What does 2D geometric information really tell us about 3D face shape?

Anil Bas and William A. P. Smith, *Member, IEEE*

encapsulate 封装 / contours 外形 / morphable 形变 / interocular 俩眼间 / adjacent 相邻

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 provide quantitative and qualitative evidence that the ambiguity exists in synthetic data and real images.

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 the photometric image formation process and to interpret appearance.

Although photometric information does provide a cue to the 3D shape of a face [1], 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 [2]. For this reason, the use of geometric information has proven very popular in 3D face reconstruction [3]–[7]. Landmark detection on highly uncontrolled face images is now a mature research field with benchmarks [8] providing an indication of likely accuracy. Landmarks are often used to initialise or constrain the fitting of 3D morphable models to images while denser 2D geometric information such as the occluding boundary are used in some of the state-of-the-art methods.

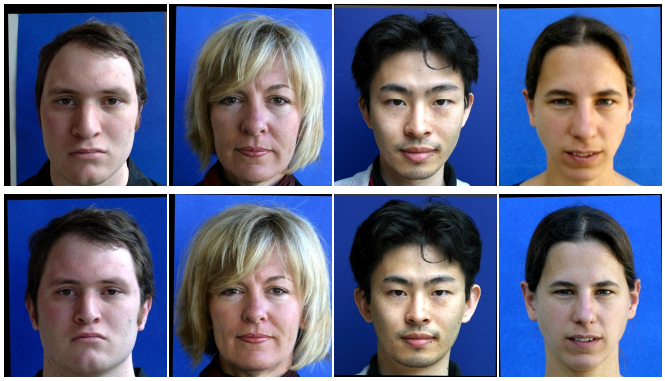


Fig. 1: Perspective transformation of real faces from the CMDP dataset [9]. The subject is the same in each column and the same camera and lighting is used. The change in viewing distance (60cm top row, 490cm bottom row) induces a significant change in projected shape.

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 [8]). 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 3D morphable model) 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 pro-

A. Bas and W. Smith are with the Department of Computer Science, University of York York, UK. e-mail: {ab1792,william.smith}@york.ac.uk

jection, 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 Figure 1. Images are taken of subjects at 60cm and 490cm. 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 ambiguity. 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 3D morphable model 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 [9].

2 Related Work

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 [10] or in CNN-based face recognition frameworks [11]. For this reason, there has been sustained interest in building feature detectors capable of accurately labelling face landmarks in uncontrolled images [8].

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 3D morphable model to these detected landmarks [3]–[6]. 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 space from 2D landmark positions is related to the problem of non-rigid structure from motion [12]. 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 3D morphable model) but only a single view of the face is available. Some work has considered other 2D shape features besides landmark points. Keller et al. [13] fit a 3D morphable model to contours (both silhouettes and inner contours due to texture, shape and shadowing). Bas et al. [7] adapt the Iterated Closest Point algorithm to fit to edge pixels with an additional landmark term. They use alternating linear least squares optimisation followed by a non-convex refinement. Although not applied to faces, Zhou et al. [14] propose a convex relaxation of the shape-from-landmarks energy.

A related problem is to describe the remaining flexibility in a statistical shape model that is partially fixed. In other words, 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. [15] show how to compute the subspace of faces with the same profile. Lüthi et al. [16] 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 an affine camera (such as scaled orthographic or "weak perspective") [3]–[6]. 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. [17] describe an algorithm for fitting a 3D morphable model 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.

Faces under perspective projection The effect of perspective transformation on face appearance has been studied from both a computational and psychological perspective previously. In art history, Latto and Harper [18] discuss how uncertainty regarding subject-artist distance when viewing a painting results in distorted perception. They show that showed that perceptions of body weight from face images are influenced by subject-camera distance. In psychology, Liu et al. [19], [20] show that human face recognition performance is degraded by perspective transformation. Perona et al. [21], [22] investigated a different effect, noting that perspective distortion influences social judgements of faces.

There have been two recent attempts to address the problem of estimating subject-camera distance from monocular, perspective views of a face [9], [23]. The idea is that the configuration of projected 2D face features conveys something about the degree of perspective transformation. Flores et al. [23] approach the problem using exemplar 3D

face models. They fit the models to 2D landmarks using the EPnP algorithm [24] and use the mean of the estimated distances as the estimated subject-camera distance. Burgos-Artizzu et al. [9] 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 in this paper 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.

Fried et al. [25] 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 [26] 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.

Other ambiguities There are other known ambiguities in the monocular estimation of 3D shape. The bas relief ambiguity [27] 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. [28] 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* [29], 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. [30] 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. Moreno-Noguer and Fua [31] use stochastic sampling to explore the set of possible solutions to nonrigid, monocular shape reconstruction. They attempt to select from within this space using additional information provided by motion or shading. Salzmann et al. [32] study the ambiguities that arise in monocular nonrigid structure from motion under perspective projection.

In an early version of this work [33], 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 for both cases. Moreover, we show how model fitting can be posed as a separable non-linear least squares problem, including solving for rotation and translation, and present much 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 3D morphable model and the scaled orthographic and pinhole projection model.

3.1 3D Morphable Model

A 3D morphable model is a deformable mesh whose vertex positions, $\mathbf{s}(\boldsymbol{\alpha})$, are determined by the shape parameters $\boldsymbol{\alpha} \in \mathbb{R}^S$. Shape is described by a linear subspace model learnt from data using PCA [34]. So, the shape of any object from the same class as the training data can be approximated as:

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

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

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 $\bar{\mathbf{s}}_i \in \mathbb{R}^3$, such that the ith vertex is given by: $\mathbf{v}_i = \mathbf{Q}_i\boldsymbol{\alpha} + \bar{\mathbf{s}}_i$.

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.

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 $\mathbf{v} = \begin{bmatrix} u & v & w \end{bmatrix}^T$ onto the 2D point $\mathbf{x} = \begin{bmatrix} x & y \end{bmatrix}^T$ 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:

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

where

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{3}$$

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 $\mathbf{r} \in \mathbb{R}^3$.

3.3 Perspective camera model

The nonlinear perspective projection of the 3D point $\mathbf{v} = \begin{bmatrix} u & v & w \end{bmatrix}^T$ onto the 2D point $\mathbf{x} = \begin{bmatrix} x & y \end{bmatrix}^T$ 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 $\mathbf{t}_{3d} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^T$ 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 paramterise 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 Figure 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 non-linear projection can be written in linear terms by using homogeneous representations $\tilde{\mathbf{v}} = \begin{bmatrix} u & v & w & 1 \end{bmatrix}^T$ and $\tilde{\mathbf{x}} = \begin{bmatrix} x & y & 1 \end{bmatrix}^T$:

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

where γ is an arbitrary scaling factor.

4 Shape-from-landmarks

In this section, we describe a novel method for fitting a 3D morphable model 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 $\mathbf{x}_i = [x_i \; y_i]^T$ $(i = 1 \dots L)$ have been observed. Without loss of generality, we assume that the ith landmark corresponds to the ith vertex in the morphable model. 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 $\mathbf{Q}$).

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.

4.1 Orthographic objective function

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

$$\varepsilon_{\text{ortho}}(\mathbf{r}, \mathbf{t}_{2d}, s, \boldsymbol{\alpha}) = \sum_{i=1}^{L} \left\| \mathbf{x}_i - \mathbf{SOP}\left[\mathbf{Q}_i \boldsymbol{\alpha} + \bar{\mathbf{s}}_i, \mathbf{R}(\mathbf{r}), \mathbf{t}_{2d}, s\right] \right\|^2. \tag{5}$$

The 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:

$$\varepsilon_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3d}, f, \boldsymbol{\alpha}) = \sum_{i=1}^{L} \left\| \mathbf{x}_i - \mathbf{pinhole}\left[\mathbf{Q}_i \boldsymbol{\alpha} + \bar{\mathbf{s}}_i, \mathbf{K}(f), \mathbf{R}(\mathbf{r}), \mathbf{t}_{3d}\right] \right\|^2. \tag{6}$$

This objective is nonlinear in all parameters and nonconvex due to the perspective projection. However, we can use the direct linear transformation (DLT) [35] 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:

$$\begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} \sim \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_i \boldsymbol{\alpha} + \bar{\mathbf{s}}_i \\ 1 \end{bmatrix}, \tag{7}$$

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

$$\begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}_\times \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{bmatrix} \mathbf{Q}_i \boldsymbol{\alpha} + \bar{\mathbf{s}}_i \\ 1 \end{bmatrix} = \mathbf{0} \tag{8}$$

where $\mathbf{0} = [0 \; 0 \; 0]^T$ and $[.]_\times$ is the cross product matrix:

$$[\mathbf{x}]_\times = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}. \tag{9}$$

This means that each landmark point 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 [36]. 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.

4.3.1 Orthographic

The orthographic objective function (5) can be written in SNLS form as

$$\varepsilon_{\text{ortho}}(\mathbf{r}, \mathbf{t}_{2\text{d}}, s, \boldsymbol{\alpha}) = \left\| \mathbf{A}(\mathbf{r}, s) \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{t} \end{bmatrix} - \mathbf{y}(\mathbf{r}, s) \right\|^2 \quad (10)$$

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

$$\mathbf{A}(\mathbf{r}, s) = s \begin{bmatrix} (\mathbf{I}_L \otimes \mathbf{PR}(\mathbf{r}))\, \mathbf{Q}_L & \mathbf{1}_L \otimes \mathbf{I}_2 \end{bmatrix}, \quad (11)$$

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

$$\mathbf{y}(\mathbf{r}, s) = s \left[(\mathbf{I}_L \otimes \mathbf{PR}(\mathbf{r}))\, \bar{\mathbf{s}} \right] - \begin{bmatrix} x_1 & y_1 & \dots & y_L \end{bmatrix}^T. \quad (12)$$

Note that this objective is exactly equivalent to the original one. The optimal solution to (10) in terms of the linear parameters is given by:

$$\begin{bmatrix} \boldsymbol{\alpha}^* \\ \mathbf{t}^* \end{bmatrix} = \mathbf{A}^+(\mathbf{r}, s)\mathbf{y}(\mathbf{r}, s) \quad (13)$$

where $\mathbf{A}^+(\mathbf{r}, s)$ is the pseudoinverse. Substituting (13) into (10) we get an equivalent objective to (5) but which depends only on the nonlinear parameters:

$$\varepsilon_{\text{ortho}}(\mathbf{r}, s) = \left\| \mathbf{A}(\mathbf{r}, s)\mathbf{A}^+(\mathbf{r}, s)\mathbf{y}(\mathbf{r}, s) - \mathbf{y}(\mathbf{r}, s) \right\|^2. \quad (14)$$

This is a nonlinear least squares problem of very low dimensionality ($[\mathbf{r}\ s]$ is only 4D) that can be solved with Gauss-Newton minimisation or similar methods.

Once optimal parameters have been obtained by minimising (14) then the parameters $\boldsymbol{\alpha}^*$ and $\mathbf{t}^*$ are obtained by (13). If we wish to impose a statistical prior on the shape parameters, e.g. in the form of Tikhonov regularisation [3] or a hard hyperbox constraint [37], this is simply introduced during the solution of (13).

4.3.2 Perspective

The perspective objective function (6), linearised via (8), can be written in SNLS form as

$$\varepsilon^{\text{DLT}}_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3\text{d}}, f, \boldsymbol{\alpha}) = \left\| \mathbf{B}(\mathbf{r}, f) \begin{bmatrix} \boldsymbol{\alpha} \\ \mathbf{t} \end{bmatrix} - \mathbf{y}(\mathbf{r}, f) \right\|^2 \quad (15)$$

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

$$\mathbf{B}(\mathbf{r}, f) = \mathbf{DE}(f)\mathbf{F}(\mathbf{r}), \quad (16)$$

with

$$\mathbf{D} = \text{diag}\left(\begin{bmatrix} \mathbf{x}_1 \\ 1 \end{bmatrix}_\times, \dots, \begin{bmatrix} \mathbf{x}_L \\ 1 \end{bmatrix}_\times \right), \quad \mathbf{E}(f) = \mathbf{I}_L \otimes \mathbf{K}(f) \quad (17)$$

and

$$\mathbf{F}(\mathbf{r}) = \begin{bmatrix} (\mathbf{I}_L \otimes \mathbf{R}(\mathbf{r}))\, \mathbf{Q}_L & \mathbf{1}_L \otimes \mathbf{I}_3 \end{bmatrix}. \quad (18)$$

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

$$\mathbf{y}(\mathbf{r}, f) = -\mathbf{D}\left(\mathbf{I}_L \otimes \mathbf{K}(f)\mathbf{R}(\mathbf{r})\right) \bar{\mathbf{s}} \quad (19)$$

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 (sum of squared perspective reprojection distances). So, we use the SNLS solution to initialise a nonlinear optimisation of the original objective over all parameters. In practice, we find that the SNLS solution is already very close to the optimum and that the subsequent nonlinear optimisation usually converges in 2-5 iterations.

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 Section 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:

$$\boldsymbol{\alpha}^*(k) = \arg_{\boldsymbol{\alpha}} \min_{\mathbf{r}, \mathbf{t}_{3\text{d}}, f, \boldsymbol{\alpha}} \varepsilon_{\text{persp}}(\mathbf{r}, \mathbf{t}_{3\text{d}}, f, \boldsymbol{\alpha}), \ \text{ s.t. } t_z = k. \quad (20)$$

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. [7] 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 [38].

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:

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

and

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

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 Section 4 (and optionally Section 5). This amounts to a shape, $\mathbf{Q}\boldsymbol{\alpha} + \bar{\mathbf{s}}$, determined by the estimated shape parameter and a pose $(\mathbf{r}, s, \mathbf{t}_{2\mathrm{d}})$ or $(\mathbf{r}, f, \mathbf{t}_{3\mathrm{d}})$ for orthographic or perspective repsectively. 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. [15]. They considered the problem of flexibility only in a 3D setting where the model is partitioned into a 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:

$$\boldsymbol{\Pi}_{\text{ortho}} = \left(\mathbf{I}_L \otimes \mathbf{P}\mathbf{R}(\mathbf{r})\right)\mathbf{Q}_L, \tag{21}$$

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 (8), leading to the following expression:

$$\boldsymbol{\Pi}_{\text{perspective}} = \mathbf{D}\left(\mathbf{I}_L \otimes \left(\mathbf{K}(f)\left[\mathbf{R}(\mathbf{r}) \quad \mathbf{t}_{3\mathrm{d}}\right]\mathbf{S}\right)\right)\mathbf{Q}_L, \tag{22}$$

where

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{23}$$

Again, $(\mathbf{r}, f, \mathbf{t}_{3\mathrm{d}})$ 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 $\mathbf{Q}\mathbf{f}$ changes as much as possible whilst the 2D projection of the landmarks, given by $\boldsymbol{\Pi}_{\text{ortho}}\mathbf{f}$ or $\boldsymbol{\Pi}_{\text{perspective}}\mathbf{f}$, changes as little as possible. This can be formulated as a constrained maximisation problem:

$$\max_{\mathbf{f}\in\mathbb{R}^S} \|\mathbf{Q}\mathbf{f}\|^2 \ \text{ subject to } \|\boldsymbol{\Pi}\mathbf{f}\|^2 = c, \tag{24}$$

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:

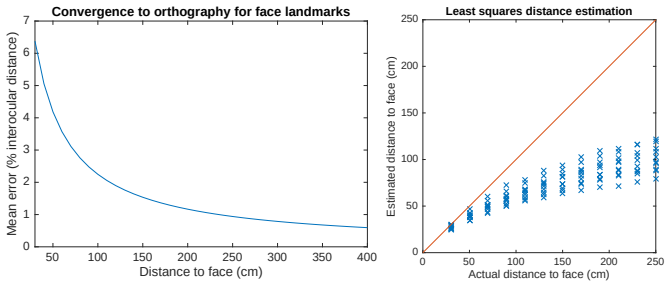
$$\mathbf{Q}^T\mathbf{Q}\mathbf{f} = \lambda\boldsymbol{\Pi}^T\boldsymbol{\Pi}\mathbf{f}. \tag{25}$$

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: $\mathbf{Q}(\boldsymbol{\alpha} + w\mathbf{f}) + \bar{\mathbf{s}}$.

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 [39] (BFM) which is a 3D morphable model 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



(a) (b)

Fig. 2: (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.

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 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 [40]) whose indices in the BFM are known. These landmarks are particularly appropriate as they were chosen so as to best measure the variability in craniofacial shape over a population. In Figure 2a, 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 metres. Hence, we experiment with distances ranging from selfie distance (30cm) 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 Figure 2b. Optimal performance would see all points falling on the diagonal red line. The distance is consistently under-estimated 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.

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

TABLE 1: Quantitative results for the perspective ambiguity on synthetic data. Each cell shows the landmark error, d_L in %, top and surface error, d_S in mm, bottom.

We now show that landmarks produced by a real 3D face shape at one distance can be explained by 3D shapes at multiple different distances. In Table 1 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 1cm on average. Note that Burgos-Artizzu et al. [9] 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 [8] 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 Figures 3 and 4 we show 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. the targets are real faces). In Figure 3, 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 Figure 4 (200cm to 60cm). 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 face under perspective projection to which we fit is shown in the first and forth columns. The fitting result under perspective projection is shown in the second to seventh 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

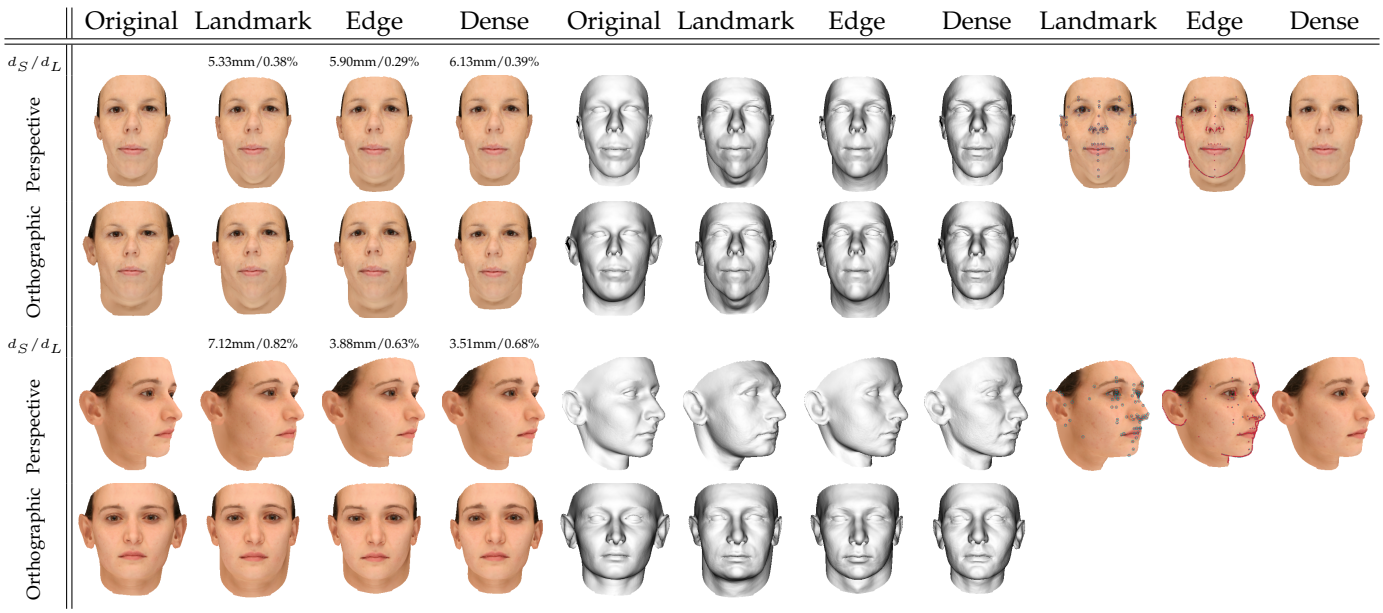


Fig. 3: Sparse and dense fitting of the synthetic images. Target at 30cm, fitted results at 120cm.

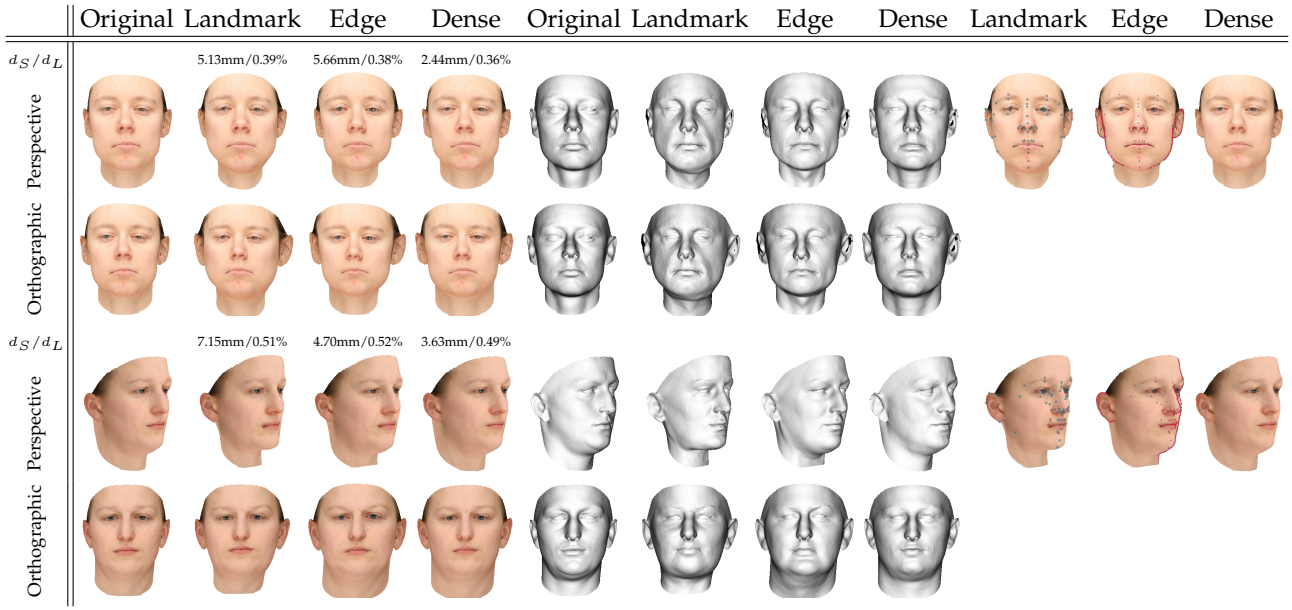


Fig. 4: Sparse and dense fitting of the synthetic images. Target at 200cm, fitted results at 60cm.

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 Figure 5 that this is indeed the case. The Caltech Multi-Distance Portraits dataset [9] 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 [8]) when they are viewed at 61cm and 488cm respectively (second and fourth image in the figure). In the third image, we blend these 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 images one and five. E.g. compare column 1 with column 4. The face in column 1 has larger ears and inner features that are more concentrated towards to the centre of the face compared to the face in column 4.

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 61cm and 488cm which is 3.11%. Next, for each identity we find the identity at the same distance with the smallest landmark

Fig. 5: 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.

error. Averaged over all identities, this gives a value of 2.86% for 61cm and 2.83% for 488cm. 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.

7.2 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 [41]), 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 Figure 6 we show a qualitative example of the flexibility modes. We fit to a real image assuming orthographic projection. We then compute the first flexibility mode and vary the shape in both directions such that the mean surface distance is 5mm. Despite the large change in the surface, the landmarks only vary by 0.94% and 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 [42] 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 [43].

Clearly, any attempt to reconstruct 3D face shape using 2D geometric information alone (such as in [3]–[7]) will be subject to the ambiguity. Hence, the range of possible solutions is 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.

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 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. [17] 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 [33].

Reproducible research

A Matlab implementation of the fitting algorithms, the scripts necessary to recreate the results in this paper and videos visualising the ambiguities will be made available at: www-users.cs.york.ac.uk/wsmith/faceambiguity.

References

[1] W. A. P. Smith and E. R. Hancock, "Recovering facial shape using a statistical model of surface normal direction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1914–1930, 2006.

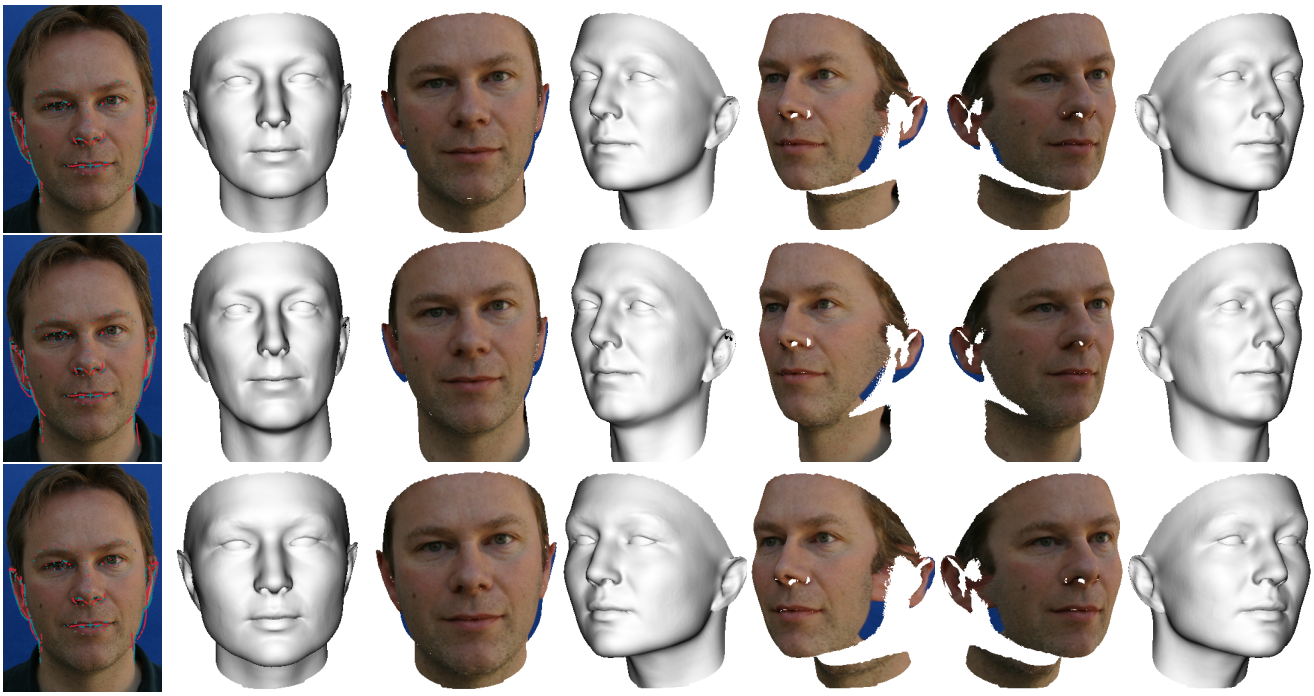


Fig. 6: Orthographic fitting with flexibility modes. 1st Row: Landmark and edge fitting under orthographic projection. 2nd/3rd Row: The first plus/minus flexibility component. Landmark distance is 0.94%, surface distance is 5 mm.

[2] E. Prados, N. Jindal, and S. Soatto, "A non-local approach to shape from ambient shading," in *Proc. SSVM*, 2009, pp. 696–708.

[3] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel, "A statistical method for robust 3D surface reconstruction from sparse data," in *Proc. 3DPVT*, 2004, pp. 293–300.

[4] O. Aldrian and W. A. P. Smith, "Inverse rendering of faces with a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1080–1093, 2013.

[5] A. Patel and W. A. P. Smith, "3D morphable face models revisited," in *Proc. CVPR*, 2009, pp. 1327–1334.

[6] R. Knothe, S. Romdhani, and T. Vetter, "Combining PCA and LFA for surface reconstruction from a sparse set of control points," in *Proc. F&G*, 2006, pp. 637–644.

[7] A. Bas, W. Smith, T. Bolkart, and S. Wuhrer, "Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences," in *Proc. ACCV Workshops*, 2016.

[8] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.

[9] X. P. Burgos-Artizzu, M. R. Ronchi, and P. Perona, "Distance estimation of an unknown person from a portrait," in *Proc. ECCV*, 2014, pp. 313–327.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. ECCV*, 1998, pp. 484–498.

[11] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, 2014, pp. 1701–1708.

[12] R. Hartley and R. Vidal, "Perspective nonrigid shape and motion recovery," in *Proc. ECCV*. Springer, 2008, pp. 276–289.

[13] M. Keller, R. Knothe, and T. Vetter, "3D reconstruction of human faces from occluding contours," in *Proc. Mirage*, 2007, pp. 261–273.

[14] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis, "3d shape estimation from 2d landmarks: A convex relaxation approach," in *Proc. CVPR*, 2015, pp. 4447–4455.

[15] T. Albrecht, R. Knothe, and T. Vetter, "Modeling the remaining flexibility of partially fixed statistical shape models," in *Proc. Workshop on the Mathematical Foundations of Computational Anatomy*, 2008.

[16] M. Lüthi, T. Albrecht, and T. Vetter, "Probabilistic modeling and visualization of the flexibility in morphable models," in *Proc. IMA Conference on Mathematics of Surfaces*, 2009, pp. 251–264.

[17] B. Amberg, A. Blake, A. Fitzgibbon, S. Romdhani, and T. Vetter, "Reconstructing high quality face-surfaces using model based stereo," in *Proc. ICCV*, 2007.

[18] R. Latto and B. Harper, "The non-realistic nature of photography: Further reasons why turner was wrong," *Leonardo*, vol. 40, no. 3, pp. 243–247, 2007.

[19] C. H. Liu and A. Chaudhuri, "Face recognition with perspective transformation," *Vision Res.*, vol. 43, no. 23, pp. 2393—2402, 2003.

[20] C. H. Liu and J. Ward, "Face recognition in pictures is affected by perspective transformation but not by the centre of projection," *Perception*, vol. 35, no. 12, pp. 1637–1650, 2006.

[21] P. Perona, "A new perspective on portraiture," *Journal of Vision*, vol. 7, no. 9, pp. 992–992, 2007.

[22] R. Bryan, P. Perona, and R. Adolphs, "Perspective distortion from interpersonal distance is an implicit visual cue for social judgments of faces," *PloS one*, vol. 7, no. 9, p. e45301, 2012.

[23] A. Flores, E. Christiansen, D. Kriegman, and S. Belongie, "Camera distance from face images," in *Proc. ISVC*, 2013, pp. 513–522.

[24] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate $O(n)$ solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[25] O. Fried, E. Shechtman, D. B. Goldman, and A. Finkelstein, "Perspective-aware manipulation of portrait photos," *ACM Trans. Graphic.*, vol. 35, no. 4, p. 128, 2016.

[26] J. Valente and S. Soatto, "Perspective distortion modeling, learning and compensation," in *Proc. CVPR Workshops*, 2015, pp. 9–16.

[27] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille, "The bas–relief ambiguity," *Int. J. Comput. Vision*, vol. 35, no. 1, pp. 33–44, 1999.

[28] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, 2001.

[29] H. Hill and V. Bruce, "A comparison between the hollow–face and 'hollow–potato' illusions," *Perception*, vol. 23, pp. 1335–1337, 1994.

[30] A. Ecker, A. D. Jepson, and K. N. Kutulakos, "Semidefinite programming heuristics for surface reconstruction ambiguities," in *Proc. ECCV*. Springer, 2008, pp. 127–140.

[31] F. Moreno-Noguer and P. Fua, "Stochastic exploration of ambiguities for nonrigid shape recovery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 463–475, 2013.

[32] M. Salzmann, V. Lepetit, and P. Fua, "Deformable surface tracking ambiguities," in *Proc. CVPR*. IEEE, 2007, pp. 1–8.

[33] W. A. P. Smith, "The perspective face shape ambiguity," in *Perspectives in Shape Analysis*, M. Breuß, F. Bruckstein, P. Maragos,

and S. Wuhrer, Eds. Springer International Publishing, 2016, pp. 299–319.
[34] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, 2003.
[35] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
[36] G. Golub and V. Pereyra, "Separable nonlinear least squares: the variable projection method and its applications," *Inverse problems*, vol. 19, no. 2, p. R1, 2003.
[37] A. Brunton, A. Salazar, T. Bolkart, and S. Wuhrer, "Review of statistical shape spaces for 3D data with comparative analysis for human faces," *Comput. Vis. Image Underst.*, vol. 128, pp. 1–17, 2014.
[38] C. Wu, D. Bradley, P. Garrido, M. Zollhöfer, C. Theobalt, M. Gross, and T. Beeler, "Model-based teeth reconstruction," *ACM Trans. Graphic.*, vol. 35, no. 6, p. 220, 2016.
[39] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. AVSS*, 2009.
[40] L. Farkas, *Anthropometry of the Head and Face*. New York: Raven Press, 1994.
[41] A. Patel and W. A. Smith, "Manifold-based constraints for operations in face space," *Pattern recognition*, vol. 52, pp. 206–217, 2016.
[42] G. Porter and G. Doran, "An anatomical and photographic technique for forensic facial identification," *Forensic Sci. Int.*, vol. 114, no. 2, pp. 97–105, 2000.
[43] K. F. Kleinberg, P. Vanezis, and A. M. Burton, "Failure of anthropometry as a facial identification technique using high-quality photographs," *J. Forensic Sci.*, vol. 52, no. 4, pp. 779–783, 2007.

Anil Bas received the B.S. degree in electronics and computer education at Kocaeli University, Turkey in 2011 and the M.S. degree in computer engineering at Selcuk University, Turkey in 2013. He is a Council of Higher Education (Turkey) Scholar and currently studying towards the Ph.D. degree in computer science at the University of York, U.K. He is a Research Associate with the Department of Computer Engineering, Marmara University, Turkey since 2012. His research interests include computer vision, face shape estimation and face analysis.

William Smith (M'08) received the B.Sc. degree in computer science, and the Ph.D. degree in computer vision from the University of York, York, U.K. He is currently a Senior Lecturer with the Department of Computer Science, University of York, York, U.K. His research interests are in shape and appearance modelling and physics-based vision. He has published more than 90 papers in international conferences and journals. He was awarded the Siemens best security paper prize at BMVC 2007 and was the finalist (UK nominee) for the ERCIM Cor Baayen award 2009. He was co-chair of BMVC 2016.