Learning Facial Expressions: From Alignment to Recognition

Daniel Gill[1,2] and Yaniv Ninio[2]

[1] Department of Statistics, The Hebrew University, Jerusalem 91905, Israel {gill@mta.ac.il}
[2] Department of Computer Science, The Academic College of Tel-Aviv Yaffo, Tel-Aviv 64044, Israel {Yaniv.Ninio@amdocs.com}

Abstract. One of the main challenges in 'real-life' object recognition applications is keeping some invariance properties such as: translation, scaling, and rotation. However, trying to maintain such invariants can impair recognition capabilities, especially when the family of objects of interest has a large shape variability. We present a general family of shape metrics that generalizes Procrustes metric and within this framework learns the desired shape metric parameters from labeled training samples. The learnt distance retains invariance properties on one hand and emphasizes the discriminative shape features on the other hand. We show how these metrics can be incorporated in multi-class classification kernel SVMs. We demonstrate the merits of this approach on multi-class facial expressions recognition using the AR dataset. The results address some questions and cautions regarding the interpretation of classification results when using still images datasets collected in a controlled lab environment and their relevance for 'real-life' applications.

1 Introduction

In recent years both academia and industry have shown growing interest in the development of computer vision systems that can locate human faces and recognize their expressions. One of the challenges of computer vision is to make computers that interact with humans in a natural way, as humans interact with each other. Since spoken language, as one of the most natural ways of humans to interact, is ambiguous, human interaction is also based on body gestures and facial expressions, which transmit implicit information and help to interpret the explicit information.

The context and the information provided implicitly are extremely important for computers to get a full understanding of what is actually transmitted in a conversation. There are many practical uses of recognizing facial expressions, such as:

- Detecting boredom, fatigue or stress while driving a vehicle and providing appropriate alerts will help to prevent many accidents from happening. Facial expression recognition systems, among others, may contribute to the development of various accident-prevention safety on-road systems.
- Automated detectors of fatigue, depression and anxiety could form another step towards personal wellness technologies. Automating such assessment becomes increasingly important in an aging population to prevent medical practitioners from becoming overburdened.

- Monitoring and interpreting facial signals can also provide important information to lawyers, police, security and intelligence agents regarding deception and true attitudes. Automated facial reaction monitoring could form a valuable toll in law enforcement as now only informal interpretations are typically used. Systems have the capability of alerting the appropriate authorities to cases of recognition of unfriendly or aggressive faces.
- Recently, facial expression recognition becomes relevant in interactive video games as well. In order to enhance the interaction and attractiveness of the games - recognition of facial expressions, such as happiness, anger, fear, surprise and others provides an interpretation of the player's mood and uses it to adjust the game correspondingly.

Many methods have been applied to expression recognition such as sequence based methods and frame based methods. Sequence based methods use temporal information extracted from a sequence of frames and frame based methods do not use temporal information and rely on a single frame. The facial features are usually divided into two categories: geometric feature-based methods and appearance-based methods. Geometric facial features are usually expressed as locations of fiducial points (around the mouth, eyes, brows, and nose). Appearance-based methods usually apply Gabor wavelets to either specific regions in the face or to the whole face. For more details see [1]. As stated above, machine analysis of facial expressions attracted the interest of many researchers. However, although humans detect and analyze faces and facial expressions in a scene with little or no effort, development of an automated system that accomplishes this task is rather difficult. In addition, one of the main problems of automatic recognition of facial expression in real-life is the inhomogeneity of scale, orientation or translation of the visual stimuli. In this paper we describe a general framework that deals with shape classification in general and apply it to facial expressions recognition in particular.

The paper is organized as follows: in the following section we describe a generalized form of the Procrustes distance for landmark-based shapes. In section 3 we describe a metric learning approach for shapes. Section 4 describes Procrustean inner product kernels to be used by kernel machines. Experimental results are described in sections 5. Finally, in section 6 we draw conclusions.

2 Shape Metric

A natural choice of landmarks is of salient points which can be identified by computer and humans. In the general case, there is a considerable loss of information by extracting only certain points, and the transformed shape cannot be restored in details from the landmarks. Yet, many essential characteristics may remain in such representation. A set of k ordered points in 2D plane can be represented as a $2k$-dimensional vector. Comparing two shapes can be based on corresponding points which are termed *homologies*. The term of distance (or similarity) between two shapes can be easily defined when using $2k$-dimensional vectors by taking only their coordinates as attributes. It is obvious that the order of the points matters. Another convenient representation is called

planar-by-complex and uses complex values to represent each 2-dimensional point, so the whole shape is represented as an $k \times 1$ 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. This framework can be extended to 3D objects in a straight forward manner, though extracting landmarks from 3D objects is more complicated.

2.1 Procrustes Shape Metric

A desired distance measure between two planar point 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: $\mathbf{x} \rightarrow \mathbf{x} - \frac{1}{k}\mathbf{1}_k^T \mathbf{x} \mathbf{1}_k$, where $\mathbf{x}^*$ denotes the complex conjugate of $\mathbf{x}$.

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

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

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.

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

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:

$$S = \sum_{i=1}^{n} \frac{\mathbf{w}_i \mathbf{w}_i^*}{\mathbf{w}_i^* \mathbf{w}_i}. \quad (3)$$

(see [2]). 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 [3]. This process is called *generalized Procrustes analysis*. Scatter analysis, using generalized Procrustes analysis handles the superimposed configurations in an Euclidean manner and provides good linear approximation of the shape space manifold in cases where the configurations variability is small.

The Procrustes superimposition tends to obtain less extreme magnitudes of point shifts. The fact that in least-squares superimposition points 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* [4]. Moreover, such variations affect the configuration's center of mass and thus affect translation as well.

2.2 General Quadratic Procrustes 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 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: $\mathbf{x} \rightarrow \mathbf{x} - \frac{1}{k}\mathbf{1}_k^T Q\mathbf{x}\mathbf{1}_k$. 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 distance*, according to matrix $Q = A^*A$, between two configurations $\mathbf{x}$ and $\mathbf{y}$ is given by:

$$\begin{aligned} d_Q^2(\mathbf{x},\mathbf{y}) &= \inf_{\beta,\vartheta,a,b} \left\| A\frac{\mathbf{y}}{\|\mathbf{y}\|_Q} - A\frac{\mathbf{x}}{\|\mathbf{x}\|_Q}\beta e^{i\vartheta} - a - bi \right\| \\ &= \left(1 - \frac{\mathbf{y}^*Q\mathbf{x}\mathbf{x}^*Q\mathbf{y}}{\mathbf{x}^*Q\mathbf{x}\mathbf{y}^*Q\mathbf{y}}\right)^{\frac{1}{2}}, \end{aligned} \tag{4}$$

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.

$$\widehat{\boldsymbol{\mu}}^Q = \arg\inf_{\boldsymbol{\mu}} \sum_{i=1}^{n} d_Q^2(\mathbf{w}_i, \boldsymbol{\mu}). \tag{5}$$

Proposition 1 *The general quadratic Procrustes mean shape is the eigenvector corresponding to the largest eigenvalue of the following matrix:*

$$S^Q = \sum_{i=1}^{n} \frac{A\mathbf{w}_i\mathbf{w}_i^*A^*}{\mathbf{w}_i^*A^*A\mathbf{w}_i}, \tag{6}$$

(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 [5–7]. 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 [8]).
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

$$\Delta_{k,l}^Q = \left(\frac{\mathbf{x}_l}{\|\mathbf{x}_l\|_Q} \widehat{\beta}_{k,l}^Q e^{i\widehat{\vartheta}_{kl}^Q} \right) - \widehat{\boldsymbol{\mu}}_k^Q \tag{7}$$

and

$$\Delta_k^Q = \widehat{\boldsymbol{\mu}}_k^Q \widehat{\beta}_k^Q e^{i\widehat{\vartheta}_k^Q} - \widehat{\boldsymbol{\mu}}^Q \tag{8}$$

where $\widehat{\beta}_{k,l}^Q, \widehat{\beta}_k^Q$ are the scaling solutions of eq. 4 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:

$$s_W^Q = \sum_{j=1}^{m} \sum_{i=1}^{n} r_{ij} d_Q^2 \left(\mathbf{w}_i, \widehat{\boldsymbol{\mu}}^Q \right) = \sum_{j=1}^{m} \sum_{i=1}^{n} r_{ij} \left(\Delta_{j,i}^Q \right)^* Q \Delta_{j,i}^Q \tag{9}$$

where

$$r_{kl} = \begin{cases} 1 \ \mathbf{x}_l \in \text{Class } k \\ 0 \ \text{Otherwise} \end{cases} \tag{10}$$

and m is the number of classes.

The between class scatter according to a matrix Q is:

$$s_B^Q = \sum_{k=1}^{m} n_k \left(\Delta_k^Q \right)^* Q \Delta_k^Q, \tag{11}$$

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:

$$Q_{opt} = \arg\max_Q \frac{s_B^Q}{s_W^Q}. \tag{12}$$

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.

An example of the differences between the classic Procrustean metric and a learnt one is shown in Figure 1. This delineation 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.

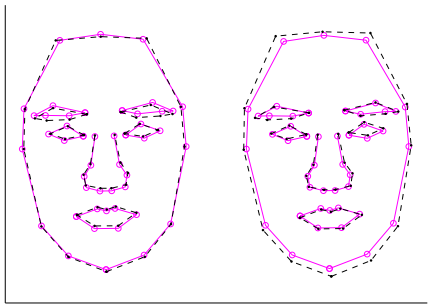


Fig. 1. Superimpositions of mean facial configurations: females (*solid line*) and males (*dashed line*) according to the full Procrustes metric (*left*) and the learnt Procrustes metric (*right*). The mean shapes are of 78 females and 32 males.

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 [9]. 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 [10]. 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 [9]). Following the polynomial and radial basis function (RBF) kernels, we propose the following kernels.

Proposition 2 *The following function is an inner product kernel for any positive integer p:*

$$k(\mathbf{x},\mathbf{y}) = \left(\frac{\mathbf{y}^* Q \mathbf{x}\mathbf{x}^* Q \mathbf{y}}{\mathbf{x}^* Q \mathbf{x}\mathbf{y}^* Q \mathbf{y}}\right)^p \tag{13}$$

For proof outline see appendix A.

Proposition 3 *The following function is an inner product kernel for any positive semi-definite matrix Q and any positive γ:*

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma\left(1-\frac{\mathbf{y}^*Q\mathbf{x}\mathbf{x}^*Q\mathbf{y}}{\mathbf{x}^*Q\mathbf{x}\mathbf{y}^*Q\mathbf{y}}\right)\right) \tag{14}$$

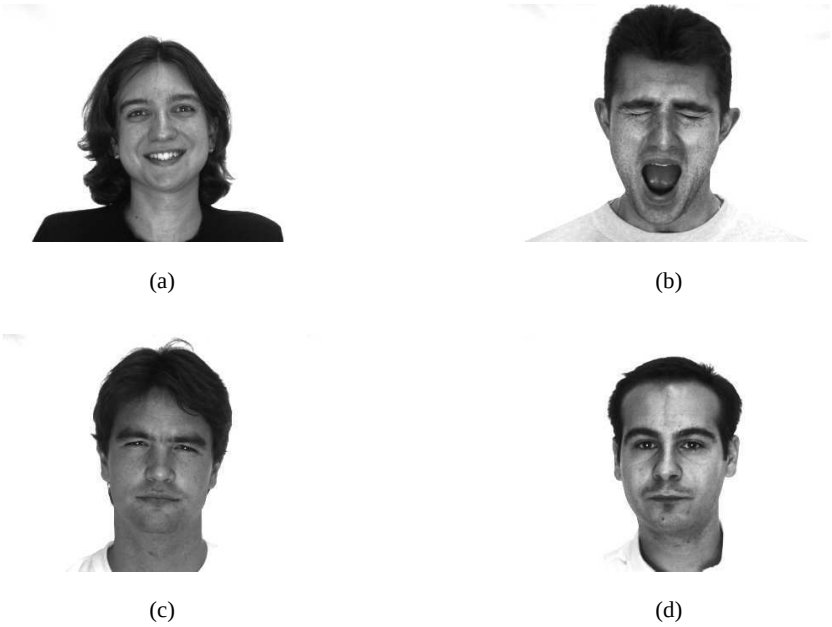
For proof see appendix B.
It should be emphasized that using Procrustean (classic or learnt) metric is not equivalent to aligning the configurations to some consensus configuration first and then using Euclidean distance, since such operation does not preserve all pairwise Procrustes distances among all the configurations at once.

4.2 Kernel design for Multi-Class Classifiers

The general framework of the general quadratic Procrustes metric was designed for many purposes. For classification tasks it has been demonstrated so far only by classifying binary classes using the pre-designed learnt kernels with SVM. The SVM classifier is basically designed for binary classification with a clear geometrical meaning of discriminating one class from another by a hyperplane that maximizes the margin. A crucial point is the choice of combining multi-class learnt metric into a kernel designed for dichotomies. Several approaches were suggested for the multiclass case where the implementation is done by combining several binary SVMs. Many studies have shown the superiority of the *one vs. one* based versions of multi-class SVMs over *one vs. all* versions. The *one vs. one* approach constructs one binary classifier for every pair of distinct classes. For each pair a distinct metric can be learnt and the classification is done according to the majority (or weighted majority) of votes of all $M(M-1)/2$ classifiers ([11],[12]) .

5 Experiments and Results

The performance of the Procrustes kernels was evaluated using the AR Face Database [13]. This database contains over 4,000 images corresponding to 126 individuals (70 men and 56 women). Images include frontal view faces with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). For our current study, we selected 904 images of four facial expressions: neutral, anger, smile and scream (226 images for each expression). Figure 2 shows examples of the four expressions. A relatively small number (36) of crucial landmarks were chosen as the size of datasets is small (the landmarks were chosen according the ease of accurate marking). The chosen landmarks are: 6 points for each eyebrow, 4 points for each eye, 16 points for the mouth (the landmarks are shown in fig. 3). The extraction of the landmarks from the facial images was done by the Bayesian Tangent Shape Model (BTSM [14]) followed by a manual correction. Six metrics (matrices) for six binary classifiers, denoted by A_{kl}, were learnt for $k \neq l$ where $k,l \in \{$*smile, scream, anger, neutral*$\}$ and $A_{kl} = A_{lk}$. The matrices learning was limited to diagonal matrices.



(a) (b)

(c) (d)

Fig. 2. The four facial expressions of the AR dataset: (a) smile, (b) scream, (c) anger, and (d) neutral

Fig. 3. The 36 facial landmarks

5.1 Experiment A

The performance of three classifiers was evaluated in this experiment:

- SVM with standard RBF kernel with square Euclidean distance in the exponent. The configurations (complex vectors) were centered and normalized but remained in their original inclination (it should be noticed that the original dataset was prepared in a way where all faces have similar size and similar inclination).
- SVM with full Procrustes distance based RBF kernel ($Q = I$).
- SVM with learnt Procrustes distance based RBF kernel (learnt Q).

The error rates were evaluated by 4-fold cross validation. The results are given in confusion matrices (Tables 1- 3).

Table 1. SVM classification results using Radial Basis Function Kernel with Euclidean distance. The mean error rate: $13.9 \pm 2.4\%$.

Actual\Predicted	Smile	Scream	Anger	Neutral
Smile	**99.6%**	0%	0%	0.4%
Scream	0.4%	**99.6%**	0%	0%
Anger	0%	0%	**68.3%**	31.7%
Neutral	1.8%	0%	21.4%	**76.8%**

Table 2. SVM classification results using Procrustes Kernel ($Q = I$). The mean error rate: $12.8 \pm 1.3\%$.

Actual\Predicted	Smile	Scream	Anger	Neutral
Smile	**99.6%**	0%	0%	0.4%
Scream	0.4%	**99.6%**	0%	0%
Anger	0%	0%	**75.2%**	24.8%
Neutral	1.3%	0%	24.3%	**74.4%**

Table 3. SVM classification results using learnt Procrustes Kernel. The mean error rate: $10.4 \pm 1.3\%$.

Actual\Predicted	Smile	Scream	Anger	Neutral
Smile	**99.6%**	0%	0%	0.4%
Scream	0%	**100%**	0%	0%
Anger	0%	0%	**77.4%**	22.6%
Neutral	1.7%	0%	16.9%	**81.4%**

It can be observed that the Procrustes kernel ($Q = I$) does not have a significant advantage over the Euclidean based RBF kernel. Three possible explanations for this are: (i) The photographs were taken in advance in artificial conditions where the location and size of the heads are similar. This fact does not provide significant advantage to the Procrustes invariance properties. (ii) It is possible that in the anger expression case people tend to tilt their heads a little bit more frequently than in the neutral case. This cue is lost in the Procrustean approach, but remains in the Euclidean RBF kernel. However, in a more realistic scenario the head's inclination is not a reliable feature. (iii) The anger and the neutral expressions are very similar (even human subjects have identification difficulties when looking at the full detailed images).
The classic Procrustes alignment tends to align the configurations in a way that minimizes the landmark distances and make the configurations to look more similar. The Procrustes learnt kernel is preferable over the two other kernels. Though it loses information conveyed by rotation (which is usually not helpful in real-life applications) it still has strong abilities to discriminate landmark based configurations.

5.2 Experiment B

The purpose of this experiment is to evaluate the contribution of the rotation feature (ϑ or $\widehat{\vartheta}^Q$ as in eq. 1 and 4 correspondently). Two kernels were evaluated:

- SVM with a kernel constructed of a product of full Procrustes distance based RBF kernel ($Q = I$) and a kernel that is based on the rotation angle between configurations: $k_1(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma_1 d_F^2(\mathbf{x},\mathbf{y})\right) \cdot \exp\left(-\gamma_2 \left(\vartheta_{\mathbf{xy}}\right)^2\right)$.
- SVM with a kernel constructed of a product of a learnt Procrustes distance based RBF kernel (learnt Q) and a kernel that is based on the learnt metric rotation angle between configurations: $k_2(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma_1 d_Q^2(\mathbf{x},\mathbf{y})\right) \cdot \exp\left(-\gamma_2 \left(\widehat{\vartheta}^Q_{\mathbf{xy}}\right)^2\right)$.

γ_1 and γ_2 are kernel hyperparameters. The error rates were evaluated by 4-fold cross validation. The results are given in confusion matrices (Tables 4 and 5).

Table 4. SVM classification results using Procrustes Kernel ($Q = I$) with rotation feature. The mean error rate: $11.1 \pm 1.3\%$.

Actual\Predicted	Smile	Scream	Anger	Neutral
Smile	**99.6%**	0%	0%	0.4%
Scream	0.4%	**99.6%**	0%	0%
Anger	0%	0%	**77.9%**	22.1%
Neutral	1.3%	0%	19.9%	**78.7%**

Indeed, the relative rotation angle between two configurations convey some cues regarding the class membership. The modified kernels accomplish significant additional improvement.

Table 5. SVM classification results using learnt Procrustes Kernel with rotation feature. The mean error rate: $9.0 \pm 1.2\%$.

Actual\Predicted	Smile	Scream	Anger	Neutral
Smile	**99.6**%	0%	0%	0.4%
Scream	0%	**100%**	0%	0%
Anger	0%	0%	**80.3%**	19.7%
Neutral	1.3%	0%	14.6%	**84.1%**

6 Conclusions

In this paper we have introduced a general framework of shape distances. Aligning configurations according to the learnt metric enables a visualization that uncovers the discriminative landmarks. Improvement in classification performance was demonstrated by using multi-class kernel SVM with a data driven kernel design. The main contribution of the learnt metric is the meaningful alignment - it is of particular importance in cases where the training sets are small. Classifiers with Euclidean distance based kernels must hold much more samples of each facial expression as they are not taking into consideration some variations in head position in space. Procrustes distance (learnt or classical) kernel based classifiers are insensitive to translation, scaling and rotation, which is a key factor in recognizing facial expressions both in still images and video sequences in 'real-life' environment. The comparison between Procrustes metric $(Q = I)$ and Euclidean metric discovers a discriminative feature - the head's angle - which is helpful when using input images taken in lab environment, but might be useless in 'real-life' conditions where there is a large variability in the face position within the frame. This property should be taken into consideration when dealing with 'real-life' systems designed by training sets produced in lab conditions.

References

1. Li, S., Jain, A., eds.: Handbook of Face Recognition. Springer (2005)
2. Kent, J.: The complex bingham distribution and shape analysis. Journal of the Royal Statistical Society, Series B **56** (1994) 285–299
3. Rholf, F., Slice, D.: Extensions of the procrustes method for the optimal superimposition of landmarks. Syst. Zool. **39** (1990) 40–59
4. Chapman, R.: Conventional procrustes approaches. Proceedings of the Michigan Morphometrics Workshop (1990) 251–267
5. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance metric learning, with application to clustering with side-information. In: Advances in Neural Information Processing Systems. Volume 18. (2004)
6. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems. Volume 18. (2004)
7. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: Advances in Neural Information Processing Systems. Volume 19. (2005)
8. Duda, R., Hart, P., Stork, D.: Pattern Classification. 2nd Ed. John Wiley & Sons (2001)
9. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press (2000)

10. 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)
11. Duan, K., Keerthi, S.: Which is the best multiclass SVM method? an empirical study. In: Multiple Classifier Systems, 6th International Workshop, MCS 2005, Seaside, CA, USA, June 13-15, 2005, Proceedings. Volume 3541 of Lecture Notes in Computer Science., Springer (2005) 278–285
12. Hsu, C., Lin, C.: A comparison of methods for multiclass support vector machines. IEEE-EC **13** (2002) 415–425
13. Martinez, A., Benavente, R. The AR face database. CVC Tech. Report **24** (1998)
14. 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:

$$k(\mathbf{x},\mathbf{y}) = \left(\frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}}\right)^p \tag{15}$$

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

$$\int\int \left(\frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}}\right)^p g(\mathbf{x})g^*(\mathbf{y})d\mathbf{x}d\mathbf{y} \geq 0 \tag{16}$$

for any function g with finite l_2 norm.
Each term of the multinomial expansion has a non-negative value:

$$(r_1, r_2, \ldots, l_1, l_2, \ldots,)! \left\| \int \left(\frac{x_1^{r_1}, x_2^{r_2} \cdots \overline{x}_1^{l_1}, \overline{x}_2^{l_2} \cdots}{\|\mathbf{x}\|^{2p}} \right) g\left(\mathbf{x}\right) d\mathbf{x} \right\|^2 \geq 0 \tag{17}$$

and hence the integral is non-negative.
Showing that:

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

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

Appendix B.

Proof:

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\gamma\left(1 - \frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}}\right)\right) = \exp\left(-\gamma\right)\exp\left(\gamma\frac{\mathbf{y}^*\mathbf{x}\mathbf{x}^*\mathbf{y}}{\mathbf{x}^*\mathbf{x}\mathbf{y}^*\mathbf{y}}\right). \tag{19}$$

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 proposition 3 we have a sum of semi-definite functions, which is also a semi-definite function.■