Exploring the Identity Manifold: Constrained Operations in Face Space

Ankur Patel and William A.P. Smith

Department of Computer Science, The University of York
{ankur,wsmith}@cs.york.ac.uk

Abstract. In this paper, we constrain faces to points on a manifold within the parameter space of a linear statistical model. The manifold is the subspace of faces which have maximally likely distinctiveness and different points correspond to unique identities. We show how the tools of differential geometry can be used to replace linear operations such as warping and averaging with operations on the surface of this manifold. We use the manifold to develop a new method for fitting a statistical face shape model to data, which is both robust (avoids overfitting) and overcomes model dominance (is not susceptible to local minima close to the mean face). Our method outperforms a generic non-linear optimiser when fitting a dense 3D morphable face model to data.

1 Introduction

Linear statistical models have been used to model variation in 2D [5] and 3D [3] shape, appearance and texture. These models are generative in nature, in the sense that instances similar to those used to train the model can be computed from a low dimensional parameter vector. Faces have proven a particularly suitable class to model using such approaches.

Perhaps the best known statistical face model is the Active Appearance Model (AAM) [5] which combines a linear model of 2D shape and 2D appearance. Rather than model appearance, the 3D Morphable Model of Blanz and Vetter [3] models the shape and texture which give rise to appearance via a model of image formation. Xiao et al.[17] have used a 3D model in conjunction with a 2D appearance model to enforce geometric constraints on the 2D shape generated.

Applying these models to face processing tasks requires a means to fit the model to observed data. This data may take many forms, such as the appearance of a face in one [3,5,17] or more [1,6] images, a noisy and incomplete 3D scan [2] or the location of a sparse set of feature points in an image [8]. Often this fitting process is underconstrained, prone to converge on local minima and computationally expensive. For these reasons, there is strong motivation for developing additional constraints to reduce the search space of the fitting process.

The most common method for learning such models from data, Principal Components Analysis (PCA), is based on the assumption that faces form a Gaussian cloud in a high dimensional space. The principal axes of this cloud

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part VI, LNCS 6316, pp. 112–125, 2010.

are estimated from a training sample, allowing any face to be approximated in terms of a small number of parameters.

Psychological results [16,11] have shown that this parameter space has an interesting perceptually-motivated interpretation: *identity* relates to direction in parameter space while *distinctiveness* is related to vector length (or equivalently distance from the mean). The reason for this is that increasing the length of a parameter vector simply exaggerates its differences from the average linearly, in other words its *features*, whereas rotating a parameter vector changes the *mix* of features present in the face. This is the justification for using angular difference in face space as a measure of dissimilarity for face recognition.

This decomposition also allows a useful probabilistic interpretation. Under the Gaussian assumption, each model parameter is independent and distributed according to a Gaussian distribution. This means that all faces lie on or near the surface of a hyper-ellipsoid in parameter space, with the probability density over the parameter vector lengths following a chi-square distribution. In other words, distinctiveness is subject to a statistical prior with the distinctiveness of most samples clustered around the expected length.

1.1 Contribution

In this paper, we use these observations to motivate a representation for faces which decomposes face appearance into identity and distinctiveness subspaces. We focus on statistical models of 3D face shape, though all of our results are equally applicable for any parametric face representation. We use ideas from differential geometry to develop tools which operate in the identity subspace, i.e. which retain constant distinctiveness. We provide empirical justification for constraining samples to have fixed distinctiveness, determined by the expected vector length.

We propose a new algorithm for fitting a statistical face model to data. Many such methods have been proposed, the details being dependent on the precise nature of the model and data. However, this inevitably involves a non-linear optimisation over the model parameters.

Examples include Cootes's [5] original algorithm for fitting AAMs to images which assumes that the relationship between error and optimal additive parameter updates is constant. Matthews and Baker's [9] inverse compositional algorithm avoided this assumption allowing faster and more robust convergence. In the domain of fitting 3D morphable models to single 2D images, Blanz and Vetter's [3] approach was to use a stochastic optimisation process in an analysis-by-synthesis framework in the hope of finding a global minimum. Careful initialisation and regularisation is required to obtain stable performance. Romdhani et al.[14] proposed an alternative approach which used additional features such as edges and specularities as part of the error term. The hope was to obtain a globally convex objective function, allowing local optimisation methods to arrive at the global optimum. All these approaches must trade off satisfaction of a model-based prior against quality of fit. To ensure robust performance, these approaches must favour the prior, resulting in model dominance.

Our approach operates via gradient descent on the manifold of equal distinctiveness. In other words, we solve for identity and assume distinctiveness takes its expected value. We show how the method naturally lends itself to a coarse-to-fine optimisation strategy and how the result avoids overfitting or local minima in which generic non-linear optimisers become stuck.

2 Statistical Modelling

Consider a sample of 3-dimensional face meshes which are in dense correspondence (i.e. the same point on every face has the same vertex index). The ith shape is represented by a vector of p vertices $\mathbf{s}_i = (x_1, y_1, z_1, \ldots, x_p, y_p, z_p) \in \mathbb{R}^{3p}$. Given m such shape vectors, we use principal components analysis to obtain an orthogonal coordinate system spanned by the m eigenvectors P_i. Any shape vector $\mathbf{s}$ may now be represented as a linear combination of the average shape and the model eigenvectors:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{i=1}^{m} c_i P_i, \tag{1}$$

where $\mathbf{c} = [c_1 \ \ldots \ c_m]^T$ is a vector of parameters. We stack the eigenvectors to form a matrix $\mathbf{P}$, such that we may write: $\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P}\mathbf{c}$. The PCA eigenvalues λ_i provide a measure of how much of the variance of the training data is captured by each eigenvector. We may choose to retain $n < m$ model dimensions, such that a certain percentage of the cumulative variance is captured. We discuss the effect of the number of model dimensions and empirically evaluate their stability in Section 2.2.

Our interest in this paper is to explore how shape samples drawn from a population distribute themselves in parameter space and how we can use this knowledge to constrain operations. We define the vector $\mathbf{b} = [c_1/\sqrt{\lambda_1} \ \ldots \ c_n/\sqrt{\lambda_n}]^T$ as the variance-normalised parameter vector. This vector is distributed according to a multivariate Gaussian with zero mean and unit variance, i.e. $\mathbf{b} \sim \mathcal{N}(0, \mathbf{I}_n)$. This is the prior constraint typically used in the model fitting process to ensure that solutions remain plausible. It is maximised by a zero vector, which corresponds to the mean sample.

However, another interpretation based on the parameter vector length is possible. The squared norm of $\mathbf{b}$ corresponds to the square of the Mahalanobis distance of $\mathbf{c}$ from the mean:

$$\|\mathbf{b}\|^2 = D_M^2(\mathbf{c}) = \sum_{i=1}^{n} \left(\frac{c_i}{\sqrt{\lambda_i}} \right)^2. \tag{2}$$

Since we assume each parameter follows a Gaussian distribution, the parenthesised terms are independent, normally distributed random variables with zero mean and unit variance. The sum of the square of such variables follows a chi-square distribution with n degrees of freedom, i.e. $\|\mathbf{b}\|^2 \sim \chi_n^2$. This distribution has expected value n and variance $2n$.

These two apparently contradictory distributions suggest that the mean face is the most probable sample but has a highly improbable vector length. For example, a model with 100 dimensions would have an expected vector length of 100 and over 99% of parameter vectors would have lengths between 70 and 130. The probability of a vector length less than 50 is negligibly small.

2.1 Identity as Direction

From the discussion above, it is clear that valid members of the class will occupy a subset of parameter space. These points will lie close to the surface of a hyperellipsoid, the diameters of which are determined by the eigenvalues of the data and the variance of the distance of samples from the manifold determined by the number of model dimensions. It is worth noting that as the number of dimensions increases, so the variance increases and the distance of samples from the manifold increases. Hence, the validity of assuming points lie on the surface of the hyperellipsoidal manifold breaks down as the number of model dimensions increases. Nevertheless, psychological results show us that the dimensionality of face space is relatively small (Meytlis and Sirovich [10] suggest 100 dimensions is sufficient, even using a crude eigenface model).

The analysis of data on a hyperellipsoidal manifold is extremely complex. Therefore, without loss of generality, we transform the manifold to a hypersphere by scaling each dimension by its corresponding standard deviation. By constraining faces to lie on the surface of this manifold, we maintain equal distinctiveness and ensure that only faces with the most probable distinctiveness can be generated. For the remainder of this paper, we therefore represent parameter vectors with squared Mahalanobis length n as unit vectors in $\mathbb{R}^n$: $\mathbf{x} = \frac{1}{\sqrt{n}} \left[\frac{c_1}{\sqrt{\lambda_1}} \ \cdots \ \frac{c_n}{\sqrt{\lambda_n}} \right]^T$.

A unit vector in n-dimensional space $\mathbf{x} \in \mathbb{R}^n$, may be considered as a point lying on the hyperspherical manifold $x \in S^{n-1}$. The two are related by $\mathbf{x} = \Phi(x)$ where $\Phi : S^{n-1} \mapsto \mathbb{R}^n$ is an embedding. If $v \in T_b S^{n-1}$ is a vector in the tangent space to S^{n-1} at a base point $b \in S^{n-1}$, the *exponential map*, denoted Exp_b of v is the point on S^{n-1} along the geodesic in the direction of v at distance $\|v\|$ from b. The inverse of the exponential map is the log map, denoted Log_b.

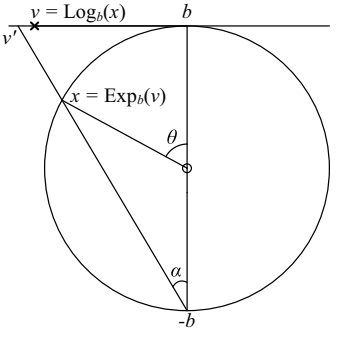

Fig. 1. Computing log and exponential maps using a stereographic projection for the S^1 manifold

The geodesic distance (i.e. angular difference) between two points on the unit hypersphere $x_1, x_2 \in S^{n-1}$ can be expressed in terms of the log map, i.e. $d(x_1, x_2) = \|\mathrm{Log}_{x_1}(x_2)\| = \arccos(\Phi(x_1) \cdot \Phi(x_2))$.

We propose a novel implementation of the exponential and log maps for a unit hypersphere which is both simple and efficient. We do so using a stereographic

projection. The log map of a point x at basepoint b is calculated as follows. We define the tangent vector $v' \in T_b S^{n-1}$ as the stereographic projection of x from $-b$ to the tangent space to S^{n-1} at b. This tangent vector has the correct direction but incorrect magnitude. To obtain the log map of x, we rescale v' giving v, such that $\|v\| = d(b, x)$. The exponential map is computed by reversing this process, i.e. by applying an inverse stereographic projection to the rescaled tangent vector. Figure 1 clarifies the geometry involved for the S^1 case.

In practice, we represent points on both the hyperspherical manifold and the tangent space as vectors embedded in $\mathbb{R}^n$. Hence, our proposed implementation of the log map of x at base point b is computed as:

$$\Phi_T \left(\mathrm{Log}_b(x) \right) = \mathbf{b} + \frac{\theta(\mathbf{v}' - \mathbf{b})}{\|\mathbf{v}' - \mathbf{b}\|}, \tag{3}$$

where $\mathbf{b} = \Phi(b)$ and $\mathbf{x} = \Phi(x)$ are both unit vectors in $\mathbb{R}^n$,

$$\mathbf{v}' = \frac{2(\mathbf{b} + \mathbf{x})}{\|\mathbf{b} + \mathbf{x}\| \cos\alpha} - \mathbf{b}, \qquad \alpha = \arccos\left(\frac{4 + \|\mathbf{b} + \mathbf{x}\|^2 - \|\mathbf{x} - \mathbf{b}\|^2}{4\|\mathbf{b} + \mathbf{x}\|} \right), \tag{4}$$

and $\theta = \arccos(\mathbf{b} \cdot \mathbf{x})$.

The result is a point in the tangent space $T_b S^{n-1}$ embedded in $\mathbb{R}^n$ according to an arbitrary embedding $\Phi_T : T_b S^{n-1} \mapsto \mathbb{R}^n$. A similar expression can be derived for the exponential map. These expressions hold for unit vectors in any number of dimensions. In the remaining sections, we use the log and exponential map to derive useful operations on the manifold.

2.2 Empirical Evaluation: χ^2 Prediction

Before we consider applications of processing data on the manifold described above, we provide some empirical assessment of how well the theoretically predicted manifold adheres to real world data. In order for all plausible data samples to lie on or near the manifold, the assumption of parameter vector lengths following the chi-squared distribution must hold. In turn, the distribution of faces along each eigenvector must follow a Gaussian distribution. In practice, these eigenvectors are estimated from a sparse sample of a high dimensional space. In the case of a dense 3D face shape model, observations typically consist of tens of thousands of vertices while the training set typically comprises only hundreds of samples.

Clearly, the validity of the estimated manifold depends on the quality of the estimated eigenvectors and therefore the size and diversity of the training set. We empirically evaluate how well unseen data adheres to our assumptions. This allows us to determine how many model dimensions can be safely retained.

Our empirical test is conducted as follows. From a pool of 100 face meshes [15], we randomly select 75. We build a PCA model and project each of the remaining 25 out-of-sample data onto the model eigenvectors. We repeat this process 80 times, giving a total of 2000 out-of-sample parameter vectors. We analyse the

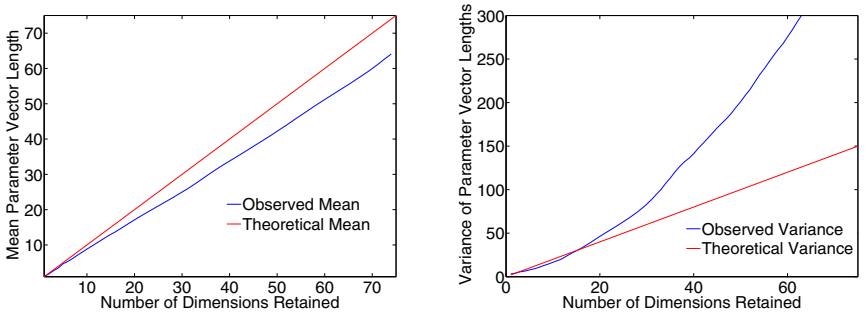

Fig. 2. Predicted versus observed mean (left) and variance (right) of out-of-sample parameter vector lengths

mean and variance of the squared-Mahalanobis length of these vectors and measure how well they agree with the predicted chi-square distribution. We would expect the mean and variance to grow linearly with the number of model dimensions retained. As can be seen in Figure 2, the observed mean lengths are close to, but smaller than, the theoretical prediction. On the other hand, the variance is only close to the predicted value for up to approximately 20 dimensions. Beyond this, the variance increases rapidly meaning many points will lie a significant distance from the manifold surface. We believe this is an effect of the sparsity of the training data. A much larger training set would allow this effect to be studied further. Nevertheless, we can see that for a modest number of dimensions, real world data does follow the statistical prediction reasonably well.

2.3 Empirical Evaluation: Manifold Approximation

The second empirical evaluation necessary to justify our approach, is to assess the error induced by forcing all samples to lie on the manifold, i.e. enforcing a hard constraint on vector length. Given an out-of-sample face, $\mathbf{s}$, the optimal parameter vector (in a least squares sense) is given by $\mathbf{c}^* = \mathbf{P}^T(\mathbf{s} - \overline{\mathbf{s}})$. Substituting $\mathbf{c}^*$ back into (1), we can obtain $\mathbf{s}^{\text{mod}}$, the shape which minimises $\|\mathbf{s}^{\text{mod}} - \mathbf{s}\|^2$. However, this shape is not constrained by the model prior and is almost always an overfit to the data. We compare this optimal model-based reconstruction to the shape, $\mathbf{s}^{\text{man}}$, obtained by projecting $\mathbf{c}^*$ to the closest point on the hyperspherical manifold:

$$\mathbf{c}^{\text{man}} = \frac{\sqrt{n}}{D_M(\mathbf{c}^*)}\mathbf{c}^*. \tag{5}$$

Over the 10 out-of-sample faces in the BFM [12] the mean Euclidian error of $\mathbf{s}^{\text{mod}}$ for a $n = 99$ parameter model was 1.128mm. By projecting to the S^{n-1} hypersphere, the mean Euclidian error of $\mathbf{s}^{\text{man}}$ increased to 1.89mm. The optimal

choice of $n-1$ dimensional subspace (with respect to Euclidian error) would be to simply retain the first $n-1$ eigenvectors of the PCA model. For our data, this gives a mean Euclidian error of only 1.133mm. However, the purpose of our choice of manifold is to enforce *plausibility*. This is reflected in the fact that error in the surface normals of the approximated faces (which in turn determines appearance), *reduces* when projecting to the manifold. For our data, the mean angular error drops from 5.92° for $\mathbf{s}^{\text{mod}}$ to 5.48° for $\mathbf{s}^{\text{man}}$. In other words, by constraining faces to be more plausible, we reduce appearance error.

3 Plausibility-Preserving Warps and Averages

Warping between faces or, more generally, computing weighted combinations of two or more faces has applications in animation and in the production of stimuli for psychological experiments [11]. The most obvious way to warp between two shapes that are in dense correspondence is to linearly warp each vertex from its position in one shape to its position in the other. Equivalently, this can be approximated by linearly warping between the two vectors of PCA parameters. However, in either case the intermediate faces will not correspond to plausible faces. Since the manifold of maximally probable distinctiveness is curved, any linear warp will include faces that do not lie on the manifold, with the least plausible face occurring halfway along the warp.

Face-antiface warps provide a particularly interesting special case. An antiface is the antipodal point of a source face on the manifold. Perceptually, antifaces appear "opposite" in some sense to the original face. The vector connecting a face to its antiface in parameter space passes through the mean. A linear warp between a face and antiface is therefore well-defined but will include implausible faces for the duration of the warp. There is a further problem with such linear warps. Psychological studies have shown that there is a perceptual discontinuity as the face trajectory crosses the mean [11]. In other words, as identity flips from face to antiface, the perceptual effect of a small movement through face space is exaggerated.

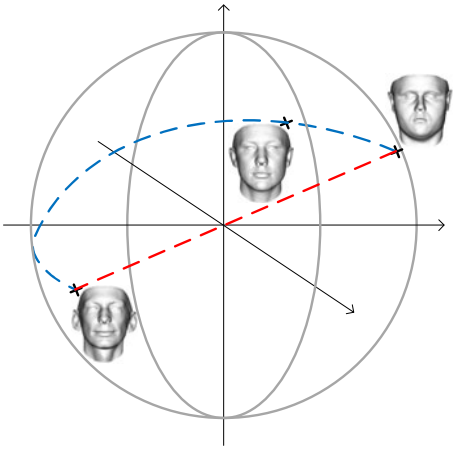

Fig. 3. Warping between face and antiface on the S^2 manifold. Linear warp is shown in red, one of the possible geodesic warps is shown in blue.

Instead, we propose warps which take place across the surface of the manifold, following the geodesic curve between the two source faces. Another way to view

these warps is as a rotation of a unit vector in $\mathbb{R}^n$. All intermediate faces in this case have equal distinctiveness and are equally plausible. In the case of antifaces, there is no single geodesic warp connecting face to antiface. In fact, there are an infinite number of valid warps, all of length π. Any such warp will smoothly vary identity from the source face to its antiface, via a series of faces with uniform distinctiveness. One way to conceptualise this is that we can set off from a point on the hyperspherical manifold in any direction and reach the antiface after travelling a distance π.

An interesting result of this observation is that we can choose any intermediate face as a target which will be visited on the warp from face to antiface. This gives us a way to specify one of the infinite face-antiface warps and may also have interesting applications in generating stimuli for psychological studies. This idea is demonstrated in Figure 3 for the S^2 manifold, which shows the difference between a plausibility-preserving and linear warp.

For a source face x_{src} and intermediate target face x_{tar}, we can define a unit vector in the tangent space, $v \in T_{x_{src}} S^{n-1}$, from x_{src} in the direction of x_{tar}: $v = \frac{\mathrm{Log}_{x_{src}}(x_{tar})}{d(x_{src}, x_{tar})}$. A geodesic warp from x_{src} to x_{tar} is therefore given by following this vector by a distance specified by the warping parameter w:

$$x_{war} = \mathrm{Exp}_{x_{src}} \left(w \frac{\mathrm{Log}_{x_{src}}(x_{tar})}{d(x_{src}, x_{tar})} \right). \tag{6}$$

When $w = 0$ we obtain the source face, i.e. $x_{war} = x_{src}$, and when $w = d(x_{src}, x_{tar})$ we obtain the target face, i.e. $x_{war} = x_{tar}$. If we set $w = \pi$ we obtain the antiface to x_{src}. Intermediate faces are obtained when $w \in (0, \pi)$.

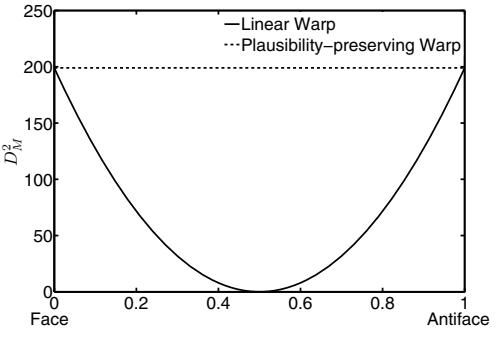

Fig. 4. Vector length or 'plausibility' is plotted throughout a warp between a face and antiface (see Figure 5)

We show an example warp from face to antiface via an intermediate target face in Figure 5 using the 199 parameter BFM [12]. Note that the effect is of smooth variation of identity, with each of the intermediate faces containing significant detail. We contrast this with a linear warp through the mean face which results in implausibly smooth intermediate faces and no transition through intermediate identities. In Figure 4 we plot the parameter vector lengths for the linear and plausibility-preserving warps.

3.1 Averages

Given $u > 2$ source faces, $x_1, \ldots, x_u \in S^{n-1}$, we wish to compute a plausible average face which captures characteristics of each of the source faces. The linear

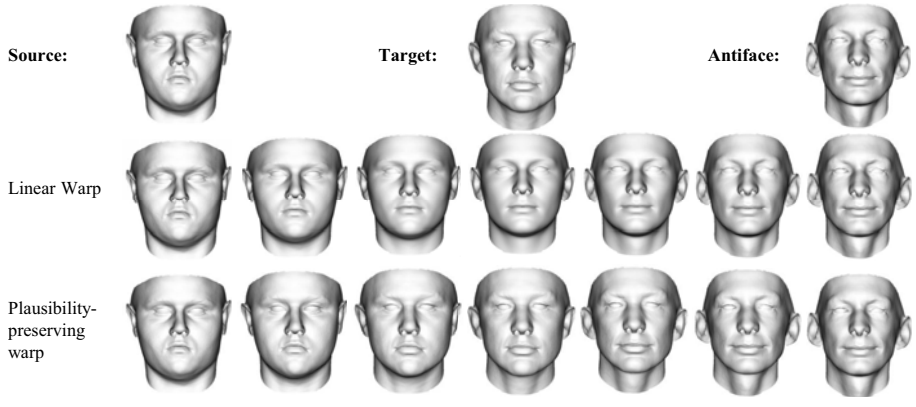

Fig. 5. Linear versus plausibility-preserving warp from face to antiface

or Euclidian mean of the parameter vectors minimises the sum of square error in $\mathbb{R}^n$ from the average to each of the source faces. This is the *extrinsic mean* and will not lie on the manifold. The result is that the face is implausibly smooth and lacking in features. We propose the use of the *intrinsic* or Karcher mean. For $u = 2$, this can be found using the warping equation given above with $w = 0.5$. For $u > 2$, this is the point $x_\mu \in S^{n-1}$ which minimises the total squared geodesic distance to each of the source faces:

$$x_\mu = \arg\min_{x \in S^{n-1}} \sum_{i=1}^{u} d(x, x_i)^2. \tag{7}$$

This point cannot be found analytically, so we solve it as an iterative optimisation using the gradient descent method of Pennec [13]. We initialise our estimate as one of the source data points, i.e. $x_\mu^{(0)} = x_1$. The estimated intrinsic mean is then iteratively updated as follows:

$$x_\mu^{(j+1)} = \mathrm{Exp}_{x_\mu^{(j)}} \left(\frac{1}{u} \sum_{i=1}^{u} \mathrm{Log}_{x_\mu^{(j)}}(x_i) \right). \tag{8}$$

This process converges rapidly, typically within 5 iterations. In Figure 6 we compare our plausibility-preserving averages with linear averaging of the 74 dimensional parameter vectors obtained using the USF data [15]. Notice that each of the Euclidian averages appears unrealistically smooth, whereas the averages computed on the manifold clearly show the presence of distinct features present in the source faces (for example, the broader nostrils of face 1 are visible in the first three averages but not the fourth).

4 Model Fitting on the Manifold of Plausible Faces

The most powerful application of the identity manifold is to use it for the purpose of constraining the process of fitting a model to data. Suppose the function

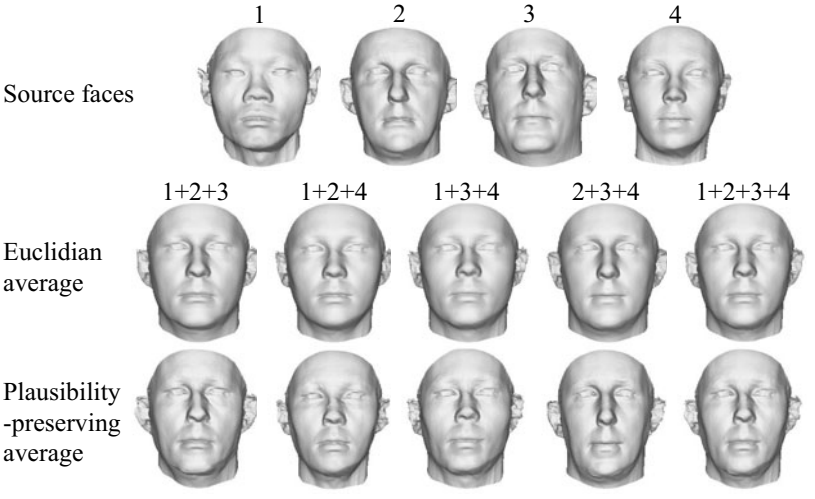

Fig. 6. Linear versus plausibility-preserving averages

$\varepsilon : S^{n-1} \mapsto \mathbb{R}$ is an objective function which evaluates the quality of fit of a face represented by a point on the plausibility manifold to some observed data. This function could take any form, for example the difference between predicted and observed appearance in an analysis-by-synthesis framework or the error between a sparse set of feature points. We pose model fitting as finding the point on the manifold which minimises this error, i.e.:

$$x^* = \underset{x \in S^{n-1}}{\arg\min}\, \varepsilon(x). \tag{9}$$

In doing so, we ensure that plausibility is enforced as a hard constraint. Note also that the optimisation is more heavily constrained since the dimensionality of the hypersphere is 1 less than the parameter space.

4.1 Local Optimisation

We can perform gradient descent on the surface of the manifold to find a local minimum in the error function. The fact that our manifold is hyperspherical has some interesting implications for such an approach. We must first compute the gradient of the objective function in terms of a vector on the tangent plane: $\nabla\varepsilon(x) \in T_x S^{n-1}$. To do so, we compute the gradient in terms of a vector in $\mathbb{R}^n$ and project the result to the tangent plane as follows:

$$\nabla\varepsilon(x) = \mathrm{Log}_x \left(\Phi^{-1} \left(\frac{\mathbf{x} - \mathbf{g}}{\|\mathbf{x} - \mathbf{g}\|} \right) \right), \tag{10}$$

where $\mathbf{x} = [x_1 \ \dots \ x_n]^T = \Phi(x)$. The gradient $\mathbf{g} = [\partial_{x_1}\varepsilon(x) \ \dots \ \partial_{x_n}\varepsilon(x)]^T$ is approximated by using finite differences to calculate the partial derivatives:

$$\partial_{x_i}\varepsilon(x) \approx \frac{\varepsilon(x_i') - \varepsilon(x)}{\epsilon}, \tag{11}$$

where $x_i' = \Phi^{-1}([x_1 \ \dots \ x_i + \epsilon \ \dots \ x_n])$.

With a means to compute the gradient, we can iteratively minimise the objective function by adapting the gradient descent algorithm to operate on the surface of a manifold:

$$x^{(t+1)} = \mathrm{Exp}_{x^{(t)}}\left(-\gamma\nabla\varepsilon(x^{(t)})\right), \tag{12}$$

where γ is the step size. Note that as γ varies, the point $\mathrm{Exp}_x\left(-\gamma\nabla\varepsilon(x)\right) \in S^{n-1}$ traces out a great circle about the hypersphere. This is the search space for the one-dimensional line search at each iteration of gradient descent.

4.2 Coarse-to-Fine Model Fitting

The difficulty with our approach is choosing an unbiased initialisation. Existing methods for fitting statistical models to data typically commence from an initialisation of the mean (i.e. zero parameter vector), e.g. [3,5]. However, this point lies far from the plausibility manifold and is therefore unsuitable in our case.

We tackle this problem and also reduce susceptibility to becoming trapped in local minima by proposing a coarse-to-fine algorithm which iteratively increases the number of model dimensions considered in the optimisation.

Consider in the simplest case a 1-dimensional model. Only two points strictly satisfy the plausibility constraint in this case and the problem therefore reduces to a binary decision:

$$\mathbf{x}^{(1)} = \begin{cases} [\ 1\ ] & \text{if } \varepsilon(\Phi^{-1}([\ 1\ ])) < \varepsilon(\Phi^{-1}([\ -1\ ])) \\ [\ -1\ ] & \text{otherwise} \end{cases}, \tag{13}$$

We use this result to initialise the solution in two dimensions, initially setting the second parameter to zero: $\mathbf{x}_{\text{init}}^{(n)} = \left[\mathbf{x}^{(n-1)} \mid 0\right]$. We then perform gradient descent, which in the two parameter case means optimising a single angular parameter. We continue this process, incrementally adding dimensions to the optimisation, each time setting the new parameter to zero and then performing gradient descent on the new manifold using this as an initialisation. Hence, the result of a local optimisation in n dimensions is used as the initialisation for optimisation in $n+1$ dimensions ensuring that the solution is already constrained to the right region of the manifold.

The nature of the hyperspherical manifold can be used to inform the step size used in the gradient descent optimisation. We assume that the result in n dimensions has restricted the solution to the correct hemisphere of the hypersphere. Travelling in the direction of the negative gradient reduces the error. To travel

in this direction whilst remaining in the same hemisphere means the maximum arc distance that can be moved is $\frac{\pi}{2}$. Hence, the result in n dimensions is given by $\mathbf{x}^{(n)} = h(d^*)$, where

$$h(d) = \mathrm{Exp}_{\Phi^{-1}(\mathbf{x}_{\mathrm{init}}^{(n)})} \left(d \frac{-\nabla\varepsilon\left(\Phi^{-1}(\mathbf{x}_{\mathrm{init}}^{(n)})\right)}{\left\|\nabla\varepsilon\left(\Phi^{-1}(\mathbf{x}_{\mathrm{init}}^{(n)})\right)\right\|} \right). \tag{14}$$

The arc distance d determines how far we travel along the great circle implied by the gradient of the objective function. Since we wish to constrain our solution to the same hemisphere, d must lie in the interval $(0, \frac{\pi}{2})$ and we hence find d^* using golden section search [7] to solve: $d^* = \arg\min_d h(d),\ 0 < d < \frac{\pi}{2}$. Multiple iterations of gradient descent can be used each time a dimension is added to the optimisation. In our results we use four iterations per dimension.

4.3 Model Fitting Example

For our experimental evaluation, we use the algorithm described above to fit our 3D morphable shape model to unseen data. We choose as an objective function the angular error between surface normals at each vertex of the model. This is an interesting choice of objective function for two reasons. First, the search landscape of the objective function is littered with local minima. Second, the fitted result is likely to have lower perceptual error than a least squares fit directly to the vertices. Whilst such a least squares fit gives minimal geometric error, the result is often a gross over-fit which does not resemble the input face. Minimising the surface normal error is a non-linear problem which is related to minimising appearance error, as undertaken by analysis-by-synthesis of image data [3].

From an input face shape, represented by p vertices, we compute surface normals at each vertex by averaging face normals of faces adjacent to the vertex. If $\mathbf{N}^i$ is the surface normal at vertex i, our objective function is the sum of squared angular errors between input and model surface normals:

$$\varepsilon(x) = \sum_{i=1}^{p} \left(\arccos(\mathbf{n}^i(\Phi(x)) \cdot \mathbf{N}^i)\right)^2, \tag{15}$$

where $\mathbf{n}^i([x_1\ \ldots\ x_n])$ is the surface normal of the ith vertex of the shape given by: $\overline{\mathbf{s}} + \mathbf{P}\mathbf{c}$, where the parameter is vector is computed by transforming the unit vector back to the hyperellipse: $\mathbf{c} = \sqrt{n}\left[x_1\sqrt{\lambda_1}\ \ldots\ x_n\sqrt{\lambda_n}\right]^T$.

We compare our manifold optimisation with direct optimisation of (15) using a generic optimiser based on the BFGS Quasi-Newton method with a cubic line search [4]. Note that the generic optimiser converges close to the mean if all parameters are optimised simultaneously. We therefore take the same coarse-to-fine approach as for the manifold fitting, whereby we iteratively increase the number of dimensions considered in the optimisation.

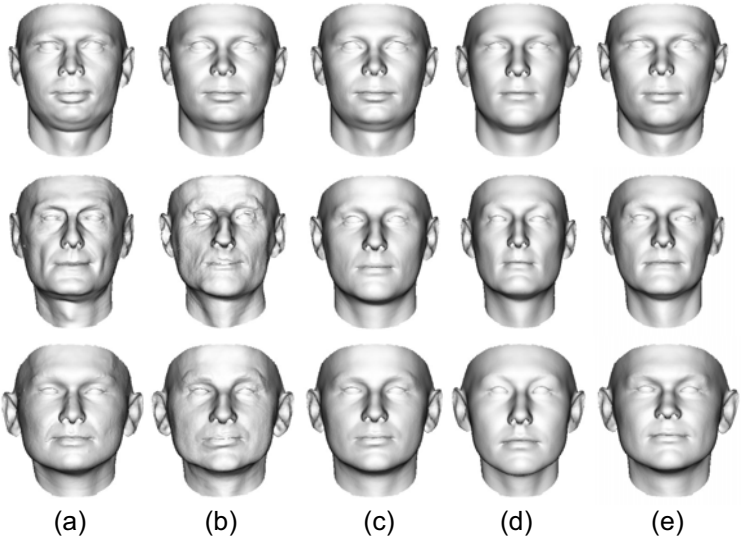

Fig. 7. Model fitting result: (a) input unseen face; (b) least squares fit to vertices; (c) parameter vector of (b) rescaled to manifold; (d) BFGS optimisation; (e) manifold optimisation. All the results are for a $n = 99$ parameter model.

In Figure 7 we show results on the BFM [12] data. Column (a) shows input faces which are not in the morphable model training set. A simple linear least squares fit of the model to the vertices of the unseen faces yields the result in column (b). Whilst this result is optimal in terms of the Euclidian error between input and reconstructed vertices, the result is an overfit and, in particular, the face in row 2 is clearly implausible. Rescaling the parameter vector obtained by least squares to the closest point on the manifold yields the result shown in column (c). While this face is now plausible, it lacks any of the distinguishing features of the input faces. Column (d) shows the result of using a generic non-linear optimiser to solve (15). Because of local minima close to the mean, these faces are implausibly smooth. Finally, our manifold fitting result is shown in column (e). Note that this result represents a trade off between over and under-fitting. The mean angular error of the surface normals for the out-of-sample faces in the BFM using (d) is 7.23°, while using the proposed method the error is 5.33°. Our result outperformed the generic non-linear optimiser for all of the BFM faces.

5 Conclusions

We have shown how a number of useful operations can be performed on the manifold of equally distinctive faces. This provides a new way to constrain operations involving the parameters of a statistical model. In particular, we have

shown how to constrain the process of fitting a model to data and how a coarse-to-fine strategy avoids local minima. Matlab implementations are available at (`http://www.cs.york.ac.uk/~wsmith/ECCV2010.html`). In future work, we intend to apply our model fitting strategy to more demanding objective functions and to experiment with other sources of data besides faces.

References

1. Amberg, B., Blake, A., Fitzgibbon, A., Romdhani, S., Vetter, T.: Reconstructing high quality face-surfaces using model based stereo. In: Proc. ICCV (2007)
2. Blanz, V., Scherbaum, K., Seidel, H.P.: Fitting a morphable model to 3D scans of faces. In: Proc. ICCV (2007)
3. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. Pattern Anal. Mach. Intell. 25(9), 1063–1074 (2003)
4. Broyden, C.G.: The convergence of a class of double-rank minimization algorithms. Journal of the Institute of Mathematics and Its Applications 6(1), 76–90 (1970)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
6. Hu, C., Xiao, J., Matthews, I., Baker, S., Cohn, J., Kanade, T.: Fitting a single active appearance model simultaneously to multiple images. In: Proc. BMVC (2004)
7. Kiefer, J.: Sequential minimax search for a maximum. Proceedings of the American Mathematical Society 4(3), 502–506 (1953)
8. Knothe, R., Romdhani, S., Vetter, T.: Combining PCA and LFA for surface reconstruction from a sparse set of control points. In: Proc. Int. Conf. on Automatic Face and Gesture Recognition, pp. 637–644 (2006)
9. Matthews, I., Baker, S.: Active appearance models revisited. Int. J. Comput. Vis. 60(2), 135–164 (2004)
10. Meytlis, M., Sirovich, L.: On the dimensionality of face space. IEEE Trans. Pattern Anal. Mach. Intell. 29(7), 1262–1267 (2007)
11. O'Toole, A.J., Vetter, T., Volz, H., Salter, E.M.: Three-dimensional caricatures of human heads: Distinctiveness and the perception of facial age. Perception 26(6), 719–732 (1997)
12. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: Proc. IEEE Intl. Conf. on Advanced Video and Signal based Surveillance (2009)
13. Pennec, X.: Probabilities and statistics on Riemannian manifolds: basic tools for geometric measurements. In: Proc. IEEE Workshop on Nonlinear Signal and Image Processing (1999)
14. Romdhani, S., Vetter, T.: Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: Proc. CVPR, vol. 2, pp. 986–993 (2005)
15. Sarkar, S.: USF humanid 3D face database (2005)
16. Valentine, T.: A unified account of the effects of distinctiveness, inversion, and race in face recognition. Quarterly Journal of Experimental Psychology A 43(2), 161–204 (1991)
17. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real–time combined 2D+3D active appearance models. In: Proc. CVPR, pp. 535–542 (2004)