Gaussian Mixture 3D Morphable Face Model

Paul Koppen\textsuperscript{a}, Zhen-Hua Feng\textsuperscript{a,b,}, Josef Kittler\textsuperscript{a}, Muhammad Awais\textsuperscript{a}, William Christmas\textsuperscript{a}, Xiao-Jun Wu\textsuperscript{b}, He-Feng Yin\textsuperscript{b}

\textsuperscript{a}Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK
\textsuperscript{b}School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China

Abstract

3D Morphable Face Models (3DMM) have been used in pattern recognition for some time now. They have been applied as a basis for 3D face recognition, as well as in an assistive role for 2D face recognition to perform geometric and photometric normalisation of the input image, or in 2D face recognition system training. The statistical distribution underlying 3DMM is Gaussian. However, the single-Gaussian model seems at odds with reality when we consider different cohorts of data, e.g. Black and Chinese faces. Their means are clearly different.

This paper introduces the Gaussian Mixture 3DMM (GM-3DMM) which models the global population as a mixture of Gaussian subpopulations, each with its own mean. The proposed GM-3DMM extends the traditional 3DMM naturally, by adopting a shared covariance structure to mitigate small sample estimation problems associated with data in high dimensional spaces. We construct a GM-3DMM, the training of which involves a multiple cohort dataset, SURREY-JNU, comprising 942 3D face scans of people with mixed backgrounds. Experiments in fitting the GM-3DMM to 2D face images to facilitate their geometric and photometric normalisation for pose and illumination invariant face recognition demonstrate the merits of the proposed mixture of Gaussians 3D face model.

Keywords: Gaussian-mixture Model, 3D Morphable Model, 3D Face Reconstruction, Face Model Fitting, Face Recognition

\textsuperscript{*}Corresponding author
Email address: z.feng@surrey.ac.uk (Zhen-Hua Feng)
1. Introduction

Face recognition technology has made an immense progress during the last decade, first thanks to the advances in face representation in the form of innovative features such as Local Binary Patterns (LBP) \cite{38, 39}, Local Phase Quantisation (LPQ) \cite{40} and Binarised Statistical Image Features (BSIF) \cite{29}, and more recently, to the capabilities of end-to-end deep learning neural networks \cite{51, 53}. As a result, for near frontal faces, captured in reasonable environmental conditions, the reported recognition rates match or exceed human performance \cite{33, 53}. However, unconstrained face recognition, characterised by extreme poses and by unfavourable illumination conditions, still poses a challenge. One of the difficulties is the lack of data deemed representative of all the appearance variations that can be encountered in realistic scenarios. In the context of the limited availability of training data that dramatically curtails the potential of machine learning technology, it is pertinent to ask what role face models can play as a source of prior knowledge that could be combined with machine learning to push the current frontiers of face recognition even further.

The problem of face modelling has been studied intensively for more than two decades. The most commonly researched have been the various variants of 2D active shape and appearance models \cite{7, 8}. Issues relating to both, model construction and model fitting, have been investigated \cite{16, 21, 24, 35, 36, 54}. To extend the capacity of 2D models so as to capture different modes of shape and appearance variations, the early frontal face 2D morphable models have been generalised to multiview \cite{9, 20} and most recently multilinear models \cite{17, 31}. However, as in the deep learning network case, the development of these multimodal 2D face models is severely limited by the lack of training data. This limiting factor applies to a much lesser extent to 3D face models \cite{2}. Implicitly, a 3D face model can render different 2D views of a face to an arbitrary range and precision of pose angles. Thus for each subject it provides information equivalent to hundreds of images of different poses. In addition, by physically separating the face model from an illumination model, it can also generate an
arbitrary number of views in different illumination conditions. In the past, the widespread adoption of 3D face models has been prevented by the cost of 3D face sensor technology. However, several 3D face models have recently been made available [4, 28, 42], the most prominent of which is the Basel face model.

Most of the 3D face models are of the morphable variety [2, 3]. Invariably, their construction involves the use of Principal Component Analysis (PCA) to decorrelate 3D face data represented in terms of a 3D mesh of spatial coordinate samples and associated RGB surface texture values. Prior to the PCA analysis, the available 3D scans are registered to a common mesh size. By virtue of the registration process, a particular vertex in the mesh has the same semantic identity for all facial scans (e.g. tip of the nose, corner of the left eye). By decorrelating the 3D face data (shape, texture), PCA determines the low dimensional subspace where the 3D face data lies, as well as the variables which represent the independent modes of face data variation. This subspace is defined by the eigenvectors associated with non-zero eigenvalues of the data covariance matrix. In the PCA space, each face instance is represented by the shape and texture parameters (coefficients of projection of the raw 3D face onto the retained eigenvectors). The name ‘morphable’ model derives from the fact that, by varying the shape parameters of the model within the range defined by the magnitude of the respective eigenvalues, the geometric form of the generated face changes and produces face morphing. Similarly, the face texture can be altered by changing the texture parameters. The ability to change the shape and texture is instrumental in fitting the 3D model to an input 2D face image to perform its frontalisation and photometric correction.

By changing the shape and texture parameters, different samples can be drawn from the morphable model. As the underlying statistics of the model is of first and second order only, the implicit assumption is that the 3D morphable model is Gaussian. If this is incorrect, the faces generated by sampling this Gaussian distribution may be unrealistic. Notwithstanding this possibility, even if the distribution is Gaussian for a homogeneous population of faces, this assumption is unlikely to hold for heterogeneous populations. Consider,
for instance, two cohorts, namely Caucasian and Chinese faces. The shapes of these two groups are clearly different. The noses of the former group protrude more from the best fitting face plane than for the Chinese cohort. Similarly, the vertical eye aperture is greater for the Caucasians than for the Chinese. There are other facial features which differ for the two groups. Thus the means of these two cohorts will be different. The within cohort variations around these means, as represented by the cohort-specific covariance matrices, may also be different. A similar list of differences could be found between, say, the African ethnicity, and the above two groups. This analysis suggests that a more appropriate model to construct is a mixture of Gaussians model, where each cohort of distinct ethnicity constitutes a mode in the distribution.

Compared with the single global Gaussian model (3DMM), a mixture model is likely to have a more compact component for each ethnic group. In contrast the range of parameter values for the global model would be relatively large, modelling not only the within class variations and their correlations, but also the differences in the cohort mean values.

In this paper we propose a Gaussian mixture 3D morphable face model (GM-3DMM) constructed using Caucasian, Chinese and African 3D face data. Each cohort has a separate mean, but we assume that the within cohort covariance matrices are common. This mitigates any small sample estimation problems arising when dealing with high dimensional data. A fusion of the within cohort covariance matrices has the additional advantage that the older subjects texture can be propagated from one cohort, when it is available, to another, where it may be lacking due to unstratified sampling problems. Drawing synthetic samples from the mixture of Gaussians is also less likely to generate phantom faces. Most importantly, the proposed mixture model has significant advantages in the context of 3D assisted 2D face recognition. We show that fitting GM-3DMM to an input 2D face image is more accurate for two reasons. First of all, the starting point of the fitting process, the appropriate cohort mean, is closer to the actual solution, and therefore the likelihood of getting stuck in a local optimum is lower. More over, as each cohort mode is more compact than the global model, the
regularisation imposed by the model on the fitting process is better targeted. We show experimentally on 2D texture face images for which we have 3D ground truth that the reconstruction error obtained using the GM-3DMM fitting is lower than its counterpart yielded by the global 3D morphable face model. We also show that the results of face recognition experiments conducted on the Multi-PIE dataset, which exhibits extreme pose and illumination variations, are superior to those achievable with 3DMM.

The paper is organised as follows. In Section 2 we present the related literature. The 3D morphable face model and its fitting to 2D images are overviewed in Section 3. The proposed Gaussian mixture 3D morphable face model is introduced in Section 4. The experiments conducted to demonstrate the merit of GM-3DMM are described in Section 5, which also presents a discussion of their results. The paper is drawn to conclusion in Section 6.

2. Related work

The generative 3D Morphable Face Model (3DMM) was first proposed by Blanz and Vetter [2]. It is constituted by two PCA-based parametric models, i.e. shape and texture models, that are trained from a set of exemplar 3D face scans. A 3DMM is able to generate realistic face instances by controlling its model parameters. In addition, lighting and camera models can be used to render such faces with appearance variations in pose and illumination. By fitting a 3DMM to a 2D face image, we can recover the 3D shape and texture information and estimate the scene properties (light and camera model parameters). Owing to these advantages, 3DMM has been widely used in many areas including, but not limited to, pattern recognition [3, 12, 22, 26, 52, 57]. For an overview of 3DMM’s applications the reader is referred to [30].

In practice, the use of a 3DMM is often limited by its representation capacity due to issues such as the size of the training set and data variety underlying the PCA-based model. Also, fitting a 3DMM to a single 2D image is very challenging. To address these issues, the state-of-the-art in 3DMM evolved on
two separate fronts: 3D face modelling and model fitting.

In 3D face modelling, the aim is to construct a 3DMM that has good representation capability as well as a compact structure. To this end, two main strategies have been investigated: 1) collecting a large number of 3D face scans with different population groups for 3DMM training; and 2) improving the underlying PCA method used for model construction.

For the former, capturing 3D face scans is very laborious; both data collection and its post-processing are tedious and time-consuming. In addition, high-quality 3D face capturing devices are relatively expensive. Notwithstanding these difficulties, a number of publicly available datasets and 3D face models have been released, e.g. the FRGC dataset [43], the Bosphorus 3D face dataset [50], the FaceWarehouse dataset [5], the Basel Face Model (BFM) [42] and the Surrey 3D face model [28]. However, both the Surrey and Basel face models were constructed using small datasets that lack diverse ethnicities. More recently, Imperial College has gained access to 10,000 3D face scans and proposed a fully automatic way to process the data and create different 3DMMs [4].

For the second strategy, instead of PCA, some other techniques have been investigated for model construction. The PCA method used in classical 3DMMs is not able to represent local facial details for both shape and texture information. To mitigate this problem, Lüthi et al. applied Gaussian Process to construct 3DMM, which was shown to exhibit better capacity in this respect [34]. Ferrari et al. used dictionary learning to form a 3D face shape model and achieved better reconstruction and fitting accuracy than the PCA-based 3DMM [19].

Once a 3DMM is constructed, we can fit it to 2D images using a model fitting algorithm. The purpose of the 3DMM fitting algorithm is to recover the 3D shape and texture information of a 2D face by solving a non-linear optimisation problem. In this process, a set of parameters, including the shape and texture model parameters as well as the parameters of the lighting model (Phong reflection model [44]) and perspective camera model, are estimated.

Classical 3DMM fitting algorithms are usually gradient-descent-based, in which the parameters are iteratively updated [2, 3, 18]. However, gradient-
descent-based approaches are easily trapped in local minima, especially when the initialised model parameters are far from the global optimum. In addition, the fitting of a 3D face model to 2D is accomplished by minimising a loss defined in 2D. In such a case, the recovered 3D face shape has to be projected into a 2D coordinate system hence the depth information is lost. Moreover, to separate the lighting and skin texture (albedo) from face appearance is ill-posed. Last, a gradient-descent-based approach has to iteratively calculate partial derivatives during the optimisation step. This expensive operation dramatically slows down 3DMM fitting and impedes its use cases in real-time applications.

To address the aforementioned issues, a number of techniques have been developed to improve the fitting accuracy and speed. For example, Romdhani and Vetter proposed to use occluding face contours and multiple image features to assist 3DMM fitting and achieved promising fitting accuracy \[49\]. The ‘linear shape and texture fitting’ (LiST) \[47\] method solves the 3DMM fitting algorithm in closed form, to improve the fitting speed. In this paper, we use the Efficient Stepwise Optimisation (ESO) fitting algorithm for our GM-3DMM fitting \[27\] as it has exhibited impressive 3DMM fitting performance in terms of both accuracy and speed.

By fitting a 3DMM to 2D face images, we are able to recover the 3D face (shape and texture) from a single 2D image. This reconstruction capability of 3DMM is very useful in face recognition. For example, the recovered 3D shape and texture model parameters are naturally robust to pose variations \[3, 25, 27\] and can be used directly for decision making. As shown in recent studies, the use of 3DMM even outperforms state-of-the-art deep neural networks in pose-invariant face recognition \[26, 27, 60, 61\]. In addition, the estimated light model can be used to deal with illumination variations. Alternatively, for a 2D query face with pose variations, we can fit a 3DMM to gallery images and render new gallery images with the same pose for face matching \[37, 45\]. Another solution for pose-invariant face recognition is to perform face frontalisation. In recent years, this strategy has been used successfully in many face matching problems \[18, 23, 53\]. In this paper, we demonstrate that our cohort-specific
GM-3DMM provides an even better basis for the above approaches.

3. Overview of 3D morphable face models

From the appearance point of view, apart from the albedo, the pertinent characteristic of 3D faces is their surface shape. A common way to represent it is in terms of a 3D surface mesh of vertices and their 3D coordinates. Given a set of vertices we can then associate with each vertex not only its geometric information (3D coordinates) but also its albedo, expressed in terms of a triplet of RGB values.

When dealing with faces, it is essential to register them to a common mesh of vertices where each vertex has a specific identity. It can be achieved by a process of registration [3, 46], which maps a canonical mesh template onto each raw 3D face image. Some examples of the original 3D face scans and the corresponding registered 3D faces are shown in Figure 1. Once 3D faces are registered, we can then build a statistical model representing the whole 3D face population.

Let the $i$th vertex $v_i$ of a registered face be located at $(x_i, y_i, z_i)$, and have the RGB colour values $(r_i, g_i, b_i)$. A registered face can be represented in terms of shape and texture as a pair of vectors:

$$s = (x_1, y_1, z_1, ..., x_n, y_n, z_n)^T, \quad t = (r_1, g_1, b_1, ..., r_n, g_n, b_n)^T$$ (1)
where \( n \) is the number of vertices.

As the samples of a face surface conveyed by its vertices are correlated, we may construct more concise statistical models by transforming the registered 3D face data into another coordinate system. A common approach is to remove redundancy by means of the Principal Component Analysis (PCA). A classical 3D morphable face model is constituted by two linked PCA-based parametric models that represent shape and skin texture properties (also known as albedo).

Let \( S \) and \( T \) denote the matrices of shape and texture principal components of dimensionality \( 3n \times n_s \) and \( 3n \times n_t \) of the respective shape and texture data covariance matrices. The decorrelated data spans a space of low dimensionality. It is defined by the number of eigenvectors that correspond to nonzero eigenvalues of the relevant data covariance matrices (the number of bases (columns) of these matrices) which is significantly smaller than the number of vertices \( n \) (\( 10^2 \) vs \( 10^4 \sim 10^5 \)). A sample from a 3D face distribution then can be represented as:

\[
s = s_0 + S\alpha, \quad t = t_0 + T\beta
\]  
(2)

where \( s_0 \) and \( t_0 \) are the mean shape and texture over all the training samples, \( \alpha = (\alpha_1, ..., \alpha_{n_s})^T \) and \( \beta = (\beta_1, ..., \beta_{n_t})^T \) are the shape and texture model coefficient vectors that have the normal distribution:

\[
\alpha \sim \mathcal{N}(0, \sigma_s), \quad \beta \sim \mathcal{N}(0, \sigma_t),
\]  
(3)

where \( \sigma_s \) and \( \sigma_t \) are the vectors of variances of the latent model shape and texture parameters. By changing \( \alpha \) and \( \beta \) we can generate, or morph, new faces. The constant part of the 3DMM consists of four components: the shape and texture bases \( S \) and \( T \), and the mean face shape \( s_0 \) and mean face texture \( t_0 \). The coefficients \( \alpha \) and \( \beta \) afford a low-dimensional coding of a 3D face.

The human face is a deformable object. The shape changes dynamically with gender, age and ethnicity. In principle, the shape model bases could capture these shape variations. However, to construct such a model would require a huge training set of 3D face images containing all the shape variations of interest. Suppose we have a large number of training samples for a specific group of
people identified by their age or gender. A more efficient solution is to transfer the intrinsic variations of one specific group to another. To this end, we propose the Gaussian-Mixture 3D Morphable face Model (GM-3DMM), that is presented in Section 4. In the remainder of this section, we first overview the 3D-2D face rendering as a prerequisite to 3DMM fitting, which is the key enabling step in applying the proposed model to the task of face recognition.

3.1. 3D-2D face rendering

Let us consider an instance of 3D face generated by the 3D morphable face model by setting the shape and texture parameters \( \alpha \) and \( \beta \) to specific values. By changing \( \alpha \) and \( \beta \) we can synthesize different 3D faces. Assuming that a 3D face is located at the origin of a coordinate system, we can render its 2D view by positioning a virtual camera in front of it at a distance \( \tau \). The 2D face image appearance captured by the camera will depend on the relative orientation of the camera coordinate system with respect to the 3D face coordinate system. This relative orientation is defined by the rotation matrix \( R \). The rotation matrix will determine which part of the synthesised 3D face is imaged by the virtual camera, in other words, the pose of the face captured in the 2D image. The camera shift and rotation, together with its focal length \( f \), specify a projection matrix that will establish the relationship between the 3D face and its 2D image.

The rendering process is schematically illustrated in Figure 2. Figure 2a illustrates the physics of the imaging process and its version providing more
detail by zooming on the camera, is shown in Figure 2b. The actual RGB intensities of a pixel recorded by the camera at viewing direction \( v \) will depend not only on the skin texture of the synthesised 3D face, but its illumination. Their magnitude is modulated by the direction \( d \) and strength of the light \( \delta \) incident on the 3D face surface, as well as the surface normal \( n \). In Figure 2a, \( r \) indicates the direction of the reflected specular light. The relationship between the pixel RGB value and the face shape, face texture and light source properties (direction and strength) is assumed to be described by the Phong model [44]. Accordingly, we assume that the scene is illuminated by a single point light source at a considerable distance to ensure that the light direction is the same at every point on the face surface.

By changing the rendering parameters we can synthesise 2D face images for each subject defined by parameters \( \alpha \) and \( \beta \) in arbitrary poses and illumination. This 2D face rendering capability has many potential applications. It offers the possibility to synthesise faces for augmenting a training set which lacks face data in certain poses and illumination conditions. Most importantly, it allows us to reconstruct the 3D face from an input 2D face image by a fitting process, which aims to determine the appropriate shape, texture, pose and illumination parameters so that the 2D face image rendered by the reconstructed model matches the input image.

### 3.2. Fitting 3DMM to 2D face images using ESO

Given a 3DMM and an input 2D face image, 3DMM fitting algorithms are able to recover the 3D shape and texture information, parametrised by \( \alpha \) and \( \beta \), of the face in the 2D image. To this end, the goal is to minimise the difference between the 2D face rendered by 3DMM and the input 2D face image by solving the optimisation problem with the loss:

\[
\min_{\alpha, \beta, \rho, \nu} \| a^f(\alpha, \rho) - a^M(\alpha, \beta, \rho, \nu) \|^2 + \lambda \zeta(\alpha, \beta)
\]

where \( a^f(\alpha, \rho) \) is the vector consisting of the RGB values of the input 2D image sampled at the vertices of the 3D face model with shape parameters \( \alpha \) projected
onto the image by the camera with parameters $\rho = \{R, \tau, f\}$. $a^{\text{M}}(\alpha, \beta, \rho, \nu)$ is the vector that contains the RGB values of the 2D face rendered by the reconstructed input image 3D model with model parameters $\alpha, \beta$, camera parameter $\rho$ and illumination controlling parameters $\nu = \{d, \delta\}$. The last term introduces regularisation to the optimised model parameters with a nonnegative penalty function $\zeta(\alpha, \beta)$, the value of which is zero at the origin.

The commonly adopted gradient descent optimisation of Eq. 4 is very time consuming [2, 3]. It is routinely accelerated by sampling the model mesh at a subset of the 3D vertices. However, this inevitably impacts negatively on the accuracy of fitting. In this work we adopt the Efficient Stepwise Optimisation (ESO) proposed in [27], which has been shown to outperform the state-of-the-art approaches including LiST [47], Zhang & Samaras [59], Aldrian & Smith [1] and MFF [49].

In the ESO fitting algorithm, the parameters needed to be optimised are divided into 5 subsets: shape, texture (albedo), camera, light direction and light strength parameters. Rather than optimising all these parameters simultaneously, ESO optimises them sequentially. In addition, ESO applies a linear approximation in each step and uses closed-form solutions to recover the respective parameters, leading to efficient 3DMM fitting. In addition, ESO uses all the vertices for model fitting, rather than a randomly selected subset, which improves the fitting accuracy.

The topology of the ESO-based 3DMM fitting process is shown in Figure 3. It splits the fitting process into two main stages, namely geometric and photometric optimisation. In each stage, the parameters of all the groups are it-
eratively and sequentially optimised until convergence is achieved. Note that the optimisation of each group of parameters is based on the assumption that those in all the other groups are fixed. In ESO, we use a set of sparse 2D facial landmarks to initialise the camera and shape model parameters, and also to constrain the model fitting procedure. The automatic facial landmark detection algorithms discussed in [12, 13] can be used for that purpose. For more details of our ESO fitting algorithm, the reader is referred to [27].

4. Gaussian mixture 3D morphable face models

In the standard 3DMM, shape and texture are each described by a single Gaussian distribution. In this section we describe the GM-3DMM: a morphable face model based on a mixture of Gaussians. The mixture components are standard 3DMM’s.

The first part of this section deals with the construction of GM-3DMMs from cohorts. Then, Section 4.2 will explain how to fit the mixture model to 2D face images. In Section 4.3 the fitting process is applied to the task of face recognition.

4.1. Model construction

The construction of the GM-3DMM is described in three parts. First we introduce a Gaussian mixture distribution and discuss how this concept relates to morphable models. Then we describe how to train a GM-3DMM from data samples. Lastly we develop a technique for learning the GM-3DMM from individual standard 3DMMs directly, without data.

Throughout this section we use the following notation: By \( X_i \) we denote the data set of cohort \( i \) (out of \( K \) cohorts). \( X \) denotes the combined set of all data. Note also that we make no distinction between shape and texture modalities as the theory applies equally to both.
4.1.1. The Gaussian mixture model

A mixture model represents the presence of subpopulations within an overall population by a weighted sum of the subpopulation distributions. Assuming the distribution of each subpopulation is Gaussian, parameterised by its mean $\mu_i$ and covariance $\Sigma_i$, the probability density of the mixture is given by

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\mu_i, \Sigma_i)$$  \hspace{1cm} (5)

where $\phi_i$ is the i-th mode mixing coefficient. In our case $\phi_i$ would reflect the frequency of occurrence of samples from the i-th sub-population.

While the cohorts of faces clearly differ in their mean, we assume that the local distributions around the mean are quite similar for all cohorts. In other words, the cohorts share a common covariance structure, $\Sigma$. Furthermore, as we are dealing with large sub-populations of roughly equal proportion, the prior probabilities $\phi_i$ can be considered equal. This simplifies the above equation to

$$p(x) = \frac{1}{K} \sum_{i=1}^{K} \mathcal{N}(\mu_i, \Sigma)$$  \hspace{1cm} (6)

Given a random sample $x$ from an unknown subpopulation, its most likely membership can be determined by finding the nearest mean under the covariance structure, which is the minimiser for the Mahalanobis distance:

$$k = \arg \min_i (x - \mu_i)^T \Sigma^{-1} (x - \mu_i)$$  \hspace{1cm} (7)

Conversely, generating a random sample from the mixture distribution involves drawing a value of $k$ at random (uniform) and then drawing $x$ from a Gaussian distribution with mean $\mu_k$ and covariance matrix $\Sigma$. Expressed in the space defined by the eigenvectors of $\Sigma$ with non-zero eigenvalues, this is equivalent to drawing a vector $\gamma \in \mathcal{N}(0, \Lambda)$ and then constructing $x$ as

$$x = \mu_k + \sum_{j=1}^{N_\gamma} \gamma_j v_j$$  \hspace{1cm} (8)

where $\Sigma V = V \Lambda$ is the eigendecomposition of $\Sigma$ with the $n_\gamma \times N_\gamma$ diagonal matrix $\Lambda$ holding the eigenvalues $\lambda_j$ associated with the $N_\gamma$ eigenvectors $v_j$ in $V$. 

14
In the context of GM-3DMM, producing a 3D face equates to drawing a value of \( k \) at random and then drawing the shape coefficients \( \alpha \sim \mathcal{N}(0, \sigma_s) \) and texture coefficients \( \beta \sim \mathcal{N}(0, \sigma_t) \) as in the 3DMM case (see Eq. 3). But the reconstruction of the 3D face from the coefficients is now based on the mean of cohort \( k \):

\[
\mathbf{s} = \mathbf{s}_0^{(k)} + \mathbf{S} \alpha, \quad \mathbf{t} = \mathbf{t}_0^{(k)} + \mathbf{T} \beta
\]

(9)

where we use the symbols for shape and texture as defined in Eq. 1–3, with the superscripted \( k \) to indicate the particular cohort mean. This produces the 3D shape \( \mathbf{s} \) and its texture \( \mathbf{t} \) sampled from the mixture distribution.

### 4.1.2. Training the GM-3DMM

The estimation of the parameters \( \mu_i \) and \( \Sigma \), can generally be performed using the EM algorithm. This applies when the cohort identities are not known. In this work, however, we construct the model by mixing cohort distributions as we have the labels. It is a mixture density composed of identifiable components.

In the most straightforward case, where we have all training data at hand, the estimation could be a simple two-stage process:

\[
\mu_k = \frac{1}{N_k} \sum_{j=1}^{N} \Delta(k - y_j) x_j \quad \forall k
\]

(10)

\[
\Sigma = \frac{1}{N - K} \sum_{j=1}^{N} (x_j - \mu_{(y_j)})(x_j - \mu_{(y_j)})^T
\]

(11)

where \( N_k \) is the number of training samples used for estimating the parameters of each component, \( N \) is the total number of samples, \( y_j \) is the class label for sample \( x_j \) and \( \Delta(\cdot) \) is the Kronecker delta function, which equals 1 if its argument is 0 and is 0 otherwise.

Alternatively, one can derive the mixture distribution from independently trained cohort distributions, each defined by its mean \( \mu_i \) and covariance matrix \( \Sigma_i \), based on \( N_i \) samples. In this case, the mixture means are the cohort means, and the mixture covariance matrix is the pooled covariance matrix over all
cohorts, given by
\[ \Sigma = \frac{1}{N - K} \sum_{i=1}^{K} (N_i - 1) \Sigma_i \] (12)

A special case arises when multiple independently trained 3DMMs describe the same cohort, i.e. the same component \( k \). This happens for example when new 3DMMs are released without their training data. In Appendix A we describe how to fuse multiple 3DMMs for the same cohort. The GM-3DMM is then obtained by pooling the cohort covariances as above.

4.1.3. Fusion of eigenvectors

A practical issue arises when we consider the size of the covariance matrix, which would take around 70GB in memory. 3DMMs avoid its computation entirely by instead working with the eigenvectors and eigenvalues, which can be estimated from the data directly (by virtue of SVD). Our aim is to accomplish a similar simplification by deriving \( \mathbf{V} \) and \( \Lambda \) of the pooled covariance in Eq. 12 directly from the cohort’s \( \mathbf{V}_i \) and \( \Lambda_i \).

We start by noting that because any set of eigenvectors \( \mathbf{V} \) is orthonormal, \( \mathbf{V}^{-1} = \mathbf{V}^T \) and so we can write \( \Sigma = \mathbf{V} \Lambda \mathbf{V}^T \). Substituting this in Eq. 12 we obtain
\[ \Sigma = \frac{1}{N - K} \sum_{i=1}^{K} (N_i - 1) \mathbf{V}_i \Lambda_i \mathbf{V}_i^T \] (13)

Now let us define a \((N - K) \times D\) matrix
\[ \mathbf{H} = [a_1 \mathbf{V}_1 \sqrt{\Lambda_1}, a_2 \mathbf{V}_2 \sqrt{\Lambda_2}, \ldots, a_K \mathbf{V}_K \sqrt{\Lambda_K}]^T \] (14)

where \( \sqrt{\Lambda_i} \) is the diagonal matrix of square roots of the eigenvalues of cohort model \( i \). Note that \( \mathbf{H}^T \mathbf{H} \) can be expressed as
\[ \mathbf{H}^T \mathbf{H} = \sum_{i=1}^{K} (a_i \mathbf{V}_i \sqrt{\Lambda_i})(a_i \sqrt{\Lambda_i} \mathbf{V}_i^T) = \sum_{i=1}^{K} a_i^2 \mathbf{V}_i \Lambda_i \mathbf{V}_i^T \] (15)

Referring to Eq. 13 provided
\[ a_i = \sqrt{(N_i - 1)/(N - K)} \] (16)
then $H^T H = \Sigma$. The SVD of $H$, i.e. $H = U \sqrt{\Lambda} V^T$, will give us the eigenvectors of $\Sigma$. As the dimensionality of $H$ is related to the size of the training set, it is generally much smaller than the size of $\Sigma$, and can be computed more efficiently than the direct eigendecomposition of sums of covariance matrices.

It should also be noted that the estimation of the cohort models parameters, $\mu_i$ and $\Sigma_i$, can be performed independently. That means the GM-3DMM can be built incrementally as more cohort models are released.

4.2. Fitting GM-3DMMs to 2D face images

In contrast to the classical 3DMM, our GM-3DMM results in a set of cohort-specific models that share the same PCA bases for each type of 3D face information (shape and texture), but have different means. This brings some advantages to fitting 3D face models to 2D images. Fitting a cohort-specific 3D face model to faces within the same cohort is easier and more accurate than fitting a global/general model. This has also been investigated and demonstrated for other computer vision tasks in [9, 14, 15, 21, 31]. The main reason is that a fitting algorithm usually starts the optimisation from the mean of the model. For a given 2D face image, the corresponding cohort mean is closer to the global optimum than that of a global model, which benefits the optimisation of model parameters. In addition, the subspace spanned by the PCA bases of a cohort-specific model is more compact, which reduces the size of the search space for parameter optimisation and the regularisation imposed by a cohort-specific model is tailored to produce better fitting results.

Despite the aforementioned advantages, fitting GM-3DMM to 2D face images is not without difficulties. The key issue is how to select the correct model for the fitting of a given 2D image. To deal with this problem, two strategies are available. The first solution is to apply a classifier that predicts the label of a 2D face for cohort-specific model selection. However, to this end, we have to design an additional model prediction stage in our pipeline that relies on the accuracy of face detection and classification methods. The alternative way is to fit all the cohort-specific models to a 2D face and choose the one with the
best fitting result as our final output. Based on our assumption, the correct cohort-specific model should produce the minimal fitting error. Accordingly, the texture residual between the input 2D face and the rendered 2D image of our reconstructed 3D face is used for model selection. The fitting result of the cohort-specific model that has the minimal norm of the residual is selected as the final output.

4.3. Face recognition based on GM-3DMM

In face recognition, 3DMMs are usually used in two different ways: i) to perform face frontalisation before the matching step, and ii) to directly use the recovered shape and albedo parameters of 2D faces for matching. The second approach is popular because the shape and albedo parameters are naturally pose- and illumination-invariant. It has achieved promising face recognition results in comparison with state-of-the-art methods including Deep Neural Networks (DNN) [26]. More specifically, given a 2D face image, 3DMM is first used to recover its 3D shape and skin texture information by estimating the model parameters ($\alpha$ and $\beta$). These are concatenated to provide a representation/features of the face, $\gamma = [\alpha^T, \beta^T]^T$. The face matching is then performed based on these features. However, this method is not applicable to the GM-3DMM. The main reason is that we have a number of cohort-specific models and different cohort models are used for fitting to 2D faces. This results in inconsistent representations across the recovered 3D shape and texture model parameters of different 2D images because the low-dimensional embeddings in a PCA-based subspace involve the different means of the cohort-specific modes of GM-3DMM.

To address this issue, we propose a multiple classifier system for our GM-3DMM-based face recognition. We first fit the $K$ cohort-specific GM-3DMM component to a given gallery set with $C$ registered subjects images $\{I_1, ..., I_C\}$. This results in $K$ representations for each subject in the gallery set, i.e. $\{\gamma^1_c, ..., \gamma^K_c\}(c = 1, ..., C)$. Given a probe image $I$, in the same manner, we fit all the $K$ cohort-specific models to it and obtain $K$ face representations
\{\hat{\gamma}^1, \ldots, \hat{\gamma}^K\}. Then, for each cohort specific model, we output a label for the probe image as:

$$label(\hat{I})^k = \arg \min_{c} D(\hat{\gamma}^k, \gamma^k_c),$$

(17)

where $D()$ is a distance measurement function. In this paper, Linear Discriminant Analysis (LDA) is used to further reduce the dimensionality of each feature vector $\gamma$ and the cosine distance is used for $D()$. For the $K$ labels of the probe image, we use majority voting to decide its final class label. If all the $K$ labels are different from each other, then the one with the minimal distance is used as the predicted class.

5. Model analysis and experimental results

5.1. SURREY-JNU 3D face dataset

The dataset of 942 3D face scans used for the construction of the GM-2DMM has been obtained by merging Surrey and JNU subsets. The Surrey subset has 168 3D faces, captured using a 3dMD Face device at the Centre for Vision, Speech and Signal Processing, University of Surrey, UK, in 2008. The JNU subset comprises 774 3D faces captured using an upgraded 3dMD Face device at the School of Internet of Things Engineering, Jiangnan University, China in 2016. A summary of the SURREY-JNU 3D dataset is presented in Table 1.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>JNU</th>
<th>Surrey</th>
<th>SURREY-JNU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>20-50</td>
<td>20-60</td>
<td>20-60</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Black/African/Caribbean</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>774</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Other ethnic group</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>Gender</td>
<td>Female</td>
<td>233</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>541</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>774</td>
<td>168</td>
<td>942</td>
</tr>
</tbody>
</table>

Table 1: A summary of the SURREY-JNU 3D face dataset
The SURREY subset includes different ethnic groups, namely Black, Asian, White and Other. The latter group includes, but is not limited to, South Asian, Latin American and Arab faces. The JNU subset contains specifically Eastern Asian faces. Figure 4 visualises the distribution of the 3D shape and texture information of all facial images in this dataset using the t-Distributed Stochastic Neighbour Embedding (t-SNE) method [55]. The cohort labels were added to the visualisation, and do not form part of the t-SNE method.

5.2. Intrinsic properties of the GM-3DMM

In this section we summarise the statistics of the GM-3DMM trained on the SURREY-JNU data set.

The amount of variance explained in the subspace defined by a PCA basis of rank $r$ is measured by the sum of eigenvalues $\sum_{i=1}^{r} \lambda_i$. Since the GM-3DMM employs a single covariance estimate, the formulation equally applies there. In this section we evaluate the models by retaining the top $r$ modes of variation that in total explain at least 98% of the data variance.

Table 2 lists the associated number of texture and shape components in the individual cohort models and the mixture model (GM-3DMM). The results show that the mixture model is an efficient representation of the cohort variations, in particular for the face shape, where we observe close to 50% compression.
Table 2: Number of principal components to explain 98% variance

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Training size</th>
<th>Texture</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black/African/Caribbean</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Asian</td>
<td>100</td>
<td>73</td>
<td>28</td>
</tr>
<tr>
<td>European</td>
<td>88</td>
<td>70</td>
<td>32</td>
</tr>
<tr>
<td>Other</td>
<td>20</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>216</strong></td>
<td><strong>165</strong></td>
<td><strong>80</strong></td>
</tr>
<tr>
<td><strong>Mixture</strong></td>
<td><strong>216</strong></td>
<td><strong>146</strong></td>
<td><strong>44</strong></td>
</tr>
</tbody>
</table>

We thus conclude that a significant amount of facial variation is shared across different ethnic groups.

The amount of cohort-specific variance explained in the GM-3DMM is measured by the projection of the cohort data onto the GM-3DMM basis vectors, $V_k$. Specifically, the vector of variances, $\sigma_k^V$, is obtained from the matrix diagonal

$$
\sigma_k^V = \text{diag}\left(\frac{1}{N_k - 1}\sum_{i=1}^{N_k} V^T_i (x_i - \mu_k)(V^T_i (x_i - \mu_k))^T\right) = \text{diag}(V^T \Sigma_k V)
$$

(18)

where we have assumed that all $x_i$ are drawn from cohort $k$. Recalling that $\Sigma_k = V_k \Lambda_k V_k^T$, the entries on the diagonal of $V^T_i V_k \Lambda_k V_k^T V_i$ describe the correlation between the bases $V$ and those of the original cohort data. A value of 0 indicates the eigenvector $v_j$ is orthogonal to cohort $k$. By contrast, 1 means the vector is completely embedded in the cohort subspace.

Figure 5 provides a cumulative plot of $\sigma_k^V$ for each cohort. On the horizontal axis we vary the number of columns of $V$. The dotted line shows the explained variance within the GM-3DMM as reference. Overall the explained cohort variance aligns closely with the mixture variance. The texture variation of “Other” seems a bit overrepresented, which could be attributed to the large variation of skin texture in this group. The underrepresentation of Black face texture is most likely due to the lack of samples and will be reevaluated when more data has been collected. In any case, at 98% variance of the mixture model, around
98% of all cohorts is included, both in shape and in texture. The mixture model thus provides good representation of all cohorts, and of all cohorts equally.

In view of 2D face image fitting and recognition, where one of the key challenges is to select the correct mixture component, we compare the GM-3DMM to the standard global 3DMM and to the individual cohort models (where covariance is estimated from the within-cohort samples only). Over five folds we draw a set of test images (3D face scans) and measure their distance to each cohort mean. The distance is measured by the Mahalanobis distance with the covariance matrix dictated by the cohort. Note that the GM-3DMM and the standard 3DMM use the same covariance for all cohorts. The feature vectors are composed of shape and texture parameters as: $[\alpha^T, \beta^T]^T$. The training sizes are the same as listed in Table 2 with test samples of size 2, 12, 12 and 4 for Black, Asian, White and Other respectively.

The results are presented in Table 3. Accuracy is computed from the confusion matrix as the sum of the elements on the diagonal divided by the sum of all elements. Higher accuracy means the model is more representative of the underlying distribution of faces.

Under the standard 3DMM, samples in the “Other” cohort are often nearer to the White cohort mean. We believe this stems from the inclusion of cross-cohort differences in the distribution, which produces an overestimate of the
Table 3: Comparison of cohort prediction under different face models.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Standard 3DMM</th>
<th>Cohort models</th>
<th>GM-3DMM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>A</td>
<td>W</td>
<td>O</td>
</tr>
<tr>
<td>Black</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Asian</td>
<td>0</td>
<td>0</td>
<td>59</td>
<td>1</td>
</tr>
<tr>
<td>White</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.91</td>
<td>0.83</td>
<td></td>
<td>0.97</td>
</tr>
</tbody>
</table>

actual variance. The separate cohort models do not provide good accuracy. The idea that face variations are shared between different ethnic groups seems validated, with the highest accuracy obtained from the joint covariance estimate in the mixture model. Notwithstanding the possibility that cohort models could estimate the covariance structure well, given enough training samples, the mixture model estimates it more efficiently and requires around a third of the number of eigenvectors (see Table 2).

5.3. Information transfer between ethnic groups

Table 1 describes the composition of the SURREY-JNU data set. While the data represents a variety of ethnic groups, the age of people is quite narrowly distributed around 20 to 25 years. A separate data set was collected under the People of the British Isles project[56]. It is a very narrow ethnic sample (only people of British descent), but has a much broader spectrum of age, ranging from 20 to 101 years (See Figure 6a).

To test the transfer of age information, we took the 3% youngest and eldest samples from the PoBI dataset. The vector of the difference between the two means (young and old), a, was used as a reference age descriptor. The quality of any model to describe age variation was then measured as the norm of the projected age vector: $\|V^T a\|$, where V are the model’s principal components. Figure 6b plots this norm for the SURREY-JNU mixture model and the fused SURREY-JNU + PoBI mixture model. Evidently, the latter model captures...
Figure 6: Age variation in the PoBI data (left), and its representation in GM-3DMMs (right).

Figure 7: Demonstrating the impact of transferring the face aging effect from one cohort to another. Top row are original images from cohorts. Bottom row are aged variants.

age variations more succinctly. Figure 7 shows that the addition of vector $a$ to existing samples is applicable across all cohorts.

5.4. Comparison on 3D-2D face fitting

By fitting a 3D face model to 2D faces, we are able to recover their 3D shape and texture information. In this section, we compare the performance of 2D image fitting between the classical 3DMM and our GM-3DMM approaches in terms of accuracy. To evaluate the accuracy of 2D image fitting, shape and texture fitting errors are measured. For the shape fitting error, we use the average distance between the recovered 3D vertices and the ground truth:

$$e_s = \frac{\sum_{i=1}^{n} \sqrt{(x'_{i} - x^*_{i})^2 + (y'_{i} - y^*_{i})^2 + (z'_{i} - z^*_{i})^2}}{n}$$

(19)
Table 4: A comparison of the model fitting errors (average vertex distance in mm) achieved with different cohort-specific 3D face models on the SURREY-JNU 3D face dataset. We use the term "*G" for a model that has been further divided into two gender-specific (female and male) models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Subset</th>
<th>Black (Bl)</th>
<th>Asian (As)</th>
<th>White (Wh)</th>
<th>Other (Ot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DMM</td>
<td></td>
<td>4.8±0.6</td>
<td>5.4±0.5</td>
<td>6.2±0.5</td>
<td>5.5±0.5</td>
</tr>
<tr>
<td>GM-3DMM (Bl)</td>
<td></td>
<td>4.3±0.5</td>
<td>6.0±0.4</td>
<td>7.0±0.3</td>
<td>6.5±0.6</td>
</tr>
<tr>
<td>GM-3DMM (As)</td>
<td></td>
<td>6.1±0.9</td>
<td>4.7±0.2</td>
<td>7.6±0.5</td>
<td>6.9±0.4</td>
</tr>
<tr>
<td>GM-3DMM (Wh)</td>
<td></td>
<td>5.1±0.3</td>
<td>7.0±0.6</td>
<td>5.5±0.2</td>
<td>5.2±0.9</td>
</tr>
<tr>
<td>GM-3DMM (Ot)</td>
<td></td>
<td>5.2±0.4</td>
<td>6.7±0.7</td>
<td>5.9±0.4</td>
<td>5.1±0.6</td>
</tr>
<tr>
<td>GM-3DMM (Bl-G)</td>
<td></td>
<td>4.3±0.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM-3DMM (As-G)</td>
<td></td>
<td>-</td>
<td>4.2±0.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM-3DMM (Wh-G)</td>
<td></td>
<td>-</td>
<td>-</td>
<td>4.8±0.3</td>
<td>-</td>
</tr>
<tr>
<td>GM-3DMM (Ot-G)</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.5±0.3</td>
</tr>
</tbody>
</table>

where \((x_i', y_i', z_i')\) and \((x_i^*, y_i^*, z_i^*)\) are the 3D coordinates of the \(i\)th vertex of the recovered 3D shape \((s')\) and the ground truth 3D shape \((s^*)\). \(n\) is the number of vertices. The RMS error of RGB values averaged over all vertices is used as the texture fitting error:

\[
e_t = \sqrt{\frac{\sum_{i=1}^{n} (r_i' - r_i^*)^2 + (g_i' - g_i^*)^2 + (b_i' - b_i^*)^2}{3n}},
\]

where \((r_i', g_i', b_i')\) and \((r_i^*, g_i^*, b_i^*)\) are the RGB values for the \(i\)th vertex of the recovered 3D texture \((t')\) and ground truth 3D texture \((t^*)\).

In our experiments, the training/test sets were formed by randomly selecting 8/2, 100/12, 88/12 and 20/4 3D faces from the Black, Asian, White and Other cohorts. We repeated this random partition 5 times and reported the average result. In each round, 216 3D scans from the SURREY-JNU dataset were used for training and other 30 were used for testing and providing ground truth 3D face information. Each test sample in the dataset has a 2D face image and the corresponding 3D face scan including its ground truth 3D shape and texture information. The 2D face images of the selected test samples were used for 3D-2D model fitting.
Table 5: A comparison of the RMS texture fitting errors (intensity) achieved by different cohort-specific 3D face models on the SURREY-JNU 3D face dataset. We use the term "*-G" for a model that has been further divided into two gender-specific (female and male) models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Subset</th>
<th>Black (Bl)</th>
<th>Asian (As)</th>
<th>White (Wh)</th>
<th>Other (Ot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3DMM</td>
<td></td>
<td>49±7</td>
<td>29±2</td>
<td>27±0</td>
<td>32±3</td>
</tr>
<tr>
<td>GM-3DMM (Bl)</td>
<td></td>
<td>22±3</td>
<td>49±4</td>
<td>53±2</td>
<td>41±8</td>
</tr>
<tr>
<td>GM-3DMM (As)</td>
<td></td>
<td>53±5</td>
<td>29±3</td>
<td>31±0</td>
<td>36±3</td>
</tr>
<tr>
<td>GM-3DMM (Wh)</td>
<td></td>
<td>56±8</td>
<td>31±2</td>
<td>25±1</td>
<td>35±5</td>
</tr>
<tr>
<td>GM-3DMM (Ot)</td>
<td></td>
<td>37±6</td>
<td>34±3</td>
<td>32±3</td>
<td>30±3</td>
</tr>
<tr>
<td>GM-3DMM (Bl-G)</td>
<td></td>
<td>28±2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM-3DMM (As-G)</td>
<td></td>
<td>-</td>
<td>23±5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GM-3DMM (Wh-G)</td>
<td></td>
<td>-</td>
<td>-</td>
<td>25±1</td>
<td>-</td>
</tr>
<tr>
<td>GM-3DMM (Ot-G)</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>32±5</td>
</tr>
</tbody>
</table>

We performed the experiments in two settings. In the first setting, the correct cohort of a given 2D face image is known and we evaluate the accuracy of fitting the appropriate GM-3DMM mixture component. In the second setting, the ethnicity is not known. We compare the performance of our proposed GM-3DMM fitting algorithm based on the proposed automatic mixture component selection strategy with the classical ESO-based fitting algorithm.

5.4.1. Setting-I

In this part, we compare our GM-3DMM with the classical 3DMM when cohort membership of the test samples is given. The shape and texture fitting results are shown in Table 4 and Table 5, respectively. As compared to the classical 3DMM method, by choosing the correct cohort-specific model, the fitting errors of our GM-3DMM are smaller. This validates our assumption that fitting the correct cohort-specific model to a 2D face image recovers the 3D shape and texture information more accurately than the classical general 3DMM.

To investigate the 3D-2D fitting accuracy of our GM-3DMM with more Gaussian mixtures, we further split each ethnicity-specific model into two
gender-specific (female and male) models, indicated by ‘*-G’ in Table 4 and Table 5. As compared with the classical 3DMM, by injecting the gender information, we have improved the shape and texture reconstruction accuracy significantly. In contrast to the ethnicity-only GM-3DMM, the ethnicity & gender GM-3DMM has better shape fitting results in terms of accuracy, as shown in Table 4. However, for texture fitting, the improvement is not consistent (Table 5). The fitting accuracy has been improved significantly only for the Asian group when we further split each ethnicity-specific model into Female and Male models. The main reason is that the diversity of appearance is larger than that of shape. As we have much smaller training sample size for the other ethnicity groups, splitting them further using gender labels may lead to the well-known small sample size problem. In consequence the mean of an ethnicity/gender cohort model is no longer representative.

5.4.2. Setting-II

In Setting-II, we constructed a three-component GM-3DMM from the Black, Asian and White cohorts. To fit the GM-3DMM to a 2D face image, the automatic model selection method was used, as introduced in Section 4.2. We compared the fitting errors of our GM-3DMM with the classical 3DMM in Table 6. Note that, to investigate the impact of the number of training samples per cohort on accuracy, we use 100/700 Asian faces for model construction. In addition, to evaluate the fitting accuracy when we have more Gaussian mixtures, we split the Asian model into Female and Male models. We did not split the other models into gender models because of the small sample size problem as discussed at the end of Section 5.4.2.

As shown in Table 6, the proposed GM-3DMM outperforms the classical 3DMM in fitting accuracy. When using more cohort-specific models by splitting the Asian group into Female and Male sub-populations, the fitting performance has been further improved in terms of accuracy. Another interesting finding is that the increase in the number of Asian faces available for training reduces the performance of the classical 3DMM. The reason is that the large number
Table 6: A comparison of the fitting accuracy between the classical 3DMM and the proposed GM-3DMM, measured in terms of shape and texture fitting errors. We used either 100 or 700 Asian faces for model construction. Two different Gaussian mixture models were constructed. Specifically, we first constructed a GM-3DMM using ethnicity labels only, designated by ‘E’. Then, we split the Asian model into two gender modes, i.e. Asian-Female and Asian-Male, designated by ‘E + G’.

<table>
<thead>
<tr>
<th>Model</th>
<th>NO. of Asian Faces</th>
<th>Fitting Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Shape (mm)</td>
</tr>
<tr>
<td>3DMM</td>
<td>100</td>
<td>5.7±0.2</td>
</tr>
<tr>
<td>GM-3DMM (E)</td>
<td>100</td>
<td>5.5±0.3</td>
</tr>
<tr>
<td>GM-3DMM (E+G)</td>
<td></td>
<td>5.4±0.3</td>
</tr>
<tr>
<td>3DMM</td>
<td>700</td>
<td>5.9±0.1</td>
</tr>
<tr>
<td>GM-3DMM (E)</td>
<td>700</td>
<td>5.4±0.3</td>
</tr>
<tr>
<td>GM-3DMM (E+G)</td>
<td></td>
<td>5.4±0.3</td>
</tr>
</tbody>
</table>

of Asian faces dominates the trained model, in which the mean and PCA bases favour the Asian group. Hence the model cannot fit the other faces very well. In contrast, by increasing the number of Asian faces in model training, the fitting errors of our GM-3DMM are slightly reduced.

5.5. Face recognition on Multi-PIE

The Multi-PIE face dataset has been widely used to benchmark face recognition algorithms in the presence of controlled pose, expression and illumination variations. Multi-PIE has more than 750,000 face images captured in 4 sessions of 337 subjects under 15 pose, 20 illumination and a range of facial expression variations. In this paper, to make a fair comparison with the state of the art results, we evaluated our GM-3DMM-based face recognition system using the commonly used protocol. The protocol was designed to test the robustness of face recognition approaches to pose and illumination variations. To this end, all the 249 subjects of the neutral expression with 7 poses (0°, ±15°, ±30° and ±45°) and all the 20 illumination variations in Session-1 were used to form the training/test subsets. In total, the subset has 34,860 (249 × 7 × 20) images. To form the training set, the first 100 subjects are used. For the 101-249 subjects,
Figure 8: Some examples of GM-3DMM fitting: (a) the input 2D face image; (b) 2D rendering of the fitted model; (c) the fitted 3D shape; (d) the fitted albedo (skin texture); (e) frontalised faces using the remapped original texture with removed lighting.

The frontal face with the neutral illumination of each subject is used as the gallery image and all the remaining images are selected as the probe images.

The face recognition rates of different algorithms on Multi-PIE are presented in Table 7. In the table, we compare our GM-3DMM-based face recognition system with a set of state-of-the-art approaches. We evaluated three different GM-3DMM based face recognition approaches:

**Frontalised** Prior to recognition, the face images are frontalised using the GM-3DMM (Figure 8 shows some examples of this). After frontalisation the CNN-based VGG-FACE model [41] and cosine distance are used for feature extraction and face recognition. The VGG-FACE model was fine-
Table 7: A comparison of the proposed GM-3DMM with state-of-the-art methods in face recognition rate (%) achieved on Multi-PIE across pose and illumination variations.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-45°</td>
</tr>
<tr>
<td>Li et al. [32]</td>
<td>63.5</td>
</tr>
<tr>
<td>DNN-RL [60]</td>
<td>67.1</td>
</tr>
<tr>
<td>DNN-CPF [58]</td>
<td>73.0</td>
</tr>
<tr>
<td>LNFF+LRA [10]</td>
<td>77.2</td>
</tr>
<tr>
<td>HPN [11]</td>
<td>71.3</td>
</tr>
<tr>
<td>U-3DMM [26]</td>
<td>73.1</td>
</tr>
<tr>
<td>ESO-3DMM [27]</td>
<td>80.8</td>
</tr>
<tr>
<td>GM-3DMM (frontalised)</td>
<td>82.7</td>
</tr>
<tr>
<td>GM-3DMM (best)</td>
<td>82.4</td>
</tr>
<tr>
<td>GM-3DMM (fusion)</td>
<td><strong>84.3</strong></td>
</tr>
</tbody>
</table>

Tuned on the training set of the frontalised Multi-PIE faces.

**Best** Face recognition in this approach was based on the extracted 3D shape and albedo parameters. Specifically, recognition was based only on the parameters extracted from one ethnicity-specific GM-3DMM, selected to yield a minimal fitting error (on the texture).

**Fusion** In this approach, too, face recognition is based on the extracted 3D shape and albedo parameters. This time however, the parameters of all the ethnicity-specific models are used for decision making, as described in Section 4.3.

The results in Table 7 show our GM-3DMM-based face recognition method outperforms both the 2D approaches, including the DNN-based face recognition algorithms, and 3DMM-based approaches. Clearly, the use of a 3D morphable model is particularly suited for solving the challenges posed by the Multi-PIE data set: extreme pose and light variations are dealt with properly in the 3D space, and effectively allow the face matching to be performed independent of these nuisance factors. This holds true for all 3DMM based approaches compared to 2D methods (see Table 7).
Comparing to the ESO-3DMM, which models the face space by a single Gaussian component, the use of multiple mixture components in the GM-3DMM further improves face recognition rates, and in fact achieves the best recognition results for all poses.

Last, the use of DNN features and our GM-3DMM-based frontalisation technique performs slightly better in face recognition accuracy than the use of the parameters of the best fitted ethnicity GM-3DMM. However, the best face recognition result is achieved when we use the proposed face recognition method that exploit all the parameters of the GM-3DMM.

5.6. A note on computational complexity

It may seem that the GM-3DMM, composed of multiple mixture components, is a computationally demanding construct. Specifically in open set face recognition, it may be necessary to fit all mixture components to an image to find the best match.

However, in many scenarios, such as face verification for border control and physical access, the identity to be verified is known, and the appropriate mixture component would be selected based on prior information. In other scenarios this information could come from a separate prediction, e.g. based on the input image. In these cases the complexity of the GM-3DMM is practically identical to that of a standard 3DMM, both in time and memory. Yet, the discriminatory power of its smaller covariance structure and the improved similarity to the cohort mean provide a clear advantage for face matching.

It is also worth pointing out that the separate mixture components share the same covariance. This means that the storage of a GM-3DMM is only larger than a standard 3DMM by at most $K - 1$ mean vectors. It also means that the fitting process could share some computations between all mixture components (matrix inversions in particular), although our current implementation fits the models sequentially.
6. Conclusion

Since their conception more than two decades ago, 3D morphable face models have attracted considerable interest because of their ability to model intrinsic properties of 3D faces. An initial exploration of the distribution of faces from a diverse group of people, however, revealed the data does not follow a unimodal distribution. This called for the extension of the 3DMM to one with multiple modes — the Gaussian Mixture 3DMM (GM-3DMM), proposed in this paper.

We constructed a GM-3DMM with the mixture components modelling 3D face images of people from a variety of ethnic groups. We also detailed the methodology necessary for building GM-3DMMs from existing 3D face models.

We conducted a number of experiments in 2D and 3D face analysis to demonstrate the merit of using the GM-3DMM as compared to a standard 3DMM or individual cohort 3DMMs. The advantages include:

- achieving better accuracy when fitting GM-3DMM to 2D face images by virtue of initialisation of the fitting process at the appropriate cohort mean and the use of tailored regularisation constraints
- mitigation of small sample problems in cohort covariance matrix estimation
- more accurate characterisation of 3D data
- transfer of learning from one cohort to another exemplified on face aging effects
- significantly better 2D face recognition results achieved on the multiPIE dataset containing extreme pose and illumination variations.

The future directions of research will aim to enhance the GM-3DMM by adding other cohorts and balancing their sizes. The proposed GM-3DMM will also be evaluated on faces-in-the-wild benchmarking datasets.
Acknowledgements

This work was supported in part by the EPSRC Programme Grant ‘FACER2VM’ (EP/N007743/1), the National Natural Science Foundation of China (61373055, 61672265) and the Natural Science Foundation of Jiangsu Province (BK20140419, BK20161135).

Appendix A. Fusion of multiple partial cohort contributions

Suppose we are given $M_j$ estimates of the distribution of one cohort population, each estimate defined by its mean $\mu_{ji}$ and covariance matrix $\Sigma_{ji}$ based on $n_{ji}$ samples. Implicitly, the total number of samples is $N_j = \sum_{j=1}^{M_j} n_{ji}$. Our aim is to find the parameterisation ($\mu_j$, $\Sigma_j$) as if we were given all $N_j$ samples.

The combined mean is simply defined as the average of the $M_j$ estimates, weighted by the number of samples available for each estimate.

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{M_j} n_{ji} \mu_{ji} \quad (A.1)$$

We note that the combined covariance matrix $\Sigma_j$ satisfies

$$(N_j - 1) \Sigma_j = \sum_{k=1}^{N_j} (x_k - \mu_j)(x_k - \mu_j)^T \quad (A.2)$$

$$= \sum_{i=1}^{M_j} \sum_{k=1}^{n_{ji}} ((x_{ki} - \mu_{ji}) + (\mu_{ji} - \mu_j)(\mu_{ji} - \mu_j))^T \quad (A.3)$$

where $x_{ki}$ denotes the $k^{th}$ sample of the $i^{th}$ sub-population of the $j^{th}$ cohort. Hence the fused $j^{th}$ cohort covariance matrix, $\Sigma_j$, can be expressed as

$$\Sigma_j = \frac{1}{N_j - 1} \sum_{i=1}^{M_j} [(n_{ji} - 1) \Sigma_{ji} + n_{ji}(\mu_{ji} - \mu_j)(\mu_{ji} - \mu_j)^T] \quad (A.4)$$

Similar to Eq. 14, in order to find the fused eigenvectors and eigenvalues of $\Sigma_j$, we can define a matrix $H_j$ based on the eigendecomposition of the $M_j$ estimates, such that $H_j^T H_j = \Sigma_j$. This time we also add the mean correction.
term defined by the scatter of the partial means (the second term of the sum in Eq. [A.4]).

\[
H_j = \left[ a_{j1} V_{j1} \sqrt{\Lambda_{j1}}, \ b_{j1}(\mu_{j1} - \mu_j), \ldots, a_{ji} V_{ji} \sqrt{\Lambda_{ji}}, \ b_{ji}(\mu_{ji} - \mu_j), \ldots \right]^T
\]

(A.5)

where \( V_{ji} \) and \( \Lambda_{ji} \) are the eigenvectors and eigenvalues corresponding to the \( i \)th sub-population of cohort \( j \), and \( a_{ji} = \sqrt{(n_{ji} - 1)/(N_j - 1)} \) and \( b_{ji} = \sqrt{n_{ji}/(N_j - 1)} \). The SVD decomposition of matrix \( H_j \) would then yield \( V_j \) and \( \Lambda_j \) to feed into the eigenvector fusion process for matrix \( \Sigma \) described in Section 4.1.3.

References


