An information-theoretic approach to face recognition from face motion manifolds

Ognjen Arandjelović [a,1], Roberto Cipolla [a,2]

[a] *University of Cambridge, Cambridge, CB2 1PZ, UK*

Abstract

In this work we consider face recognition from *Face Motion Manifolds* (FMMs). The use of the Resistor-Average Distance (RAD) as a dissimilarity measure between densities confined to FMMs is motivated in the proposed information-theoretic approach to modelling face appearance. We introduce a kernel-based algorithm that makes use of the simplicity of the closed-form expression for RAD between two Gaussian densities, while allowing for modelling of complex and nonlinear, but intrinsically low-dimensional manifolds. Additionally, it is shown how geodesically local FMM structure can be modelled, naturally leading to a stochastic algorithm for generalizing to unseen modes of data variation. Recognition performance of our method is demonstrated experimentally and is shown to exceed that of state-of-the-art algorithms. Recognition rate of 98% was achieved on a database of 100 people under varying illumination.

Key words: face recognition, face motion manifolds, kernel, resistor-average distance

1 Introduction

Important practical applications of automatic face recognition (AFR) have made it a very popular research area in the last three decades, see [1, 2, 3, 4] for surveys. Most

[1] E-mail: oa214@eng.cam.ac.uk
[2] E-mail: cipolla@eng.cam.ac.uk

of the methods developed deal with *single-shot* recognition. In controlled imaging conditions (lighting, pose and/or occlusions) many have demonstrated good (nearly perfect) recognition results [4]. On the other hand, single-shot face recognition in uncontrolled, or loosely controlled conditions still poses a significant challenge [4].

The nature of many practical applications is such that more than a single image of a face is available. In surveillance, for example, the face can be tracked to provide a temporal sequence of a moving face. In access control use of face recognition the user may be assumed to be cooperative and hence can be instructed to move the head in front of a fixed camera. Regardless of the setup in which multiple images of a face are acquired, it is clear that this abundance of information can be used to achieve greater robustness of face recognition by resolving some of the inherent ambiguities of the single-shot recognition problem.

In the next section we briefly review relevant literature on face recognition from video. Section 3 introduces the concept of classification using Kernel RAD, followed by a section in which we show how errors in the face registration process can be modelled and incorporated in the described recognition framework. Section 5 describes the pipeline we used to extract and preprocess images of faces from realistic video sequences. Experimental evaluation of the proposed method and its comparison to state-of-the-art methods in the literature is reported in Section 6. We conclude the paper with a discussion of the results and an outline of promising directions for future research.

2 Related previous work

Single-shot face recognition is a well established research area. Algorithms such as Bayesian Eigenfaces [5, 6], Fisherfaces [7, 4], Elastic Bunch Graph Matching [8, 9] or the 3D Morphable Model [10, 11] have demonstrated good recognition results when illumination and pose variations are not large. However, all existing single-shot methods suffer from the limited ability to generalize to unseen illumination conditions or pose.

Compared to single-shot recognition, face recognition from video is a relatively new area of research. Most of the existing algorithms perform recognition from image *sequences*, using the temporal component to enforce prior knowledge on likely head movements. In the algorithm of of Zhou et al. [12] the joint probability distribution of identity and motion is modelled using sequential importance sampling, yielding the recognition decision by marginalization. In [13] Lee et al. approximate face manifolds by a finite number of infinite extent subspaces and use temporal information to robustly estimate the operating part of the manifold.

There are fewer methods that recognize from manifolds without the associated or-

dering of face images, which is the problem we address in this paper. Two algorithms worth mentioning are the Mutual Subspace Method (MSM) of Yamaguchi et al. [14, 15] and the Kullback-Leibler divergence based method of Shakhnarovich et al. [16].

In MSM, infinite extent linear subspaces are used to compactly characterize face sets i.e. the manifolds that they lie on. Two sets are then compared by computing the first three principal angles between corresponding principal component analysis (PCA) subspaces [14]. Varying recognition results were reported using MSM, see [14, 16, 17, 15]. The major limitation of MSM is its simplistic modelling of manifolds of face appearance variations. Their high nonlinearity (see Figure 1) invalidates the assumption that data is well described by linear subspaces. More subtly, the nonlinearity of modelled manifolds means that the PCA subspace estimates are very sensitive to the particular choice of training samples. For example, in the original paper [15] in which face motion videos were used, the estimates are sensitive to the extent of rotation in a particular direction. Finally, MSM does not have a meaningful probabilistic interpretation.

The Kullback-Leibler (KL) divergence based method [16] is founded on information-theoretic grounds. In the proposed framework, it is assumed that i-th person's face patterns are distributed according to $p_i(\mathbf{x})$. Recognition is then performed by finding $p_j(\mathbf{x})$ that best explains the set of input samples – quantified by the Kullback-Leibler divergence. The key assumption in their work, that makes divergence computation tractable, is that face patterns are normally distributed i.e. $p_i(\mathbf{x}) = \mathcal{N}(\bar{\mathbf{x}}_i, \mathbf{C}_i)$. This is a crude assumption (see Figure 1), which explains the somewhat poor results reported with this method [17]. KL divergence was also criticized for being asymmetric [18] (also see Section 3.1.1).

3 Recognition using statistical models of FMMs

Assuming that the AFR system user performs random head motion in front of the camera, anatomical constraints of the head and the constraints of the imaging setup make certain head poses more likely than others. This motivates the interpretation of face video sequences that we employ in this work – as sets of independently and identically distributed (i.i.d) samples from the corresponding probability density functions (see Figure 1). An attractive feature of this approach is that it inherently encapsulates a statistical interpretation of video sequences and naturally lends itself to probabilistic modelling of noise and outliers.

Formalizing the above, we assume that an image $\mathbf{x}$ of subject i's face is drawn from the probability density $p_F^{(i)}(\mathbf{x})$ within the face space, and embedded in the image space by means of a mapping function $f^{(i)} : \mathbb{R}^d \rightarrow \mathbb{R}^D$. The resulting point in the D-dimensional space is further perturbed by noise drawn from a noise distribution

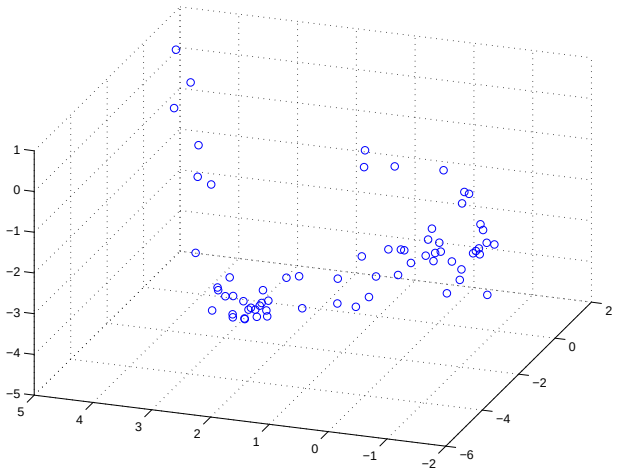


Fig. 1. *A typical face manifold of head motion (significant pitch and yaw, some roll) around the fronto-parallel face. Shown is a projection to the first 3 principal components. The manifold can be seen to be smooth and intrinsically low-dimensional, but highly nonlinear.*

p_n (note that the noise operates in the image space) to form the observed image $\mathbf{X}$. Therefore the distribution of the observed face images of the subject i is given by:

$$p^{(i)}(\mathbf{X}) = \int p_F^{(i)}(\mathbf{x}) p_n(f_i(\mathbf{x}) - \mathbf{X}) d\mathbf{x} \tag{1}$$

Note that both the manifold embedding function f and the density p_F on the manifold are subject-specific, as denoted by the superscripts, while the noise distribution p_n is assumed to be common for all subjects.

3.1 Dissimilarity between manifolds

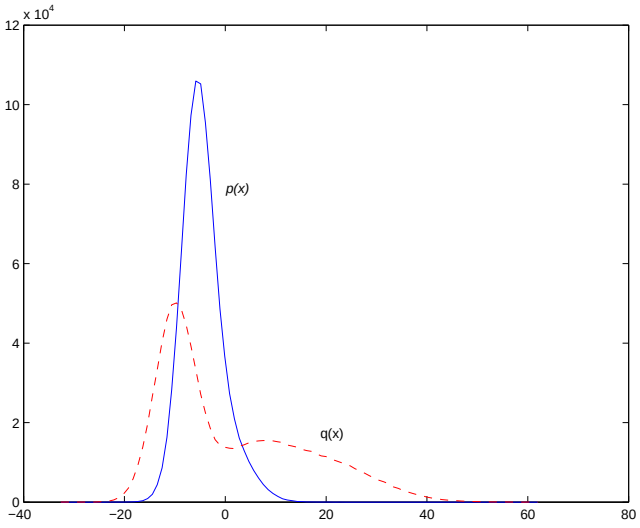
3.1.1 Kullback-Leibler divergence

One of the best known dissimilarity measures between probability density functions (pdfs) is the Kullback-Leibler (KL) divergence, sometimes also called the Mutual Entropy. It is defined as [19]:

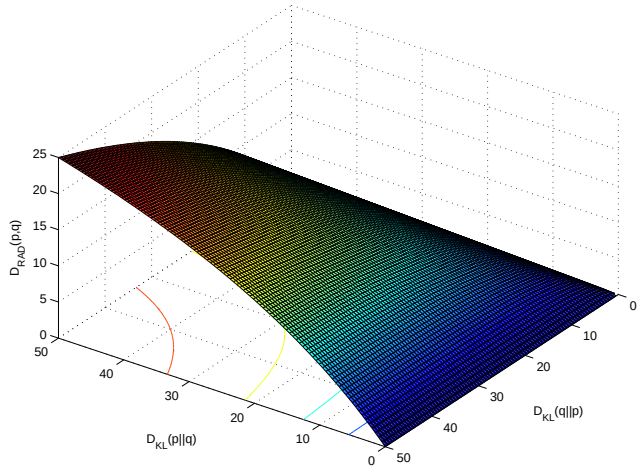
$$D_{KL}(p||q) \doteq \int p(\mathbf{x}) \log_2 \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) d\mathbf{x} \tag{2}$$

It is nonnegative and equal to zero iff $p(\mathbf{x}) \equiv q(\mathbf{x})$. Note that it is also asymmetrical.

The appeal of KL divergence stems from its information theory founded approach to quantifying how well a particular pdf $q(\mathbf{x})$ describes samples from anther pdf $p(\mathbf{x})$. To gain the intuition behind this divergence, consider the form of (2). The integrand $p(\mathbf{x}) \log_2 (p(\mathbf{x})/q(\mathbf{x}))$ can be seen to have a large value when $p(\mathbf{x})$ is significant and $p(\mathbf{x}) \gg q(\mathbf{x})$. Therefore, the regions of the integration space that



(a)



(b)

Fig. 2. *A 1D illustration of asymmetry of KL divergence (a).* $D_{KL}(q||p)$ *is an order of magnitude greater than* $D_{KL}(p||q)$ *– the "wider" distribution* $q(\mathbf{x})$ *explains the "narrower"* $p(\mathbf{x})$ *better than the other way round. In (b),* $D_{RAD}(p,q)$ *is plotted as a function of* $D_{KL}(p||q)$ *and* $D_{KL}(q||p)$.

produce a large contribution to $D_{KL}(p||q)$ are those that are well explained by $p(\mathbf{x})$, but not by $q(\mathbf{x})$ (note that owing to the asymmetry of the expression the converse is not true). This makes KL divergence suitable in cases when it is known a priori that one of the densities $p(\mathbf{x})$ or $q(\mathbf{x})$ describes a wider range of data variation than the other (e.g. as in [20]), see Figure 2 (a). However, in the proposed recognition framework, this is not the case – pitch and yaw changes are expected to be the dominant modes of variation in both training and novel data. Additionally, exact head poses assumed by the user are expected to somewhat vary from sequence to sequence and the robustness to variations not seen in either is desired. This motivates the use of a symmetric "distance" measure.

3.1.2 Resistor-Average distance.

In this paper we use the Resistor-Average distance (RAD) as a measure of dissimilarity between two probability densities. It is defined as:

$$D_{RAD}(p,q) \doteq \left[D_{KL}(p||q)^{-1} + D_{KL}(q||p)^{-1}\right]^{-1} \tag{3}$$

Much like the KL divergence from which it is derived, it is nonnegative and equal to zero iff $p(\mathbf{x}) \equiv q(\mathbf{x})$, but unlike it, it is symmetric. Another important property of the Resistor-Average distance is that when two classes of patterns $\mathcal{C}_p$ and $\mathcal{C}_q$ are distributed according to, respectively, $p(\mathbf{x})$ and $q(\mathbf{x})$, $D_{RAD}(p,q)$ reflects the error rate of the Bayes-optimal classifier between $\mathcal{C}_p$ and $\mathcal{C}_q$ [18].

To see in what manner RAD differs from the KL divergence, it is instructive to consider two special cases: when divergences in both directions between two pdfs are approximately equal and when one of them is much greater than the other:

- $D_{KL}(p||q) \approx D_{KL}(q||p) \equiv D$

$$D_{RAD}(p,q) \approx D \tag{4}$$

- $D_{KL}(p||q) \gg D_{KL}(q||p)$ or
  $D_{KL}(p||q) \ll D_{KL}(q||p)$

$$D_{RAD}(p,q) \approx \min\left(D_{KL}(p||q), D_{KL}(q||p)\right) \tag{5}$$

It can be seen that RAD very much behaves like a smooth min of $D_{KL}(p||q)$ and $D_{KL}(q||p)$, also illustrated in Figure 2 (b).

3.2 Estimating RAD for Nonlinear Densities

Following the choice of the Resistor-Average distance as a means of quantifying the similarity of manifolds, we turn to the question of estimating this distance for two arbitrary, nonlinear face manifolds. For a general case there is no closed-form expression for RAD. However, when $p(\mathbf{x})$ and $q(\mathbf{x})$ are two normal distributions [21]:

$$\begin{aligned} D_{KL}(p||q) =& \frac{1}{2}\log_2\left(\frac{|\mathbf{\Sigma}_q|}{|\mathbf{\Sigma}_p|}\right) + \\ & \frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}_p\mathbf{\Sigma}_q^{-1} + \mathbf{\Sigma}_q^{-1}(\bar{\mathbf{x}}_q - \bar{\mathbf{x}}_p)(\bar{\mathbf{x}}_q - \bar{\mathbf{x}}_p)^T\right) - \frac{D}{2} \end{aligned} \tag{6}$$

where D is the dimensionality of data, $\bar{\mathbf{x}}_p$ and $\bar{\mathbf{x}}_q$ data means, and $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_q$ the corresponding covariance matrices.

To achieve both expressive modelling of nonlinear FMMs as well as an efficient procedure for comparing them, in the proposed method a nonlinear projection of data using Kernel Principal Component Analysis (Kernel PCA) is performed first. We shown that with an appropriate choice for the kernel type and bandwidth, the assumption of normally distributed face patterns in the projection space produces good KL divergence estimates. With the reference to (1), an FMM is effectively unfolded from the embedding image space.

3.3 Kernel principal component analysis

PCA is a technique in which an orthogonal basis transformation is applied such that the data covariance matrix $\mathbf{C} = \langle(\mathbf{x}_i - \langle\mathbf{x}_j\rangle)(\mathbf{x}_i - \langle\mathbf{x}_j\rangle)^T\rangle$ is diagonalized. When data $\{\mathbf{x}_i\}$ lies on a linear manifold, the corresponding linear subspace is spanned by the dominant (in the eigenvalue sense) eigenvectors of $\mathbf{C}$. However, in the case of nonlinearly distributed data, PCA does not capture the true modes of variation well.

The idea behind KPCA is to map data into a high-dimensional space in which it *is* approximately linear – then the true modes of data variation can be found using standard PCA. Performing this mapping explicitly is prohibitive for computational reasons and inherently problematic due to the "curse of dimensionality". This is why a technique known as the "kernel trick" is used to implicitly realize the mapping. Let function $\mathbf{\Phi}$ map the original data from input space to a high-dimensional pattern space in which it is (approximately) linear, $\mathbf{\Phi} : \mathbb{R}^D \rightarrow \mathbb{R}^\Delta$, $\Delta \gg D$. In KPCA the choice of mappings Φ is restricted to the set such that there is a function k (the kernel) such that:

$$\mathbf{\Phi}(\mathbf{x}_i)^T\mathbf{\Phi}(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

In this case, the principal components of the data in $\mathbb{R}^\Delta$ space can be found by performing computations in the input, $\mathbb{R}^D$ space only.

Assuming zero-centred data in the feature space (for information on centring data in the feature space as well as a more detailed treatment of KPCA see [22]), the problem of finding principal components in the feature space is equivalent to solving the eigenvalue problem:

$$\mathbf{K}\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{8}$$

where $\mathbf{K}$ is the kernel matrix:

$$\mathbf{K}_{j,k} = k(\mathbf{x}_j, \mathbf{x}_k) = \mathbf{\Phi}(\mathbf{x}_j)^T\mathbf{\Phi}(\mathbf{x}_k) \tag{9}$$

The projection of a data point $\mathbf{x}$ to the i-th kernel principal component is computed

using the following expression [22]:

$$a_i = \sum_{m=1}^{N} u_i^{(m)} k(\mathbf{x}_m, \mathbf{x}) \tag{10}$$

3.4 Combining RAD and kernel PCA

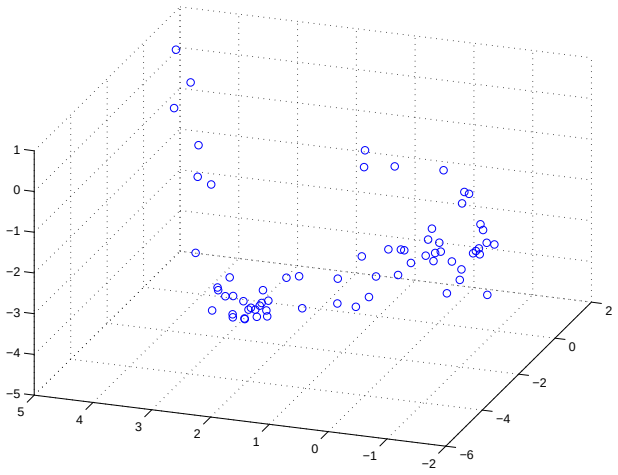
The variation of face patterns is highly nonlinear (see Figure 3 (a)), making the task of estimating RAD between two sparsely sampled face manifolds in the image space hard. The approach taken in this work is that of mapping the data from the input, image space into a space in which it lies on a nearly linear manifold. As before, we would not like to compute this mapping explicitly. Also, note that the inversions of data covariance matrices and the computation of their determinants in the expression for the KL divergence between two normal distributions (6) limit the maximal practical dimensionality of the pattern space.

In our method both of these problems are solved using Kernel PCA. The key observation is that regardless of how high the pattern space dimensionality is, the data has covariance in at most N directions, where N is the number of data points. Therefore, given two data sets of faces, each describing a smooth manifold, we first find the kernel principal components of their union. After dimensionality reduction is performed by projecting the data onto the first M kernel principal components, the RAD between the two densities, each now assumed Gaussian, is computed. Note that the implicit nonlinear map is different for each data set pair. The importance of this can be seen by noticing that the intrinsic dimensionality of the manifold that *both* sets lie on is lower than of the manifold that all data in a database lie on, resulting in its more accurate "unfolding", see Figure 3 (b).

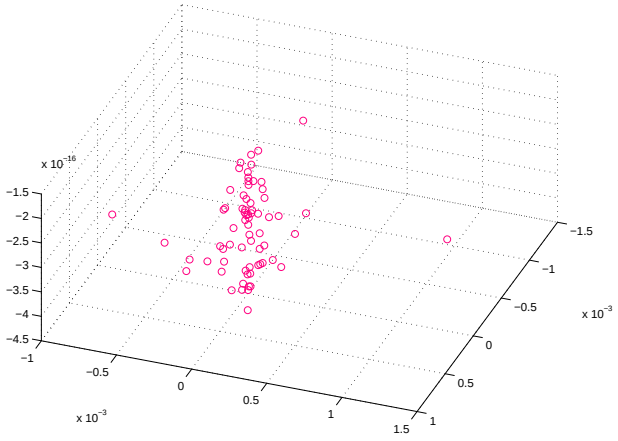
We estimate covariance matrices in the Kernel PCA space using Probabilistic PCA (PPCA) [23]. In short, probabilistic PCA is an extension of the traditional PCA that recovers parameters of a linear generative model of data (i.e. the full corresponding covariance matrix), with the assumption of isotropic Gaussian noise: $\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T + \sigma\mathbf{I}$. Note the model of noise density in (1) that this assumption implies: $g^{(i)}(p_n(\mathbf{x})) \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$, where $g^{(i)}\left(f^{(i)}(\mathbf{x})\right) = \mathbf{x}$.

4 Synthetically repopulating FMMs

For most applications, due to the practical limitations in the data acquisition process, AFR algorithms have to work with sparsely populated face manifolds. Furthermore, some modes of data variation may not be present in full. Specifically, in the AFR for authentication setup considered in this work, the practical limits on



(a)



(b)

Fig. 3. *A typical face motion manifold in the input, image space exhibits high nonlinearity (a). The "unfolded" manifold is shown in (b). It can be seen that Kernel PCA captures the modes of data variation well, producing a Gaussian-looking distribution of patterns, confined to a roughly 2-dimensional space (corresponding to the intrinsic dimensionality of the manifold). In both (a) and (b) shown are projections to the first three principal components.*

how long the user can be expected to wait for verification, as well as how controlled his motion can be required to be, limit the possible variations that are seen in both training and novel video sequences. Finally, the noise in the face localization process (see Section 5) increases the dimensionality of the manifolds faces lie on, effectively resulting in even less densely populated manifolds. For a quantitative insight, it is useful to mention that the face appearance variations present in a typical video sequence used in this paper typically lie on a manifold of intrinsic dimensionality of 3-7, with 85 samples on average.

In this work, FMMs are synthetically repopulated in a manner that achieves both higher manifold sample density, as well as some generalization to unseen modes of variation (see work by Martinez [24], and Sung and Poggio [25] for related

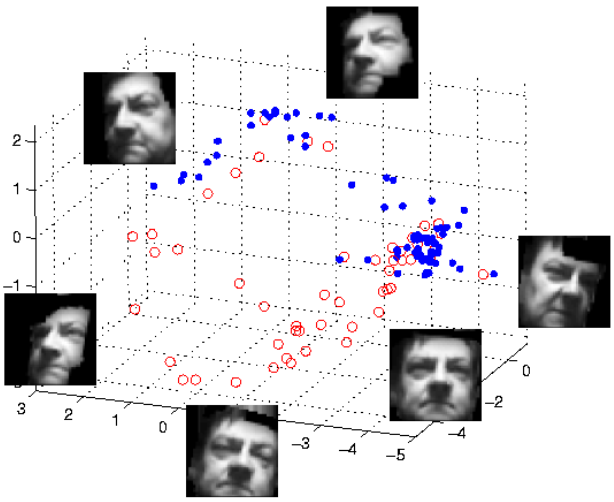


Fig. 4. *The original, input data (dots) and the result of stochastically repopulating the corresponding manifold (circles). A few samples from the dense result are shown as images, demonstrating that the proposed method successfully captures and extrapolates the most significant modes of data variation.*

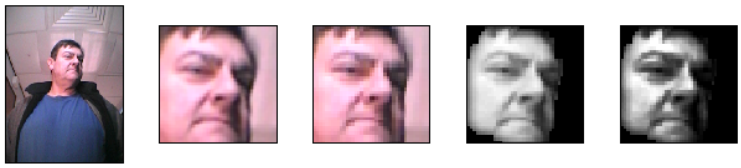
approaches). To this end, we use domain-specific knowledge to learn face transformations in a more sophisticated way than could be realized by simple interpolation and extrapolation.

Given an image of a face, $\mathbf{x}$, we stochastically repopulate its geodesic neighbourhood by a set of novel images $\{\mathbf{x}_j^S\}$. Under the assumption that the embedding function $f^{(i)}$ in (1) is smooth, geodesically close images correspond to small changes in the imaging parameters (e.g. yaw or pitch). Therefore, using the first-order Taylor approximation of the effects of a projective camera, the face motion manifold is locally similar to the the *affine warp* manifold of $\mathbf{x}$. The proposed algorithm then consists of random draws of a face image $\mathbf{x}$ from the data, stochastic perturbation of $\mathbf{x}$ by a set of affine warps $\{\mathbf{A}_j\}$ and finally, the augmentation of data by the results of the warps – see Algorithm 2. Writing the affine warp matrix decomposed to rotation and translation, skew and scaling:

$$\mathbf{A} = \begin{pmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & k & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1+s_x & 0 & 0 \\ 0 & 1+s_y & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{11}$$

in the proposed method, affine transformation parameters θ, t_x and t_y, k, and s_x and s_y are drawn from zero-mean Gaussian densities.

Fig. 5. *Typical outliers present in our data.*



(a) (b) (c) (d) (e)

Fig. 6. *Intermediate results in the face localization and normalization pipeline employed in our AFR system. (a) Original input frames with resolution of* 320×240 *pixels. (b) Output of the face detector with average bounding box size of* 75×75 *pixels. (c) Face images after background removal and feathering. (d) Face images after resizing to the uniform scale of* 20×20 *pixels (e) The final face images after histogram equalization.*

4.1 Outlier rejection

In most cases, automatic face detection in cluttered scenes will result in a considerable number of incorrect localizations – *outliers*. Typical outliers produced by the face detector employed in this paper (see Section 5) can be seen in Figure 5.

Note that due to the complexity of face manifolds, outliers cannot be easily removed in the input space. On the other hand, outlier rejection after Kernel PCA-based manifold "unfolding" is trivial. However, a way of computing the kernel matrix robust to the presence of outliers is needed. To this end, our algorithm uses RANSAC [26] with the underlying Kernel PCA model. The application of RANSAC in the proposed framework is summarized in Algorithm 1. Finally, the recognition method proposed in this paper is in fullness shown in Algorithm 2.

5 Automatic preprocessing of face images

We use the now acclaimed Viola–Jones cascaded detector [27] for localization of faces in cluttered images. Figure 8 shows examples of input frames, and Figure 6 (b) shows a few examples of the correctly detected faces.

5.1 Background removal

The bounding box of a detected face typically contains a portion of the background. The removal of the background is beneficial both because the it can contain signif-

Algorithm 1 RANSAC Kernel PCA

Input: set of observations $\{\mathbf{x}_i\}$, KPCA space dimensionality D
Output: kernel principal components $\{\mathbf{u}_i\}$

1: **Initialize best minimal sample**
Valid sample set $\mathcal{B} = \emptyset$

2: **RANSAC iteration**
for $it = 0$ to $LIMIT$ **do**

3: **Random sample draw**
Random samples $\{y_i\} \overset{D}{\leftarrow} \{x_i\}$

4: **Kernel PCA**
$\{u_i\}$ = KPCA($\{y_i\}$)

5: **Nonlinear projection**
$\{x_i^P\} \overset{\{\mathbf{u}_i\}}{\longleftarrow} \{x_i\}$

6: **Consistent data**
$\mathcal{B}_{it} = |filter(D_{MAH}(x_i, 0) < T)|$

7: **Update best minimal sample**
$|\mathcal{B}_{it}| > |\mathcal{B}| \; ? \; \mathcal{B} = \mathcal{B}_{it}$

8: **end for**

9: **Kernel PCA using best minimal sample**
$\{u_i\}$ = KPCA($\mathcal{B}$)

icant clutter and also depending on the environment in which specific people were imaged, there is the danger of learning to discriminate based on the background, rather than face appearance. This is achieved by *set-specific* skin colour segmentation: Given a set of images from the same subject, we construct colour histograms for that subject's face pixels and for the near-face background pixels in that set. Note that the classifier here is tuned for the given subject *and* the given background environment. The face pixels are collected by taking the central portion of the few most symmetric images in the set (assumed to correspond to the frontal face images); the background pixels are collected from the 10 pixel-wide strip around the face bounding box provided by the face detector. After classifying each pixel within the bounding box independently, we smooth the result using a simple 2-pass algorithm that enforces the connectivity constraint on the face and boundary regions, see Figure 6 (d).

Algorithm 2 Robust Kernel RAD

‘

Input: sets of observations $\{\mathbf{a}_i\}$, $\{\mathbf{b}_i\}$

Output: $D_{RAD}(\{\mathbf{a}_i\}, \{\mathbf{b}_i\})$

1: **Inliers with RANSAC**
$\mathcal{V} = \{\mathbf{a}_i^V\}, \{\mathbf{b}_i^V\} = \text{RANSAC}(\{\mathbf{a}_i\}, \{\mathbf{b}_i\})$

2: **Synthetic data**
$\mathcal{S} = \{\mathbf{a}_\mathbf{i}^\mathbf{S}\}, \{\mathbf{b}_\mathbf{i}^\mathbf{S}\} = \text{perturb}\Big(\langle \mathbf{a}^\mathbf{V} \rangle, \langle \mathbf{b}^\mathbf{V} \rangle\Big)$

3: **RANSAC Kernel PCA**
Principal components $\{\mathbf{u}_i\} = \text{KPCA}(\mathcal{V} \cup \mathcal{S})$

4: **Nonlinear projection**
$\{\mathbf{a}_i^P\}, \{\mathbf{b}_i^P\} \xleftarrow{\{\mathbf{u}_i\}} (\mathcal{V}, \mathcal{S})$

5: **Closed-form RAD**
$D_{RAD}(\{\mathbf{a}_i\}, \{\mathbf{b}_i\})$

6 Empirical evaluation

We compared the recognition performance of the the following methods [3] :

- KL divergence-based algorithm of Shakhnarovich et al. (Simple KLD) [16],
- Simple RAD (based on Simple KLD),
- Kernelized Simple KLD algorithm (Kernel KLD),
- Kernel RAD,
- Robust Kernel RAD,
- Mutual Subspace Method (MSM) [15],
- Majority vote using Eigenfaces, and
- Nearest Neighbour (NN) in the set distance sense; that is, achieving $\min_{\mathbf{x} \in S_0} \min_{\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|_2$.

In the all KLD and RAD-based methods, 85% of data energy was explained by the principal subspaces. In non-kernelized algorithms this typically resulted in the

[3] Methods were reimplemented through consultation with authors.

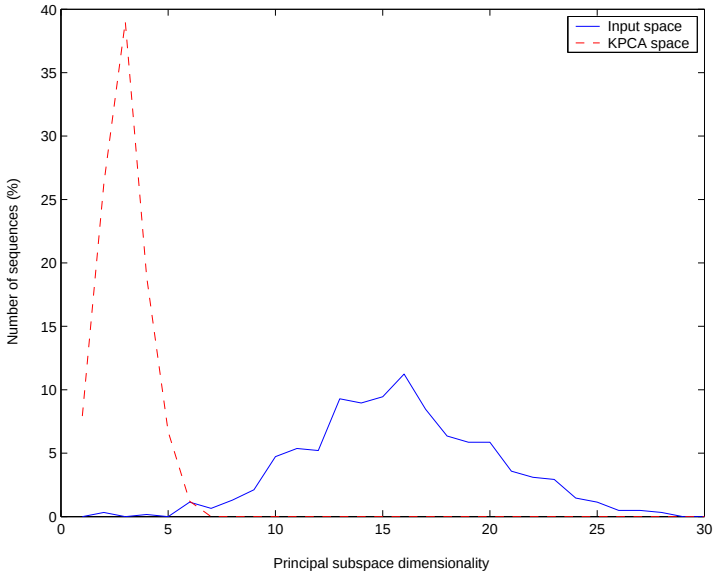


Fig. 7. *Histograms of the dimensionality of the principal subspace in kernelized (dotted line) and non-kernelized (solid line) KL divergence-based methods, across the evaluation data set. The corresponding average dimensionalities were found to be* ~ 4 *and* ~ 16*. The large difference illustrates the extent of nonlinearity of Face Motion Manifolds.*

principal subspace dimensionality of 16, see Figure 7. In MSM, first 3 principal angles were used for recognition, while the dimensionality of PCA subspaces describing the data was set to 9 [15]. In the Eigenfaces method, the 150-dimensional principal subspace used explained $\sim 95\%$ of data energy. A 20-dimensional nonlinear projection space was used in all kernel-based methods with the RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp -\gamma(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)$. The optimal value of parameter γ was learnt by optimizing the recognition performance on a 20 person training data set. Note that people from this set were not included in the evaluation reported in Section 6.2. We used $\gamma = 0.380$ for greyscale images normalized to have pixel values in the range $[0.0, 1.0]$.

6.1 Data

The evaluation of methods in this paper was done on a database with 100 individuals of varying age (see Table 1) and race, and equally represented genders. For each individual in the database we collected a training and a test video sequence of the person's face in random motion, sampled at 10fps. The motion was only loosely controlled, most sequences containing significant yaw and pitch, and some roll. Illumination conditions were mildly different in training and test sequences, see Figures 8 and 9.

Table 1
The distribution of age for the database used in the experiments.

Age	18-25	26-35	36-45	46-55	65+
Percentage	29%	45%	15%	7%	4%

Fig. 8. *Frames from a typical video sequence used for method evaluation in this paper.*

Fig. 9. *A single face motion sequence after face localization and preprocessing.*

6.2 Results

The performance of evaluated recognition algorithms is summarized in Table 2. The results suggest a number of conclusions.

Firstly, note the relatively poor performance of the two nearest neighbour-type methods – the Set NN and the Majority vote using Eigenfaces. These can be considered as a proxy for gauging the difficulty of the recognition task, seeing that both

Table 2
Results of the comparison of our novel algorithm with existing methods in the literature. Shown is the identification rate in %.

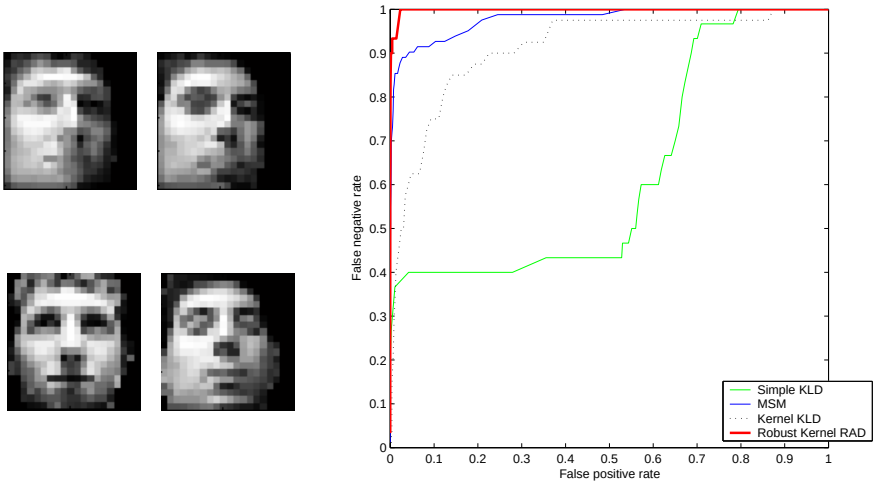
	DB 1	DB 2	Avg.
Robust Kernel RAD	98	98	98
MSM	89	88	88
Kernel RAD	90	87	88
Kernel KLD	78	80	79
Set Nearest Neighbour	70	73	72
Majority Vote w/ Eigenfaces	72	70	71
Simple KLD	40	63	52

can be expected to perform relatively well if the imaging conditions are not greatly different in training and test data sets. Inspection of the incorrect recognitions of these methods offered an interesting insight in a particular weakness of these algorithms, see Figure 10 (a). This reaffirms the conclusion of [28], showing that it is not only changes in illumination that are problematic, but that there are also certain intrinsically difficult imaging configurations.

The Simple KLD method consistently achieved the poorest results on our database. We believe that the likely reason for this is the high nonlinearity of face manifolds corresponding to training sets used, caused by near, office lighting used to vary the illumination conditions. This is supported by the dramatic and consistent increase in the recognition performance with kernelization. This result confirms the first premise of this work, showing that sophisticated face manifold modelling is indeed needed to accurately describe variations that are expected in realistic imaging conditions. Furthermore, the improvement observed with the use of Resistor-Average distance suggests its greater robustness with respect to unseen variations in face appearance, compared to the KL divergence. Kernel RAD performance is comparative to that of MSM, which ranked second-best. The best performing algorithm was found to be the proposed Robust Kernel RAD. Significant improvement in recognition ($\sim 10\%$) with synthetic manifold repopulating was found. ROC curves corresponding to the methods that best illustrate the contributions of this paper are shown in Figure 10 (b), with Robust Kernel RAD achieving an Equal Error Rate of 2%.

7 Summary and conclusions

In this paper we introduced a novel approach to face recognition from Face Motion Manifolds. In the proposed algorithm the Resistor-Average distance computed on



(a) (b)

Fig. 10. *(a) Receiver Operator Characteristic (ROC) curves of the Simple KLD, MSM, Kernel KLD and the proposed Robust Kernel RAD methods. The latter can be seen to exhibit superior performance, achieving an Equal Error Rate of 2%. (b) The most common failure mode of NN-type recognition algorithms is caused by "hard" illumination conditions and head poses. The two top images show faces of two different people that due to severe illumination and half-profile head orientation look very similar (see [28]) – Set NN incorrectly classified these fames as belonging to the same person. Information from other frames (e.g. the two images on the bottom) is not used for a more robust similarity measure.*

nonlinearly mapped data using Kernel PCA is used as a dissimilarity measure between distributions of face appearance, derived from video. A data-driven method for generalizing to unseen modes of variation is described, resulting in stochastic manifold repopulating. Finally, the proposed concepts were empirically evaluated on a database with 100 individuals and mild illumination variation. Our method consistently achieved a high recognition rate, on average correctly recognizing in 98% of the cases and outperforming state-of-the-art algorithms in the literature.

The findings of this paper suggest a number of promising research directions. Firstly, we intend to explore other means of modelling FMMs, with emphasis on methods that provide a more principled way of dealing with noise and can deal with manifolds of higher intrinsic dimensionality. Additionally, the results presented in this paper suggest that *a priori* learning of reliability of head poses and illumination conditions could be used for information fusing from multiple frames resulting in a probabilistic estimate on the confidence of a recognition decision.

Acknowledgements

We would like to thank the Toshiba Corporation for their kind support for our research, the people from the University of Cambridge Engineering Department who

volunteered to have their face videos entered in our face database and Trinity College, Cambridge.

References

[1] W. A. Barrett, A survey of face recognition algorithms and testing results., Systems and Computers 1 (1998) 301–305.

[2] R. Chellappa, C. L. Wilson, S. Sirohey, Human and machine recognition of faces: A survey., Proceedings of the IEEE 83 (5) (1995) 705–740.

[3] T. Fromherz, P. Stucki, M. Bichsel, A survey of face recognition., MML Technical Report. (97.01).

[4] W. Zhao, R. Chellappa, A. Rosenfeld, P. J. Phillips, Face recognition: A literature survey., UMD CFAR Tech. Report CAR-TR-948.

[5] R. Gross, J. She, J. F. Cohn, Quo vadis face recognition., Workshop on Empirical Evaluation Methods in Computer Vision 1 (2001) 119–132.

[6] B. Moghaddam, W. Wahid, A. Pentland, Beyond eigenfaces - probabilistic matching for face recognition, IEEE International Conference on Automatic Face and Gesture Recognition (1998) 30–35.

[7] W. S. Yambor, Analysis of PCA-based and fisher discriminant-based image recognition algorithms., Master's thesis, Colorado State University (2000).

[8] D. S. Bolme, Elastic bunch graph matching., Master's thesis, Colorado State University (2003).

[9] B. Kepenekci, Face recognition using gabor wavelet transform., Ph.D. thesis, The Middle East Technical University (2001).

[10] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model., IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (9) (2003) 1063–1074.

[11] S. Romdhani, V. Blanz, T. Vetter, Face identification by fitting a 3D morphable model using linear shape and texture error functions., In Proc. IEEE European Conference on Computer Vision (2002) 3–19.

[12] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video., Computer Vision and Image Understanding 91 (1) (2003) 214–245.

[13] K. Lee, M. Yang, D. Kriegman, Video-based face recognition using probabilistic appearance manifolds., In Proc. IEEE Conference on Computer Vision and Pattern Recognition.

[14] K. Fukui, O. Yamaguchi, Face recognition using multi-viewpoint patterns for robot vision., Int'l Symp. of Robotics Research.

[15] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence., IEEE International Conference on Automatic Face and Gesture Recognition (10) (1998) 318–323.

[16] G. Shakhnarovich, J. W. Fisher, T. Darrel, Face recognition from long-term observations., In Proc. IEEE European Conference on Computer Vision

(2002) 851–868.
[17] L. Wolf, A. Shashua, Learning over sets using kernel principal angles, Journal of Machine Learning Research 4 (10) (2003) 913–931.
[18] D. H. Johnson, S. Sinanović, Symmetrizing the Kullback-Leibler distance., Technical report, Rice University.
[19] T. M. Cover, J. A. Thomas, Elements of Information Theory., Wiley, 1991.
[20] O. Arandjelović, G. Shakhnarovich, J. Fisher, R. Cipolla, T. Darrell, Face recognition with image sets using manifold density divergence., In Proc. IEEE Conference on Computer Vision and Pattern Recognition.
[21] S. Yoshizawa, K. Tanabe, Dual differential geometry associated with Kullback-Leibler information on the gaussian distributions and its 2-parameter deformations., SUT Journal of Mathematics 35 (1) (1999) 113–137.
[22] B. Schölkopf, A. Smola, K. Müller, Kernel principal component analysis., Advances in Kernel Methods - SV Learning (1999) 327–352.
[23] M. E. Tipping, C. M. Bishop, Probabilistic principal component analysis, Journal of the Royal Statistical Society 3 (61) (1999) 611–622.
[24] A. M. Martinez, Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class., IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (6) (2002) 748–763.
[25] K. K. Sung, T. Poggio, Example-based learning for view-based human face detection., IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 39–51.
[26] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography., IEEE Transactions on Computers 24 (6) (1981) 381–395.
[27] P. Viola, M. Jones, Robust real-time face detection., International Journal of Computer Vision 57 (2) (2004) 137–154.
[28] T. Sim, S. Zhang, Exploring face space., IEEE Workshop on Face Processing in Video (2004) 84.