What Does 2D Geometric Information Really Tell Us About 3D Face Shape?

Anil Bas[1,2] · William A. P. Smith[1]

Abstract

A face image contains geometric cues in the form of configurational information and contours that can be used to estimate 3D face shape. While it is clear that 3D reconstruction from 2D points is highly ambiguous if no further constraints are enforced, one might expect that the face-space constraint solves this problem. We show that this is not the case and that geometric information is an ambiguous cue. There are two sources for this ambiguity. The first is that, within the space of 3D face shapes, there are flexibility modes that remain when some parts of the face are fixed. The second occurs only under perspective projection and is a result of perspective transformation as camera distance varies. Two different faces, when viewed at different distances, can give rise to the same 2D geometry. To demonstrate these ambiguities, we develop new algorithms for fitting a 3D morphable model to 2D landmarks or contours under either orthographic or perspective projection and show how to compute flexibility modes for both cases. We show that both fitting problems can be posed as a separable nonlinear least squares problem and solved efficiently. We demonstrate both quantitatively and qualitatively that the ambiguity is present in reconstructions from geometric information alone but also in reconstructions from a state-of-the-art CNN-based method.

Keywords 3D morphable face model · Shape ambiguity · Perspective projection · Landmarks

1 Introduction

A 2D image of a face contains various cues that can be exploited to estimate 3D shape. In this paper, we explore to what degree 2D geometric information allows us to estimate 3D face shape. This is sometimes referred to as "configurational" information and includes the relative layout of features (usually encapsulated in terms of the position of semantically meaningful landmark points) and contours (caused by occluding boundaries or texture edges). The advantage of using such cues is that they provide direct information about the shape of the face, without having to model

✉ Anil Bas
   ab1792@york.ac.uk

   William A. P. Smith
   william.smith@york.ac.uk

[1] Department of Computer Science, University of York, York, UK

[2] Department of Computer Engineering, Marmara University, Istanbul, Turkey

the photometric image formation process and to interpret appearance.

Although photometric information does provide a cue to the 3D shape of a face (Smith and Hancock 2006), it is a fragile cue because it requires estimates of lighting, camera properties and reflectance properties making it difficult to apply to "in the wild" images. Moreover, in some conditions, the shape-from-shading cue may be entirely absent. Perfectly ambient light cancels out all shading other than ambient occlusion which provides only a very weak shape cue (Prados et al. 2009). For this reason, the use of geometric information has proven very popular in 3D face reconstruction (Blanz et al. 2004; Aldrian and Smith 2013; Patel and Smith 2009; Knothe et al. 2006; Cao et al. 2014a; Bas et al. 2016). Landmark detection on highly uncontrolled face images is now a mature research field with benchmarks (Sagonas et al. 2016) providing an indication of likely accuracy. Landmarks are often used to initialise or constrain the fitting of 3D morphable models (3DMMs) to images while denser 2D geometric information such as the occluding boundary are used in some of the state-of-the-art methods.

Fig. 1 Perspective transformation of real faces from the CMDP dataset (Burgos-Artizzu et al. 2014). The subject is the same in each column and the same camera and lighting is used. The change in viewing distance (60 cm top row, 490 cm bottom row) induces a significant change in projected shape (Color figure online)

In this paper we show that 2D geometric information only provides a partial constraint on 3D face shape. In other words, face landmarks or occluding contours are an ambiguous shape cue. Rather than try to explain 2D geometric data with a single, best fitting 3D face, we seek to recover a subspace of possible 3D face shapes that are consistent with the 2D data. "Consistent" here means that the model explains the data within the tolerance with which we can hope to locate these features within a 2D image. For example, state-of-the-art automatic face landmarking provides a mean landmark error under 4.5% of interocular distance for only 50% of images [according to the second conduct of the 300 Faces in the Wild challenge (Sagonas et al. 2016)]. We show how to compute this subspace and show that it contains very significant shape variation. The ambiguity arises for two reasons. The first is that, within the space of possible faces (as characterised by a 3DMM) there are degrees of flexibility that do not change the 2D geometric information when projection parameters are fixed (this applies to both orthographic and perspective projection). The second is caused by the nonlinear effect of perspective.

When a human face is viewed under perspective projection, its 2D shape varies with the distance between the camera and subject. The effect of perspective transformation is to distort the relative distances between facial features and can be quite dramatic. When a face is close to the camera, it appears taller and slimmer with the features closest to the camera (nose and mouth) appearing relatively larger and the ears appearing smaller and partially occluded. As distance increases and the shape converges towards the orthographic projection, faces appear broader and rounder with ears that protrude further. We show some examples of this effect in Fig. 1. Images are taken at 60 cm and 490 cm. Each face is cropped and rescaled such that the interocular distance is the same. The distortion caused by perspective transformation is clearly visible. This effect leads to the second ambigu-

ity. Namely that, two different (but natural) 3D face shapes viewed at different distances can give rise to the same 2D geometric features.

In order to demonstrate both ambiguities, we propose novel algorithms for fitting a 3DMM to 2D geometric information and extracting the subspace of possible 3D shapes. Our contribution is to observe that, under both orthographic and perspective projection, model fitting can be posed as a separable nonlinear least squares optimisation problem that can be solved efficiently without requiring any problem specific optimisation method, initialisation or parameter tuning. In addition, we use real face images to verify that the ambiguity is present in actual faces. We show that, on average, 2D geometry is more similar between different faces viewed at the same distance than it is between the same face viewed at different distances. We present quantitative and qualitative results on synthetic 2D geometric data created by projection of real 3D scans. We also present qualitative results on real images from the Caltech Multi-Distance Portraits (CMDP) dataset (Burgos-Artizzu et al. 2014).

2 Related Work

2.1 3D Face Shape from 2D Geometric Information

Facial landmarks, i.e. points with well defined correspondence between identities, are used in a number of ways in face processing. Most commonly, they are used for registration and normalisation, as is done in training an Active Appearance Model (Cootes et al. 1998) or in CNN-based face recognition frameworks (Taigman et al. 2014). For this reason, there has been sustained interest in building feature detectors capable of accurately labelling face landmarks in uncontrolled images (Sagonas et al. 2016).

Motivated by the recent improvements in the robustness and efficiency of 2D facial feature detectors, a number of researchers have used the position of facial landmarks in a 2D image as a cue for 3D face shape. In particular, by fitting a 3DMM to these detected landmarks (Blanz et al. 2004; Aldrian and Smith 2013; Patel and Smith 2009; Knothe et al. 2006). All of these methods assume an affine camera and hence the problem reduces to a multilinear problem in the unknown shape and camera parameters. The problem of interpreting 3D face shape from 2D landmark positions is related to the problem of non-rigid structure from motion (Hartley and Vidal 2008). However, in that case, the basis set describing the non-rigid deformations is unknown but multiple views of the deforming object are available. In our case, the basis set is known (it is "face space" - represented here by a 3DMM) but only a single view of the face is available. Some work has considered other 2D shape features besides landmark points. Keller et al. (2007) fit a 3DMM to contours

(both silhouettes and inner contours due to texture, shape and shadowing). Bas et al. (2016) adapt the Iterated Closest Point algorithm to fit to edge pixels with an additional landmark term. They use alternating linear least squares followed by a non-convex refinement. Although not applied to faces, Zhou et al. (2015) propose a convex relaxation of the shape-from-landmarks energy. Several recent works (Cao et al. 2013, 2014a; Saito et al. 2016) use landmark fitting to generate ground truth to train a direct image-to-shape parameters regressor. Again, the landmark fitting optimisation is performed using alternating minimisation, this time under perspective projection with a given focal length. Interestingly, Cao et al. (2014a) explicitly note that varying the focal length leads to different shapes and use binary search to find the one that gives lowest residual error.

A related problem is to describe the remaining flexibility in a statistical shape model that is partially fixed. If the position of some points, curves or subset of the surface is known, the goal is to characterise the space of shapes that approximately fit these observations. Albrecht et al. (2008) show how to compute the subspace of faces with the same profile. Lüthi et al. (2009) extended this approach into a probabilistic setting.

The vast majority of 2D face analysis methods that involve estimation of 3D face shape or fitting of a 3D face model assume a linear camera (such as scaled orthographic/weak perspective or affine) (Blanz et al. 2004; Aldrian and Smith 2013; Patel and Smith 2009; Knothe et al. 2006). Such a camera does not introduce any nonlinear perspective transformation. While this assumption is justified in applications where the subject-camera distance is likely to be large, any situation where a face may be viewed from a small distance must account for the effects of perspective (particularly common due to the popularity of the "selfie" format). For this reason, in this paper we consider both orthographic and perspective camera models.

We emphasise that we study the ambiguities only in a monocular setting and, for the perspective case, assuming no geometric calibration. Multiview constraints would reduce or remove the ambiguity. For example, Amberg et al. (2007) describe an algorithm for fitting a 3DMM to stereo face images. In this case, the stereo disparity cue used in their objective function conveys depth information which helps to resolve the ambiguity. However, note that even here, their solution is unstable when camera parameters are unknown. They introduce an additional heuristic constraint on the focal length, namely they restrict it to be between 1 and 5 times the sensor size.

2.2 Deep Model-Based Face Analysis

While the methods above rely on explicit features such as detected landmarks, state of the art methods for 3DMM fitting use deep convolutional neural networks (CNNs) that can learn to exploit any combination of features. Typically, these methods train a CNN to regress 3DMM parameters directly from an input image using a variety of different forms of supervision. Tran et al. (2017) perform supervised, discriminative training by first running a multi-image fitting method (Piotraschke and Blanz 2016) on sets of images of the same person and then training the network to predict these parameters from single images. Their multi-image fitting method is based on weighted averaging of single image fits that are themselves initialised by landmark fitting. This initial landmark fit is subject to the ambiguities described in this paper, though the subsequent use of appearance-based losses may not be. However, the latest state-of-the-art in analysis-by-synthesis based fitting suggests that even using dense appearance information the ambiguity may still exist. Schönborn et al. (2017) use a sampling approach based on Markov Chain Monte Carlo to estimate the full posterior distribution using a hybrid loss including landmarks and appearance error. They note a very high posterior standard deviation in estimated distance from the camera concluding that the ambiguity under perspective cannot be resolved.

The latest state-of-the-art in regression-based fitting Sanyal et al. (2019) relies entirely on landmark reprojection error, again subject to the ambiguities we describe. Tewari et al. (2017) propose to use a model-based decoder (differentiable renderer) such that the estimated shape, texture, pose and illumination parameters can be rendered back into an image and a self-supervised appearance loss computed. We draw particular attention to the fact that this method incorporates a landmark loss. The appearance loss only provides a useful gradient for training when already close to a good solution, so the landmark loss is essential to coarsely train the network. This loss is subject to exactly the ambiguities we describe in this paper. In addition, during training, the learning rate on the Z translation (i.e. subject-camera distance) is set three orders of magnitude lower than all other parameters. In other words, the network essentially learns to reconstruct faces assuming a fixed face distance. The idea of self-supervision has been extended in a number of ways. Tran and Liu (2018) make the 3DMM itself learnable. Tewari et al. (2018) learn a corrective space to add details not captured by the model. Genova et al. (2018) learn to regress from face identity parameters to 3DMM parameters such that the rendered face encodes to similar identity parameters to the original image.

CNNs have also been used to directly estimate *correspondence* between a 3DMM and a 2D face image, without explicitly estimating 3DMM shape parameters or pose. Unlike landmarks, this correspondence is dense, providing a 2D location for every visible vertex. This was first proposed by Güler et al. (2017) who use a fully convolutional network and pose the continuous regression task as a coarse to fine classification task. Yu et al. (2017) take a similar approach

but go further by using the correspondences to estimate 3D face shape by fitting a 3DMM. Wu et al. (2018) learn this fitting process as well. Sela et al. (2017) take a multitask learning approach by training a CNN to predict both correspondence and facial depth. In all cases, this estimated dense correspondence provides an ambiguous shape cue, exactly as we describe in this paper.

2.3 Faces under Perspective Projection

The effect of perspective transformation on face appearance has been studied from both a computational and psychological perspective previously. In psychology, Liu and Chaudhuri (2003) and Liu and Ward (2006) show that human face recognition performance is degraded by perspective transformation. Perona (2007) and Bryan et al. (2012) investigated a different effect, noting that perspective distortion influences social judgements of faces. In art history, Latto and Harper (2007) discuss how uncertainty regarding subject-artist distance when viewing a painting results in distorted perception. They show that perceptions of body weight from face images are influenced by subject-camera distance.

There have been two recent attempts to address the problem of estimating subject-camera distance from monocular, perspective views of a face (Flores et al. 2013; Burgos-Artizzu et al. 2014). The idea is that the configuration of projected 2D face features conveys something about the degree of perspective transformation. Flores et al. (2013) approach the problem using exemplar 3D face models. They fit the models to 2D landmarks using perspective-n-point (Lepetit et al. 2009) and use the mean of the estimated distances as the estimated subject-camera distance. Burgos-Artizzu et al. (2014) on the other hand work entirely in 2D. They present a fully automated process for estimating 2D landmark positions to which they apply a linear normalisation. Their idea is to describe 2D landmarks in terms of their offset from mean positions, with the mean calculated either across views at different distances of the same face, or across multiple identities at the same distance. They can then perform regression to relate offsets to distance. They compare performance to humans and show that they are relatively bad at judging distance given only a single image.

Our results highlight the difficulty that both of these approaches face. Namely that many interpretations of 2D facial landmarks are possible, all with varying subject-camera distance. We approach the problem in a different way by showing how to solve for shape parameters when the subject-camera distance is known. We can then show that multiple explanations are possible. The perspective ambiguity is hinted at in the literature, e.g. Booth et al. (2018) state "we found that it is beneficial to keep the focal length constant in most cases, due to its ambiguity with t_z", but never explored in a rigorous manner.

Fried et al. (2016) explore the effect of perspective in a synthesis application. They use a 3D head model to compute a 2D warp to simulate the effect of changing the subject-camera distance, allowing them to approximate appearance at any distance given a single image. Valente and Soatto (2015) also proposed a method to warp a 2D image to compensate for perspective. However, their goal was to improve the performance of face recognition systems that they showed are sensitive to such transformations.

Schumacher and Blanz (2012) investigate ambiguities from a perceptual point of view. They explore whether, after seeing a frontal view, participants accept a 3D reconstruction as the correct profile as often as they do for the original profile. It shows that human observers consider the reconstructed shape equally plausible as ground truth, even if it differs significantly from ground truth and even if choices include the original profile of the face.

2.4 Other Ambiguities

There are other known ambiguities in the monocular estimation of 3D shape. The bas relief ambiguity (Belhumeur et al. 1999) arises in photometric stereo with unknown light source directions. A continuous class of surfaces (differing by a linear transformation) can produce the same set of images when an appropriate transformation is applied to the illumination and albedo. For the particular case of faces, Georghiades et al. (2001) resolve this ambiguity by exploiting the symmetries and similarities in faces. Specifically they assume: bilateral symmetry; that the forehead and chin should be at approximately the same depth; and that the range of facial depths is about twice the distance between the eyes.

In the *hollow face illusion* (Hill and Bruce 1994), shaded images of concave faces are interpreted as convex faces with inverted illumination. The illusion even holds when the hollow face is moving, with rotations being interpreted in reverse. This is a binary version of the bas relief ambiguity occurring when both convex and concave faces are interpreted as convex so as to be consistent with prior knowledge.

More generally, ambiguities in surface reconstruction have been considered in a number of settings. Ecker et al. (2008) consider the problem of reconstructing a smooth surface from local information that contains a discrete ambiguity. The ambiguities studied here are in the local surface orientation or gradient, a problem that occurs in photometric shape reconstruction. Salzmann et al. (2007) study the ambiguities that arise in monocular nonrigid structure from motion under perspective projection.

Like us, Moreno-Noguer and Fua (2013) also explore ambiguities in shape-from-landmarks in the context of objects represented by a linear basis (in their case, nonrigid deformations of an object rather than the space of faces). However, unlike in this paper, they assume that the intrinsic

camera parameters are known. Hence, they do not model the perspective ambiguity that we describe (in which a change in distance is compensated by a change in focal length). Different to our flexibility modes, instead of analytically deriving a subspace, they use stochastic sampling to explore the set of possible solutions. They attempt to select from within this space using additional information provided by motion or shading.

In an early version of this work (Smith 2016), we considered only the effect of perspective and assumed that rotation and translation were fixed. Here we go further by also considering orthographic projection and showing how to compute flexibility modes. Moreover, we show how model fitting can be posed as a separable nonlinear least squares problem, including solving for rotation and translation, and present more comprehensive experimental results. Finally, we consider not only landmarks but also show how to fit to contours where model-image correspondence is not known.

3 Preliminaries

Our approach is based on fitting a 3DMM to 2D landmark observations under either orthographic or perspective projection. Hence, we begin by describing the 3DMM and the scaled orthographic and pinhole projection model. We provide the definition of symbols in Table 1.

3.1 3D Morphable Model

A 3DMM is a deformable mesh whose vertex positions, $\varsigma(\boldsymbol{\alpha})$, are determined by the shape parameters $\boldsymbol{\alpha} \in \mathbb{R}^S$. Shape is described by a linear subspace model learnt from data using principal component analysis (PCA) (Blanz and Vetter 2003). So, the shape of any object from the same class as the training data can be approximated as:

$$\varsigma(\boldsymbol{\alpha}) = \mathbf{Q}\boldsymbol{\alpha} + \bar{\varsigma}, \tag{1}$$

where the vector $\varsigma(\boldsymbol{\alpha}) \in \mathbb{R}^{3N}$ contains the coordinates of the N vertices, stacked to form a long vector: $\varsigma = [u_1, v_1, w_1, \ldots, u_N, v_N, w_N]^{\mathrm{T}}$, $\mathbf{Q} \in \mathbb{R}^{3N \times S}$ contains the S retained principal components and $\bar{\varsigma} \in \mathbb{R}^{3N}$ is the mean shape. Hence, the ith vertex is given by: $\mathbf{v}_i = [\varsigma_{3i-2}, \varsigma_{3i-1}, \varsigma_{3i}]^{\mathrm{T}}$.

For convenience, we denote the sub-matrix corresponding to the ith vertex as $\mathbf{Q}_i \in \mathbb{R}^{3 \times S}$ and the corresponding vertex in the mean face shape as $\bar{\varsigma}_i \in \mathbb{R}^3$, such that the ith vertex is given by: $\mathbf{v}_i = \mathbf{Q}_i\boldsymbol{\alpha} + \bar{\varsigma}_i$.

Since the morphable model that we use has meaningful units (i.e. it was constructed from scans where vertex positions were recorded in metres) we do not need a scale parameter to transform from model to world coordinates.

Table 1 Definition of symbols

Symbol	Description	Object type
N	No. 3D vertices	$\in \mathbb{Z}$
S	No. model dimensions	$\in \mathbb{Z}$
L	No. 2D landmarks	$\in \mathbb{Z}$
$\mathbf{Q}$	Principal components	$\in \mathbb{R}^{3N \times S}$
$\boldsymbol{\alpha}$	Shape parameter vector	$\in \mathbb{R}^S$
$\bar{\varsigma}$	Mean face shape	$\in \mathbb{R}^{3N}$
$\mathbf{v}_i$	ith 3D point (vertex)	$\in \mathbb{R}^3$
$\mathbf{x}_i$	ith 2D point	$\in \mathbb{R}^2$
$\mathbf{P}$	Orthographic projection matrix	$\in \mathbb{R}^{2 \times 3}$
$\mathbf{R}$	Rotation matrix	$\in SO(3)$
$\mathbf{r}$	Axis-angle vector	$\in \mathbb{R}^3$
$\mathbf{t}$	Translation vector	$\in \mathbb{R}^2$ or $\mathbb{R}^3$
s	Scale	$\in \mathbb{R}_{>0}$
f	Focal length	$\in \mathbb{R}_{>0}$
$\mathbf{K}$	Camera intrinsics	$\in \mathbb{R}^{3 \times 3}$
ε	Objective function	$\in \mathbb{R}_{\geq 0}$
$\mathbf{d}$	Vector of residuals	$\in \mathbb{R}^{2L}$ or $\mathbb{R}^{3L}$
$\mathbf{I}_n$	Identity matrix	$\in \{0, 1\}^{n \times n}$
$\mathbf{1}_n$	Column vector of ones	$\in \{1\}^n$
$\mathbf{J}$	Jacobian matrix	$\in \mathbb{R}^{2L \times 4}$ or $\mathbb{R}^{3L \times 4}$
t_z	Face-camera distance	$\in \mathbb{R}_{>0}$
k	Threshold value	$\in \mathbb{R}_{>0}$
$\boldsymbol{\Pi}$	2D Projection	$\in \mathbb{R}^{2L \times S}$ or $\mathbb{R}^{3L \times S}$
$\mathbf{f}$	Flexibility modes	$\in \mathbb{R}^S$
λ_i	ith eigenvalue	$\in \mathbb{R}$
$\mathcal{B}$	Occluding boundary vertices	$\subset \{1, \ldots, N\}$
$\otimes$	Kronecker product	Operator

3.2 Scaled Orthographic Projection

The scaled orthographic, or weak perspective, projection model assumes that variation in depth over the object is small relative to the mean distance from camera to object. Under this assumption, the projection of a 3D point $\mathbf{v} = [u, v, w]^{\mathrm{T}}$ onto the 2D point $\mathbf{x} = [x, y]^{\mathrm{T}}$ is given by $\mathbf{x} = \mathbf{SOP}[\mathbf{v}, \mathbf{R}, \mathbf{t}_{2\mathrm{d}}, s] \in \mathbb{R}^2$ which does not depend on the distance of the point from the camera, but only on a uniform scale s given by the ratio of the focal length of the camera and the mean distance from camera to object:

$$\mathbf{SOP}[\mathbf{v}, \mathbf{R}, \mathbf{t}_{2\mathrm{d}}, s] = s\mathbf{PRv} + s\mathbf{t}_{2\mathrm{d}} \tag{2}$$

where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

is a projection matrix and the pose parameters $\mathbf{R} \in SO(3)$, $\mathbf{t}_{2d} \in \mathbb{R}^2$ and $s \in \mathbb{R}^+$ are a rotation matrix, 2D translation and scale respectively. In order to constrain optimisation to valid rotation matrices, we parameterise the rotation matrix by an axis-angle vector $\mathbf{R}(\mathbf{r})$ with $\mathbf{r} \in \mathbb{R}^3$. The conversion from an axis-angle representation to a rotation matrix is given by:

$$\mathbf{R}(\mathbf{r}) = \cos\theta \mathbf{I} + \sin\theta \left[\bar{\mathbf{r}}\right]_\times + (1 - \cos\theta)\bar{\mathbf{r}}\bar{\mathbf{r}}^\mathrm{T}, \tag{3}$$

where $\theta = \|\mathbf{r}\|$ and $\bar{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|$ and

$$\left[\mathbf{a}\right]_\times = \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}$$

is the cross product matrix.

3.3 Perspective Camera Model

The nonlinear perspective projection of the 3D point $\mathbf{v} = [u, v, w]^\mathrm{T}$ onto the 2D point $\mathbf{x} = [x, y]^\mathrm{T}$ is given by the pinhole camera model $\mathbf{x} = \mathbf{pinhole}[\mathbf{v}, \mathbf{K}, \mathbf{R}, \mathbf{t}_{3d}] \in \mathbb{R}^2$ where $\mathbf{R} \in SO(3)$ is a rotation matrix and $\mathbf{t}_{3d} = [t_x, t_y, t_z]^\mathrm{T}$ is a 3D translation vector which relate model and camera coordinates (the extrinsic parameters). The matrix:

$$\mathbf{K} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

contains the intrinsic parameters of the camera, namely the focal length f and the principal point (c_x, c_y). We assume that the principal point is known (often the centre of the image is an adequate estimate) and parameterise the intrinsic matrix by its only unknown $\mathbf{K}(f)$. Note that varying the focal length amounts only to a uniform scaling of the projected points in 2D. This corresponds exactly to the scenario in Fig. 1. There, subject-camera distance was varied before rescaling each image such that the interocular distance was constant, effectively simulating a lack of calibration information. This nonlinear projection can be written in linear terms by using homogeneous representations $\tilde{\mathbf{v}} = [u, v, w, 1]^\mathrm{T}$ and $\tilde{\mathbf{x}} = [x, y, 1]^\mathrm{T}$:

$$\gamma\tilde{\mathbf{x}} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t}_{3d} \end{bmatrix} \tilde{\mathbf{v}}, \tag{4}$$

where γ is an arbitrary scaling factor.

4 Shape-from-Landmarks

In this section, we describe a novel method for fitting a 3DMM to a set of 2D landmarks. Here, "landmarks" can be interpreted quite broadly. It simply means a point for which both the 2D position and the corresponding vertex in the morphable model are known. Later, we will relax this requirement by showing how to establish these correspondences for points on the occluding boundary that do not have clear semantic meaning in the way that a typical landmark does.

We assume that L 2D landmark positions $\mathbf{x}_i = [x_i, y_i]^\mathrm{T}$ ($i = 1 \ldots L$) have been observed. Without loss of generality, we assume that the ith landmark corresponds to the ith vertex in the morphable model.

The objective is to find the shape, pose and camera parameters that, when projected to 2D, minimise the sum of squared distances over all landmarks. We introduce objective functions for the orthographic and perspective cases and then show how they can be expressed as separable nonlinear least squares problems. Figure 2 provides an overview of estimating shape from geometric information.

4.1 Orthographic Objective Function

In the orthographic case, we seek to minimise the following objective function:

$$\begin{aligned} \varepsilon_{\mathrm{ortho}}(\mathbf{r}, \mathbf{t}_{2d}, s, \boldsymbol{\alpha}) \\ = \mathbf{d}_{\mathrm{ortho}}(\mathbf{r}, \mathbf{t}_{2d}, s, \boldsymbol{\alpha})^\mathrm{T} \mathbf{d}_{\mathrm{ortho}}(\mathbf{r}, \mathbf{t}_{2d}, s, \boldsymbol{\alpha}), \end{aligned} \tag{5}$$
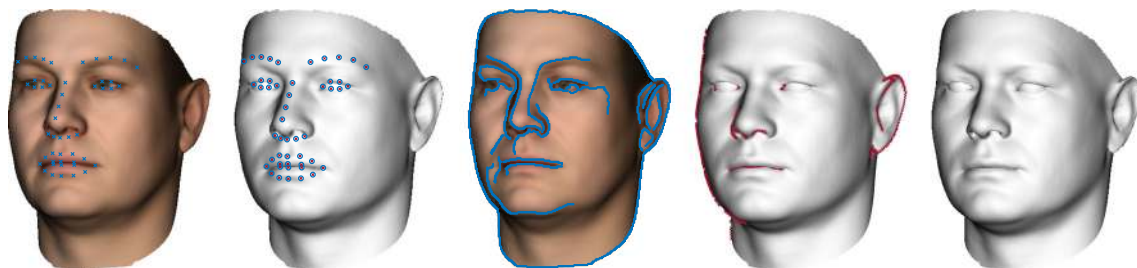

Fig. 2 Overview of estimating shape from geometric information. From left to right: input image with landmarks; shape-from-landmarks (Sect. 4) with image landmarks shown as red crosses and projected model landmarks shown as blue circles; input image with edge pixels shown in blue; shape-from-contours (Sect. 5) with occluding boundary vertices labelled with red crosses; final reconstruction (Color figure online)

where the vector of residuals $\mathbf{d}_{\text{ortho}}(\mathbf{r}, \mathbf{t}_{2d}, s, \boldsymbol{\alpha}) \in \mathbb{R}^{2L}$ is given by:

$$
\begin{aligned}
&\mathbf{d}_{\text{ortho}}(\mathbf{r}, \mathbf{t}_{2d}, s, \boldsymbol{\alpha}) \\
&= \begin{bmatrix} \mathbf{x}_1 - \mathbf{SOP}\left[\mathbf{Q}_1\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_1, \mathbf{R}(\mathbf{r}), \mathbf{t}_{2d}, s\right] \\ \vdots \\ \mathbf{x}_L - \mathbf{SOP}\left[\mathbf{Q}_L\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_L, \mathbf{R}(\mathbf{r}), \mathbf{t}_{2d}, s\right] \end{bmatrix}.
\end{aligned} \tag{6}
$$

These residuals are linear in the shape parameters, translation vector and scale but nonlinear in the rotation vector. Previous work has treated this as a multilinear optimisation problem and used alternating coordinate descent. Instead, we observe that the problem can be treated as linear in the shape and translation parameters simultaneously and nonlinear in scale and rotation.

4.2 Perspective Objective Function

In the perspective case, we seek to minimise the following objective function:

$$
\begin{aligned}
&\varepsilon_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3d}, f, \boldsymbol{\alpha}) \\
&= \mathbf{d}_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3d}, f, \boldsymbol{\alpha})^{\mathrm{T}} \mathbf{d}_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3d}, f, \boldsymbol{\alpha}),
\end{aligned} \tag{7}
$$

where the vector of residuals $\mathbf{d}_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3d}, f, \boldsymbol{\alpha}) \in \mathbb{R}^{2L}$ is given by:

$$
\begin{aligned}
&\mathbf{d}_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3d}, f, \boldsymbol{\alpha}) \\
&= \begin{bmatrix} \mathbf{x}_1 - \mathbf{pinhole}\left[\mathbf{Q}_1\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_1, \mathbf{K}(f), \mathbf{R}(\mathbf{r}), \mathbf{t}_{3d}\right] \\ \vdots \\ \mathbf{x}_L - \mathbf{pinhole}\left[\mathbf{Q}_L\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_L, \mathbf{K}(f), \mathbf{R}(\mathbf{r}), \mathbf{t}_{3d}\right] \end{bmatrix}.
\end{aligned} \tag{8}
$$

These residuals are nonlinear in all parameters and nonconvex due to the perspective projection. However, we can use the direct linear transformation (DLT) (Hartley and Zisserman 2003) to transform the problem to a linear one. The solution of this easier problem provides a good initialisation for nonlinear optimisation of the true objective.

From (1) and (4) we have a linear similarity relation for each landmark point:

$$
\begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} \sim \mathbf{K}\begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}\begin{bmatrix} \mathbf{Q}_i\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_i \\ 1 \end{bmatrix}, \tag{9}
$$

where $\sim$ denotes equality up to a non-zero scalar multiplication. We rewrite as a collinearity condition:

$$
\begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}_{\times} \mathbf{K}\begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix}\begin{bmatrix} \mathbf{Q}_i\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_i \\ 1 \end{bmatrix} = \mathbf{0} \tag{10}
$$

where $\mathbf{0} = [0\ 0\ 0]^{\mathrm{T}}$. This means that each landmark yields three equations that are linear in the unknown shape parameters $\boldsymbol{\alpha}$ and the translation vector $\mathbf{t}_{3d}$.

4.3 Separable Nonlinear Least Squares

We now show that both objective functions can be written in a separable nonlinear least squares (SNLS) form, i.e. a form that is linear in some of the parameters (including shape) and nonlinear in the remainder. This special form of least squares problem can be solved more efficiently than general least squares problems and may converge when the original problem would diverge (Golub and Pereyra 2003). SNLS problems are solved by optimising a nonlinear least squares problem only in the nonlinear parameters, hence the problem dimensionality is reduced and the number of parameters that require initial guesses reduced. For convenience, henceforth we denote by $\mathbf{Q}_L \in \mathbb{R}^{3L \times S}$ the submatrix of $\mathbf{Q}$ containing the rows corresponding to the L landmarks (i.e. the first $3L$ rows of $\mathbf{Q}$).

4.3.1 Orthographic

The vector of residuals (6) in the orthographic objective function (5) can be written in SNLS form as

$$
\mathbf{d}_{\text{ortho}}(\mathbf{r}, \mathbf{t}_{2d}, s, \boldsymbol{\alpha}) = \mathbf{A}(\mathbf{r}, s)\begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{t}_{2d} \end{bmatrix} - \mathbf{y}(\mathbf{r}, s) \tag{11}
$$

where $\mathbf{A}(\mathbf{r}, s) \in \mathbb{R}^{2L \times S+2}$ is given by

$$
\mathbf{A}(\mathbf{r}, s) = s\left[(\mathbf{I}_L \otimes [\mathbf{PR}(\mathbf{r})])\,\mathbf{Q}_L\ \mathbf{1}_L \otimes \mathbf{I}_2\right], \tag{12}
$$

and $\mathbf{y}(\mathbf{r}, s) \in \mathbb{R}^{2L}$ is given by

$$
\mathbf{y}(\mathbf{r}, s) = s\,(\mathbf{I}_L \otimes [\mathbf{PR}(\mathbf{r})])\,\bar{\mathbf{s}} - [x_1, y_1, \ldots, y_L]^{\mathrm{T}}, \tag{13}
$$

where $\mathbf{I}_L$ is the $L \times L$ identity matrix and $\mathbf{1}_L$ is the length L vector of ones.

Note that the vector of residuals in (11) is exactly equivalent to the original one in (6). The optimal solution to the original objective function (5) in terms of the linear parameters is given by:

$$
\begin{bmatrix} \boldsymbol{\alpha}^* \\ \mathbf{t}_{2d}^* \end{bmatrix} = \mathbf{A}^+(\mathbf{r}, s)\mathbf{y}(\mathbf{r}, s) \tag{14}
$$

where $\mathbf{A}^+(\mathbf{r}, s)$ is the pseudoinverse. Substituting (14) into (11) we get a vector of residuals that is exactly equivalent to (6) but which depends only on the nonlinear parameters:

$$
\mathbf{d}_{\text{ortho}}(\mathbf{r}, s) = \mathbf{A}(\mathbf{r}, s)\mathbf{A}^+(\mathbf{r}, s)\mathbf{y}(\mathbf{r}, s) - \mathbf{y}(\mathbf{r}, s). \tag{15}
$$

Substituting this into (5), we get an equivalent objective function, $\varepsilon_{\mathrm{ortho}}(\mathbf{r}, s)$, again depending only on the nonlinear parameters. This is a nonlinear least squares problem of very low dimensionality ([$\mathbf{r}$ s] is only 4D). We solve this using the trust-region-reflective algorithm for which we require $\mathbf{J}_{\mathbf{d}_{\mathrm{ortho}}}(\mathbf{r}, s) \in \mathbb{R}^{2L \times 4}$, the Jacobian of the residual function. In Appendix A, we analytically derive $\mathbf{J}_{\mathbf{d}_{\mathrm{ortho}}}$. Although computing these derivatives is quite involved, in practice it is still faster than using finite difference approximations. Once optimal parameters have been obtained by minimising $\varepsilon_{\mathrm{ortho}}(\mathbf{r}, s)$ then the parameters $\boldsymbol{\alpha}^*$ and $\mathbf{t}^*$ are obtained by (14).

If we wish to impose a statistical prior on the shape parameters we can use Tikhonov regularisation, as in (Blanz et al. 2004), during the solution of (14).

4.3.2 Perspective

The perspective residual function (8), linearised via (10), can be written in SNLS form as

$$\mathbf{d}_{\mathrm{persp}}^{\mathrm{DLT}}(\mathbf{r}, \mathbf{t}_{3\mathrm{d}}, f, \boldsymbol{\alpha}) = \mathbf{B}(\mathbf{r}, f) \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{t}_{3\mathrm{d}} \end{bmatrix} - \mathbf{z}(\mathbf{r}, f) \quad (16)$$

where $\mathbf{B}(\mathbf{r}, f) \in \mathbb{R}^{3L \times S+3}$ is given by:

$$\mathbf{B}(\mathbf{r}, f) = \mathbf{D}\mathbf{E}(f)\mathbf{F}(\mathbf{r}), \quad (17)$$

with

$$\mathbf{D} = \mathrm{diag}\left(\begin{bmatrix} \mathbf{x}_1 \\ 1 \end{bmatrix}_\times, \ldots, \begin{bmatrix} \mathbf{x}_L \\ 1 \end{bmatrix}_\times \right), \quad \mathbf{E}(f) = \mathbf{I}_L \otimes \mathbf{K}(f)$$

and

$$\mathbf{F}(\mathbf{r}) = \begin{bmatrix} (\mathbf{I}_L \otimes \mathbf{R}(\mathbf{r})) \, \mathbf{Q}_L & \mathbf{1}_L \otimes \mathbf{I}_3 \end{bmatrix}.$$

The vector $\mathbf{z}(\mathbf{r}, f) \in \mathbb{R}^{3L}$ is given by:

$$\mathbf{z}(\mathbf{r}, f) = -\mathbf{D}\left(\mathbf{I}_L \otimes [\mathbf{K}(f)\mathbf{R}(\mathbf{r})]\right) \bar{\mathbf{s}}$$

Exactly as in the orthographic case, we can write optimal solutions for the linear parameters in terms of the nonlinear parameters and solve a 4D nonlinear minimisation problem in $(\mathbf{r}, f)$. In contrast to the orthographic case, this objective is not equivalent to minimisation of the original objective, i.e. the sum of squared perspective reprojection distances in (7). So, we use the SNLS solution to initialise a nonlinear least squares optimisation of the original objective over all parameters, again using trust-region-reflective. In practice, we find that the SNLS solution is already very close to the optimum and that the subsequent nonlinear least squares optimisation usually converges in 2-5 iterations, shown in Fig. 3b.
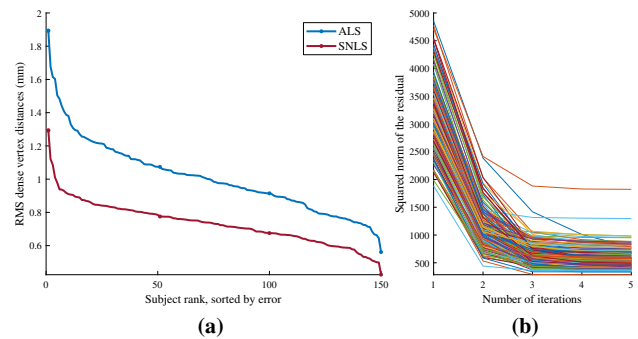

Fig. 3 **a** Quantitative comparison between alternating linear least squares (ALS) and separable nonlinear least squares (SNLS) on 150 subjects in the Facewarehouse dataset. The average dense surface error is 1.01 mm for ALS and 0.73 mm for SNLS. **b** Convergence rates of nonlinear least squares optimisation (Color figure online)

4.4 Perspective Ambiguities

Solving the optimisation problems above yields a least squares estimate of the pose and shape of a face, given 2D landmark positions. In Sect. 6, we show that for both orthographic and perspective cases, with pose fixed there remain degrees of flexibility that allow the 3D shape to vary without significantly increasing the objective value. However, for the perspective case there is an additional degree of freedom related to the subject-camera distance, i.e. t_z. If, instead of allowing t_z to be optimised along with other parameters, we fix it to some chosen value k, then we can obtain different shape and pose parameters:

$$\boldsymbol{\alpha}^*(k) = \arg_{\boldsymbol{\alpha}} \min_{\mathbf{r}, \mathbf{t}_{3\mathrm{d}}, f, \boldsymbol{\alpha}} \varepsilon_{\mathrm{persp}}(\mathbf{r}, \mathbf{t}_{3\mathrm{d}}, f, \boldsymbol{\alpha}), \quad \text{s.t. } t_z = k.$$

Given 2D landmark observations, we therefore have a continuous (nonlinear) space of solutions $\boldsymbol{\alpha}^*(k)$ as a function of subject-camera distance. This is the perspective face shape ambiguity. If the mean reprojection error with a value of k other than the optimal one is still smaller than the tolerance of our landmark detector, then shape recovery is ambiguous.

5 Shape-from-Contours

In order to extend the method in the previous section to also exploit contour information, we follow Bas et al. (2016) and use an iterated closest edge fitting strategy. We assume that manually provided or automatically detected landmarks are available and we initialise by fitting to these using the method in the previous section. Next, we alternate between establishing correspondences and refitting as follows:

1. Compute occluding boundary vertices for current shape and pose estimate and project to 2D.

2. Correspondence is found between edges detected in the image and the projection of model vertices that lie on the occluding boundary. This is done in a nearest neighbour fashion with some filtering for robustness.

3. With the correspondences to hand, edge vertices can be treated like landmarks with known correspondence and the method from the previous section applied to refit the model (initialising with the nonlinear parameters obtained in the previous iteration and retaining the original landmarks).

These three steps are iterated to convergence.

In detail, we begin by labelling a subset of pixels as edges, stored in the set $\mathcal{E} = \{(x, y)|(x, y) \text{ is an edge}\}$. In practice, we compute edges by applying the Canny edge detector with a fixed threshold to the input image. More robust performance would be obtained by using a problem-specific edge detector such as boosted edge learning. This was recently done for fitting a morphable tooth model to contours in uncontrolled images (Wu et al. 2016).

Model contours are computed based on the pose and shape parameters as the occluding boundary of the 3D face. The set of occluding boundary vertices, $\mathcal{B}(\boldsymbol{\alpha}, \mathbf{r}, \mathbf{t}, s)$ (for the orthographic case), are defined as those lying on a mesh edge whose adjacent faces have a change of visibility. This definition encompasses both outer (silhouette) and inner (self-occluding) contours. In addition, we check that potential edge vertices are not occluded by another part of the mesh (using z-buffering) and we ignore edges that lie on a mesh boundary since they introduce artificial edges. In this paper, we deal only with occluding contours (both inner and outer). If texture contours were defined on the surface of the morphable model, it would be straightforward to include these in our approach.

We find the set of edge/contour pairs, $\mathcal{N}$, that are mutual nearest neighbours in a Euclidean distance sense in 2D, i.e. $(i^*, (x^*, y^*)) \in \mathcal{N}$ if:

$$(x^*, y^*) \\ = \underset{(x,y)\in\mathcal{E}}{\arg\min} \, \|[x \, y]^{\mathrm{T}} - \mathbf{SOP}\left[\mathbf{Q}_{i*}\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_{i*}, \mathbf{R}(\mathbf{r}), \mathbf{t}_{2\mathrm{d}}, s\right]\|^2$$

and

$$i^* \\ = \underset{i\in\mathcal{B}(\boldsymbol{\alpha},\mathbf{r},\mathbf{t},s)}{\arg\min} \, \|[x^* \, y^*]^{\mathrm{T}} - \mathbf{SOP}\left[\mathbf{Q}_{i}\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}_{i}, \mathbf{R}(\mathbf{r}), \mathbf{t}_{2\mathrm{d}}, s\right]\|^2.$$

Using mutual nearest neighbours makes the method robust to contours that are partially missed by the edge detector. The perspective case is identical except that the pinhole projection model is used. The correspondence set can be further filtered by excluding some proportion of pairs whose distance is largest or pairs whose distance exceeds a threshold.

6 Flexibility Modes

We now assume that a least squares model fit has been obtained using the method in Sect. 4 (and optionally Sect. 5). This amounts to a shape, $\mathbf{Q}\boldsymbol{\alpha} + \bar{\boldsymbol{\varsigma}}$, determined by the estimated shape parameters and a pose $(\mathbf{r}, s, \mathbf{t}_{2\mathrm{d}})$ or $(\mathbf{r}, f, \mathbf{t}_{3\mathrm{d}})$ for orthographic or perspective respectively. We now show that there are remaining modes of flexibility in the model fit. Keeping pose parameters fixed, we wish to find perturbations to the shape parameters that change the projected 2D geometry as little as possible (i.e. minimising the increase in the reprojection error of landmark vertices) while changing the 3D shape as much as possible.

Our approach to computing these flexibility modes is an extension of the method of Albrecht et al. (2008). They considered the problem of flexibility only in a 3D setting where the model is partitioned into a disjoint fixed part and a flexible part. We extend this so that the constraint on the fixed part acts in 2D after orthographic or perspective projection while the flexible part is the 3D shape of the whole face.

In the orthographic case, we define the 2D projection of the principal component directions for the L landmark vertices as:

$$\boldsymbol{\Pi}_{\mathrm{ortho}} = (\mathbf{I}_L \otimes (\mathbf{PR}(\mathbf{r}))) \, \mathbf{Q}_L, \qquad (18)$$

where $\mathbf{r}$ is the rotation vector that was estimated during fitting. Intuitively, we seek modes that move the landmark vertices primarily along the projection axis, which depends only on the rotation, and therefore do not move their 2D projection much. Hence, the flexibility modes do not depend on the scale or translation of the fit or even the landmark positions. For the perspective case, we again use the DLT linearisation in (10), leading to the following expression:

$$\boldsymbol{\Pi}_{\mathrm{persp}} = \mathbf{D} \left(\mathbf{I}_L \otimes \left(\mathbf{K}(f) \left[\mathbf{R}(\mathbf{r}) \, \mathbf{t}_{3\mathrm{d}} \right] \mathbf{S} \right) \right) \mathbf{Q}_L, \qquad (19)$$

where

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Again, $\mathbf{r}$, f and $\mathbf{t}_{3\mathrm{d}}$ are the rotation vector, focal length and translation that were estimated during fitting. By using the DLT linearisation, the intuition here is that we want the camera rays to the landmark vertices to remain as parallel as possible with the homogeneous vectors representing the observed landmarks.

Concretely, we seek flexibility modes, $\mathbf{f} \in \mathbb{R}^S$, such that $\mathbf{Qf}$ changes as much as possible whilst the 2D projection of the landmarks, given by $\boldsymbol{\Pi}_{\text{ortho}}\mathbf{f}$ or $\boldsymbol{\Pi}_{\text{persp}}\mathbf{f}$, changes as little as possible. This can be formulated as a constrained maximisation problem:

$$\max_{\mathbf{f} \in \mathbb{R}^S} \|\mathbf{Qf}\|^2 \text{ subject to } \|\boldsymbol{\Pi}\mathbf{f}\|^2 = c, \tag{20}$$

where $\boldsymbol{\Pi}$ is one of the projection matrices and $c \in \mathbb{R}^+$ controls how much variation in the 2D projection is allowed (this value is arbitrary since it does not appear in the subsequent flexibility mode computation). Introducing a Lagrange multiplier and differentiating with respect to $\mathbf{f}$ yields:

$$\mathbf{Q}^{\text{T}}\mathbf{Qf} = \lambda \boldsymbol{\Pi}^{\text{T}}\boldsymbol{\Pi}\mathbf{f}. \tag{21}$$

This is a generalised eigenvalue problem whose solution is a set of flexibility modes $\mathbf{f}_1, \ldots, \mathbf{f}_S$ along with their corresponding generalised eigenvalue $\lambda_1, \ldots, \lambda_S$, sorted in descending order. Therefore, $\mathbf{f}_1$ is the flexibility mode that changes the 3D shape as much as possible while minimising the change to the projected 2D geometry. If a face was fitted with shape parameters $\boldsymbol{\alpha}$ then its shape is varied by adjusting the weight w in: $\mathbf{Q}(\boldsymbol{\alpha} + w\mathbf{f}) + \bar{\boldsymbol{\varsigma}}$.

We can truncate the number of flexibility modes by setting a threshold k_1 on the mean Euclidean distance by which the surface should change and testing whether the corresponding change in mean landmark error is less than a threshold k_2. We retain only those flexibility modes where this is the case.

7 Experimental Results

We now present experimental results to demonstrate the ambiguities that arise in estimating 3D face shape from 2D geometry. We make use of the Basel Face Model (Paysan et al. 2009) (BFM) which is a 3DMM comprising 53,490 vertices and which is trained on 200 faces. We use the shape component of the model only. The model is supplied with 10 out-of-sample faces which are scans of real faces that are in correspondence with the model. We use these for quantitative evaluation on synthetic data. Unusually, the model does not factor out scale, i.e. faces are only aligned via translation and rotation. This means that the vertex positions are in absolute units of distance. This allows us to specify camera-subject distance in physically meaningful units. For all fittings we use Tikhonov regularisation with a low weight. For sparse (landmark) fitting, where overfitting is more likely, we use $S = 70$ dimensions and constrain parameters to be within $k = 2$ standard deviations of the mean. For dense fitting, we use all $S = 199$ model dimensions and constrain parameters to be $k = 3$ standard deviations of the mean.

We make use of two quantitative error measures in our evaluation. For data with ground truth 3D, d_S is the mean Euclidean distance between the ground truth and reconstructed surface after aligning with Procrustes analysis. d_L is the mean distance between observed landmarks and the corresponding projection of the reconstructed landmark vertices, expressed as a percentage of the interocular distance.

7.1 SNLS Fitting

In Sect. 4.3 we introduced a novel formulation of 3DMM fitting under orthographic and perspective projection using SNLS. Although our goal in this paper is to investigate ambiguities in the 3D interpretation of 2D geometry and not to advance the state of the art in 3DMM fitting, we nevertheless begin by demonstrating that our SNLS formulation is indeed superior to alternating least squares (ALS) as used in previous work (Bas et al. 2016; Zhu et al. 2015; Aldrian and Smith 2013; Cao et al. 2013, 2014a; Saito et al. 2016). In order to evaluate in a realistic setting, we require images with corresponding ground truth 3DMM fits. For this reason, we use the Facewarehouse dataset and model (Cao et al. 2014b). We use leave-one-out testing, building each model on 149 subjects and testing on the remaining one and use the 74 landmarks provided with the dataset. For this evaluation we test only the orthographic setting. Figure 3a shows the mean Euclidean distance between dense ground truth and estimated face surface in mm after Procrustes alignment. We do not use any regularisation for either algorithm and therefore do not need to choose the weight parameter. For all subjects SNLS achieves a lower error, on average reducing it by about 30%.

As a second experiment, we provide a quantitative fitting comparison on synthetic face images in various poses (rotations of $0°$, $\pm15°$ and $\pm30°$ about the vertical axis) which are rendered in orthographic projection from the out-of-sample faces supplied by the BFM. We use the algorithm of Zhu et al. (2015) with the landmarks detected by the automatic method of Zhu and Ramanan (2012). The fitting method and the landmark detector are both publicly available. Table 2 reports the mean Euclidean distance between ground truth and estimated face surface in mm after Procrustes alignment. This shows that our SLNS optimisation provides better overall performance and superior results for all poses.

7.2 Perspective Ambiguity

We begin by investigating the perspective ambiguity using synthetic data. We use the out-of-sample BFM scans to create input data by choosing pose parameters and projecting the faces to 2D. For sparse landmarks, we use the 70 anthropometric landmarks [due to (Farkas 1994)] whose indices in the BFM are known. These landmarks are particularly appropriate as they were chosen so as to best measure the variability
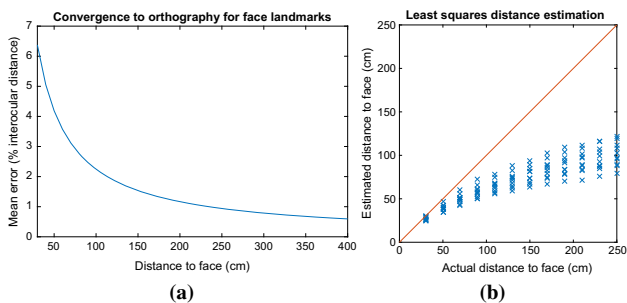
Fig. 4 **a** Mean landmark error (y axis) between perspective and orthographic projection, averaged over 10 BFM scans, as subject-camera distance (x axis) is varied. **b** Subject-camera distance estimation by least squares optimisation (Color figure online)

Table 2 Quantitative comparison between Zhu et al. (2015) and SNLS on synthetic data with automatically detected landmarks

Method	Rotation angle					Mean
	−30°	−15°	0°	15°	30°	
Zhu et al. (2015)	4.63	5.09	4.19	5.22	4.92	4.81
SLNS (ours)	4.53	4.29	4.16	3.99	4.07	4.21

Each cell shows the mean euclidean vertex distance for related pose in mm

Table 3 Quantitative results for the perspective ambiguity on synthetic data

Actual distance (cm)	Fitting distance (cm)				
	30	60	120	240	Ortho
30	0.21	0.24	0.26	0.27	0.28
	7.23	9.70	13.07	14.55	14.47
60	0.30	0.26	0.27	0.27	0.28
	8.07	6.29	6.60	6.99	7.48
120	0.37	0.29	0.28	0.28	0.28
	9.52	6.17	5.38	5.39	5.62
240	0.42	0.32	0.29	0.29	0.28
	10.16	6.72	5.59	5.37	5.38
Ortho	0.47	0.35	0.31	0.30	0.29
	11.02	7.43	6.01	5.54	5.29

Each cell shows the landmark error, d_L in %, top and surface error, d_S in mm, bottom

in craniofacial shape over a population. In Fig. 4a, we show over what range of distances perspective transformation has a significant effect on 2D face geometry. For each face, we project the 70 landmarks to 2D under perspective projection and measure d_L with respect to the orthographic projection of the landmarks. As t_z increases, the projection converges towards orthography and the error tends to zero. The landmark error falls below 1% when the distance is around 2.5 m. Hence, we experiment with distances ranging from selfie distance (30 cm) up to this distance.

Our first evaluation of the perspective ambiguity is based on estimating the subject-camera distance as one of the parameters in the least squares fitting process. We use the out-of-sample BFM scans as target faces, vary the subject-camera distance and project the 70 Farkas landmarks to 2D under perspective projection. We use a frontal pose ($\mathbf{r} = [0\ 0\ 0]$) and arbitrarily set the focal length to $f = 1$. We initialise the optimisation with the correct focal length and rotation, giving it the best possible chance of estimating the correct distance. We plot estimated versus ground truth distance in Fig. 4b. Optimal performance would see all points falling on the diagonal red line. The distance is consistently underestimated and the mean percentage error in the estimate is 42%. It is clear that the 2D landmarks alone do not contain enough information to accurately estimate subject-camera distance as part of the model fitting process.

We now show that landmarks produced by a real 3D face shape at one distance can be explained by 3D shapes at mul-

tiple different distances. In Table 3 we show quantitative results. Each row of the table corresponds to a distance at which we place each of the BFM scans in a frontal pose before projecting to 2D. We then fit to these landmarks with the subject-camera distance assumed to be the value shown in the column. The results show that we are able to explain the data almost as well at the wrong distance as the correct one but the 3D shape is very different, differing by over a 1 cm on average. Note that Burgos-Artizzu et al. (2014) found that the difference between landmarks on the same face placed by two different humans was typically 3% of the interocular distance. Similarly, the 300 faces in the wild challenge (Sagonas et al. 2016) found that even the best methods did not obtain better than 5% accuracy for more than 50% of the landmarks. Hence, the difference between target and fitted landmarks is substantially smaller than the accuracy of either human or machine placed landmarks. Importantly, this means that the fitting energy could not be used to resolve the ambiguity. The residual difference between target and fitted landmarks is too small to meaningfully choose between the two solutions.

We now show qualitative examples from the same experiment. In Fig. 5 we show orthographic renderings of perspective fits to the face shown in the first column. In the first row, the target landmarks were generated by viewing the face at 30 cm, in the second row the face was at 120 cm. In each column we show fitting results at different distances. In the final column we show the landmarks of the real face (circles) overlaid with the landmarks from the fitted faces (dots) showing that highly varying 3D faces can produce almost identical 2D landmarks.

In Figs. 6 and 7 we go further by showing the results of fitting to sparse 2D landmarks (the Farkas feature points), landmarks/edges and all vertices for 4 of the BFM scans (i.e.
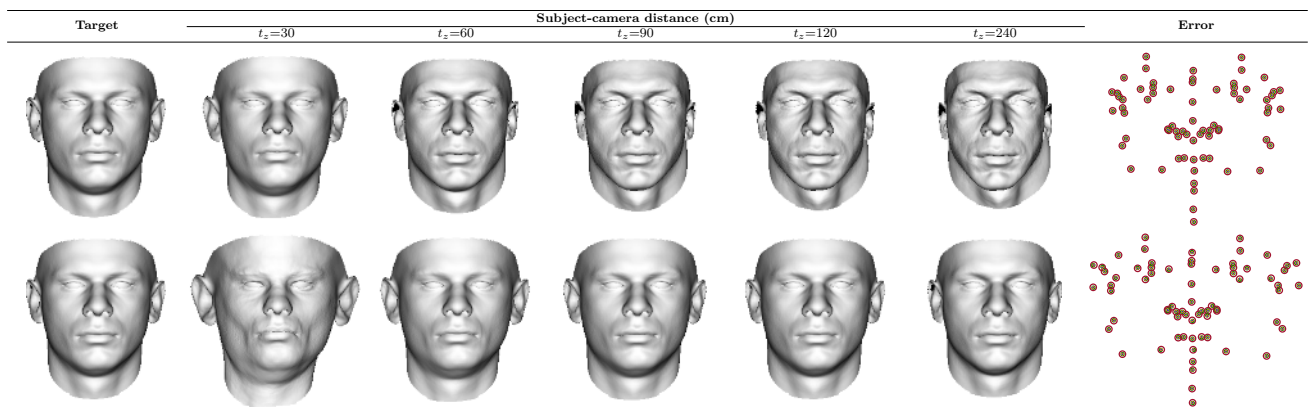
Fig. 5 Qualitative perspective face shape ambiguity. There is a subspace of possible 3D face shapes with varying subject-camera distance within the landmark tolerance. Target face is at 30 cm (first row) and 120 cm (second row) (Color figure online)
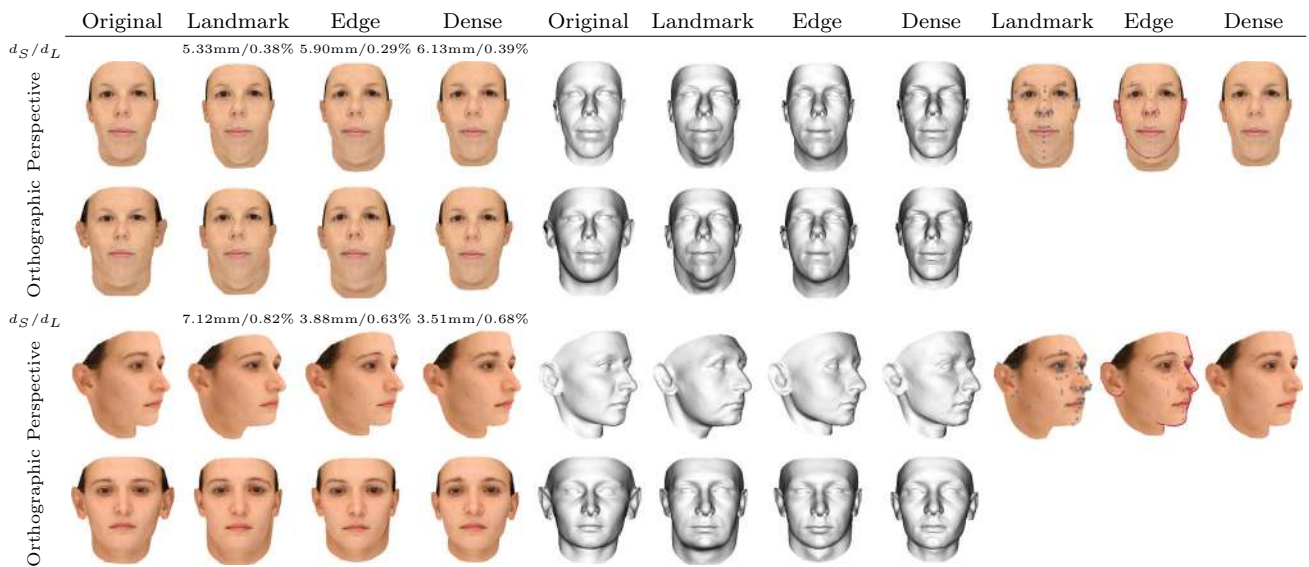

Fig. 6 Sparse and dense fitting of the synthetic images. Target at 30 cm, fitted results at 120 cm (Color figure online)

the targets are real faces). In Fig. 6, the target face is close to the camera ($t_z = 30$ cm) and we fit the model at a far distance ($t_z = 120$ cm). This configuration is reversed in Fig. 7 (200 cm to 60 cm). Since we are only interested in the spatial configuration of features in the image, we show both target and fitted mesh with the texture of the real target face. The target perspective projection to which we fit is shown in the first and fifth columns. The fitting result under perspective projection is shown in the second to fourth columns and sixth to eight columns. To enable comparison between the target and fitted faces, we render them under orthographic projection in rows two and four respectively. The landmarks from the target (plotted as blue circles) and fitted (shown as red dots) face are shown under perspective projection in the column nine. We illustrate edge correspondence (model contours) between faces in the tenth column. In the last column, we average the target and fitted face texture from the dense

fitting result, showing that there is no visible difference in the 2D geometry of these two images.

The implication of these results is that, in a sample of real faces, we might expect that two different identities with different face shapes could give rise to approximately the same 2D landmarks when viewed from different distances. We show in Fig. 8 that this is indeed the case. The Caltech Multi-Distance Portraits dataset (Burgos-Artizzu et al. 2014) contains images of 53 subjects viewed at 7 different distances. 55 landmarks are placed manually on each face image. We search for pairs of faces whose landmarks (when viewed at different distances) are close in a Procrustes sense. Despite the small sample size, we find a pair of faces whose mean landmark error is 2.48% [i.e. they are within the expected accuracy of a landmark detector (Sagonas et al. 2016)] when they are viewed at 61 cm and 488 cm respectively (second and fourth image in the figure). In the third image, we blend these
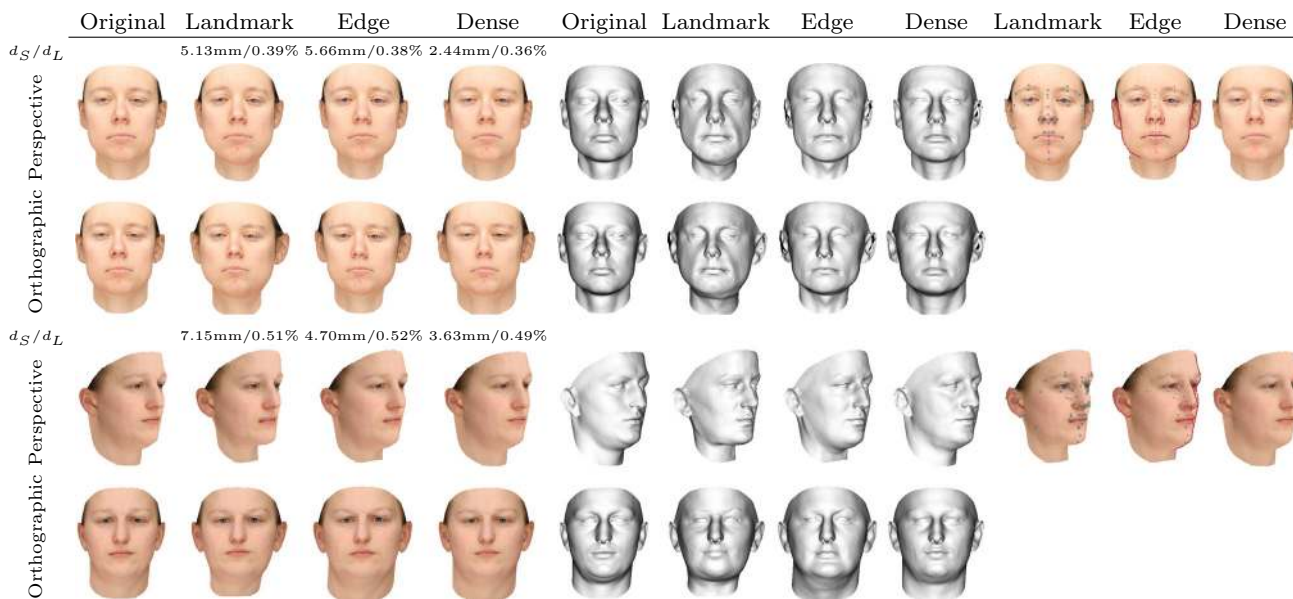
	Original	Landmark	Edge	Dense	Original	Landmark	Edge	Dense	Landmark	Edge	Dense
d_S/d_L		5.13mm/0.39%	5.66mm/0.38%	2.44mm/0.36%							

	Original	Landmark	Edge	Dense	Original	Landmark	Edge	Dense	Landmark	Edge	Dense
d_S/d_L		7.15mm/0.51%	4.70mm/0.52%	3.63mm/0.49%							

Fig. 7 Sparse and dense fitting of the synthetic images. Target at 200 cm, fitted results at 60 cm (Color figure online)

Fig. 8 Perspective ambiguity in real faces. Two faces are shown at two different distances. The blend in the middle shows that their 2D geometry is similar when viewed at very different distances (Color figure online)

two images to show that their 2D features indeed align well. To highlight that their face shape is in fact quite different, we show their appearance with distances reversed in columns one and five (allowing direct comparison between columns one and four or two and five). E.g. compare column one with column four. The face in column one has larger ears and inner features that are more concentrated towards to the centre of the face compared to the face in column four.

The CMDP data can also be used to demonstrate a surprising conclusion. For all 53 subjects, we compute the mean landmark error between the same identity at 61 cm and 488 cm which is 3.11%. Next, for each identity we find the identity at the same distance with the smallest landmark error. Averaged over all identities, this gives a value of 2.86% for 61 cm and 2.83% for 488 cm. We therefore conclude that 2D geometry between different identities at the same distance is more similar than between the same identity at different distances. If the number of identities was increased, the size of this effect would likely increase since the chance of finding closely matching different identity pairs would increase.

7.3 Beyond Geometric Cues

The fitting methods we propose in this paper use only explicit geometric cues, i.e. landmarks and contours. State-of-the-art CNN-based methods can exploit any 3D shape cues such as shading, texture, shadows or context from external face features such as hair or clothes or even from background objects. One might suppose that these additional cues resolve the ambiguity we describe. However, we now show that this is not the case. We used the pre-trained network of Tran et al. (2017) which is publicly available. This network is trained discriminatively to regress the same 3DMM parameters from different images of the same person. If the training set contained distance variation, then it would be hoped that the network would learn invariance to perspective ambiguities. We ran the network on images of 53 subjects viewed at closest and farthest distances from the CMDP dataset (Burgos-Artizzu et al. 2014). We begin by evaluating the invariance of the shape reconstructions to changes in distance by measuring the mean Euclidean distance after Procrustes alignment between all pairs of 3D reconstructions. This is a standard metric for comparing 3D face reconstructions, e.g. Sanyal et al. (2019); Feng et al. (2018). These comparisons provides a 106×106 distance matrix. One would expect that the shape difference of the same subject viewed at two different distances would be the lowest. However, for the majority of identities, this is not the case. In Fig. 9a we show the distance matrix (same identity in consecutive positions) and in Fig. 9b we binarise this by choosing the best matching shape for each row. Perfect performance would yield 2×2 blocks along the diagonal. We show two examples from this experiment in Fig. 10. These results show that Tran et al.
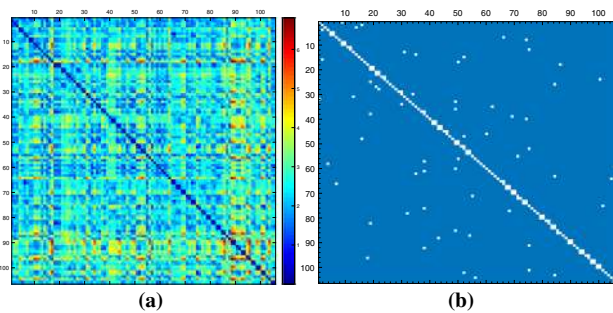
Fig. 9 **a** Heat map and **b** binarised distance matrix visualising similarity between subjects viewed at two different (closest and farthest) distances. We measured the distances between 3D surfaces acquired by running pre-trained Tran et al. (2017) on real images from the CMDP dataset. One would expect 2×2 blocks of white on the diagonal if the network is performing perfectly (Color figure online)
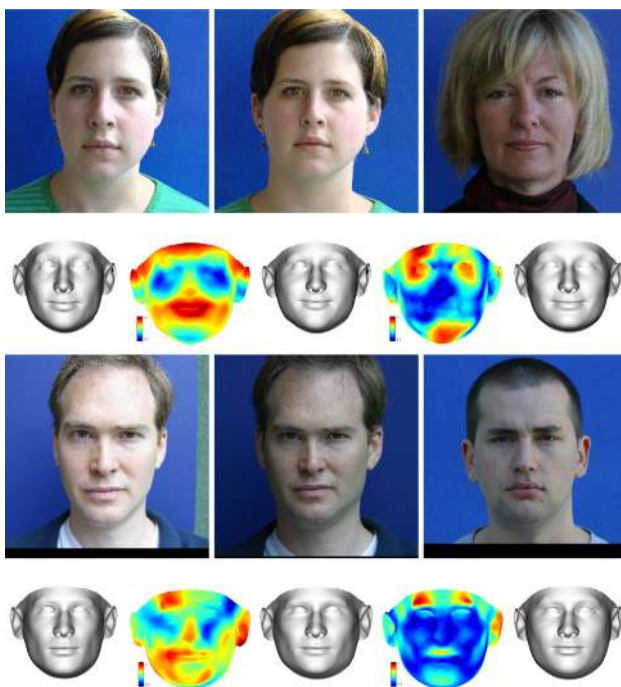

Fig. 10 Tran et al. (2017) regresses face shapes that are more different for for the same face viewed at different distances (2nd Row: 2.62 mm, 4th Row: 2.5 mm) than for different identities at the same distance (2nd Row: 1.79 mm, 4th Row: 1.26 mm) (Color figure online)

(2017) has not learnt invariance to perspective transformation in terms of the metric difference between the shapes themselves.

Another hypothesis is that the shape parameters themselves estimated by Tran et al. (2017) may be discriminative across distance for the purposes of recognition. We compute the normalised dot product distance for each shape vector at one distance against all shape vectors at the other distance. This allows us to compare the discriminativeness of the parameters under perspective transformation. We compare against our perspective fitting with either unknown or
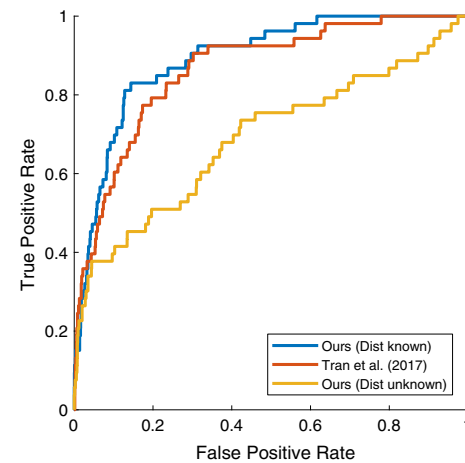

Fig. 11 ROC curves of Tran et al. (2017) and our method in the distance known and unknown settings on the CMDP dataset (Color figure online)

known subject-camera distance and show ROC curves for all three methods in Fig. 11. The area under curve (AUC) values for Tran et al. (2017) and our method with known distances and unknown distances are 0.866, 0.892 and 0.690, respectively. Using only geometric information and with unknown distance, it is clear that the estimated shape and hence parameters are ambiguous and perform poorly for recognition. Tran et al. (2017) has clearly learnt some invariance to distance but performance is still far from perfect on what is a fairly trivial dataset in the context of face recognition. With distance known (and hence the ambiguity avoided), even using only very sparse geometric information we obtain the best performance.

7.4 Flexibility Modes

We now explore the flexibility that remains when a model has been fitted to 2D geometric information. There is a surprising amount of remaining flexibility. Using the 70 Farkas landmark points under orthographic projection in a frontal pose, the BFM has around 50 flexibility modes that change the 3D shape by $k_1 = 2$ mm while inducing a mean change in landmark position of less than $k_2 = 2$ pixels. Restricting consideration to those flexibility modes where the shape parameter vector remains "plausible" [i.e. stays within 3 standard deviations of the expected Mahalanobis length (Patel and Smith 2016)], the number reduces to 7. This still means that knowing the exact 2D location of 70 landmark points only reduces the space of possible 3D face shapes to a 7D subspace of the morphable model.

In Figs. 12 and 13 we show qualitative examples of the flexibility modes. We fit to a real image under both orthographic and perspective projection. We then compute the first flexibility mode and vary the shape in both directions such that the mean surface distance is 10 mm. Despite the
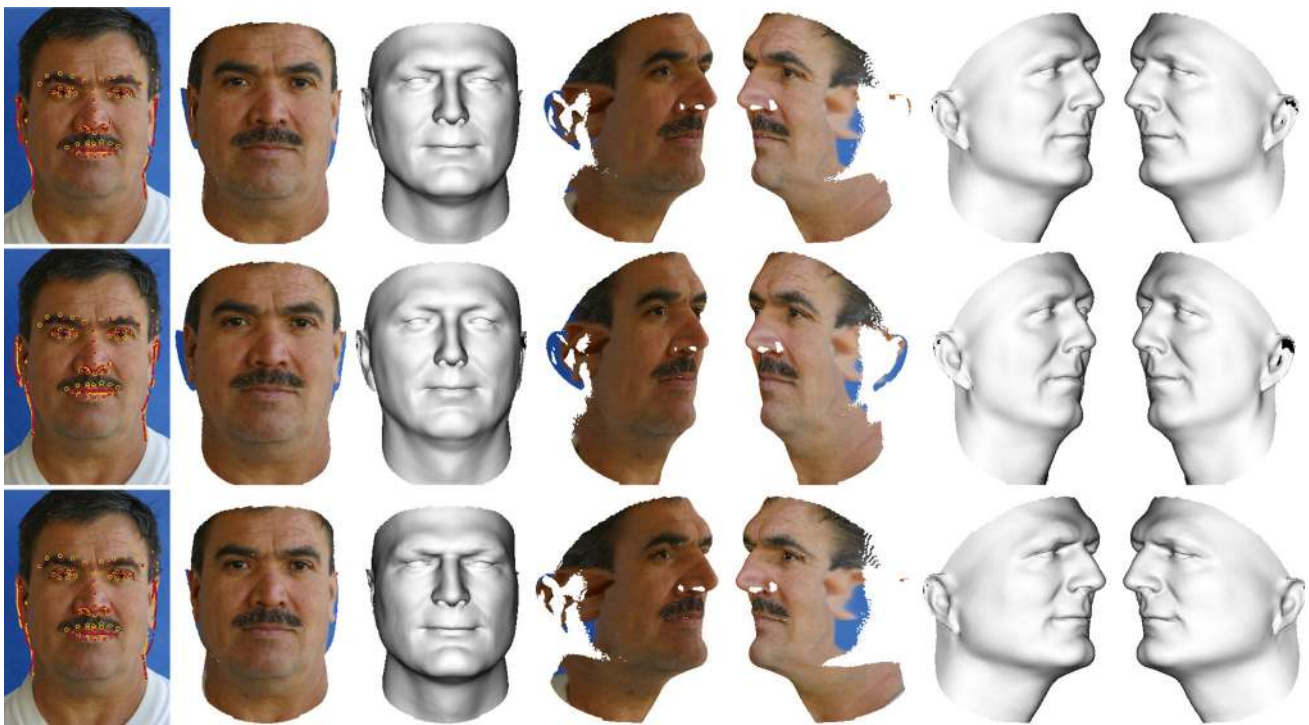
Fig. 12 Orthographic fitting with flexibility modes. 1st Row: landmark and edge fitting. 2nd/3rd Row: the first plus and minus flexibility components. Landmark distance is 1.14% and surface distance is 10 mm (Color figure online)
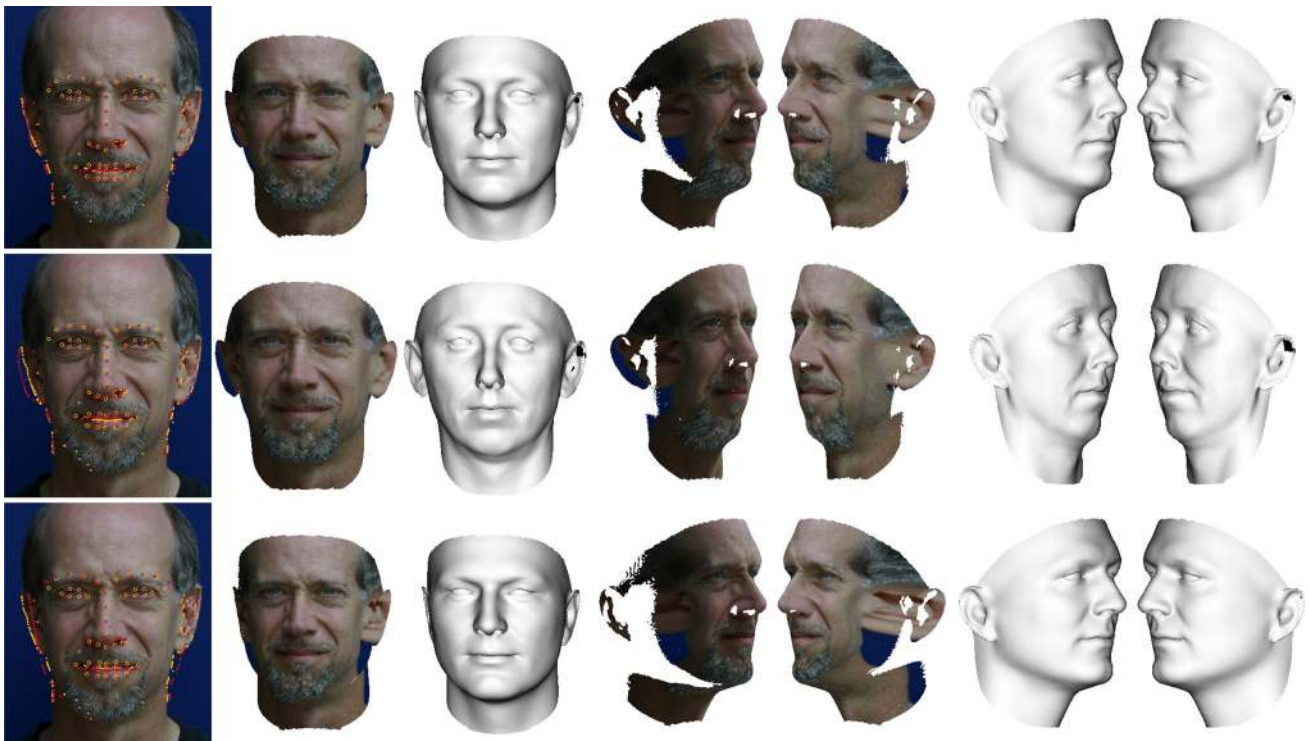

Fig. 13 Perspective fitting with flexibility modes. 1st Row: landmark and edge fitting. 2nd/3rd Row: the first plus and minus flexibility components. Landmark distance is 1.79% and surface distance is 10 mm (Color figure online)

large change in the surface, the landmarks only vary by 1.14% for orthographic and 1.79% for perspective fitting. The correspondence when the texture is sampled onto the mesh remains similar. In other words, three very different surfaces provide plausible 3D explanations of the 2D data.

8 Conclusions

In this paper we have studied ambiguities that arise when 3D face shape is estimated from monocular 2D geometric information. We have shown that 2D geometry (either sparse landmarks, semi-dense contours or dense vertex information) can be explained by a space of possible faces which vary significantly in 3D shape. We consider it surprising that the natural variability in face shape should include variations consistent with perspective transformation and that there are degrees of flexibility in face shape that have only a small effect on 2D geometry when pose is fixed. There are a number of interesting implications of these ambiguities.

In forensic image analysis, metric distances between features have been used as a way of comparing the identity of two face photographs. For example, Porter and Doran (2000) normalise face images by the interocular distance before using measurements such as the width of the face, nose and mouth to compare identities. We have shown that, after such normalisation, all distances between anthropometric features can be equal (up to the accuracy of landmarking) for two very different faces. This casts doubt on the use of such techniques in forensic image analysis and perhaps partially explains the studies that have demonstrated the weakness of these approaches (Kleinberg et al. 2007).

Clearly, any attempt to reconstruct 3D face shape using 2D geometric information alone [such as in (Blanz et al. 2004; Aldrian and Smith 2013; Patel and Smith 2009; Knothe et al. 2006; Bas et al. 2016)] will be subject to the ambiguity. Hence, the range of possible solutions is large and the likely accuracy low. If estimated 3D face shape is to be used for recognition, then the dissimilarity measure must account for the ambiguities we have described. On the other hand, CNN-based methods that learn to exploit any combination of features cannot necessarily overcome this uncertainty, as our results show. We believe that discriminative methods will require richer training data (either synthetic or real) containing significant variation in subject-camera distance, including small distances. Typically, there has been a reliance on web-crawled image databases, mainly of celebrities. These do not usually contain images at selfie distance and so new databases may be required.

For some face analysis problems, the purpose of fitting a statistical shape model is simply to establish correspondence.

For example, it may be that face texture will be processed on the surface of the mesh, or that correspondence is required in order to compare different face textures for recognition. In such cases, these ambiguities are not important. Any solution that fits the dense 2D shape features (i.e. any from within the space of solutions described by the ambiguity) will suffice to correctly establish correspondence.

There are many ways in which the work can be extended. First, our model fitting approach could be cast in probabilistic terms. By seeking the least squares solution, we are obtaining the maximum likelihood explanation of the data under an assumption of Gaussian noise on the 2D landmarks. Our flexibility modes capture the likely parts of the posterior distribution but a fully probabilistic setting would allow the posterior to be explicitly modelled and uncertainty quantified. Second, it would be interesting to investigate whether additional cues resolve the ambiguities. For example, an interesting follow-up to the work of Amberg et al. (2007) would be to investigate whether there is an ambiguity in uncalibrated *stereo* face images. Alternatively, we could investigate whether photometric cues (shading, shadowing and specularities) or statistical texture cues help to resolve the ambiguity. In the case of shading, it is not clear that this will be the case. Assuming illumination is unknown, it is possible that a transformation of the lighting environment could lead to shading which is consistent with (or at least close to) that of the target face (Smith 2016).

Reproducible Research

A Matlab implementation of the fitting algorithms, the scripts necessary to recreate the results in this paper and videos visualising the ambiguities is available at: http://www-users.cs.york.ac.uk/wsmith/faceambiguity. For the purposes of creating the images in this paper, we developed a full featured off-screen renderer in Matlab. We make this publicly available at: https://github.com/waps101/MatlabRenderer.

Appendix A SNLS Derivatives

Here we provide all of the derivatives required to optimise the SNLS objective functions. Specifically, we show how to compute the Jacobian matrices of the residual functions for

the orthographic (11) and perspective case linearised via the DLT (16).

Matrix Derivative Identities

he following identities are used in our derivations.

The derivatives of the axis-angle to rotation matrix function in (3) are given by Gallego and Yezzi (2015):

$$\frac{\partial \mathbf{R}}{\partial r_i} = \begin{cases} [\mathbf{e}_i]_\times & \text{if } \mathbf{r} = \mathbf{0} \\ \frac{r_i[\mathbf{r}]_\times + [\mathbf{r} \times (\mathbf{I} - \mathbf{R}(\mathbf{r}))\mathbf{e}_i]_\times}{\|\mathbf{r}\|^2} \mathbf{R}(\mathbf{r}) & \text{otherwise} \end{cases}$$

where $\mathbf{e}_i$ is the ith vector of the standard basis in $\mathbb{R}^3$.

The scalar derivative of the Kronecker product is:

$$\frac{\partial (\mathbf{X} \otimes \mathbf{Y})}{\partial x} = \frac{\partial \mathbf{X}}{\partial x} \otimes \mathbf{Y} + \mathbf{X} \otimes \frac{\partial \mathbf{Y}}{\partial x}.$$

For the special case involving the identity matrix, i.e. where $\mathbf{X} = \mathbf{I}$, this simplifies to:

$$\frac{\partial (\mathbf{I} \otimes \mathbf{Y})}{\partial x} = \mathbf{I} \otimes \frac{\partial \mathbf{Y}}{\partial x}.$$

The scalar derivative of the pseudoinverse $\mathbf{A}^+(x)$ of $\mathbf{A}$ at x is given by:

$$\frac{\partial \mathbf{A}^+}{\partial x} = -\mathbf{A}^+ \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^+ + \mathbf{A}^+ \mathbf{A}^{+\mathrm{T}} \frac{\partial \mathbf{A}^\mathrm{T}}{\partial x} (\mathbf{I} - \mathbf{A}\mathbf{A}^+) \\ + (\mathbf{I} - \mathbf{A}^+\mathbf{A}) \frac{\partial \mathbf{A}^\mathrm{T}}{\partial x} \mathbf{A}^{+\mathrm{T}} \mathbf{A}^+$$

Orthographic Case

The derivatives of the matrix $\mathbf{A}(\mathbf{r}, s)$ are given by:

$$\frac{\partial \mathbf{A}}{\partial s} = \left[(\mathbf{I}_L \otimes \mathbf{P}\mathbf{R}(\mathbf{r})) \mathbf{Q}_L \; \mathbf{1}_L \otimes \mathbf{I}_2 \right],$$
$$\frac{\partial \mathbf{A}}{\partial r_i} = \left[s \left(\mathbf{I}_L \otimes \mathbf{P}\frac{\partial \mathbf{R}}{\partial r_i} \right) \mathbf{Q}_L \; \mathbf{0}_{2L \times 2} \right].$$

The derivatives of the vector $\mathbf{y}(\mathbf{r}, s)$ are given by:

$$\frac{\partial \mathbf{y}}{\partial s} = (\mathbf{I}_L \otimes \mathbf{P}\mathbf{R}(\mathbf{r})) \bar{\mathbf{s}}$$
$$\frac{\partial \mathbf{y}}{\partial r_i} = s \left[\left(\mathbf{I}_L \otimes \mathbf{P}\frac{\partial \mathbf{R}}{\partial r_i} \right) \bar{\mathbf{s}} \right].$$

From the components above we can compute the derivatives of the residual function:

$$\frac{\partial \mathbf{d}_{\text{ortho}}}{\partial s} = \left(\mathbf{A}(\mathbf{r}, s) \frac{\partial \mathbf{A}^+}{\partial s} + \frac{\partial \mathbf{A}}{\partial s} \mathbf{A}^+(\mathbf{r}, s) \right) \mathbf{y}(\mathbf{r}, s) \\ + \mathbf{A}(\mathbf{r}, s) \mathbf{A}^+(\mathbf{r}, s) \frac{\partial \mathbf{y}}{\partial s} - \frac{\partial \mathbf{y}}{\partial s},$$
$$\frac{\partial \mathbf{d}_{\text{ortho}}}{\partial r_i} = \left(\mathbf{A}(\mathbf{r}, s) \frac{\partial \mathbf{A}^+}{\partial r_i} + \frac{\partial \mathbf{A}}{\partial r_i} \mathbf{A}^+(\mathbf{r}, s) \right) \mathbf{y}(\mathbf{r}, s) \\ + \mathbf{A}(\mathbf{r}, s) \mathbf{A}^+(\mathbf{r}, s) \frac{\partial \mathbf{y}}{\partial r_i} - \frac{\partial \mathbf{y}}{\partial r_i}.$$

Finally, the Jacobian, $\mathbf{J}_{\mathbf{d}_{\text{ortho}}}(\mathbf{r}, s)$, is obtained by stacking these four vectors into a $2L \times 4$ matrix:

$$\mathbf{J}_{\mathbf{d}_{\text{ortho}}}(\mathbf{r}, s) = \left[\frac{\partial \mathbf{d}_{\text{ortho}}}{\partial r_1} \; \frac{\partial \mathbf{d}_{\text{ortho}}}{\partial r_2} \; \frac{\partial \mathbf{d}_{\text{ortho}}}{\partial r_3} \; \frac{\partial \mathbf{d}_{\text{ortho}}}{\partial s} \right].$$

Perspective Case

The derivatives of the matrix $\mathbf{B}(\mathbf{r}, f)$ are given by:

$$\frac{\partial \mathbf{B}}{\partial f} = \mathbf{D} \frac{\partial \mathbf{E}}{\partial f} \mathbf{F}(\mathbf{r}) \quad \text{and} \quad \frac{\partial \mathbf{B}}{\partial r_i} = \mathbf{D}\mathbf{E}(f) \frac{\partial \mathbf{F}}{\partial r_i},$$

where

$$\frac{\partial \mathbf{E}}{\partial f} = \mathbf{I}_L \otimes \frac{\partial \mathbf{K}}{\partial f} \quad \text{and} \quad \frac{\partial \mathbf{F}}{\partial r_i} = \left[\left(\mathbf{I}_L \otimes \frac{\partial \mathbf{R}}{\partial r_i} \right) \mathbf{Q}_L \; \mathbf{0}_{3L \times 3} \right],$$

and

$$\frac{\partial \mathbf{K}}{\partial f} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The derivatives of the vector $\mathbf{z}(\mathbf{r}, f)$ are given by:

$$\frac{\partial \mathbf{z}}{\partial f} = -\mathbf{D} \left(\mathbf{I}_L \otimes \left[\frac{\partial \mathbf{K}}{\partial f} \mathbf{R}(\mathbf{r}) \right] \right) \bar{\mathbf{s}},$$
$$\frac{\partial \mathbf{z}}{\partial r_i} = -\mathbf{D} \left(\mathbf{I}_L \otimes \left[\mathbf{K}(f) \frac{\partial \mathbf{R}}{\partial r_i} \right] \right) \bar{\mathbf{s}}.$$

From the components above we can compute the derivatives of the residual function:

$$\frac{\partial \mathbf{d}_{\text{persp}}^{\text{DLT}}}{\partial f} = \left(\mathbf{B}(\mathbf{r}, f) \frac{\partial \mathbf{B}^+}{\partial f} + \frac{\partial \mathbf{B}}{\partial f} \mathbf{B}^+(\mathbf{r}, f) \right) \mathbf{z}(\mathbf{r}, f) \\ + \mathbf{B}(\mathbf{r}, f) \mathbf{B}^+(\mathbf{r}, f) \frac{\partial \mathbf{z}}{\partial f} - \frac{\partial \mathbf{z}}{\partial f},$$
$$\frac{\partial \mathbf{d}_{\text{persp}}^{\text{DLT}}}{\partial r_i} = \left(\mathbf{B}(\mathbf{r}, f) \frac{\partial \mathbf{B}^+}{\partial r_i} + \frac{\partial \mathbf{B}}{\partial r_i} \mathbf{B}^+(\mathbf{r}, f) \right) \mathbf{z}(\mathbf{r}, f) \\ + \mathbf{B}(\mathbf{r}, f) \mathbf{B}^+(\mathbf{r}, f) \frac{\partial \mathbf{z}}{\partial r_i} - \frac{\partial \mathbf{z}}{\partial r_i}.$$

Finally, the Jacobian, $\mathbf{J}_{\mathbf{d}_{\text{persp}}^{\text{DLT}}}(\mathbf{r}, f)$, is obtained by stacking these four vectors into a $3L \times 4$ matrix:

$$\mathbf{J}_{\mathbf{d}_{\text{persp}}^{\text{DLT}}}(\mathbf{r}, f) = \begin{bmatrix} \frac{\partial \mathbf{d}_{\text{persp}}^{\text{DLT}}}{\partial r_1} & \frac{\partial \mathbf{d}_{\text{persp}}^{\text{DLT}}}{\partial r_2} & \frac{\partial \mathbf{d}_{\text{persp}}^{\text{DLT}}}{\partial r_3} & \frac{\partial \mathbf{d}_{\text{persp}}^{\text{DLT}}}{\partial f} \end{bmatrix}.$$

References

Albrecht, T., Knothe, R., & Vetter, T. (2008). Modeling the remaining flexibility of partially fixed statistical shape models. In *Proceedings of workshop on the mathematical foundations of computational anatomy*.

Aldrian, O., & Smith, W. A. P. (2013). Inverse rendering of faces with a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(5), 1080–1093.

Amberg, B., Blake, A., Fitzgibbon, A., Romdhani, S., & Vetter, T. (2007). Reconstructing high quality face-surfaces using model based stereo. In *Proceedings of ICCV*.

Bas, A., Smith, W., Bolkart, T., & Wuhrer, S. (2016). Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *Proceedings of ACCV workshops* (pp. 377–391).

Belhumeur, P. N., Kriegman, D. J., & Yuille, A. L. (1999). The bas-relief ambiguity. *International Journal of Computer Vision*, *35*(1), 33–44.

Blanz, V., Mehl, A., Vetter, T., & Seidel, H. P. (2004). A statistical method for robust 3D surface reconstruction from sparse data. In *Proceedings of 3DPVT* (pp. 293–300).

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

Booth, J., Roussos, A., Ververas, E., Antonakos, E., Ploumpis, S., Panagakis, Y., et al. (2018). 3D reconstruction of "in-the-wild" faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2638–2652.

Bryan, R., Perona, P., & Adolphs, R. (2012). Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces. *PLoS ONE*, *7*(9), e45,301.

Burgos-Artizzu, X. P., Ronchi, M. R., & Perona, P. (2014). Distance estimation of an unknown person from a portrait. In *Proceedings of ECCV* (pp. 313–327).

Cao, C., Hou, Q., & Zhou, K. (2014a). Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics*, *33*(4), 43.

Cao, C., Weng, Y., Lin, S., & Zhou, K. (2013). 3D shape regression for real-time facial animation. *ACM Transactions on Graphics*, *32*(4), 41.

Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2014b). Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, *20*(3), 413–425.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. In *Proceedings of ECCV* (pp. 484–498).

Ecker, A., Jepson, A. D., & Kutulakos, K. N. (2008). Semidefinite programming heuristics for surface reconstruction ambiguities. In *Proceedings of ECCV* (pp. 127–140).

Farkas, L. (1994). *Anthropometry of the head and face*. New York: Raven Press.

Feng, Z. H., Huber, P., Kittler, J., Hancock, P., Wu, X. J., Zhao, Q., Koppen, P., & Rätsch, M. (2018). Evaluation of dense 3D reconstruction from 2D face images in the wild. In *Proceedings of FG* (pp. 780–786)

Flores, A., Christiansen, E., Kriegman, D., & Belongie, S. (2013). Camera distance from face images. In *Proceedings of ISVC* (pp. 513–522).

Fried, O., Shechtman, E., Goldman, D. B., & Finkelstein, A. (2016). Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics*, *35*(4), 128.

Gallego, G., & Yezzi, A. (2015). A compact formula for the derivative of a 3-D rotation in exponential coordinates. *Journal of Mathematical Imaging and Vision*, *51*(3), 378–384.

Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., & Freeman, W. T. (2018). Unsupervised training for 3D morphable model regression. In *Proceedings of CVPR*.

Georghiades, A., Belhumeur, P., & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 643–660.

Golub, G., & Pereyra, V. (2003). Separable nonlinear least squares: The variable projection method and its applications. *Inverse Problems*, *19*(2), R1.

Güler, R. A., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., & Kokkinos, I. (2017). DenseReg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of CVPR*.

Hartley, R., & Vidal, R. (2008). Perspective nonrigid shape and motion recovery. In *Proceedings of ECCV* (pp. 276–289).

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Hill, H., & Bruce, V. (1994). A comparison between the hollow-face and 'hollow-potato' illusions. *Perception*, *23*, 1335–1337.

Keller, M., Knothe, R., & Vetter, T. (2007). 3D reconstruction of human faces from occluding contours. In *Proceedings of mirage* (pp. 261–273).

Kleinberg, K. F., Vanezis, P., & Burton, A. M. (2007). Failure of anthropometry as a facial identification technique using high-quality photographs. *Journal of Forensic Sciences*, *52*(4), 779–783.

Knothe, R., Romdhani, S., & Vetter, T. (2006). Combining PCA and LFA for surface reconstruction from a sparse set of control points. In *Proceedings of FG* (pp. 637–644).

Latto, R., & Harper, B. (2007). The non-realistic nature of photography: Further reasons why turner was wrong. *Leonardo*, *40*(3), 243–247.

Lepetit, V., Moreno-Noguer, F., & Fua, P. (2009). EPnP: An accurate $O(n)$ solution to the PnP problem. *International Journal of Computer Vision*, *81*(2), 155–166.

Liu, C. H., & Chaudhuri, A. (2003). Face recognition with perspective transformation. *Vision Research*, *43*(23), 2393–2402.

Liu, C. H., & Ward, J. (2006). Face recognition in pictures is affected by perspective transformation but not by the centre of projection. *Perception*, *35*(12), 1637–1650.

Lüthi, M., Albrecht, T., & Vetter, T. (2009). Probabilistic modeling and visualization of the flexibility in morphable models. In *Proceedings of IMA conference on mathematics of surfaces* (pp. 251–264).

Moreno-Noguer, F., & Fua, P. (2013). Stochastic exploration of ambiguities for nonrigid shape recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(2), 463–475.

Patel, A., & Smith, W. A. (2016). Manifold-based constraints for operations in face space. *Pattern Recognition*, *52*, 206–217.

Patel, A., & Smith, W. A. P. (2009). 3D morphable face models revisited. In *Proceedings of CVPR* (pp. 1327–1334).

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *Proceedings of AVSS*.

Perona, P. (2007). A new perspective on portraiture. *Journal of Vision*, *7*(9), 992–992.

Piotraschke, M., & Blanz, V. (2016). Automated 3D face reconstruction from multiple images using quality measures. In *Proceedings of CVPR* (pp. 3418–3427).

Porter, G., & Doran, G. (2000). An anatomical and photographic technique for forensic facial identification. *Forensic Science International*, *114*(2), 97–105.

Prados, E., Jindal, N., & Soatto, S. (2009). A non-local approach to shape from ambient shading. In *Proceedings of SSVM* (pp. 696–708).

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, *47*, 3–18.

Saito, S., Li, T., & Li, H. (2016). Real-time facial segmentation and performance capture from RGB input. In *Proceedings of ECCV* (pp. 244–261).

Salzmann, M., Lepetit, V., & Fua, P. (2007). Deformable surface tracking ambiguities. In *Proceedings of CVPR* (pp. 1–8).

Sanyal, S., Bolkart, T., Feng, H., & Black, M. J. (2019). Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of CVPR*. https://ringnet.is.tue.mpg.de/. Accessed 22 Jul 2019.

Schönborn, S., Egger, B., Morel-Forster, A., & Vetter, T. (2017). Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, *123*(2), 160–183.

Schumacher, M., & Blanz, V. (2012). Which facial profile do humans expect after seeing a frontal view? A comparison with a linear face model. *ACM Transactions on Applied Perception*, *9*(3), 11.

Sela, M., Richardson, E., & Kimmel, R. (2017). Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of ICCV*, (pp. 1585–1594).

Smith, W. A. P. (2016). The perspective face shape ambiguity. In M. Breuß, F. Bruckstein, P. Maragos, & S. Wuhrer (Eds.), *Perspectives in shape analysis* (pp. 299–319). Cham: Springer.

Smith, W. A. P., & Hancock, E. R. (2006). Recovering facial shape using a statistical model of surface normal direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(12), 1914–1930.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of CVPR* (pp. 1701–1708).

Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., & Theobalt, C. (2018). Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Proceedings of CVPR*.

Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Pérez, P., & Theobalt, C. (2017). Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of ICCV* (pp. 3735–3744).

Tran, A. T., Hassner, T., Masi, I., & Medioni, G. (2017). Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of CVPR* (pp. 1493–1502).

Tran, L., & Liu, X. (2018). Nonlinear 3D face morphable model. In *Proceedings of CVPR*.

Valente, J., & Soatto, S. (2015). Perspective distortion modeling, learning and compensation. In *Proceedings of CVPR workshops* (pp. 9–16).

Wu, C., Bradley, D., Garrido, P., Zollhöfer, M., Theobalt, C., Gross, M., et al. (2016). Model-based teeth reconstruction. *ACM Transactions on Graphics*, *35*(6), 220.

Wu, Y., Ha, L. A. V., Xu, X., & Kakadiaris, I. A. (2018). Convolutional point-set representation: A convolutional bridge between a densely annotated image and 3D face alignment. arXiv:180306542

Yu, R., Saito, S., Li, H., Ceylan, D., & Li, H. (2017). Learning dense facial correspondences in unconstrained images. In *Proceedings of CVPR* (pp. 4723–4732).

Zhou, X., Leonardos, S., Hu, X., & Daniilidis, K. (2015). 3D shape estimation from 2D landmarks: A convex relaxation approach. In *Proceedings of CVPR* (pp. 4447–4455).

Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of CVPR* (pp. 787–796).

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of CVPR* (pp. 2879–2886).