has to be fulfilled for a function to be a dot product in some high dimensional space (see Mercer's theorem [13]). Following the polynomial and radial basis function (RBF) kernels, we propose the following kernels.

**Claim 3.** *The following function is an inner product kernel for any positive integer p:*

$$k(\mathbf{x}, \mathbf{y}) = \left( \frac{\mathbf{y}^* Q \mathbf{x} \mathbf{x}^* Q \mathbf{y}}{\mathbf{x}^* Q \mathbf{x} \mathbf{y}^* Q \mathbf{y}} \right)^p \tag{17}$$

For proof outline see appendix A.

**Claim 4.** *The following function is an inner product kernel for any positive semi-definite matrix $Q$ and any positive $\gamma$:*

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\gamma\left(1 - \frac{\mathbf{y}^*Q\mathbf{x}\mathbf{x}^*Q\mathbf{y}}{\mathbf{x}^*Q\mathbf{x}\mathbf{y}^*Q\mathbf{y}}\right)\right) \tag{18}$$

For proof see appendix B.

## 5   Experimental Results

The main role of the general quadratic Procrustes metric described in the previous section is to align configurations in a way that reveals the discriminative features. Most of the datasets we examined were taken from the *shapes package* (http://www.maths.nott.ac.uk/personal/ild/shapes/), and they consist of organismal forms data. We used six datasets where the samples are represented by configurations of 2D landmarks, and each dataset is made of two categories. The datasets are: gorilla midline skull data (8 landmarks 30 females and 29 males), chimpanzee skull data (8 landmarks, 26 females and 28 males), orang utan skull data (8 landmarks 30 females and 30 males), mouse vertebrae (6 landmarks, 23 large mice and 23 small mice), landmarks taken in the near midline from MR images of the brain (13 landmarks 14 subjects diagnosed with schizophrenia and 14 normal subjects). In addition, we used a facial dataset consists of landmarks taken from frontal images of 32 males and 78 females - all with neutral expression. The extraction of the landmarks from the facial images was done by the Bayesian Tangent Shape Model (BTSM) [15].

Figure 2 uncovers discriminative landmarks in facial configurations means. The general quadratic Procrustes mean shape of females' faces is fitted using the learnt metric (general quadratic Procrustes fit) onto the general quadratic Procrustes mean shape of males' faces. It is evident that the learnt metric reveals differences between the two classes. The males' mandibles tend to be larger, and their forehead hairlines tend to be higher than those of females. These differences are not revealed when using the standard Procrustes metric.

The contribution of the Procrustes kernels and the learnt metric was evaluated by the leave-one-out error rate of three classifiers:

- SVM with standard RBF kernel where the input configurations are pre-processed by generalized Procrustes analysis onto the training samples full Procrustes mean shape.
- SVM with full Procrustes distance based RBF kernel ($Q = I$).
- SVM with learnt Procrustes distance based RBF kernel (learnt $Q$).

The leave-one-out error rates are given in Table 1. The results demonstrate two things: (i) The Procrustes kernel is preferable over the general Procrustes analysis followed by standard Euclidean based kernel (ii) The learnt metric improves the classifier performance.
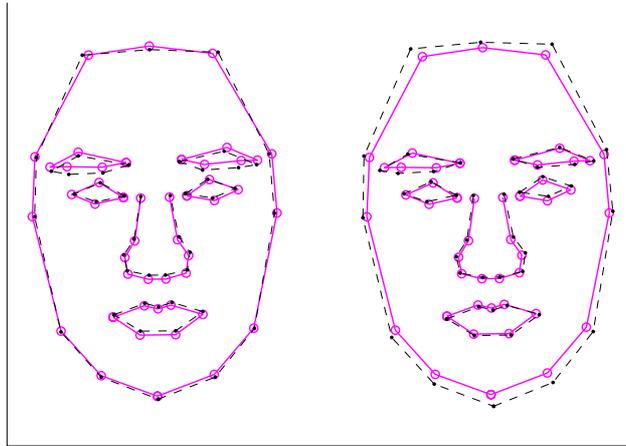
**Fig. 2.** Superimpositions of mean facial configurations: females (*solid line*) and males (*dashed line*) according to the full Procrustes metric (*left*) and the learnt Procrustes metric (*right*)

**Table 1.** Leave-One-Out error rates of the SVM classifiers

| Dataset | Standard RBF | Procrustes Kernel ($Q = I$) | Learnt Procrustes Kernel |
|---|---|---|---|
| Gorilla Skulls | 3.39% | 3.39% | **0%** |
| Mouse Vertebrae | 6.52% | 4.35% | **2.17%** |
| Orang Utan Skulls | 11.11% | 5.56% | **3.70%** |
| Faces | 12.73% | 11.82% | **10.91%** |
| Chimpanzee Skulls | 31.48% | 31.48% | **25.93%** |
| Schizophrenia | 32.14% | 32.14% | **28.57%** |

## 6    Discussion and Conclusions

We have presented an algorithm for learning shape distances, generalizing the Procrustes distance. In the two-classes case, the learnt metric induces a configuration superimposition where weights are assigned to the landmarks according to their discriminative role. Aligning configurations according to the learnt metric enables a visualization that uncovers the discriminative landmarks. Substantial improvement in classification performance was demonstrated by using Procrustes kernel (which keeps the pairwise full Procrustes distances between shapes, where generalized Procrustes analysis does not) and became even more pronounced when plugging in the learnt metric. The main contribution of the learnt metric is the meaningful alignment - it is of particular importance in cases where the training sets are small. Euclidean related kernels cannot learn translation, scaling, and rotation invariants from small data sets. Many types of devices for measuring 3D coordinates are in a wide-spread use: computed tomography

(CT), optical scans of surfaces (laser scanners), etc. All the methods discussed here can easily be extended to handle 3D configurations.

## Acknowledgements

## References

1. Bookstein, F.: Morphometric Tools for Landmark Data: Geometry and Biology. Cambridge University Press, Cambridge (1991)
2. Kendall, D.: Shape manifolds, procrustean metrics, and complex projective spaces. Bull. London Math. Soc. 16, 81–121 (1984)
3. Li, S., Jain, A. (eds.): Handbook of Face Recognition. Springer, Heidelberg (2005)
4. Dryden, I., Mardia, K.: Statistical Shape Analysis, 1st edn. Wiley, Chichester (1998)
5. Kent, J.: The complex bingham distribution and shape analysis. Journal of the Royal Statistical Society Series B 56, 285–299 (1994)
6. Rholf, F., Slice, D.: Extensions of the procrustes method for the optimal superimposition of landmarks. Syst. Zool. 39, 40–59 (1990)
7. Siegel, A., Benson, R.: A robust comparison of biological shapes. Biometrics 38, 341–350 (1982)
8. Chapman, R.: Conventional procrustes approaches. In: Proceedings of the Michigan Morphometrics Workshop, pp. 251–267 (1990)
9. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance metric learning, with application to clustering with side-information. Advances in Neural Information Processing Systems 18 (2004)
10. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. Advances in Neural Information Processing Systems 18 (2004)
11. Globerson, A., Roweis, S.: Metric learning by collapsing classes. Advances in Neural Information Processing Systems 19 (2005)
12. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. John Wiley & Sons, Chichester (2001)
13. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge (2000)
14. Aha, D.: Feature weighting for lazy learning algorithms. In: Liu, H., Motoda, H. (eds.) Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer, Norwell, MA (1998)
15. Zhou, Y., Gu, L., Zhang, H.J.: Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. In: CVPR (2003)

## Appendix A

**Proof Outline:** First we show that the following function is an inner product kernel for any positive integer p:

$$k(\mathbf{x}, \mathbf{y}) = \left( \frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}} \right)^p \tag{19}$$

We have to show that this kernel satisfies Mercer's theorem. This is done by proving that:

$$\int \int \left( \frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}} \right)^p g(\mathbf{x})g^*(\mathbf{y})d\mathbf{x}d\mathbf{y} \geq 0 \tag{20}$$

for any function $g$ with finite $l_2$ norm.

Each term of the multinomial expansion has a non-negative value:

$$(r_1, r_2, \ldots, l_1, l_2, \ldots, )! \left\| \int \left( \frac{x_1^{r_1}, x_2^{r_2} \cdots \overline{x}_1^{l_1}, \overline{x}_2^{l_2} \cdots}{\|\mathbf{x}\|^{2p}} \right) g(\mathbf{x}) \, d\mathbf{x} \right\|^2 \geq 0 \tag{21}$$

and hence the integral is non-negative.

Showing that:

$$k(\mathbf{x}, \mathbf{y}) = \left( \frac{\mathbf{y}^*Q\mathbf{x}\mathbf{x}^*Q\mathbf{y}}{\mathbf{x}^*Q\mathbf{x}\mathbf{y}^*Q\mathbf{y}} \right)^p \tag{22}$$

Satisfies Mercer's theorem is done in a similar way by using eigen-decomposition the non-negativity of $Q$'s eigenvalues. ∎

## Appendix B

**Proof**

$$k(\mathbf{x}, \mathbf{y}) = \exp\left( -\gamma \left( 1 - \frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}} \right) \right) = \exp(-\gamma) \exp\left( \gamma \frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}} \right). \tag{23}$$

The first factor on the right side is positive and the second factor can be arbitrarily close approximated by polynomial of the exponent with positive coefficients, thus using claim 3 we have a sum of semi-definite functions, which is also a semi-definite function. ∎