Face Recognition Using a Unified 3D Morphable Model

Face recognition using a unified 3D morphable model

Guosheng Hu[1], Fei Yan[3], Chi-Ho Chan[3], Weihong Deng[4], William Christmas[3], Josef Kittler[3], Neil M. Robertson[1,2]

[1]Anyvision, Queen's Road, Belfast, UK, BT39DT (www.anyvision.co)
[2]ECIT, Queen's University of Belfast, UK, BT39DT
[3]CVSSP, University of Surrey, Guildford, UK, GU27XH
[4]Beijing University of Posts and Telecommunications, Beijing, China, 100876

Abstract. We address the problem of 3D-assisted 2D face recognition in scenarios when the input image is subject to degradations or exhibits intra-personal variations not captured by the 3D model. The proposed solution involves a novel approach to learn a subspace spanned by perturbations caused by the missing modes of variation and image degradations, using 3D face data reconstructed from 2D images rather than 3D capture. This is accomplished by modelling the difference in the texture map of the 3D aligned input and reference images. A training set of these texture maps then defines a perturbation space which can be represented using PCA bases. Assuming that the image perturbation subspace is orthogonal to the 3D face model space, then these additive components can be recovered from an unseen input image, resulting in an improved fit of the 3D face model. The linearity of the model leads to efficient fitting. Experiments show that our method achieves very competitive face recognition performance on Multi-PIE and AR databases. We also present baseline face recognition results on a new data set exhibiting combined pose and illumination variations as well as occlusion.

Keywords: 3D morphable model, face recognition

1 Introduction

3D-assisted 2D face recognition has been attracting increasing attention because it can be used for pose-invariant face matching. This requires fitting a 3D face model to the input image, and using the fitted model to align the input and reference images for matching. As 3D facial shapes are intrinsically invariant to pose and illumination, the fitted shape also provides an invariant representation that can be used directly for recognition. The use of a face prior has been demonstrated to offer impressive performance on images of faces subject to a wide pose variations, even outperforming deep learning [1] [2].

Most popular are 3D morphable face models which represent 3D face images in a PCA subspace. 3D face models proposed in the literature can capture and represent different modes of variations. Some focus solely on 3D shape (3DSM) [3], [4], [5], [6]. Others (3DMM) model also the skin texture [7], [8], [9], or even face expression (E-3DMM) [10], [11]. When fitting 3DMM to an input image, it is essential to estimate the scene illumination, as skin texture and lighting are intrinsically entwined, and need to be separated.

The problem of 3D model to the 2D image fitting becomes challenging when the input image exhibits intra-personal variations not captured by the 3D model, or the image is corrupted in some way. In this work, we use the term 'intra-personal' to represent any variations which are not inter-personal ones (facial shape and texture). We assume that fitting the shape would be affected to a lesser extent if the automatic landmarking procedure used is robust to shape variations and to occlusion. However, fitting the skin texture using 3DMM or E-3DMM would become problematic, if the domain of the input data has changed. The problem associated with the missing modes of variation could be rectified by enhancing the 3D face model. However this would require collecting relevant 3D face data, a labour-intensive task which would often be impracticable. In any case, this approach would not be appropriate for dealing with other image degradation effects, such as occlusion or image compression artefacts.

The aim of this paper is to develop techniques that can harness the benefits of 3D models in 2D face recognition when the input image is corrupted, e.g. by occlusion, or when it exhibits intra-personal variations which cannot be explicitly synthesised by the models. We address the problem by learning directly from 2D face data the subspace spanned by the missing modes of variation in the surface texture space superimposed on the 3D face shape structure. This is accomplished by estimating the pose of the input image and the face shape from the detected landmarks. The difference of the aligned input and reference images is used to construct a surface texture map. A training set of these texture maps then defines the perturbation space which can be represented using PCA bases. Assuming that the image perturbation subspace is orthogonal to the 3D face model space, then these additive components can be recovered from an unseen input image, resulting in an improved fit of the 3D face model.

We refer to this proposed method as unified 3DMM (U-3DMM). Unlike the existing 3DMMs, U-3DMM models additional modes of variations in a unified linear framework, which can generalise also to occlusion. In addition, fitting U-3DMM to 2D images is very efficient. It involves first estimating the perturbation component of the input image. Once this component is removed, the core 3D face model fitting is a linear estimation problem. Last, the training set for U-3DMM is much easier to collect than that for 3DMMs.

We conduct an extensive evaluation of U-3DMM on databases which contain diverse modes of variation and perturbation. Experiments show the face recognition rates of U-3DMM are very competitive to state-of-the-art methods. We also present baseline face recognition results on a new dataset including combined pose, illumination and occlusion variations. The datasets and features extracted by U-3DMM will be made publicly available.

The contributions can be summarised as:

- U-3DMM augments the core 3D face model by an additional PCA subspace, a perturbation subspace. Specifically, we project 2D images to 3D space via geometric fitting. Then, in the 3D space, the difference of two images (one being a reference and the other exhibiting additional variations) works as a training sample to learn the perturbation part of U-3DMM. This process is detailed in Section 4.3 and Fig. 4. The linear model of these supplementary variations is generic. The framework can model any variation(s), e.g. occlusion, if appropriate training data is available.

– It is an open problem to achieve an accurate and efficient fitting for 3DMMs. Unlike non-linear models such as Phong illumination model used by 3DMMs, the linear perturbation model of U-3DMM can be fitted very efficiently.
– Large number of 3D faces used to train inter- and intra-personal variations are expensive to collect. In comparison, the proposed method uses 2D images, which are much cheaper and easier to acquire, to train diverse variations in the U-3DMM framework.

The paper is organised as follows. In Section 2, we present the related work. The 3DMM and its fitting problem are formulated in Section 3. Section 4 details our methodology. The proposed algorithm is evaluated in Section 5. Section 6 draws conclusions.

2 Related work

In this section, we discuss the current state-of-the-art. We first introduce various 3D models and fitting strategies, then the motivation of this work is discussed.

2.1 3D face models

3D face modeling is an active research field with many applications. The biometrics community uses 3D models to improve face recognition performance. In the Graphics and animation community, 3D models are used to reconstruct facial details such as wrinkles. In this work, we mainly focus on the 3D models used for biometrics, namely face recognition. These 3D models are classified into three categories: 3D shape model (3DSM), 3D Morphable Model (3DMM) and extended 3DMM (E-3DMM).

3DSM solves the pose problem using either pose normalisation (PN) [6] or pose synthesis (PS) [3], [4], [5]. For the PN method, input images of arbitrary poses are converted to a canonical (frontal) view via a 3D model, then traditional 2D face matchers are used for recognition. On the other hand, PS methods synthesise multiple virtual images with different poses for each gallery image. Only virtual images with similar pose to the probe are chosen for matching. However, these models can only explicitly model one intra-personal variation (pose).

Unlike the 3DSM, the 3D morphable model (3DMM) [12], [7] consists of not only a shape model but a texture model learned from a set of 3D exemplar faces. The traditional 3DMMs [12], [7] can explicitly model pose and illumination variations. Pose is estimated by either a perspective camera [12], [7] or an affine camera [13], [8], and illumination is modelled by either Phong model [12], [7] or Spherical Harmonic model [13], [8].

In addition to pose and illumination variations, the extended 3DMM [10],[11],[14],[15] (E-3DMM) can model facial expressions. Specifically, the authors collected large number of 3D scans with diverse expressions to train a shape model which can capture both facial shape and expression variations. Experiments show E-3DMM achieves promising face recognition performance in the presence of pose and expression variations. The very recent work [14] uses E-3DMM to improve the accuracy of the facial landmark detection.

2.2 Fitting

3DMM and E-3DMM can recover the pose, shape, facial texture and illumination from a single image via a fitting process. The fitting is mainly conducted by minimising the RGB value differences over all the pixels in the facial area between the input image and its model-based reconstruction. As the fitting is an ill-posed problem, it is difficult to achieve an efficient and accurate fitting. To improve the fitting performance, many methods have been proposed.

The first fitting method is a Stochastic Newton Optimisation (SNO) [7]. To reduce the computational cost, SNO randomly samples a small subset of the model vertices to construct the fitting cost function. However this small subset does not capture enough information of the whole face, leading to inferior fitting. The Inverse Compositional Image Alignment (ICIA) algorithm [16],[17], a gradient-based method, modifies the cost function so that the Jacobian matrix becomes constant. Thus, the Jacobian matrix does not need to be updated in every iteration, improving the efficiency. The efficiency is also the driver behind the linear shape and texture fitting algorithm (List) [18]. List constructs linear systems for shape and texture optimisations, and it uses gradient-based methods to optimise pose and illumination. Multi-Feature Fitting (MFF) [19] is an accurate fitting strategy. MFF extracts many complementary features, such as edge and specular highlight, from the input image to constrain the fitting, leading to a smoother cost function. A recent work [8] is an efficient fitting strategy. Specifically, a probabilistic model [8] incorporating model generalisation error is used to estimate shape. To model specularity reflectance, [8] first projects the fitting cost function into a specularity-free space to model diffuse light. After that, the results are projected back to the original RGB colour space to model specularity. Two more recent works [20], [21] use image local features for fitting, achieving promising results.

2.3 Motivation

Although 3DMM and its variants (3DSM and E-3DMM) model pose, illumination and expression, they do not explicitly model other intra-personal variations, which limits their applications. Many of the existing 3DMMs model the intra-personal variations in a non-linear fashion, making the fitting a difficult problem. In comparison, we propose a unified linear framework which can model many more intra-personal variations. In addition, the linearity nature of our framework leads to a very efficient and accurate fitting.

3 Traditional 3D morphable model

In this section, the traditional 3DMMs [12],[7] and the fitting problem are formulated. To construct a 3DMM, the registered 3D facial scans including shape and texture are needed. Let the ith vertex of a registered face be located at (x_i, y_i, z_i) and have grey value g_i. Then the shape and texture can be represented as $\mathbf{s}' = (x_1, y_1, z_1, ..., x_n, y_n, z_n)^T$ and $\mathbf{t}' = (g_1, g_2, ..., g_n)^T$, respectively. Symbol n is the number of vertices of a registered face. PCA is then applied to m example faces $\mathbf{s}'$ and $\mathbf{t}'$ separately to express shape

s and texture **t** as:

$$\mathbf{s} = \mathbf{s}_0 + \mathbf{S}\boldsymbol{\alpha}, \quad \mathbf{t} = \mathbf{t}_0 + \mathbf{T}\boldsymbol{\beta} \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^{3n}$ and $\mathbf{t} \in \mathbb{R}^n$. $\mathbf{s}_0$ and $\mathbf{t}_0$ are the mean shape and texture of m training faces respectively. The columns of **S** and **T** are eigenvectors of shape and texture covariance matrices respectively. The free coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ constitute low-dimension codings of **s** and **t**, respectively.

3DMM can recover the 3D shape, texture, pose, and illumination from a single image via a fitting process. The fitting is conducted by minimising the intensity differences between the input and model reconstructed images. To perform such a minimisation, the 3DMM has to be aligned to the input image by projecting the 3D vertices of $\mathbf{s}(\boldsymbol{\alpha})$ to a 2D image plane via a camera model parameterised by ρ. Then we define $\mathbf{a}^M$ and $\mathbf{a}^I$: 1) $\mathbf{a}^M$ is a vector concatenating the pixel values generated by the vertices of a 3DMM. The value of $\mathbf{a}^M$ is determined by facial texture and illumination. In common with [19],[7], the texture is represented by $\mathbf{t}(\boldsymbol{\beta})$ and the illumination is modelled by the Phong reflection with parameter μ. 2) Based on the current alignment determined by $\boldsymbol{\alpha}$ and ρ, the vertices of a 3DMM find the nearest corresponding pixels of a 2D input image. The corresponding pixel values are concatenated as a vector $\mathbf{a}^I$. Therefore, $\mathbf{a}^M$ and $\mathbf{a}^I$ depend on $\{\boldsymbol{\beta}, \mu\}$ and $\{\boldsymbol{\alpha}, \rho\}$, respectively. The fitting can be formulated:

$$\min_{\boldsymbol{\alpha},\rho,\boldsymbol{\beta},\mu} \|\mathbf{a}^I(\boldsymbol{\alpha}, \rho) - \mathbf{a}^M(\boldsymbol{\beta}, \mu)\|^2 \tag{2}$$

In common with [12], [7], $\mathbf{a}^M$ is formulated as:

$$\mathbf{a}^M = \underbrace{(\mathbf{t}_0 + \mathbf{T}\boldsymbol{\beta})}_{inter-personal} .* \underbrace{(l_a\mathbf{I} + l_d\mathbf{N}\mathbf{d}) + \mathbf{e}}_{illumination} \tag{3}$$

where.$*$ denotes element-wise multiplication; l_a and l_d are the strengths of ambient and directed light; $\mathbf{I}$ is a vector with all entries equal to 1; $\mathbf{N} \in \mathbb{R}^{n\times 3}$ is stacked by the surface normal at each vertex; $\mathbf{d} \in \mathbb{R}^3$ denotes light direction; $\mathbf{e}$ is stacked by the specular reflectance e of every vertex: $e = k_s\langle \mathbf{v}, \mathbf{r}\rangle^\tau$. k_s is a constant for specularity; $\mathbf{v}$ and $\mathbf{r}$ denote the viewing and reflection directions respectively. τ denotes the coefficient of shininess. Then Eq. (2) can be rewritten as:

$$\min_{\phi} \| \underbrace{\mathbf{a}^I(\boldsymbol{\alpha}, \rho)}_{input} - \underbrace{((\mathbf{t}_0 + \mathbf{T}\boldsymbol{\beta}). * (l_a\mathbf{I} + l_d\mathbf{N}\mathbf{d}) + \mathbf{e})}_{reconstruction} \|^2 \tag{4}$$

where $\phi = \{\boldsymbol{\alpha}, \rho, \boldsymbol{\beta}, l_a, l_d, \mathbf{d}, \tau\}$. This is a difficult non-linear optimisation problem due to (1) the exponential form of e and (2) the element-wise multiplication. For different optimisation strategies, refer to Section 2.2.

4 Unified 3D morphable model (U-3DMM)

We propose a unified 3D morphable model (U-3DMM), which linearly models inter- and intra-personal variations. Inter-personal variations, which are usually used to model

identity, are discriminative between different people. In comparison, intra-personal variations are caused by various other random factors such as illumination and occlusion. Inter- and intra-personal variations jointly determine the observed images as shown in Fig. 1. In this section, first, the construction of our U-3DMM is described. Next, an efficient fitting strategy is detailed. Finally, we propose a method to train intra-personal variations using 2D images.

4.1 Model

Like 3DMM, U-3DMM consists of shape and texture models. The shape model is exactly the same as **s** in Eq. (1). Here we only focus on the texture part of U-3DMM.

Motivation and Assumption The existing 3DMMs model the relationship between inter- and intra-personal variations in a non-linear fashion, for example, the element-wise multiplication operation between inter-personal and illumination in Eq. (3). There are two weaknesses of this nonlinear relationship: 1) it does not generalise well because different relationships should be found to handle different intra-personal variations. For example, the Phong model can only model illumination. 2) The nonlinearity causes difficulties of optimisation. To solve these two problems, we assume an input face is equal to the sum of inter- and intra-personal variations following [22],[23]:

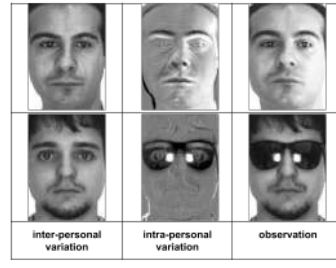

Fig. 1: Intra- and inter-personal variations. The images in the 2nd column are obtained by subtracting the ones in the 3rd column from those in the 1st column with an offset 128

$$\mathbf{a} = \mathbf{a}^{inter} + \mathbf{a}^{intra} \tag{5}$$

where $\mathbf{a}$ is a face, i.e either $\mathbf{a}^M$ or $\mathbf{a}^I$. $\mathbf{a}^{inter}$ and $\mathbf{a}^{intra}$ are the inter- and intra-peronal parts respectively. The effectiveness of this assumption has been validated in [22],[23]. Specifically, this assumption is successfully used for metric learning in [22] and sparse representation-based classification [23], respectively. The former greatly improves the generalisation capacity of the learned metric and the latter solves the single training sample problem. In the field of 3D modeling, this assumption enables 3DMM to model various intra-personal varations in a unified framework. In addition, it leads to an efficient and accurate fitting detailed in Section 4.2.

Modeling Instead of a non-linear relationship in Eq. (3), the reconstructed texture of U-3DMM is linearly modelled as the sum of two parts in Eq. (5). Each part is modelled linearly. To train these two parts separately, training data $\mathbf{t}'$ and $\mathbf{u}'$ are used: $\mathbf{t}'$, which is the same as in Section 3, captures the identity facial texture information; $\mathbf{u}'$ represents one training sample of texture in 3D that captures intra-personal variation such

as expression. $\mathbf{u}'$ has the same dimension as $\mathbf{t}'$ and it is organised in the same order in 3D space as $\mathbf{t}'$. $\mathbf{u}'$ can be any type of intra-personal variation. The generation of $\mathbf{u}'$ will be detailed in Section 4.3. PCA is applied to m samples $\mathbf{t}'$ and p samples $\mathbf{u}'$ separately to generate the inter- and intra-personal subspaces $\mathbf{T}$ and $\mathbf{U}$ respectively. Thus the U-3DMM reconstructed texture $\mathbf{a}^M$ is formulated as:

$$\mathbf{a}^M = \underbrace{\mathbf{t}_0 + \mathbf{T}\boldsymbol{\beta}}_{inter-personal} + \underbrace{\mathbf{U}\boldsymbol{\gamma}}_{intra-personal} \tag{6}$$

$\mathbf{T}, \mathbf{t}_0$ and $\boldsymbol{\beta}$ have the same meaning as in Eq. (1). The inter-personal part is the same as that of 3DMM in Eq. (3). The columns of $\mathbf{U} \in \mathbb{R}^{n \times p}$ are the eigenvectors of the intra-personal variation covariance matrix. $\boldsymbol{\gamma}$ is a free parameter that determines the intra-personal variations. It is assumed that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have Gaussian distributions:

$$p(\boldsymbol{\beta}) \sim \mathcal{N}(0, \boldsymbol{\sigma}_t) \tag{7}$$

$$p(\boldsymbol{\gamma}) \sim \mathcal{N}(\boldsymbol{\gamma}_0, \boldsymbol{\sigma}_u) \tag{8}$$

where the value of $\boldsymbol{\gamma}_0$ is computed by projecting the mean of all the training samples $\mathbf{u}'$ to PCA space $\mathbf{U}$, $\boldsymbol{\sigma}_t = (\sigma_{1,t}, ..., \sigma_{m-1,t})^T$, $\boldsymbol{\sigma}_u = (\sigma_{1,u}, ..., \sigma_{p,u})^T$, and $\sigma_{i,t}^2$ and $\sigma_{i,u}^2$ are the ith eigenvalues of inter- and intra-personal variation covariance matrices respectively.

Advantages The main advantage of U-3DMM is that it can generalise well to diverse intra-personal variations. Table 1 shows that U-3DMM has better generalisation capacity than the existing 3D models. This advantage results from the unified intra-personal part in Eq. (6) which can model more intra-personal variations than the existing 3D models. In addition, compared with the complicated non-linear inter-personal and illumination modeling in Eq. (3), we explicitly linearise the inter- and intra-personal parts in two PCA spaces.

Table 1: Generalisation capacity of 3D models

Method	Pose	Illumination	Expression	Occlusion	Other
3DSM [6, 3]	✓				
3DMM [7, 8]	✓	✓			
E-3DMM [10, 11]	✓	✓	✓		
U-3DMM	✓	✓	✓	✓	✓

4.2 Fitting

By virtue of a fitting process, U-3DMM can recover the pose, 3D facial shape, facial texture and intra-personal variations from an input image as shown in Fig. 2. Linearly

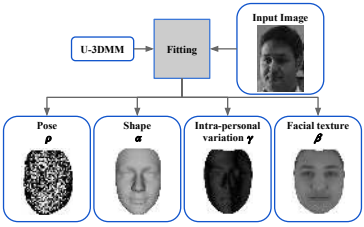

Fig. 2: Input and output of a fitting

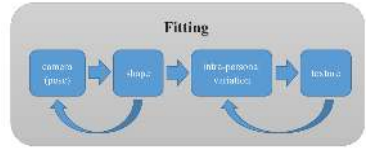

Fig. 3: Topology of fitting algorithm

separating intra- and inter-personal parts allows us to achieve an efficient fitting. Based on Eq. (6), the fitting problem of U-3DMM is formulated as:

$$\min_{\boldsymbol{\alpha},\rho,\boldsymbol{\beta},\boldsymbol{\gamma}} \| \underbrace{\mathbf{a}^I(\boldsymbol{\alpha},\rho)}_{input} - \underbrace{(\mathbf{t}_0 + \mathbf{T}\boldsymbol{\beta} + \mathbf{U}\boldsymbol{\gamma})}_{reconstruction} \|^2 \tag{9}$$

Compared with Eq. (4), clearly, the reconstruction part is linear. To solve this fitting problem, we propose a fitting strategy, which sequentially optimises pose (ρ), shape ($\boldsymbol{\alpha}$), intra-personal ($\boldsymbol{\gamma}$) and facial texture ($\boldsymbol{\beta}$) parameters in four separate steps. Closed-form solutions can be obtained for each of these steps. These parameters are optimised by iterating two sequences of steps in turn as shown in Fig. 3. In each step, only one group of parameters are estimated, and the others are regarded as constant.

Pose and shape estimations In the first two steps, pose (ρ) and shape ($\boldsymbol{\alpha}$) are optimised by solving linear systems using the method in [8]. Specifically, motivated by the fact that the pose and shape variations cause the facial landmarks to shift, ρ and $\boldsymbol{\alpha}$ are estimated by minimising the distance between the landmarks of the input images and those reconstructed from the model. The cost functions for ρ and $\boldsymbol{\alpha}$ are linear [8], and so ρ and $\boldsymbol{\alpha}$ have closed-form solutions. Once ρ and $\boldsymbol{\alpha}$ are estimated, the correspondence between the vertices of the model and pixels of the input images is established.

Intra-personal variation estimation The cost function in Eq. (9) is used to estimate $\boldsymbol{\gamma}$,. In this step, $\mathbf{a}^I(\boldsymbol{\alpha},\rho)$ in Eq. (9) is constant since ρ and $\boldsymbol{\alpha}$ have already been recovered in the first two steps. To avoid over-fitting, a regularisation term based on Eq. (8) is used to constrain the optimisation. Therefore, the optimisation problem is defined as:

$$\min_{\boldsymbol{\gamma}} \|(\mathbf{a}^I - \mathbf{t}_0 - \mathbf{T}\boldsymbol{\beta}) - \mathbf{U}\boldsymbol{\gamma}\|^2 + \lambda_1 \|(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0)./\boldsymbol{\sigma}_u\|^2 \tag{10}$$

The closed-form solution is $\boldsymbol{\gamma} = (\mathbf{U}^T\mathbf{U} + \boldsymbol{\Sigma}_u)^{-1}(\mathbf{U}^T(\mathbf{a}^I - \mathbf{t}_0 - \mathbf{T}\boldsymbol{\beta}) + \lambda_1(\boldsymbol{\gamma}_0./\boldsymbol{\sigma}_u^2))$ where $\boldsymbol{\Sigma}_u = \text{diag}(\lambda_1/\sigma_{1,u}^2, ..., \lambda_1/\sigma_{p,u}^2)$, ./ denotes element-wise division, and λ_1 is a weighting parameter for the regularisation term. Note that $\boldsymbol{\beta}$ is set to $\mathbf{0}$ in the first iteration: in other words the mean facial texture $\mathbf{t}_0$ is used as the initial estimate of the reconstructed image. In subsequent iterations, $\boldsymbol{\beta}$ is replaced by the estimate recovered in the previous iteration.

Facial texture estimation Having obtained an estimate of $\{\rho, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$, $\boldsymbol{\beta}$ can be recovered in the final step. Similar to Eq. (10), the cost function for estimating $\boldsymbol{\beta}$ is defined as:

$$\min_{\boldsymbol{\beta}} \|(\mathbf{a}^I - \mathbf{t}_0 - \mathbf{U}\boldsymbol{\gamma}) - \mathbf{T}\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}./\boldsymbol{\sigma}_t\|^2 \tag{11}$$

The closed-form solution is: $\boldsymbol{\beta} = (\mathbf{T}^T\mathbf{T} + \boldsymbol{\Sigma}_t)^{-1}\mathbf{T}^T(\mathbf{a}^I - \mathbf{t}_0 - \mathbf{U}\boldsymbol{\gamma})$, where λ_2 is a free weighting parameter and $\boldsymbol{\Sigma}_t = \text{diag}(\lambda_2/\sigma_{1,t}^2, ..., \lambda_2/\sigma_{m-1,t}^2)$

4.3 Intra-personal variation data collection

An important prerequisite of building the U-3DMM is to collect intra-personal variation data, i.e $\mathbf{u}'$. The straightforward approach would be to collect enough 3D scans to capture all types of intra-personal variations. However, such 3D data collection is very expensive. In comparison, it is much easier and cheaper to collect 2D image data which covers such variations. Motivated by this, we propose a method to use 2D images to generate 3D intra-personal variation $\mathbf{u}'$.

The outline of our method is illustrated in Fig. 4. Assume that we have two facial images of the *same* person: one without intra-personal variations $\mathbf{a}_c$ and the other $\mathbf{a}_e$ with such variation, e.g. illumination variation in Fig. 4. In the real world, it is easy to collect this type of image pairs from the internet or from publicly available face databases. To project $\mathbf{a}_e$ and $\mathbf{a}_c$ to 3D space, the correspondence between them and the shape model of U-3DMM has to be created first. Like Section 4.2, such a correspondence can be created via geometric fitting, i.e the pose and shape fitting. By virtue of this correspondence, the intensities of $\mathbf{a}_e$ and $\mathbf{a}_c$ can be associated with the 3D vertices of the shape model, generating 3D data $\mathbf{u}_e$ and $\mathbf{u}_c$.. In Eq. (6), the reconstructed image is computed as a sum of inter- and intra-personal variations. We then define the intra-personal variation $\mathbf{u}'$ as the difference between $\mathbf{u}_e$ and $\mathbf{u}_c$:

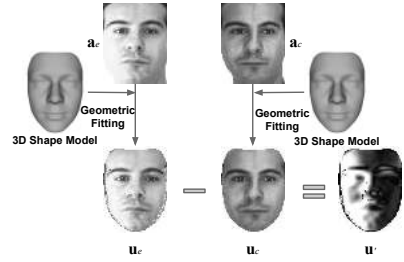

Fig. 4: 3D intra-personal variation data generation. $\mathbf{a}_c$ and $\mathbf{a}_e$ are one 2D image pair without and with intra-personal variation; They are projected to 3D space to reconstruct $\mathbf{u}_c$ and $\mathbf{u}_e$; $\mathbf{u}'$ is the generated 3D data.

$$\mathbf{u}' = \mathbf{u}_e - \mathbf{u}_c \tag{12}$$

The samples of $\mathbf{u}'$ are projected to PCA space to obtain $\mathbf{U}$ of Eq. (6).

Invisible regions Due to self-occlusions and pose variations, some facial parts of the 2D images ($\mathbf{a}_e$ and $\mathbf{a}_c$) are not visible. In this work, the pixel values of $\mathbf{u}_e$ and $\mathbf{u}_c$ corresponding to the self-occluded parts of $\mathbf{a}_e$ and $\mathbf{a}_c$ are set to 0. Although those invisible parts are set to 0 for some training images, the same parts are visible for some other training images under different poses. Therefore, training images of different poses are complementary to model the intra-personal variation part.

5 Experiments

Face recognition aims to reduce the impact of intra-personal variations but keep the discriminative inter-personal information. Thus, we remove the intra-personal variations estimated during fitting and keep the shape and texture for face recognition.

In common with [7],[8],[18], $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are concatenated as a facial feature for face recognition. Cosine similarity and nearest neighbour classifier are used. Landmarks are manually assigned for the initialisation of U-3DMM fitting. To demonstrate the effectiveness of U-3DMM, we compare U-3DMM with the state of the art. To make an extensive comparison, we implemented a very effective 3DMM using multiple feature fitting [19], Sparse Representation Classification (SRC) [24], Extended SRC (ESRC) [23]. The recognition rates of other methods are cited from their papers. We evaluated these methods on Multi-PIE [25], AR [26], and a new synthetic database. Labeled Faces in the Wild (LFW) [27] is another popular face dataset, however, most subjects in LFW have only one image. As U-3DMM needs image pairs of the same subject to train intra-personal term, LFW is not appropriate to evaluate our method and is not used in our experiment.

5.1 Pose, occlusion and illumination variations

U-3DMM is the first 3D approach to explicitly model combined pose, occlusion and illumination variations. In this section, U-3DMM is compared with state-of-the-art.

Database and protocol To our knowledge, there is no database containing large pose, occlusion and illumination variations. Nevertheless, the Multi-PIE database [25] contains two out of three variations, i.e. pose and illumination. We add random occlusions to Multi-PIE images to synthesise a dataset containing all these variations. To simulate real occlusions, the synthetic ones have various sizes and locations within a face.

We generate random occlusions on the facial images. First, we detect the facial area, the width and height of which are denoted as **W** and **H**. Then, a uniformly distributed random coordinate (x, y) in the facial area is generated. Last, the width and height (w and h) of the occlusion are produced by $\{w, h\} = \{\mathbf{W}, \mathbf{H}\} \times rand(0.2, 0.5)$, where $rand$ denotes a uniformly distributed random number generator. $(0.2, 0.5)$ is the range of the random numbers. Hence, the occlusion area of one image can be represented as (x, y, w, h).

A subset of Multi-PIE containing four typical illuminations and four typical poses is used. The four illuminations are left, frontal and right lighting, and ambient lighting (no directed lighting) with lighting IDs 02, 07, 12 and 00. The four poses are frontal and left-rotated by angles 15°, 30°, 45° with pose IDs 051, 140, 130 and 080. Random occlusions are applied to these images. To train the intra-personal variation part of U-3DMM, a subset of 37 identities (from ID-251 to ID-292) in session II is used. The test set contains all the subjects (from ID-001 to ID-250) in session I. In the test set, the frontal images with ambient lighting and without occlusion are the gallery images and the others are the probe. Both training and test sets contain various pose, illumination and occlusion variations.

Results Both 3D shape and texture parameters (α and β) of U-3DMM are discriminative. It is interesting to explore the impact of them on the performance. From Table 2, the face recognition rates when using texture information only is much higher than that when using shape only, indicating that texture is more discriminative than shape. For different illuminations, the performance when using shape does not vary greatly, compared with using texture only. Clearly, facial texture is more sensitive to illumination variations than shape. It is also observed that combining shape and texture by concatenating α and β consistently outperforms either one of them. In all the following evaluations, we use both shape and texture.

We compare qualitatively the reconstruction performance of 3DMM and U-3DMM. Facial textures reconstructed by β are shown in Fig. 5. Clearly, 3DMM suffers from the over-fitting problem caused by occlusion while U-3DMM can reconstruct the facial feature more accurately.

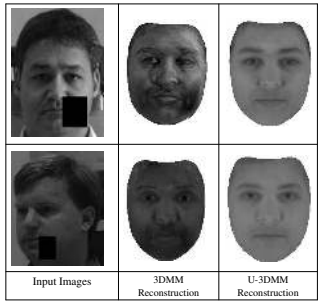

Fig. 5: Frontal texture reconstructions

Parameter	ambient light	left light	frontal light	right light
shape (α)	32.1	31.3	31.2	32.1
texture (β)	59.6	43.4	27.9	48.2
shape+texture	64.2	48.5	33.9	53.3

Table 2: Average recognition rate (%) of U-3DMM over all the poses and occlusions per illumination

To demonstrate the effectiveness of U-3DMM, it is compared with state-of-the-art methods: SRC [24], ESRC [23] and 3DMM [19]. Fig. 6 illustrates how the recognition performance varies with illumination over poses and occlusions. Our U-3DMM outperforms the other three methods because U-3DMM can effectively handle pose, illumination and occlusion simultaneously. By comparison, SRC and ESRC do not handle the pose problem and 3DMM does not explicitly model occlusion. SRC is worst because it suffers from the problem of having just 'single image per subject in gallery' [24],[23]. In the case of 'frontal light', all the methods work worse than the other three illuminations. The inferior performance results from the fact that the illumination effects between the gallery (ambient lighting) and probe (frontal lighting) are larger than the other illumination conditions.

Fig. 7 shows how recognition rates, averaged over illuminations and occlusions, vary with pose. All the face recognition rates decrease with the increase of pose variations, showing that pose variations present a challenging problem. U-3DMM works much better than the others due to its strong intra-personal variation modeling capacity. In the case of frontal pose (rotation angle is 0°) which means only illumination and occlusion are presented, ESRC achieves promising recognition rates because it can explicitly model illumination and occlusion. U-3DMM outperforms ESRC because U-3DMM can extract discriminative shape information for recognition while ESRC cannot.

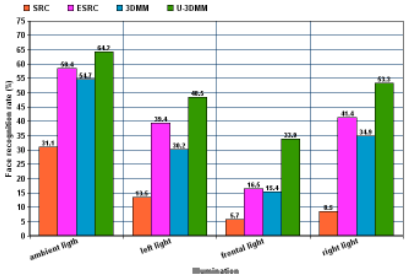

Fig. 6: Average recognition rate over all the poses and occlusions per illumination

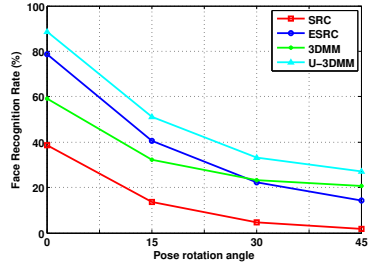

Fig. 7: Average recognition rate over all the illuminations and occlusions per pose

5.2 Pose and illumination variations

Face recognition in the presence of pose variation (PFR) and combining pose and illumination variations (PIFR) is very challenging. Extensive research has been conducted to solve PFR and PIFR problems. In this section, U-3DMM is compared with state-of-the-art methods on Multi-PIE database.

Database and protocol Following the existing work, two settings (*Setting*-I and *Setting*-II) are used for PFR and PIFR respectively. *Setting*-I uses a subset in session 01 consisting of 249 subjects with 7 poses and 20 illumination variations. The images of the first 100 subjects constitute the training set. The remaining 149 subjects form the test set. In the test set, the frontal images under neutral illumination work as the gallery and the remaining are probe images. *Setting*-II uses the images of all the 4 sessions (01-04) under 7 poses and only neutral illumination. The images from the first 200 subjects are used for training and the remaining 137 subjects for testing. In the test set, the frontal images from the earliest session work as gallery, and the others are probes.

Table 3: Recognition rate (%) across poses on Multi-PIE

	Method	-45°	-30°	-15°	+15°	+30°	+45°	Avg
2D	GMA [28]	75.0	74.5	82.7	92.6	87.5	65.2	79.6
	DAE [29]	69.9	81.2	91.0	91.9	86.5	74.3	82.5
	SPAE [30]	84.9	92.6	96.3	95.7	94.3	84.4	91.4
3D	Asthana [6]	74.1	91.0	95.7	95.7	89.5	74.8	86.8
	MDF [31]	78.7	94.0	99.0	98.7	92.2	81.8	90.7
	E-3DMM (HDF) [15]	97.4	99.5	99.5	99.7	99.0	96.7	**98.6**
	U-3DMM (PCA)	91.2	95.7	96.8	96.9	95.3	90.9	94.5
	U-3DMM (HDF)	96.5	98.4	99.2	98.9	97.9	96.1	**97.8**

Pose-robust face recognition Table 3 compares our U-3DMM with the state-of-the-art. SPAE[30] works best among all the 2D methods due to its strong non-linear modeling capacity. E-3DMM [15] and U-3DMM outperform another two 3D methods ([6] and [31]), as E-3DMM [15] and U-3DMM can model both pose and facial shape rather than pose only by [6, 31]. E-3DMM only reports the results using High-Dimensional gabor Feature (HDF) [32] rather than PCA coefficients (α and β). For fair comparison, we also extracted the HDF feature from pose-normalised rendered images as follows: First, an input image is aligned to U-3DMM via geometric fitting. Second, the intensity value of each vertex of U-3DMM is assigned by the value of the corresponding pixel of the input image. The values of invisible parts of U-3DMM are assigned with the values from symmetry visible vertices. Last, a frontal face image is rendered using the obtained intensity values of U-3DMM. U-3DMM (HDF) works much better than U-3DMM (PCA) because 1) the HDF feature can capture both global and local facial information, in comparison with only global information captured by PCA coefficients; 2) HDF uses more invariant Gabor feature than pixel values which are actually coded by PCA coefficients. Our U-3DMM (HDF) works slightly worse than E-3DMM (HDF), however, U-3DMM has advantages over E-3DMM: 1) E-3DMM itself can only model pose and expression, however, U-3DMM can explicitly model more intra-personal variations; 2) The expression part of E-3DMM is trained using 3D faces with various expressions, while U-3DMM is trained using easily-collected 2D images; 3) E-3DMM estimates the depth of background, leading to extra computational costs but being useless for improving face recognition rates; while U-3DMM does not.

Pose- and illumination-robust face recognition As shown in Table 4, the subspace method [33] works much worse than the others. The deep learning methods, i.e FIP (face identity-preserving) [2], RL (FIP reconstructed features) [2] and MVP (multi-view perceptron) [34], achieve promising results. U-3DMM outperforms deep learning methods (RL, FIP, and MVP) because 3D methods intrinsically model pose and illumination. Apart from worse performance, deep learning methods share the difficulty of designing a 'good' architecture because 1) there is no theoretical guide and 2) the large number of free parameters are hard to tune.

Table 4: Recognition rate (%) averaging 20 illuminations on Multi-PIE

Method		-45°	-30°	-15°	0°	+15°	+30°	+45°	Avg
Subspace Learning	Li [33]	63.5	69.3	79.7	N/A	75.6	71.6	54.6	69.1
Deep Learning	RL [2]	66.1	78.9	91.4	94.3	90.0	82.5	62.0	80.7
	FIP [2]	63.6	77.5	90.5	94.3	89.8	80.0	59.5	79.3
	MVP [34]	75.2	83.4	93.3	95.7	92.2	83.9	70.6	84.9
3D Method	U-3DMM	73.1	86.9	93.3	99.7	91.3	81.2	69.7	**85.0**

5.3 Other intra-personal variations

To further validate the effectiveness of U-3DMM, we evaluate it on the AR database [26].

Database and protocol The AR database contains more than 4000 frontal images of 126 subjects with variations in expressions, illuminations and occlusions. To train the intra-personal component of U-3DMM, we use 10 randomly chosen subjects (5 male and 5 female) in Session 1 and with 13 images per subject. Following [10], we randomly chose 100 subjects in Session 1 for testing. In the test set, the neutral images work as gallery and the others are probe.

Results We compare U-3DMM with 3DMM, E-3DMM [10], SRC, and ESRC . In the presence of illumination, occlusion or both, U-3DMM works much better than SRC and ESRC. This conclusion is consistent with that drawn in Section 5.1. For expression variations, our U-3DMM outperforms SRC and ESRC, but works worse than E-3DMM. Note that E-3DMM uses a commercial SDK to extract facial feature and the authors [10] do not report face recognition rates using only shape and texture coefficients that our U-3DMM uses. In addition, our U-3DMM has two advantages over E-3DMM [10]: 1) U-3DMM can potentially model any variations, while E-3DMM is designed specifically to capture pose and expression variations; 2) U-3DMM is more efficient than E-3DMM.

Table 5: Recognition rate (%) evaluated on AR database

Method	Expression	Illumination	Occlusion	Illu.+Occl.	Time
SRC [24]	80.7	71.3	42.5	23.5	-
ESRC [23]	94.3	98.7	80.5	74.5	-
E-3DMM [10]	99.0	-	-	-	slow [1]
3DMM [19]	89.3	99.1	74.5	71.8	23s
U-3DMM	95.3	98.7	92.0	85.5	0.98

[1] The computational complexity of E-3DMM is the same as 3DMM.

6 Conclusions

We propose the U-3DMM, which provides a generic linear framework to model complicated intra-personal variations. The linearity of U-3DMM leads to an efficient and accurate fitting. The experimental results demonstrate that U-3DMM achieves very competitive face recognition rates against the state-of-the-art.

Acknowledgements

This work was sponsored by EPSRC project ‘Signal processing in a networked battlespace’ under contract EP/K014307/1, ‘FACER2VM’ under EP/N007743/1, NSFC project under Grant 61375031 and 61573068. The support from EPSRC and the MOD University Defence Research Collaboration (UDRC) in Signal Processing is gratefully acknowledged.

References

1. Hu, G., Chan, C.H., Yan, F., Christmas, W., Kittler, J.: Robust face recognition by an albedo based 3d morphable model. In: Biometrics (IJCB), 2014 IEEE International Joint Conference on, IEEE (2014) 1–8
2. Zhu, Z., Luo, P., Wang, X., Tang, X.: Deep learning identity preserving face space. In: Proc. ICCV. Volume 1. (2013) 2
3. Niinuma, K., Han, H., Jain, A.K.: Automatic multi-view face recognition via 3d model based pose regularization. In: Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on, IEEE (2013) 1–8
4. Prabhu, U., Heo, J., Savvides, M.: Unconstrained pose-invariant face recognition using 3d generic elastic models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(10) (2011) 1952–1961
5. Zhang, X., Gao, Y., Leung, M.K.: Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. Information Forensics and Security, IEEE Transactions on **3**(4) (2008) 684–697
6. Asthana, A., Marks, T.K., Jones, M.J., Tieu, K.H., Rohith, M.: Fully automatic pose-invariant face recognition via 3d pose normalization. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 937–944
7. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(9) (2003) 1063–1074
8. Aldrian, O., Smith, W.A.: Inverse rendering of faces with a 3d morphable model. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(5) (2013) 1080–1093
9. Rodriguez, J.T.: 3D Face Modelling for 2D+3D Face Recognition. PhD thesis, Surrey University, Guildford, UK (2007)
10. Chu, B., Romdhani, S., Chen, L.: 3d-aided face recognition robust to expression and pose variations. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE (2014) 1907–1914
11. Amberg, B., Knothe, R., Vetter, T.: Expression invariant 3d face recognition with a morphable model. In: Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on, IEEE (2008) 1–6
12. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. (1999) 187–194
13. Zhang, L., Samaras, D.: Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(3) (2006) 351–363
14. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. arXiv preprint arXiv:1511.07212 (2015)
15. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 787–796
16. Romdhani, S., Vetter, T.: Efficient, robust and accurate fitting of a 3d morphable model. In: Computer Vision. Proceedings. 9th IEEE International Conference on, IEEE (2003) 59–66
17. Kang, B., Byun, H., Kim, D.: Multi-resolution 3d morphable models and its matching method. In: Pattern Recognition, 19th International Conference on, IEEE (2008) 1–4
18. Romdhani, S., Blanz, V., Vetter, T.: Face identification by fitting a 3d morphable model using linear shape and texture error functions. ECCV 2002 (2006) 3–19
19. Romdhani, S., Vetter, T.: Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 2., IEEE (2005) 986–993

20. Huber, P., Feng, Z.H., Christmas, W., Kittler, J., Rätsch, M.: Fitting 3d morphable models using local features. arXiv preprint arXiv:1503.02330 (2015)
21. Zhu, X., Yan, J., Yi, D., Lei, Z., Li, S.Z.: Discriminative 3d morphable model fitting. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. Volume 1., IEEE (2015) 1–8
22. Chen, D., Cao, X., Wang, L., Wen, F., Sun, J.: Bayesian face revisited: A joint formulation. In: Computer Vision–ECCV 2012. Springer (2012) 566–579
23. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(9) (2012) 1864–1870
24. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(2) (2009) 210–227
25. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28**(5) (2010) 807–813
26. Martinez, A.M.: The ar face database. CVC Technical Report **24** (1998)
27. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst (2007)
28. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2160–2167
29. Bengio, Y.: Learning deep architectures for ai. Foundations and trends® in Machine Learning **2**(1) (2009) 1–127
30. Kan, M., Shan, S., Chang, H., Chen, X.: Stacked progressive auto-encoders (spae) for face recognition across poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 1883–1890
31. Li, S., Liu, X., Chai, X., Zhang, H., Lao, S., Shan, S.: Morphable displacement field based image matching for face recognition across pose. In: Computer Vision–ECCV 2012. Springer (2012) 102–115
32. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 3025–3032
33. Li, A., Shan, S., Gao, W.: Coupled bias-variance tradeoff for cross-pose face recognition. Image Processing, IEEE Transactions on **21**(1) (2012) 305–315
34. Zhu, Z., Luo, P., Wang, X., Tang, X.: Deep learning multi-view representation for face recognition. arXiv preprint arXiv:1406.6947 (2014)