Inverse Rendering of Faces with a 3D Morphable Model

Oswald Aldrian

Submitted for the degree of Doctor of Philosophy

Department of Computer Science

The University of York

October 2012

In memory of my father †

Abstract

In this thesis, we present a complete framework to inverse render faces with a 3D Morphable Model. By decomposing the image formation process into a geometric and photometric part, we are able to state the problem as a multilinear system which can be solved accurately and efficiently. As we treat each contribution as independent, the objective function is convex in the parameters and a globally optimal solution can be found. We start by recovering 3D shape using a novel algorithm which incorporates generalisation errors of the model obtained from empirical measurements. The algorithm is extended so it can efficiently deal with mixture distributions. We then describe three methods to recover facial texture, and for the second and third, diffuse lighting, specular reflectance and camera properties from a single image. These methods make increasingly weak assumptions and can all be solved in a linear fashion. We further modify our framework so it accounts for global illumination effects. This is achieved by incorporating statistical models for ambient occlusion and bent normals into the image formation model. We show that solving for ambient occlusion and bent normal parameters as part of the fitting process improves the accuracy of the estimated texture map and illumination environment. We present results on challenging data, rendered under complex natural illumination with both specular reflectance and occlusion of the illumination environment. We evaluate our findings on publicly available datasets, where we are able to obtain state-of-the-art results. Finally, we present a practical method to synthesise a larger population from a small training-set and show how the new instances can be used to build a flexible PCA model.

Contents

List of Figures

List of Tables

Acknowledgements

I like to thank my supervisor Dr William Smith for his guidance and continuous encouragement. The many rewarding and insightful discussions have been very helpful for my personal and professional development. It was a privilege to work alongside him.

I also like to thank my assessor Prof. Edwin Hancock for monitoring my progress and providing alternative views on various matters.

A further thanks goes to my external examiner Prof. David Marshall.

I am particularly grateful to the Engineering and Physical Sciences Research Council (EPSRC) for funding this work.

And last but not least, I like to thank my family for offering their continuous support.

Declaration

I declare that the work in this thesis has solely been conducted by myself. Work of other authors is clearly referenced and acknowledged. Large parts of this thesis are based on publications by the author. Publications that originated during the course of this thesis are listed at the end of Section 1.1.

Nomenclature

Symbol	Description
$\mathbf{a}$	Shape parameter vector
$\mathbf{b}$	Texture parameter vector
$\sigma_{s,i}$	Eigenvalue for i'th principal shape component
$\sigma_{t,i}$	Eigenvalue for i'th principal texture component
$\mathbf{V}_i$	Eigenvector for i'th shape mode
$\mathbf{T}_i$	Eigenvector for i'th texture mode
$\bar{\mathbf{v}}$	Mean shape
$\bar{\mathbf{t}}$	Mean texture
$\mathbf{v}$	Shape instance
$\mathbf{t}$	Texture instance
ρ	Diffuse albedo
ω	Light source direction
ν	Viewing direction
$\mathcal{H}$	Spherical harmonic (SH) basis functions
$\mathcal{U}$	Modified SH basis functions (For RGB colour)
$\mathcal{S}$	Modified SH basis functions (Reflected about viewing direction ν)
$\mathbf{l}$	Diffuse lighting parameter vector
$\mathbf{x}$	Specular lighting parameter vector

General notations

Symbol	Description
a	Scalars are denoted as normal face letters
$\mathbf{a}$	Vectors are denoted as lower case bold face letters (If not indicated otherwise, they are column vectors)
$\mathbf{A}$	Matrices are denoted as upper case bold face letters
$\mathcal{A}$	Calligraphic letters represent either tensors or SH basis functions (depending on context)
$\mathbf{I}$	Identity matrix of appropriate size
$\mathbf{a}^{\mathrm{T}}\mathbf{b}$, $\mathbf{a} \cdot \mathbf{b}$	Scalar product between vector $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a} \times \mathbf{b}$	Cross product between vector $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a}. * \mathbf{b}$	Element-wise multiplication between vector $\mathbf{a}$ and $\mathbf{b}$
$\|\mathbf{a}\|$	Euclidian (L_2) norm of vector $\mathbf{a}$

Chapter 1

Introduction

Inverse rendering is the process of estimating contributing factors to the image formation process. The estimated quantities can be used to infer high-level knowledge about the scene content. The most common form of measurements available for vision applications are 2D photographs. Video sequences can be seen as an ordered collection of 2D photographs. The most difficult case is having a single photograph of an arbitrary object with unknown reflectance properties under unconstrained illumination and pose. The problem is further hindered by allowing arbitrary background, noisy measurements and no knowledge of the image capturing device. The solution to this setting is referred to as the *holy grail of computer vision* and has yet to be addressed within the community. The only clue that this might be feasible is the fact that humans perform remarkably well on this task [1]. Large parts of the human brain are dedicated to the vision system [2]. Besides the "dedicated hardware", humans possess a large database of objects and are capable of extrapolating and generalising their knowledge to unknown scenarios. An ability that further grows with experience.

Computer vision systems on the other hand are not even close to the performance of humans on high-level tasks. Artificial vision systems unfold their strength on quantitative evaluations. Examples include detecting changes in colour at different locations, a task known to be very difficult for humans [3]. Generally speaking, computers are good in performing low-level calculations and a researcher has to exploit this strength in order to perform well on high-level tasks. A further advantage of artificial systems is *precise memory*. We use the

phrase *precise memory* to not undermine the extraordinary memory capabilities of humans. However, a computer could store photographs (or identities) of every person on earth, leaving organisational issues aside. Humans are capable of memorising a few hundred or thousand individuals, at most [4]. Humans take a long time to build up memory and their knowledge can not be easily transferred to other humans or technical systems.

Addressing the vision problem as a whole (by emulating human vision) has turned out to be rather elusive. The amount of variation and ambiguities present in photographs inhibit the success of general purpose vision systems, at least at present. What has proven successful are problem-specific approaches. Researchers throughout the vision community have built models of objects like human organs [5; 6], vehicles [7; 8] or faces [9], to name just a few. In this thesis we focus on the object class human faces.

Given a photograph we want to know "what can be seen in that photograph?". Computer vision can be seen as the inverse of computer graphics. In computer graphics we would like to synthesise photorealistic images (or sequences of images) by defining objects (shape attributes), reflectance properties, illumination, light transport and other properties. In computer vision our aim is to estimates these quantities from either multiple images under different view points [10], or images from the same viewpoint under different illumination conditions [11], from an ordered collection of images (video sequence), or as mentioned previously, from a single photograph.

Computer vision is a quantitative discipline. On a higher abstraction level, Image Understanding or Scene Analysis uses these quantitative measurements to provide a semantic description of the scene. An image is represented as a set of features. The features can take the form of the raw measurements (RGB values). As a preprocessing step, the image can be filtered to reduce noise or enhance certain image characteristics (like edges). Features should be easy to identify, robust to extract and have the same meaning across the object of interest. We distinguish two kind of features:

Image Based The features are directly inferred from the image(s). In other words, the output is a function of the input image only. Image based methods are widely used for low-level tasks like filtering, detecting corners

and edges or morphological operations [12]. The majority of early computer and machine vision methods fall into this category. High-level tasks, like recognising and tracking objects, possibly through 3D scenes, can hardly be addressed without additional knowledge.

Model Based Most state-of-the art vision algorithms deploy some sort of model [13]. Model is an abstract term and can either refer to a class of objects (like human faces), but also to less concrete things like noise, background or other types of uncertainty. Loosely speaking a model refers to some sort of prior knowledge or assumptions. In terms of usage, they work well within their scope of assumptions. As soon as these are violated, model-based methods tend to breakdown.

Face recognition has attracted over 4 decades of research attention. It started with bottom-up strategies that extract edges, contours or silhouettes. The fragments were connected locally to "synthesise" a 2D structure, and high-level knowledge inferred from these structures were geometrical relations like: height/width of the face, distance between eyes, nose, mouth and alike [14]. Such approaches have largely been unsuccessful in unconstrained conditions, due to the drastic uncertainty present in real photographs. Feature selection is arbitrary, their detection prone to errors and the overall performance not satisfactory [15].

Combining image based and model-based methods has attracted researchers in recent years. Smith and Hancock [16; 17] combined classical shape from shading with a statistical model of surface normals. They show how the statistical model can be used to provide reasonable estimates for shadow regions. Other examples include usage of model-based contours and silhouettes [18] or shape priors.

A major challenge is to increase the generalisation ability of the model. A straightforward way to do so is increasing the number of training examples. This however is not always possible. Consider for example building a 3D human skull model from CT data [19; 20]. The acquisition of each training sample is associated with exposure to damaging radiation. Therefore, they are only performed when medically necessary. As a result, many "training examples" might show unusual deformations (caused by accidents) and are not representative of the normal population.

But there remains a different challenge, even when training examples can be obtained more easily. Consider the example of facial texture. Features like freckles and moles, even when present in the training-set, are unlikely to appear in exactly the same form in the test-set. Such features will eventually average out. As part of this thesis, we incorporate knowledge of how a shape model generalises to unseen data into a fitting algorithm and show how class specific priors reduce inference error when observations are sparse. A promising approach of learning flexible PCA models from small datasets is presented in Appendix A.

At the core of inverse rendering is the image formation process, or as in most practical cases, an approximation to it. A useful separation of constituting factors is into scene and identity parameters. In this thesis, identity is defined by the persons 3D shape (expression neutral) and diffuse albedo. Among others, these assumptions are violated by the following factors:

- Ageing
- Weight gain and loss
- Facial hair
- Makeup
- Accessories like piercings or tattoos
- Sun tan

Some of these can be modelled as external factors, and indeed the ultimate goal of inverse rendering is modelling each contributing factor separately and independently. In this thesis we consider pose, illumination and imaging device properties as purely extrinsic factors. Other sources of variability like expression are more difficult to categorise. Psychologists agree that the same expression can be mimicked by different persons [21]. On the other hand, expressions are coupled with the underlying skeletal structure and are therefore identity dependent. When separation into intrinsic and extrinsic factors is not clear, a possible solution is to model a "subspace within a subspace" [22]. On a larger scale, the problem can be addressed using multilinear methods [23; 24] at the cost of requiring an exponentially growing training-set. Subspace identity models [22] have also been

applied to deal with pose and illumination. We believe however, that if a factor can be modelled independently, it should be done.

Inverse rendering is interesting in itself. It is a challenging task and requires careful design. This is particularly true in the case of faces, as humans are natural experts in judging the result. The accuracy of a result is usually measured by some sort of reconstruction error. Overall, inverse rendering of faces does not differ from other ill-posed problems. But the fact that humans can instantly judge whether a result looks "good or not", makes the value of the objective function a secondary measure. As such, inverse rendering of faces is more challenging than other high dimensional problems where visual results *"plotting the data"* is less intuitive.

There are plenty of interesting applications once parameters are extracted. Identity parameters can be used for access control to buildings or ATM machines. Because of their compact representation, they can be used for scene coding (images, videos or 3D scenes), with applications in storage and low bitrate communication. Altering scene parameters allow for face relighting and pose change which is of interest to the computer games and movie industry. Other applications are in the area of human computer interaction or video driven animation.

1.1 Contributions

Both shape and reflectance properties contribute to appearance. Solving for both simultaneously leads to a non-convex optimisation problem [25] which is notoriously difficult to solve. In this thesis, we instead propose estimating shape using geometric features alone (e.g. the position of sparse feature points or silhouettes and edges) in a manner which is independent of illumination and reflectance effects. With a shape estimate to hand, we are then able to derive linear methods for reflectance and illumination analysis. This thesis makes several contributions in the area of 3D face modelling and fitting. The major theme of the thesis is the decomposition of the inverse rendering pipeline into distinct components which are formulated independently. The image formation process is described as an interaction of these components in either additive or multiplicative fashion and takes the form of a multi-affine system. Each component is stated as a convex

objective function which can be solved efficiently to a global optimum assuming independence of contributing factors. In real world scenarios, these assumptions are often violated. We adequately address the fact of the ill-posed nature by using prior terms for shape texture and illumination which guide the solution to a plausible one. In particular, we present the most general fitting algorithm for 3D morphable models which takes arbitrary complex illumination and global shading effects into account. Moreover, our approach is intuitive and highly efficient. The novelties proposed in this thesis have lead to the following publications:

Journal Paper

- Inverse Rendering of Faces with a 3D Morphable Model

  O. Aldrian and W.A.P. Smith. in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, (accepted)

Conference Papers

- Inverse Rendering of Faces on a Cloudy Day

  O. Aldrian and W.A.P. Smith. in Proc. ECCV 2012

- Model-based Ambient Occlusion for Inverse Rendering

  O. Aldrian and W.A.P. Smith. in Proc. ICIAR 2012

- Inverse Rendering with a Morphable Model: A Multilinear Approach

  O. Aldrian and W.A.P. Smith. in Proc. BMVC 2011

- Inverse Rendering in SUV Space with a Linear Texture Model

  O. Aldrian and W.A.P. Smith. in ICCV 2011, Workshop on Color and Photometry in Computer Vision

- A Linear Approach of 3D Face Shape and Texture Recovery using a 3D Morphable Model

  O. Aldrian and W.A.P. Smith. in Proc. BMVC 2010

- Learning the Nature of Generalisation Errors in a 3D Morphable Model

  O. Aldrian and W.A.P. Smith. in Proc. ICIP 2010

1.2 Thesis structure

Face recognition (verification) under unconstrained conditions is still an unsolved problem [26]. In this thesis, we further push the boundaries towards that goal. This section lists the achievements we made towards enlarging the domain of applicability of face recognition and image understanding on a chapter by chapter basis.

In Chapter 2 we review the relevant literature in face modelling in a broader context. We work our way from 2D approaches to 3D extensions. The type of model used in this thesis is a 3D morphable model, which is explained in greater detail. Subsequently, we discuss different methods for fitting models to image data. This is followed by a focussed literature review of work that is directly relevant to this thesis.

In Chapter 3 we propose a novel method to reconstruct 3D shape from a sparse set of feature points using and iterative algorithm. We incorporate knowledge inferred from out-of-sample data into the reconstruction algorithm and show how this can be used to reduce generalisation error. We show how to do this using a linear method which achieves an accuracy which is competitive with far more complex and computationally expensive state-of-the-art analysis-by-synthesis approaches [27]. Inference error can further be reduced by using class specific priors, in particular when observations are sparse.

Chapter 4 focusses on texture and illumination modelling. We present three efficient approaches which allow recovery of texture model parameters (and in the second and third cases, diffuse and specular reflectance properties) under unknown, arbitrarily complex illumination. The methods make increasingly weaker assumptions at the expense of a slight increase in complexity. Nevertheless, all three allow the global optimum to be obtained. The first two approaches exploit photometric invariants. The first assumes that reflectance consists purely of diffuse Lambertian and that the light sources in the scene are all of the same, known colour. In this case, texture can be recovered using linear least squares. The second method relaxes the reflectance assumption and allows for an additive specular term. We use a specular invariant colour subspace and again assume that all light sources in the scene have the same colour. We use spherical har-

monics to model the illumination environment which leads to a bilinear system in the texture parameters and spherical harmonic coefficients. The global optimum can be obtained iteratively using alternating least squares. We also estimate the specular reflectance function by fitting a higher order spherical harmonic basis to the specular difference image obtained by subtracting the estimated diffuse image from the input. The last method makes no assumption about the light source colours yet still allows arbitrarily complex environment illumination. Using the same assumptions about specular reflectance, this leads to a multilinear system in the texture parameters, spherical harmonic coefficients and specular reflection parameters. We propose two ways to regularise the problem by encouraging the environment to be grey or simple. Our approach makes the weakest assumptions of any available algorithm for fitting a morphable model.

In Chapter 5 we modify the image formation process and incorporate global illumination effects. We propose four efficient methods using ambient occlusion and bent normals. Ambient occlusion is a phenomenon observed when non-convex objects are illuminated by ambient light. Bent normals differ from surface normals and point in the direction of least occlusion with respect to the upper hemisphere. By doing so, we circumvent a systematic error inflicted by inverse spherical harmonic lighting of non-convex objects. The four methods are of increased expressiveness in terms of approximating global shading effects. We build statistical models of ambient occlusion and bent normals and model the parameters of the training samples jointly with shape attributes (3D vertices and surface normals). This allows to predict ambient occlusion and bent normals given a shape estimate. To account for the error we make in shape estimation, we further relax the static relationship and estimate ambient occlusion parameters as part of the fitting process. The methods increasingly reduce modelling error for diffuse texture and the lighting environment and are formulated such that in each case a unique solution is obtained.

In Appendix A we investigate ways to address problems with linear 3DMMs. We show how a larger population can by synthesised from a small training-set. The new samples are locally consistent with what has been observed in the training-set. We use these shapes to construct a 3D morphable model and test its generalisation ability to unseen data.

Chapter 2

Literature Review

This chapter reviews and summarises face recognition and modelling, starting with eigenfaces introduced by Turk and Pentland [28] until current state-of-the art 4D morphable models [29]. The models we discuss are generative and model a probability distribution over the data. This is in contrast to discriminative methods, which infer the world state directly from the observations. The main disadvantage of discriminative algorithms is, that they can not be used for synthesis tasks. In other words, they can not be used for face modelling. It is also far more difficult to incorporate prior knowledge into discriminative methods [13].

The first part of this chapter reviews models and fitting strategies at a high level. In the second part of the chapter (starting from Section 2.6), we review state-of-the art approaches to shape, reflectance and illumination modelling. Depending on the model, the construction process can be as simple as applying principal components analysis (PCA) to a set of 2D images showing faces in the same pose [28]; the other extreme requires high resolution 3D meshes of a large number of the same and different individuals with varying expressions to be brought into dense correspondence [30]. This is a notoriously difficult problem with an active research community. One of the challenges lays in the fact, that the definition of the problem is ill-posed in itself [31]. After the training phase, the models can be used to synthesise new faces, for example virtual characters in video games or movies. A different application is in forensics, where they can assist in creating identities of suspects. The second application domain of statistical models is image analysis, where the model is used to constrain the set of plausible

solutions. This approach is widely adopted for image segmentation. Regardless of the domain, model-based image analysis require a fitting strategy. The fitting strategy depends partly on the type of model which is used. Nevertheless there is a lot of common ground across models, and in the second part of this chapter we provide an overview of model independent fitting principles.

2.1 2D Models

A most basic face model was introduced by Turk and Pentland [28] in 1991. Although statistical face modelling started earlier (see for example [32]), [28] is widely regarded as the inauguration of face modelling. The model is built from a set of k 2D frontal images. The images are of the same resolution (e.g. 256×256), with the face centre-aligned. Each image $\mathbf{I}_i \in \mathbb{R}^{65536}$ is treated as a one-dimensional feature vector. A mean face $\mu = \sum_{i=1}^{k} \mathbf{I}_i$ is computed and subtracted from each instance. Decomposing the mean-free sample matrix $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$ results in the eigenfaces $\mathbf{U}$ with their corresponding variance $\mathrm{diag}\{\mathbf{\Sigma}\}$. Despite its many limitations, which we will discuss in detail shorty, the paper was widely accepted within the community. Many of the more recent methods are based on this, so called, top-down approach. The following lists some of the limitations of the eigenfaces approach.

- Frontal only, Or profile only etc.
- Features not in correspondence
- Neglecting image formation process (3D)
- Images taken with different cameras
- Illumination not accounted for
- Does not account for expressions
- Assuming faces are unimodal distributed
- Non-probabilistic
- Purely model-based

Many of the above mentioned drawbacks have been addressed in subsequent work. It is also important to mention, that some of the problems (for example illumination or pose invariance) can be approached in different ways. The

aim of this chapter is to summarise the most important developments that have taken place within the last decade. The material is treated conceptually, using mathematical rigour only when necessary.

An intuitive approach to extend [28], is to build separate models for different pose angles or illumination conditions. Besides being a very ad-hoc solution, the number of required training examples grows rapidly and there remains the problem of "in between" variation. Models that address different types of variation at the same time are often termed: "Models for Style and Identity". In general, there exist two types of interaction between the modes (additive and multiplicative). Additive methods, like Linear Discriminat Analyses [33] can often be reduced to a single Factor Analyses [22] model, by modelling the parameters jointly. Multiplicative interaction can be stated as multilinear systems. Note, that bilinear systems are just a particular member of multilinear systems and will not be treated separately. Both types of models are used throughout this work and will be explained in greater detail in subsequent chapters.

The face-space spanned by the eigenfaces (including the mean face) show a characteristic blurriness. There are several reasons for this fact; however the most prominent is lack of feature correspondence. The pixel locations in the training-set do not correspond to unique facial features. The problem is further propagated by the lack of correspondence in the test-set. Feature correspondence in 2D has been addressed by Cootes et al. with the introduction of Active Shape Models [34]. A more intuitive name is "Point Distribution Model". The idea is to model the variability of a class of object which can be described by a particular set of features. Each feature corresponds to a unique attribute of the object. During the learning phase, each training example has to be labelled (which can be assisted by an automatic feature detector). Because Active Shape Models (as introduced in [34]) only define a set of sparse 2D spatial locations, they are not suitable for modelling photo realistically appearing human faces.

The merging of eigenfaces with point distribution models has resulted in the development of "Active Appearance Models (AAM)" [35] [36], and indeed the core principle of this idea is visible in the most sophisticated face models seen today. AAMs are very popular in computer vision. They can be constructed from 2D datasets which are widely available [37; 38; 39]. For a detailed description of

2D datasets, we refer the interested reader to Gross [40].

Evaluating and comparing recognition experiments is non-trivial. As an example, assume we have a dataset of n photographs acquired at research institute α under setup (pose, illumination, etc.) β with equipment γ. Now we train a model on a subset of n (for instance 90%) and use the remaining 10% as a test-set. The reported performance of a particular algorithm can be very high. It is possible however, that the reason therefore is that the training-set and test-set are to "similar". In a more realistic approach, we want to train our model on *Set 1* $= \{\alpha_1, \beta_1, \gamma_1\}$ and test the performance on *Set 2* $= \{\alpha_2, \beta_2, \gamma_2\}$, where $\alpha_1 \neq \alpha_2, \beta_1 \neq \beta_2, \gamma_1 \neq \gamma_2$. Under these assumptions, many of the 2D approaches perform poorly.

Many of the aforementioned limitations can be addressed by switching to a 3D face model [41]. The following section describes this in greater detail.

2.2 3D Morphable models

The core model used in this thesis is a 3D Morphable Model (3DMM). This section shortly reviews the construction process of 3DMMs. For a detailed description see Blanz [42; 43]. A 3DMM is a generative model which models 3D shape and diffuse albedo in a low dimensional linear subspace. They represent the state-of-the-art in 3D face modelling and recognition. However, their construction is time consuming and labour intensive. The training data should represent the target population as accurate as possible with respect to different variations like gender, age, ethnical origin, weight etc. Indeed, many researchers argue: "Its only weakness is the requirement of the 3D models" [44]. Due to this fact, Paysan et al. made their 3DMM, The "Basel Face Model" (BFM) publicly available for non-commercial purposes [27]. This offers new opportunities for the community. However, the published model does not include individual scans, but only PCA components. The only way the BFM can be used to build ones own statistical model or attribute specific priors, is by sampling from the model. Most of the experiments conducted in this theses use the BFM.

2.2.1 3D Data acquisition

Constructing a 3DMM requires training data which represents the target population as well as possible. 3D shape and texture can be acquired with different modalities. They can be divided into passive and active techniques. Important features are, among others, acquisition time and accuracy. Acquisition time is crucial as movement of the head leads to inaccuracies and artefacts. This is particularly the case when scanning expressions, due to the difficulty of holding the same expression for a longer time.

Passive techniques, like stereo vision, shape-from-shading or shape-from-motion usually lack adequate accuracy for 3D face modelling [45]. Spatial resolution for these techniques are in the mm range. Active techniques differ from passive ones, due to emission and reception of signals. Especially for poorly textured regions like the cheeks or forehead, active techniques outperform passive ones. We distinguish two types of active techniques :

Time-of-flight: Measure the time of an emitted signal after it is reflected from the object. As face scanning is performed on a very low range of depths, sensors require nanosecond timing for accurate surface reconstruction.

Triangulation: Work by projecting a structured light pattern onto the scene. Depth can be reconstructed by triangulation between projector-to-camera or camera-to-camera. Accuracy in μm can be achieved for short range measurements.

With current available devices, active systems based on triangulation offer higher accuracy compared to systems based on time-of-flight [45]. Regardless of the data acquisition technique used, the scans have to be processed to remove scanning artefacts like holes and spikes. Another preprocessing step is the removal of regions which are not of interest (for instance: the back of the head or regions behind the ears), which is often done manually [9; 42; 43]

2.2.2 Constructing a face-space

3DMMs are constructed from a set of face scans that are in *dense correspondence* (see Figure 2.1). Solving the correspondence problem is a key step in morphable

model construction. The difficulty arises mostly for unstructured regions like forehead and cheeks, where meaningful correspondence is hard to define (even for human experts). Standard optical flow algorithms have turned out to be unsuitable (at least without modification). Recent approaches based on group-wise mesh processing have shown promise [46]. Solving the correspondence problem is an active research area and not addressed within this thesis. We refer the interested reader to [31; 47; 48].

When solved adequately, every face mesh has the same topology (number and order of vertices) and can be embedded in a vector space. Every vertex in each scan corresponds to the same anatomical landmarks (see Figure 2.1) and is associated with an RGB colour. Each mesh consists of p vertices and is written as a vector $\mathbf{v} = [x_1\ y_1\ z_1\ \ldots\ x_p\ y_p\ z_p]^{\mathrm{T}} \in \mathbb{R}^n$, where $n = 3p$. Applying principal components analysis to the data matrix formed by stacking the m meshes yields $m-1$ eigenvectors $\mathbf{V}_i$, their corresponding variances $\sigma_{s,i}^2$ and the mean shape $\bar{\mathbf{v}}$. An equivalent model is constructed for surface texture (or more precisely, diffuse albedo). Any face can be approximated as a linear combination of the modes of variation:

$$\mathbf{v} = \bar{\mathbf{v}} + \sum_{i=1}^{m-1} a_i \mathbf{V}_i, \qquad \mathbf{t} = \bar{\mathbf{t}} + \sum_{i=1}^{m-1} b_i \mathbf{T}_i,$$

where $\mathbf{a} = [a_1\ \ldots\ a_{m-1}]^{\mathrm{T}}$ and $\mathbf{b} = [b_1\ \ldots\ b_{m-1}]^{\mathrm{T}}$ are vectors of shape and texture parameters respectively. Representing faces as a decorrelated subspace model also allows for dimensionality reduction by discarding lower order principal components, which are likely to model noise in the training data. For convenience, we also define the variance-normalised shape/texture parameter vectors as: $\mathbf{c}_s = [a_1/\sigma_{s,1}\ \ldots\ a_{m-1}/\sigma_{s,m-1}]^{\mathrm{T}}$ and $\mathbf{c}_t = [b_1/\sigma_{t,1}\ \ldots\ b_{m-1}/\sigma_{t,m-1}]^{\mathrm{T}}$. The coefficients $c_{s,i}$ and $c_{t,i}$ are normally distributed with zero mean and unit variance. Figure 2.2 shows the variation of the first two shape and texture principal components for $+3\sigma_s/\sigma_t$ and $-3\sigma_s/\sigma_t$.

3DMMs should be restrictive such that unlikely faces are rarely instantiated. To do so, the statistics of the training data can be used to calculate the likelihood

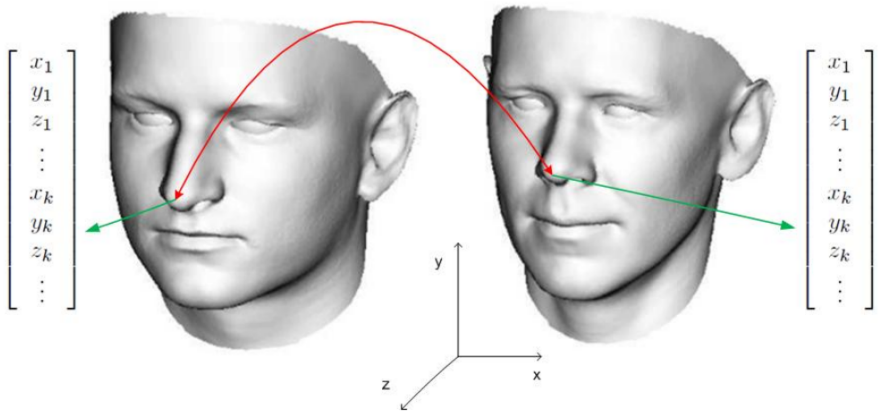

Figure 2.1: Correspondence is achieved by assigning the vertex coordinates an indexed geometry. This allows certain features of the face (e.g the tip of the nose) always to have the same vertex index (e.g $k = 35345$), although they have different x, y, z values.

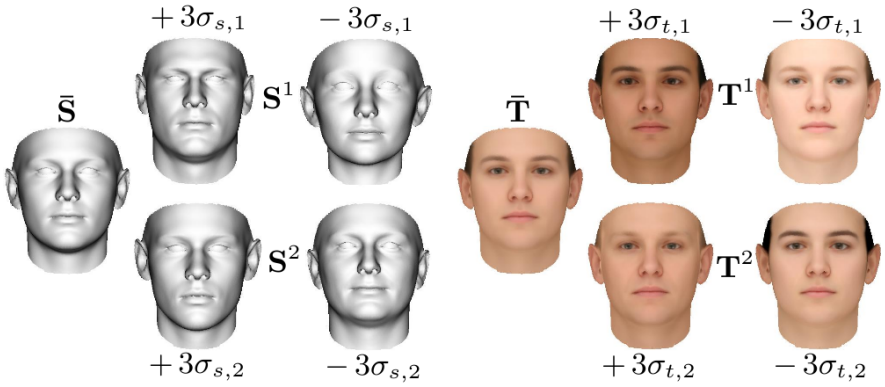

Figure 2.2: The mean shape and texture together with the variation of the first two principal components by adding $\pm 3\sigma_{s,t}$.

for a particular shape and texture:

$$p(\mathbf{v}) \sim e^{-\frac{1}{2}\sum_i \frac{c_{s,i}^2}{\sigma_{s,i}^2}}, \quad p(\mathbf{t}) \sim e^{-\frac{1}{2}\sum_i \frac{c_{t,i}^2}{\sigma_{t,i}^2}}.$$

Experiments on real world data have shown that assuming Gaussian distributions of faces yields good results when sufficient information is available at inference stage. However, it is counter-intuitive to assume the face-space to be Gaussian distributed. Models build on this principle assign the "mean face" the highest

likelihood. The mean face however appears gender neutral. And the probability of observing a gender neutral face in practice is very low. Patel and Smith [49] make a different assumption. They observed that the length of the parameter vectors follows a Chi-square distribution. In other words, realistically appearing faces lay on a spherical manifold. In their model, the mean face (which has vector length zero) is unlikely, and faces are forced to possess a certain level of distinctiveness.

2.2.3 Segmented morphable model

A 3DMM constructed from m face scans, can make use of $m-1$ principal components for face modelling. In some cases, e.g. if the 3DMM is constructed from a low number of samples, this might not be sufficient for accurate modelling. A trivial way to overcome this problem is to increase the training data. However this might not be always feasible. A different way is to segment the model into distinct (non-overlapping) regions. In [9] this was firstly done for nose, eyes, mouth and the remaining face. Faces can now be instantiated using $N_S = 4(m-1)$ coefficients. Segmentation can be performed for shape and texture. Segmenting the morphable model increases flexibility but makes the fitting process more complex and time consuming.

2.2.4 The Basel Face Model

Building a morphable model is time consuming and labour intensive. To spur research in the community, Paysan et al. [27] have made their 3DMM available to the public. Statistical models can only synthesise what has previously be seen in the training data. The BFM was trained on 3D face scans from 200 people (half male, half female). The age per individual ranges between 8 to 62 years, with an average age of 24.97 years. The weight of the subjects lays between 40 and 123 kilogram. The average weight is 66.48 kilogram.

At the time the BFM was published, there existed only two comparable morphable models. The MPI-M [9] and the USF-M [50]. We decided to use the BFM for our experiments. According to [27], their model is superior to the other two. It is constructed using an ABW-3D structured light system with an ac-

quisition time of $\sim 1s$. For comparison, a Cyberware 3030 laser scanner takes $\sim 15s$ to acquire a scan. This results in lower scanning artefacts. The resolution of the model is 53490 vertices, where each vertex i is represented as a six tuple $(x_i, y_i, z_i, r_i, g_i, b_i)$. It is claimed, that the registration algorithm used to construct the model outperforms existing methods. There are other features which made the BFM the model of our choice. It is complemented by (among others):

- Segmentation mask for 4 regions:
  1. mouth
  2. eyes
  3. nose
  4. the remaining part of the face
- Principal directions for gender, age, height and weight
- Position and index of 70 Farkas [51] feature points
- A list of symmetric vertices
- Ten registered out-of-sample scans
- 270 Renderings of the scans (incl. visible Farkas points)
- Recognition results on CMU-PIE database

A drawback to mention is, that the training data is not provided with the model. This would have been advantageous for our experiments in Section 3.3, where we explore how class specific priors and multimodal distributions can reduce inference error, when observations are sparse.

2.2.5 Image formation process

A vertex i is represented as a 3D coordinate $[x_i \ y_i \ z_i]^{\mathrm{T}}$ and a corresponding texture (diffuse albedo) $[r_i \ g_i \ b_i]^{\mathrm{T}}$. Both can be embedded in a reference frame as a function of 2D (u, v) coordinates, see Figure 2.3.

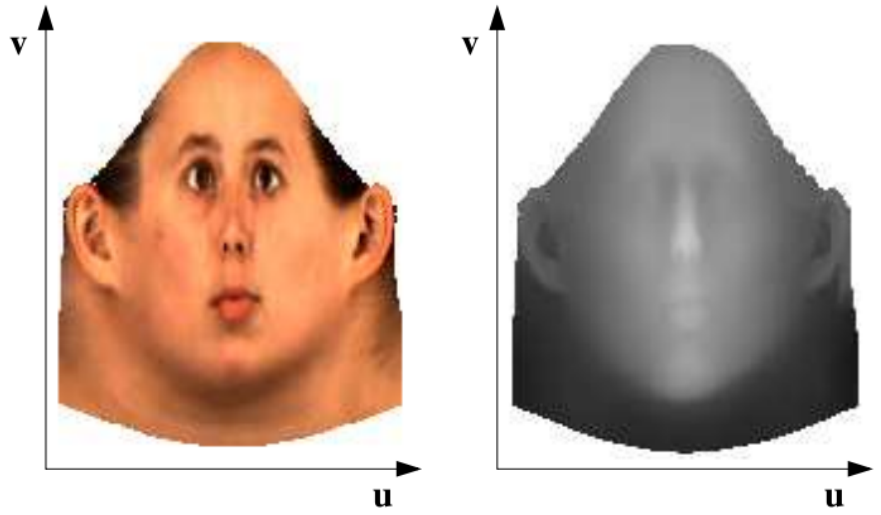

Figure 2.3: Texture and 3D shape represented in a (u, v) reference frame. The image is courtesy of [43]. Representation shown uses cylindrical embedding. Others are possible.

Shape and camera modelling

To form an image, we map the 3D vertices from the reference frame (u, v) to the image frame (x, y). This can be done via a camera projection matrix $\mathbf{C} \in \mathbb{R}^{3 \times 4}$. We commonly distinguishes three types of camera matrices:

Orthographic Camera: Is an affine camera which performs a parallel projection. The camera centre is assumed at infinity. The scaling factor for the x and y component is the same.

Weak Perspective Camera: Is also an affine camera which performs a parallel projection. The camera centre is assumed at infinity. The scaling factor for the x and y component differ from each other.

Projective Camera: A full projective camera, as modelled by a pinhole camera, performs a perspective projection. The distance from the image plane to the camera centre, the focal length f determines the level of perspective distortion.

In each case, the camera matrix performs rotations ϕ, θ and γ which correspond to a vertical, a horizontal and a rotation around the camera axis. The

camera matrix also accounts for scaling and translation of the object. In [52], the image formation process is referred to as *Shape Projection* and denoted by the vector valued function: $(x_i, y_i) = \mathbf{p}(u_i, v_i, \mathbf{a}, \rho)$, where $\mathbf{a}$ is the shape parameter. The vector $\rho = [f, \phi, \theta, \gamma, t_x, t_y, \mathbf{t}_w^{\mathrm{T}}]^{\mathrm{T}}$ contains the projection parameters, where t_x and t_y defines the image plane position of the optical axis, and $\mathbf{t}_w$ is a 3D translation. Projecting a 3D object to a 2D view-plane must account for vertices which are not visible after the projection. Two types of occlusion are possible:

Back-facing vertices: Vertices which are back-facing with respect to the observer (camera) are culled. Back-facing vertices can easily be identified by evaluating the inner product $(\mathbf{n} \cdot \mathbf{l} > 0)$ between the normals, $\mathbf{n}$, and the viewing direction, $\mathbf{l}$.

Self Occlusion: Accounts for vertices that are not visible because they are occluded by other parts of the object. Identifying these vertices is computational more expensive than in the previous case. Common ways to tackle this problem use a Z-Buffer [53], which can be efficiently implemented in hardware.

The domain of vertices that are visible after the projection is denoted by $\Omega(\mathbf{a}, \rho) \in (u, v)$. The aforementioned process is used to project a 3D object to a 2D viewing plane. In order to construct a 3D model (u, v)-space from a 2D image (x, y)-space the process has to be inverted. This is referred to as *Inverse Shape Projection*: $(u_i, v_i) = \mathbf{p}^{-1}(x_i, y_i, \mathbf{a}, \rho)$ and maps an image point to the reference frame [43; 52].

Illumination and texture modelling

An observed pixel of a facial photograph is not only a function of the persons diffuse albedo. Depending on the type, number, direction and colour of the light sources, the pixels vary. Illumination can have drastic effects on the appearance of facial images. Ho and Kriegman [54] demonstrate an example, where the same person is illuminated in four different lighting conditions (*Set 1*). In the same demonstration, four different persons are illuminated in the same lighting condition (*Set 2*). Taking the sum of squared differences of pixel intensities, any

image in *Set 1* compared to any other image in *Set 1* yields a higher error than comparing with any image in *Set 2*. This can be interpreted in two ways. Firstly, how significant the presence of illumination is on the appearance of faces. And secondly, that using the L_2 norm as means of comparison for identities in images might not be the best measure.

One way to address the problem is to transform images into an illumination invariant representation. The mapping however is not bijective per-se and results in loss of information. A more promising way to tackle the problem is to model illumination explicitly. In 3D face modelling two different approaches are commonly used.

Physical Modelling: This approach emerged from the graphics community. Physical modelling is the intuitive way of modelling illumination effects. Objects and light sources are distributed in a 3D space. This allows modelling of cast and attached shadows as well as specular highlights very accurately.

Statistical Modelling: A different approach models illumination in a low dimensional subspace. This approach treats a Lambertian surface as a low pass filter which turns high frequency components into shading effects. The principle is similar to a Fourier transform, however the orthogonal basis functions are functions defined on a sphere.

Interaction of light with skin is a very complex phenomenon that depends on many factors. Human skin is composed of many layers. This leads to subsurface scattering which is further dependent on wavelength. Appearance of human skin depends on biological factors like melanin and haemoglobin levels, among others. Skin composition varies spatially, which makes the modelling process even more difficult. Because of the difficulty, the face recognition community has focused on simpler approaches like the Lambertian and the Phong reflectance model. The Lambertian model models the interaction of a distant light source with surface geometry:

$$I(\omega) = \rho \max(\omega \cdot \mathbf{n}, 0),$$

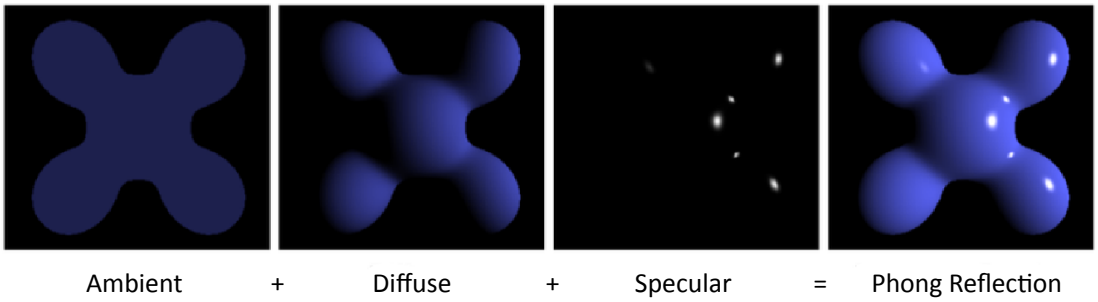

Figure 2.4: Comparison of an geometric object rendered using ambient diffuse and specular reflection. In combination this is known as Phong reflectance (Image courtesy of Brad Smith).

where ω is the strengths and direction of the light source and $\mathbf{n}$ the surface normal. The inner product of the the two factors is scaled by the albedo value ρ. Although pixel intensity in the Lambertian model is independent of viewing direction, which does not allow modelling of specular highlights, the model is used frequently because of its efficiency. A more accurate approach, which further models specular highlights and ambient light is the Phong model. Here, the intensity of an observed pixel is view-point dependent. Intensity values are modelled as:

$$I(\omega) = a\rho + \rho \max(\omega \cdot \mathbf{n}, 0) + s(\omega, \mathbf{n}, \nu), \tag{2.1}$$

where a is the intensity of ambient light and the term $s(\omega, \mathbf{n}, \nu)$ models specular highlights. As opposed to the Lambertian model, surface patches that are in shadow are illuminated by ambient light. The specular colour only depend on the colour of the light source. Figure 2.4 shows a geometric object rendered using ambient, diffuse and specular reflection.

2.2.6 Fitting algorithms

For graphics applications 3DMMs can be used to synthesise photorealistic 3D scenes or 2D images. The user provides scene transformation parameters, light sources, etc. to create the images. In inverse rendering, we are given a photograph and estimate all parameters (intrinsic and extrinsic) using a fitting strategy.

According to [55] fitting algorithms can be evaluated according four major characteristics:

Accuracy Fitting accuracy might be the most crucial feature and can be assessed by different measures. If the identity of the person is known, accuracy can be measured on the entire mesh, or on the identity parameters **a** and **b**. A common quantitative measure is the L_2 difference. Other accuracy measures include angular distance, which is a more qualitative measures.

Efficiency Some applications require short computation times. For instance: access control to buildings / restricted areas or identification at country borders. Efficiency is one of the major limitations of currently available fitting algorithms.

Robustness In some real worlds scenarios, humans can not be assumed to be cooperative. Often parts of the face are occluded or poorly illuminated. In addition, accessories like glasses, facial hair or cosmetics, can hamper the recognition process.

Automation An ideal fitting algorithm would not require human intervention. Faces and facial feature should be extracted and located automatically. If not present in the fitting algorithm itself, feature detectors can be used as preprocessing step.

A further characteristic is the *domain of convergence* [55]. This property is related to the "landscape" of the cost function to be optimised. Many fitting algorithms introduced for morphable models follow an analyses-by-synthesis framework. Here shape and texture, in conjunction with scene parameters are optimised within the same objective function, which is littered with local minima. The non-convex objective function still has a global minimum. However, in order for the optimisation algorithm to find this extrema, we must initialise close to this point. So generally speaking, the domain of convergence is small. A better way to formulate the problem would be according to a convex rule set, also known as disciplined convex programming [56]. Here, the building blocks of the objective function are convex, and only operators which preserve convexity are applied thereon.

2.2.7 Summary

The advantages that come with morphable models are associated with an expensive construction process. The awareness of the community for this problem has led research institutes to release parts of their models to the public to drive and enhance further advances. Skin reflectance is a complicated process, often approximated by simpler models like the Lambertian or Phong model. The effect of illumination as a source of variation in images is striking. Modelling illumination accurately is crucial. 3D morphable models reveal their full strength, when they are embedded in fitting algorithms where intrinsic parameters are estimated together with scene parameters like camera properties and illumination. Fitting algorithms are trade-offs between different characteristics. As a rule of thumb, highly accurate fitting algorithms might be low in efficiency, and highly efficient fitting algorithms lack accuracy.

2.3 3DMM Extensions

Progressing from 2D models to 3D models have shown to improve reliability of face recognition. The advantage lays in the fact, that the results can be pose and illumination normalised. This makes comparison of identity parameters trivial. The concept can be further extended to other sources of variability. As a motivating example, we use facial expressions.

We discuss four approaches to address 3D assisted face recognition when facial expressions are present in the input images. We will later discuss how the four approaches readily translate to other sources of variation, in our case illumination.

Expressions are an important source of human communication, and convey information about a persons emotional state. Ekman et al. [57] identify 6 universally accepted expressions (anger, disgust, fear, happiness, sadness, surprise). Note, that when expressions are modelled implicitly, they can be altered or transferred to a different individual, by keeping other parameters, like identity and pose constant [58].

2.3.1 Physical modelling

Rigid transformations of faces can well be modelled physically via a camera matrix (see Section 2.2.5). On the other hand, non-rigid shape deformations conform one of the major problems in face recognition. Facial Action Coding System (FACS) [59] models expressions physically using 46 action units, which are associated with facial muscular structure. This approach is widely used for character animation in the games and movies industry. Physical models have the advantage, that they make sense intuitively. In other words, they model the underlying cause of appearance. Expressions are caused by muscle action which causes skin movement. Although reasonable for modelling expressions in 3D it is difficult to come up with a fitting strategy for 2D images.

2.3.2 Invariance

The problem of expression invariance was also addressed by Bronstein et al. [60]. They make use of a concept called "Canonical Position", which transforms the 3D shape into a state that preserves identity. Recognition is then performed in the new representation, by comparing features. The disadvantage is, that expressions can not be transferred or altered as they are not modelled explicitly. A further disadvantage is, that the transformation is usually not bijective and results in loss of information.

2.3.3 Multiplicative

A different way to model expressions is using statistical frameworks. We discuss the multiplicative approach first, which are also termed multilinear models [61; 62; 63]. A multidimensional matrix is called a tensor. A model which accounts for identity and expression can be represented as a tensor of order three. The first mode holds the vertices, the second mode holds identity and the third mode expressions. The first mode does not model a source of variation, but requires 3D shapes to be represented in a common vector space (they are in dense correspondence).

A tensor requires a full dataset for each attribute, as opposed to the additive

example discussed in the next section (2.3.4). In multilinear models, every expression must be present for every single subject. The dataset increases rapidly if more modes of variation are taken into account. When data is missing, the gaps have to be filled. A demonstration of video driven animation with multilinear models (including the missing data problem), was demonstrated by Vlasic et al. [24]. Although an experienced observer can make out the manipulation, the approach looks very promising.

Variation of attributes independently requires decomposition of the tensor. Naturally, decomposing a rank-2 tensor (Matrix) is carried out via singular value decomposition (SVD). An extension to higher order tensors, is the so called N-mode SVD. The data tensor $\mathcal{D}$ is decomposed into a core tensor $\mathcal{C}$ and orthogonal matrices $\mathbf{M}_x$ which transform the core tensor according to the $\times_x$ mode of variation:

$$\mathcal{D} = \mathcal{C} \times_1 \mathbf{M}_n \times_2 \mathbf{M}_e.$$

Here $\mathbf{M}_n$ represents the space of identities (expression neutral) and $\mathbf{M}_e$ the space of expressions. A particular instance with identity $\mathbf{a}_n$ and expression $\mathbf{a}_e$ can be composed via a tensor vector product:

$$f(\mathbf{a}_n, \mathbf{a}_e) = \mathcal{C} \times_1 \mathbf{a}_n \times_2 \mathbf{a}_e.$$

Compared to matrix SVD, N-mode SVD, does not result in an optimal solution and further refinement is required [24]. Unfortunately, many matrix SVD properties do not hold for N-mode SVD. Tensors can easily be extended to higher orders, for example including illumination [64].

2.3.4 Additive

Amberg et al. [65] approached the problem of expression invariant face recognition in an additive way. They built a statistical model of 270 subjects scanned in neutral expression. In addition, a subset of the 270 persons was scanned in various different expressions. Applying PCA to the expression neutral data leads

to the well know approach:

$$f(\mathbf{a}_n) = \mu + \mathbf{M}_n\mathbf{a}_n.$$

A separate statistical model is built from the expression scans. This is done by subtracting neutral scans from the expression scans (for matching subjects only). PCA is then applied to the data matrix, which yields a statistical expression model centred around the mean (neutral) expression. The expression eigenmodes are denoted, $\mathbf{M}_e$. A new face with identity vector $\mathbf{a}_n$ and expression vector $\mathbf{a}_e$ can now be instantiated from the model as follows:

$$f(\mathbf{a}_n, \mathbf{a}_e) = \mu + \mathbf{M}_n\mathbf{a}_n + \mathbf{M}_e\mathbf{a}_e. \tag{2.2}$$

This formulation allows for an implicit separation of identity and expression. For systems based on neutral data, expression normalisation can be performed as a preprocessing step. The assumption, that expressions can be transferred between individuals is strictly speaking not correct. Nevertheless experiments showed good recognition rates on faces with expressions. On neutral faces, recognition rates are lower compared to systems designed for neutral expressions. Expression normalised faces in [65] often lack their distinctive features and appear "mean-like". This might be a general problem of PCA based approaches. If the information inferred from the 2D image is low, the models depend on prior knowledge and solutions close to the mean face are always preferred.

2.3.5 Summary

We briefly outlined four different ways how to address expression invariant face recognition. On a conceptual level, the same ideas can be applied to different sources of variation, for instance illumination. Physical models are suited for simple situations, for example single point light source. For complex illumination, physical models are exhaustive. The problem can be completely side-stepped by transferring the image into an illumination invariant. The choice between multiplicative or additive is not as trivial. Sometimes it is determined by the physical interaction. Take for instance Equation 2.1. Texture and diffuse lighting

are coupled in a multiplicative way, whereas ambient and specular parts are added.

2.4 Joint instance modelling

The morphable model models shape and texture separately. It makes sense to assume that the two instances are not correlated (even though we do not know, whether this is correct or not). Other instances however might be amenable to joint modelling. Paysan et al. [66] made an attempt to learn the relationship between skull-shapes and face-shapes. They use three sets of training data. The first set comprises 20 CT scans of skulls. The second training-set consists of 3D scans of 840 subjects, which is labelled with attributes like gender, age, weight, height. The third training-set consists of 23 MRI head scans where both skull and face are visible. This set is used to calculate the relationship between both instances. The relationship between skull and face is ambiguous. Skull shapes for given individuals remain mainly constant over time. Appearance on the other hand can vary due to age, weight, facial hair or cosmetics. Prior knowledge about attributes can be used to direct fitting into the principal space learnt from the labelled scans.

2.5 Lightstage

3D geometry obtained by laser scanners or structured light scanners offer limited resolution (Global shape is represented accurately, but subtle details are missed out). A different method to acquire geometry and reflectance was introduced by Debevec et al. [67] and also used by Ma et al. [68]. The method estimates diffuse and specular normals from polarised spherical gradient illumination patterns. Experiments have shown, that for scattering materials, like human skin the geometry can be best recovered using specular normals.

Specular normals and diffuse normals are separated using polarisers. Polarised light is scattered in various directions when penetrating different skin layers; which results in loss of polarisation. Specular reflections on the other hand do not change polarisation (skin penetration is only minimal). Polarising filters placed

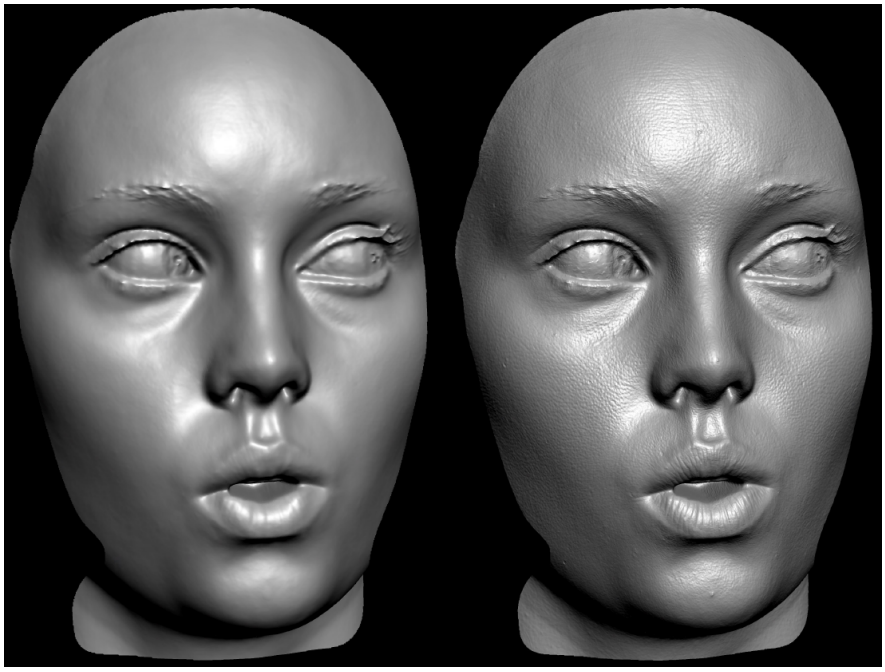

Figure 2.5: This illustration shows 3D geometry recovered from the diffuse normals only (left) and combined diffuse / specular normals (right) of actress Emily O'Brien. The combined method shows significantly more detail [69].

in front of the camera are used to separate diffuse and specular components. Figure 2.5 shows an example of 3D geometry rendered using diffuse normals and specular normals [69].

The polarised spherical gradient illumination patterns are produced by a light-stage system. When all LED's are turned on, the image is illuminated equally from all directions and should result in a "perfectly unshaded" diffuse albedo image. However there remains a shading effect which can not be eliminated. The darker regions, around the eyes and the nose in Figure 2.6 are caused by ambient occlusion (AO). The principle of AO was first described by Langer et al. [70]. AO is a global shading method and view point independent. The level of occlusion at each point in the image is a function of the entire objects' geometry.

For graphics applications, the effect is often desired and adds realism to renderings. In face recognition applications they form a source of variation which has not been taken into account so far.

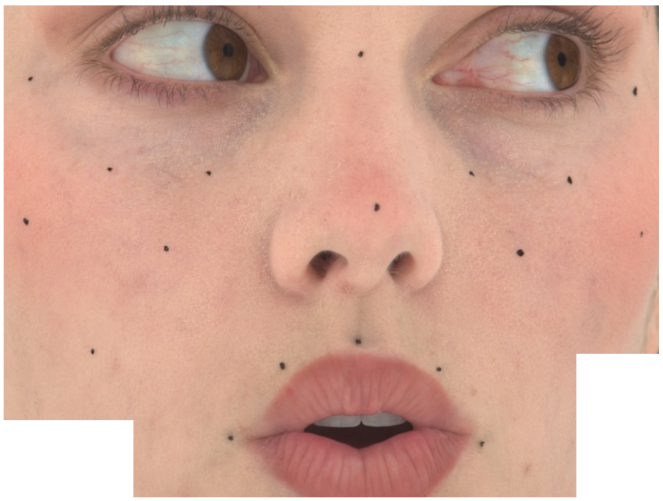

Figure 2.6: A diffuse all image obtained via Light-Stage with all LED's switched on [69].

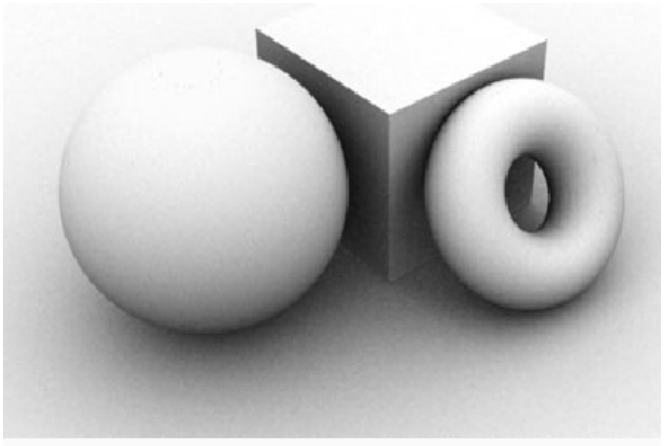

Figure 2.7: Ambient occlusion of a scene containing three geometric objects. The objects can easily be identified by only observing the shading patterns (Image courtesy of *www.renderwiki.com*).

Ambient occlusion does convey significant shape information. Figure 2.7 shows AO of three geometric objects. The objects can easily be recognised by human observers. This suggests that ambient occlusion could be used in a similar fashion than shape-from-shading.

2.6 State of the art

Inverse rendering aims to recover object and scene properties (geometry, reflectance and illumination) from image data. This problem is well understood and methods exist for the case of one unknown (e.g. illumination estimation with known geometry and reflectance [71]). Perhaps the best known result is that the appearance of a convex Lambertian object under arbitrary illumination can be efficiently described using a low dimensional spherical harmonic basis. This representation has found wide application in both graphics and vision.

However, the problem becomes ill-posed when two or more properties are unknown. For example, Ramamoorthi et al. [72] point out that it is not possible to distinguish between low-frequency texture and lighting effects. They suggest that this ambiguity can only be resolved by using active methods or making assumptions about the expected characteristics of the texture and lighting. The latter of these alternatives is exactly the idea we consider in this thesis, namely by restricting our consideration to the class of human faces.

Whether in 2D (eye-centre-aligned [32] or shape-free, warped images [73]) or 3D (depth maps [74], fields of surface normals [17] or meshes in dense correspondence [9]), human faces have been shown to be highly amenable to description using a linear statistical model. 2D approaches model appearance directly, with the training data capturing both extrinsic scene properties (such as illumination and camera parameters) and intrinsic face properties (geometry and reflectance properties). Separating these effects is a challenge at the statistical modelling stage [75]. On the other hand, 3D approaches use face shape and reflectance data collected using face capture devices, allowing face intrinsics to be modelled directly. At the stage of fitting the model to image data, the forward rendering process must be simulated and extrinsic parameters estimated as part of the fitting process [25].

In the context of inverse rendering, such statistical models provide a useful constraint and make it possible to solve problems which would be ill-posed in the general case. We focus on the most difficult case where geometry, reflectance, illumination and camera properties are all unknown and only a single image is available. In this setting, the problem is underconstrained. For example, a red

observation may be caused by red skin colour, red illumination, an increased sensitivity in the camera's red channel or a combination of those three factors. Nevertheless, with appropriate regularisation we are able to obtain accurate solutions without having to make unrealistic assumptions about reflectance properties or the complexity of the illumination environment.

In this section we discuss relevant previous work in the area of inverse rendering, statistical face modelling, face shape estimation and recognition. We use this work to motivate the methods we present in this thesis.

2.6.1 Inverse rendering

In the context of arbitrary objects, inverse lighting for a Lambertian surface was considered by Marschner and Greenberg [71]. Ramamoorthi and Hanrahan [72] used spherical harmonics to describe the reflected light field as a convolution of lighting and reflectance. They present a signal processing framework for a variety of inverse rendering problems under the assumption of known geometry. Hertzmann and Seitz [76] showed how the use of a reference object of known shape and with similar reflectance properties as the object under study could be used for unambiguous, non-Lambertian photometric stereo. Goldman et al. [77] extended this approach by using a library of fundamental materials. They are able to estimate shape and spatially varying reflectance though they require many images under known, varying illumination.

In the context of faces, Marschner et al. [78] used geometrically and photometrically calibrated images of human faces taken from a variety of viewing directions and under varying illumination directions. Combined with a laser range scanned model of the subject under study, dense measurements of the BRDF could be made. Georghiades [79] incorporated the Torrance and Sparrow [80] model of reflectance into an extended uncalibrated photometric stereo algorithm. This allowed accurate shape and reflectance parameters to be recovered from multiple images of various objects, including faces. Fuchs et al. [81] fitted a morphable model to multiple face images, providing point-to-point correspondence for a number of viewing conditions. They experimented with fitting a number of analytical reflectance models to the observed data. Both of these approaches

required multiple images under varying illumination conditions.

2.6.2 Statistical face modelling

Since the late 80s, there has been an interest in learning the subspace of face images or shapes ("face-space") from a representative training sample. In 2D, the seminal eigenfaces paper [28] popularised the idea of applying PCA to a sample of face images. This provides a compact, parametric representation which is useful for recognition and classification or can be used generatively to synthesise appearances. The 3D analog was proposed by Atick et al. [82]. However, in both cases the problem of dense correspondence was not considered. Instead, the face images or surfaces were roughly aligned by a global transformation. Hence, the models described mis-registrations as well as identity variation.

The correspondence problem was addressed in 2D by Craw and Cameron [73] who used manually-labelled landmark points to warp images to the mean shape before applying PCA. This was developed further by Cootes et al. [35] who modelled both 2D shape and appearance in their widely used Active Appearance Model framework. More recently, groupwise alignment has been used to automatically register samples of images using both stochastic [83] and minimum description length [84] approaches. In 3D, Blanz and Vetter [25] used a version of optical flow applied to cylindrical parameterisations of the face surfaces to establish dense correspondence between each face and a chosen template face (i.e. a pairwise approach). A groupwise approach was proposed by Sidorov et al. [46] based on establishing a common embedding across all training samples.

2.6.3 Face shape and reflectance estimation

Model-based approaches to face shape estimation have grown in popularity over the last decade. Under the assumption of frontal pose, constant albedo, known point light source and Lambertian reflectance, Atick et al. [82] fit their 3D statistical face model to images using a gradient-descent based optimisation of the shape parameters. Blanz and Vetter [9; 25] substantially relaxed these assumptions by incorporating a statistical model of texture and estimating pose, illumination and camera parameters in addition to shape and texture parameters in a non-linear

optimisation. The method required careful initialisation and relied on a stochastic optimisation procedure to avoid local minima. This is highly computationally expensive and gives no guarantee that the global minimum will be obtained.

A number of alternative approaches have been considered. Romdhani et al. [85] proposed a linear approach for computing an incremental update to the shape and texture parameters given dense measurements of residual errors provided by optical flow. Their iterative approach requires nonlinear optimisation of pose and illumination parameters and the overall objective is therefore nonlinear. Romdhani and Vetter [55] introduced an efficient and accurate fitting algorithm which uses features derived from the input image such as edges and specular highlights in combination with image intensity values. The overall cost function is smoother and therefore easier to optimise. Moghaddam et al. [18] focussed on a geometric cue by fitting a morphable model to face silhouettes observed from multiple directions. Similarly, Blanz et al. [86] showed how to fit a morphable model to a sparse sample of feature point positions. Their approach required careful selection of a global regularisation parameter and required iterative re-estimation of the perspective projection parameters. Knothe et al. [87] considered the problem of model dominance and used local feature analysis to locally improve the fit of the model to a set of sparse feature points.

The most similar work in spirit to that presented here is due to Zhang and Samaras [88]. They construct a statistical model of harmonic images (low dimensional subspace derived from surface normals and albedo), registered to a morphable face shape model. At the expense of stricter assumptions about reflectance (specularities are neglected), they are able to fit their model under arbitrary illumination by estimating the parameters of the spherical harmonic image model and illumination parameters. Their approach does not link the global shape obtained by the morphable model to the normals of the harmonic images. Moreover, the harmonic images contain directional and quadratic terms which cannot be efficiently modelled by a linear approach. Shim et al. [89] model specularities in 2D using an empirical PCA model built from training data. They use the first principal component to model specular reflections. Empirical models are trained to specific imaging conditions and lack the ability to generalise to images taken under different conditions. Therefore, an analytical approach to

modeling specularities is highly desirable.

There have been a number of attempts to use shading information for face shape recovery. Zhao and Chellappa [90] showed how to estimate face shape under the assumption of bilateral symmetry (i.e. frontal pose). Dovgard and Basri [91] extended this by incorporating a statistical surface model though they retained the strict assumptions about reflectance and illumination. This method allowed recovery of albedo without a statistical model. Smith and Hancock [17] combined a statistical model constructed in the surface normal domain with a classical shape-from-shading constraint in an iterative framework. Under frontal pose, their method was able to estimate finescale surface detail not captured by the statistical model based on shading information. This approach was extended to incorporate non-Lambertian reflectance models and colour images [92].

Under strict assumptions about reflectance properties and illumination, classical shape-from-shading has been applied to the problem of face shape recovery. For example, Prados and Faugeras [93] use a perspective projection and assume that the viewer and light source are co-located. Under these conditions they use viscosity solutions to derive a provably convergent shape-from-shading algorithm which is able to obtain coarse face shape estimates. Kemelmacher and Basri [94] use a single 3D reference face model which is molded to match an observed face in order to estimate face shape, illumination and albedo from one image.

2.6.4 Pose and illumination insensitive face recognition

Face recognition under extreme variations of illumination and pose has presented a serious research challenge. Appearance-based approaches [95; 96; 97] do not aim to recover intrinsic facial features from an image, but rather model the image variability caused by changes in illumination. The advantage here is that the basis set can be used in a generative manner to synthesise photorealistic images under arbitrary and possibly extreme lighting conditions. The drawback of these approaches is that they either require multiple training images (typically 7-9) or knowledge of the underlying shape and reflectance information (which may be recovered from the multiple training images). Similar work using bootstrap image sets has shown that similar performance can be obtained using a single training

image [98]. The assumption is that the training images containing illumination variation need only be class-specific, i.e. images of faces, without requiring multiple images of the particular subject to be recognised. Nishino et al. [99] use a similar approach, but explicitly estimate the illumination direction based on the light reflected by the eye. They then model the variation in appearance for a particular illumination direction as locally linear and obtain good recognition results.

For pose variations, there were attempts to extend early methods to multiple views. For example, Pentland et al. [100] constructed view-based eigenfaces and Cootes et al. [101] extended the Active Appearance Model to account for variation in pose by building a separate model for each of a number of different poses. Georghiades et al. [95] used their few-to-many approach to recover a 3D model from a sample of training images using a variant of photometric stereo. They were able to synthesise views of face under novel lighting and pose given as few as three images of the face taken under variable lighting. Blanz et al. [102] used their morphable model framework to correct for variations in pose. Having fitted the model to an image, any novel pose can be rendered under any arbitrary lighting conditions. Occluded areas of the input face are implicitly recovered when estimating the face shape parameters that most closely match the visible areas of the face.

2.6.5 Global shading

The appearance of a face in an image is determined by a combination of intrinsic and extrinsic factors. The intrinsic properties of a face include its shape and reflectance properties (which vary spatially, giving rise to parameter maps or, in the case of diffuse albedo, texture maps). The extrinsic properties of the image include illumination conditions, camera properties and viewing conditions. Inverse rendering seeks to recover intrinsic properties from an image of an object. These can subsequently be used for recognition or re-rendering under novel pose or illumination.

The forward rendering process is very well understood and physically-based rendering tools allow for photorealistic rendering of human faces. The inverse

process on the other hand is much more challenging. Perhaps the best known results apply to convex Lambertian objects. In this case, reflectance is a function solely of the local surface normal direction and irradiance (even under complex environment illumination) can be accurately described using a low dimensional spherical harmonic approximation [96]. This observation underpins the successful appearance-based approaches to face recognition.

However, faces are not globally convex and it has been shown that occlusions of the illumination environment play an important role in human perception of 3D shape [103]. Prados et al. [104] have shown how shading caused by occlusion under perfectly ambient illumination can be used to estimate 3D shape. In this thesis we take a step towards incorporating global illumination effects into the inverse rendering process. This is done in the context of fitting a 3D morphable face model, so the texture is subject to a global statistical constraint.

We use a model which incorporates ambient occlusion [105] and bent normals [106] into the image formation process. This is an approximation to the rendering equation that is popular in graphics, because it can be precomputed and subsequently used in real-time rendering applications. Both properties are a function of the 3D shape of an object.

2.7 Conclusions

Statistical models have dominated the face analysis literature over the last decade. This is because of the robustness and flexibility they offer in a number of real world problem settings. The advantage of a 3D model is that it explicitly separates intrinsic face properties from those related to the specific conditions present in an observed image. This has led to state-of-the-art performance in face recognition under varying pose and illumination [26]. However, morphable models have not been widely adopted and are rarely used in preference to their 2D counterparts. This is because of the difficulties involved in fitting such models to images. Existing methods are slow, prone to becoming stuck in local minima, sensitive to parameter tuning and require the fitting algorithm to be heavily engineered towards specific imaging environments. In particular, no existing method allows for both arbitrarily complex illumination and non-Lambertian reflectance. Fur-

ther, we present the first fitting algorithm for morphable models that take global illumination into account. This makes our method the most general to date and the first to guarantee globally optimal solutions in both the estimated shape and texture. Moreover, we formulate every step of our method as linear, which makes it highly efficient.

Chapter 3

Shape and Pose

In this chapter, we present a novel algorithm for shape parameter estimation under unknown pose given the 2D coordinates of a sparse set of feature points. We make the assumption that global shape is sufficiently accurately determined by the feature point locations (as long as they do not deviate too far from their true position) and a shape prior. We incorporate knowledge inferred from out-of-sample data into the reconstruction algorithm and show how this can be used to reduce generalisation error. We then extend our method so that it can efficiently be used with mixture distributions.

3.1 Shape model fitting

The novelty of our approach is to evaluate how the model generalises to different feature point locations and integrate this knowledge into the fitting process. This helps to prevent overfitting and ensures that errors are penalised in an appropriate way. The problem is algebraically formulated such that solutions to the unknowns can be obtained in closed-form. Shape and pose are geometric entities and independent of facial texture, illumination and photometric camera properties. As opposed to analysis-by-synthesis methods (which address image formation as a whole), we treat geometric and photometric parts separately. By doing so, we avoid the problem of a highly non-convex optimisation littered with local minima. With a shape estimate at hand, we have access to surface nor-

mals which influence the photometric part. In order to obtain a linear solution, we decompose the geometric problem into two steps which can be iterated and interleaved:

1. Estimation of the camera projection matrix, $\mathbf{C}$, using known 3D-2D correspondences.
2. Estimation of 3D shape parameters, $\mathbf{a}$, using a known camera projection matrix.

This constitutes to a bilinear system in the unknown parameters, which can be solved using alternating least squares. Because the shape model is affine (mean shape is factored) a unique solution exists when assuming independence of pose and shape. We initialise by using the mean shape to compute an initial estimate of the camera projection matrix, $\mathbf{C} \in \mathbb{R}^{3\times4}$. With this to hand, shape parameters can be recovered using only matrix multiplications. By using the recovered shape to re-estimate the camera matrix, we can iterate the process which typically converges in ≤ 5 iterations.

3.1.1 Estimating the camera projection matrix

We represent 2D locations of feature points in the image, $\mathbf{x}_i \in \mathbb{R}^3$, and corresponding 3D locations of the feature points within the model, $\mathbf{X}_i \in \mathbb{R}^4$, as homogeneous coordinates. To estimate the camera projection matrix, we require normalised versions: $\tilde{\mathbf{x}}_i = \mathbf{T}\mathbf{x}_i$ and $\tilde{\mathbf{X}}_i = \mathbf{U}\mathbf{X}_i$, where $\mathbf{T} \in \mathbb{R}^{3\times4}$ and $\mathbf{U} \in \mathbb{R}^{4\times4}$ are similarity transforms which translate the centroid of the image/model points to the origin and scale them such that the RMS distance from the origin is $\sqrt{2}$ for the image points and $\sqrt{3}$ for the model points. In our approach we assume an affine camera and compute the normalised projection matrix, $\tilde{\mathbf{C}} \in \mathbb{R}^{3\times4}$, using the *Gold Standard Algorithm* [107]. Given $N \geq 4$ model to image point correspondences $\mathbf{X}_i \leftrightarrow \mathbf{x}_i$, we determine the maximum likelihood estimate of $\tilde{\mathbf{C}}$ which minimises: $\sum_i \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{C}}\tilde{\mathbf{X}}_i\|^2$, subject to the affine constraint $\tilde{\mathbf{C}}_3 = [0\ 0\ 0\ 1]$. Each

point correspondence contributes to the following $2N \times 8$ system of equations:

$$\begin{bmatrix} \tilde{\mathbf{X}}_1^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \tilde{\mathbf{X}}_1^{\mathrm{T}} \\ \vdots & \vdots \\ \tilde{\mathbf{X}}_N^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \tilde{\mathbf{X}}_N^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{C}}_1^{\mathrm{T}} \\ \tilde{\mathbf{C}}_2^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{x}}_{1,1} \\ \tilde{\mathbf{x}}_{1,2} \\ \vdots \\ \tilde{\mathbf{x}}_{N,1} \\ \tilde{\mathbf{x}}_{N,2} \end{bmatrix}.$$

We solve this system using least squares and obtain the camera matrix by performing the following de-normalization step: $\mathbf{C} = \mathbf{T}^{-1}\tilde{\mathbf{C}}\mathbf{U}$.

3.1.2 Modelling feature point variance

In order to explain the difference between observed and modelled feature point positions in an image, we model two sources of variance. By having an explicit model of variance, we negate the need for an ad-hoc regularisation weight parameter. The first source of variance is the generalisation error of the morphable model. This describes how feature points deviate from their true position in 3D when the optimal model parameters are used to describe a face. Generalisation error is spatially varying, i.e. some regions of the face are harder to generalise to than others, and it is this affect that is captured by having per-feature point variance. The second source of variance is the 2D pixel noise, this is related to the accuracy with which the feature points can be marked up in 2D.

Given an out-of-sample face mesh $\mathbf{v}_i$ (i.e. a face that was not used to train the statistical model), we project onto the model to obtain the closest (in a least squares sense) possible approximation: $\mathbf{v}_i' = \mathbf{V}\mathbf{V}^{\mathrm{T}}(\mathbf{v}_i - \overline{\mathbf{v}}) + \overline{\mathbf{v}}$. The vector of element-wise errors is given by: $\mathbf{e}_i = \mathbf{abs}\,(\mathbf{v}_i - \mathbf{v}_i')$. We define $\hat{\mathbf{e}}_i$ as the vector formed by sub-selecting the elements of $\mathbf{e}_i$ which correspond to the N sparse feature points. From a sample of k such out-of-sample faces, we can now compute the standard deviation associated with each coordinate of the feature points: $\sigma_{3D,j} = \frac{1}{k}\sum_{i=1}^{k} \hat{\mathbf{e}}_{i,j}$, where $\sigma_{3D,j} \in \mathbb{R}^3$. This provides an empirical means to predict how a feature point is likely to vary from its true position due to generalisation errors. The units of $\sigma_{3D,j}$ are mm. The result can be used for 3D - 3D

reconstruction. By defining matrix $\mathbf{\Sigma} = \mathrm{diag}(\sigma_{3D,j}^{-1})$, the objective $\mathbb{E}$ becomes:

$$\mathbb{E} = (\mathbf{Va} + \bar{\mathbf{v}} - \mathbf{y})^{\mathrm{T}} \mathbf{\Sigma}^{\mathrm{T}} \mathbf{\Sigma} (\mathbf{Va} + \bar{\mathbf{v}} - \mathbf{y}).$$

The equation can be brought into a standard form: $(\mathbf{Ax} + \mathbf{b})^{\mathrm{T}} \mathbf{\Omega} (\mathbf{Ax} + \mathbf{b})$, by setting $\mathbf{A} = \mathbf{V}$, $\mathbf{x} = \mathbf{a}$, $\mathbf{b} = \bar{\mathbf{v}} - \mathbf{y}$ and $\mathbf{\Omega} = \mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}$ ($\mathbf{\Omega}$ is a positive semidefinite matrix $\in \mathbb{R}^{3N \times 3N}$). This allows for a very efficient solution with respect to the unknown parameters, $\mathbf{x}$. The following shows a step-by-step derivation and includes an optional regularisation term λ:

$$\begin{aligned}
\mathbb{E} &= (\mathbf{Ax} + \mathbf{b})^{\mathrm{T}} \mathbf{\Omega} (\mathbf{Ax} + \mathbf{b}) + \lambda \|\mathbf{x}\| \\
&= \left[(\mathbf{Ax})^{\mathrm{T}} \mathbf{\Omega} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \right] (\mathbf{Ax} + \mathbf{b}) + \lambda \|\mathbf{x}\| \\
&= (\mathbf{Ax})^{\mathrm{T}} \mathbf{\Omega} \mathbf{Ax} + (\mathbf{Ax})^{\mathrm{T}} \mathbf{\Omega} \mathbf{b} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{Ax} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{b} + \lambda \|\mathbf{x}\| \\
&= \mathbf{x}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{Ax} + \left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{b} \right)^{\mathrm{T}} \mathbf{x} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{Ax} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{b} + \lambda \|\mathbf{x}\|.
\end{aligned}$$

The solution to the quadratic form is found by solving the equation $\frac{\mathrm{d}\,\mathbb{E}}{\mathrm{d}\,\mathbf{x}} = 0$:

$$\begin{aligned}
\frac{\mathrm{d}\,\mathbb{E}}{\mathrm{d}\,\mathbf{x}} &= 2\mathbf{x}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + \left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{b} \right)^{\mathrm{T}} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + 2\lambda \mathbf{x}^{\mathrm{T}} = 0 \\
&= 2\mathbf{x}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + \mathbf{b}^{\mathrm{T}} \left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \right)^{\mathrm{T}} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + 2\lambda \mathbf{x}^{\mathrm{T}} \\
&= 2\mathbf{x}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + 2\mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + 2\lambda \mathbf{x}^{\mathrm{T}} \\
&= \mathbf{x}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + \lambda \mathbf{x}^{\mathrm{T}} + \mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} \\
&= \left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} \right)^{\mathrm{T}} \mathbf{x} + \lambda \mathbf{x} + \left(\mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} \right)^{\mathrm{T}} \\
&= \mathbf{A}^{\mathrm{T}} \left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \right)^{\mathrm{T}} \mathbf{x} + \lambda \mathbf{x} + \mathbf{A}^{\mathrm{T}} \left(\mathbf{b}^{\mathrm{T}} \mathbf{\Omega} \right)^{\mathrm{T}} \\
&= \left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + \lambda \mathbf{I} \right) \mathbf{x} + \mathbf{A}^{\mathrm{T}} \mathbf{\Omega}^{\mathrm{T}} \mathbf{b},
\end{aligned}$$

and the solution vector $\mathbf{x}$ is given by: $\mathbf{x} = -\left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega} \mathbf{A} + \lambda \mathbf{I} \right)^{-1} \left(\mathbf{A}^{\mathrm{T}} \mathbf{\Omega}^{\mathrm{T}} \mathbf{b} \right)$.

Feature point variance in the image plane

In order to predict how this results in variation in the image plane, we project the variances to 2D, in units of pixels. This is illustrated in Figure 3.1. We associate six auxiliary points (in homogeneous coordinates) with each vertex of interest. The vertex is centred at the origin and the distances $\|\mathbf{a} - \mathbf{b}\|$, $\|\mathbf{c} - \mathbf{d}\|$ and $\|\mathbf{e} - \mathbf{f}\|$ correspond to the 3D standard deviation: $\sigma_{3D,j,x}$, $\sigma_{3D,j,y}$ and

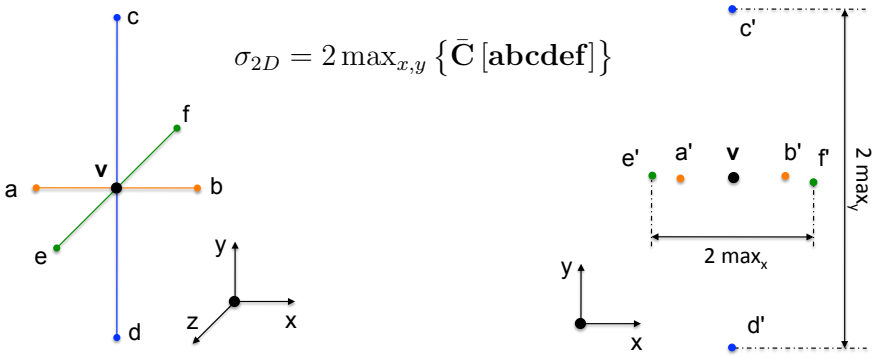

Figure 3.1: Example of projecting 3D standard deviation to the image plane. The camera matrix $\bar{\mathbf{C}}$ performs a -70 degrees rotation around the y-axes, and a minor tilt around the x-axes.

$\sigma_{3D,j,z}$. We define $\bar{\mathbf{C}} \in \mathbb{R}^{3\times4}$ as the camera projection matrix without translational components. This is required because the variances are with respect to the feature point position and do not need globally translating. Our final 2D variances are given by the sum of the projected 2D variances and a 2D pixel error, η^2, which models error in feature point markup: $\sigma^2_{2D,j} = 2\max_{x,y}\left\{\bar{\mathbf{C}}\left[\mathbf{abcdef}\right]_j\right\}^2 + \eta^2$. We use a value of $\eta = \sqrt{3}$ pixels in our experiments.

3.1.3 A probabilistic approach

The 3D shape parameters are obtained using a probabilistic approach which follows that of Blanz et al. [86]. However, our derivation is more complex as we allow different 2D variances, $\sigma^2_{2D,i}$, for each feature point. Our aim is to find the most likely shape vector $\mathbf{c}_s$ given an observation of N 2D feature points in homogeneous coordinates: $\mathbf{y} = [x_1\ y_1\ 1\ \ldots\ x_N\ y_N\ 1]^{\mathrm{T}}$ and taking the model prior into account. From Bayes' rule we can state: $P(\mathbf{c}_s|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{c}_s)\cdot p(\mathbf{c}_s)$. The coefficients are normally distributed with zero mean and unit variance, i.e. $\mathbf{c}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$. The probability of observing a given $\mathbf{c}_s$ is: $p(\mathbf{c}_s) = \nu\cdot\exp\left(-\frac{1}{2}\|c_s\|^2\right)$, where ν is a normalisation constant. The conditional likelihood of data $\mathbf{y}$ is given by:

$$P(\mathbf{y}|\mathbf{c}_s) = \prod_{i=1}^{3N} \nu\cdot\exp\left(-\frac{\left[y_{m2D,i} - y_i\right]^2}{2\sigma^2_{2D,i}}\right).$$

Here, $y_{m2D,i}$ are the homogeneous coordinates of the 3D feature points projected to 2D. To do so, we construct a matrix $\hat{\mathbf{V}} \in \mathbb{R}^{3N \times m-1}$ by subselecting the rows of the eigenvector matrix $\mathbf{V}$ associated with the N feature points. The matrix is further modified by inserting a row of zeros after every third row of $\mathbf{V}$, resulting in matrix: $\hat{\mathbf{V}}_h \in \mathbb{R}^{4N \times m-1}$. We form a block diagonal matrix $\mathbf{P} \in \mathbb{R}^{3N \times 4N}$ in which the camera matrix, $\mathbf{C}$, is placed on the diagonal:

$$\mathbf{P} = \begin{bmatrix} \mathbf{C} & & \\ & \ddots & \\ & & \mathbf{C} \end{bmatrix}.$$

Finally, we can define the 2D points obtained by projecting the 3D model points given by $\mathbf{c}_s$ to 2D: $y_{m2D,i} = \mathbf{P}_i \cdot (\hat{\mathbf{V}}_h \mathbf{c}_s + \bar{\mathbf{v}})$, where $\mathbf{P}_i$ is the ith row of $\mathbf{P}$. Substituting into Bayes' rules, we arrive at our conditional probability:

$$P(\mathbf{c_s}|\mathbf{y}) = \nu \cdot \exp\left(-\sum_{i=1}^{3N} \frac{[y_{m2D,i} - y_i]^2}{2\sigma_{2D,i}^2} - \frac{1}{2}\|\mathbf{c}_s\|^2 \right),$$

which can be maximised by minimising the exponent:

$$\mathbb{E} = -2 \cdot \log P(\mathbf{c}_s|\mathbf{y}) = \sum_{i=1}^{3N} \frac{[y_{m2D,i} - y_i]^2}{\sigma_{2D,i}^2} + \|\mathbf{c}_s\|^2. \tag{3.1}$$

To simplify, we bring Equation 3.1 into standard form and solve using the derivation in Section 3.1.2. The variances are rewritten as $\mathbf{\Sigma} = \text{diag}(\sigma_{2D,i}^{-1})$ and $\mathbf{\Omega} = \mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}$. We set $\mathbf{A} = \mathbf{P}\hat{\mathbf{V}}_h$, $\mathbf{b} = \mathbf{P}\bar{\mathbf{v}} - \mathbf{y}$ and $\mathbf{x} = \mathbf{c}_s$.

3.2 Experiments

In this section we present a comprehensive experimental evaluation of our method. We evaluate the accuracy of our shape reconstruction algorithm from 3D feature points and compare it with regularised least squared and probabilistic PCA. We then focus on 3D shape reconstruction from 2D feature points. We use a subset of 70 of the anthropometric landmarks suggested by Farkas [51], and compare the results with ground truth data and a recently published state-of-the-art fitting

method [55]. Note that feature point saliency is an interesting topic in itself. Recent work has attempted to learn the most salient feature points from data [108].

3.2.1 Synthetic data

The BFM is supplied with a database of synthetic images spanning 10 out-of-sample identities in 9 pose angles and 3 illumination conditions (270 renderings) with known ground truth. These 10 meshes are also used to empirically estimate feature point variance. We compare our results against the state-of-the-art *Multi feature fitting algorithm* [55], for which shape and texture coefficients are provided with the model. We use the 99 most significant modes for shape reconstruction.

3D–3D Shape reconstruction

To begin with, we show shape 3D - 3D reconstruction results from feature points $\mathcal{F}$ (which are a subset of the Farkas feature points [51]) for out-of-sample faces. We compare the proposed method against regularised least squares with isotropic variance (RLS), and probabilistic PCA (PPCA) [109]. For each method 99 shape modes are used. To make our results comparable with RLS, we assign the mean 3D variance of the feature points to the method. We then find the optimal regularisation factor λ, which minimised the mean squared error for all ten faces via a line search. This ensures that differences between the proposed method and RLS is influenced by relative weightings only. In general, the result of RLS is independent of which constant is chosen for σ^2; as long as one uses the optimal value of λ. Comparing with PPCA is interesting, since the variance for this method is also based on an optimality criterion. Table 3.1 shows quantitative results for the three methods. We show results for mean squared error (MSE) and (eigenvalue normalised) angular distance (AE). The second measure is an approximation to comparing identity only. Qualitative comparison is shown in Figure 3.2. Reconstructions of face 006 and 022 show an unusual high error. The reason therefore seems to be a mis-registration between the faces and the model.

Face :	001	002	006	014	017	022	052	053	293	323	mean
Mean squared Euclidian error: $\times 10^{12} \mu m^2$											
Prop.	0.1154	0.2136	1.9752	0.3392	0.2078	1.8462	0.2066	0.1740	0.0779	0.0917	0.5247
RLS	0.1422	0.3142	3.0120	0.2969	0.2861	1.7245	0.2217	0.2425	0.3942	0.1072	0.6741
PPCA	0.1420	0.3918	5.9639	0.4150	0.2602	2.3073	0.1495	0.1692	0.1668	0.1219	1.0088
Eigenvalue normalised angular distance:											
Prop.	46.67	54.64	49.63	49.75	47.77	57.34	45.88	50.22	39.60	45.96	48.77
RLS	52.12	53.31	65.70	57.73	53.11	69.37	53.65	53.40	44.36	48.12	55.09
PPCA	50.24	64.94	68.42	53.41	46.48	60.85	38.86	50.27	49.17	47.75	53.04

Table 3.1: Shape reconstruction errors for 10 out of sample faces. We compare our method to regularised least squares and probabilistic PCA.

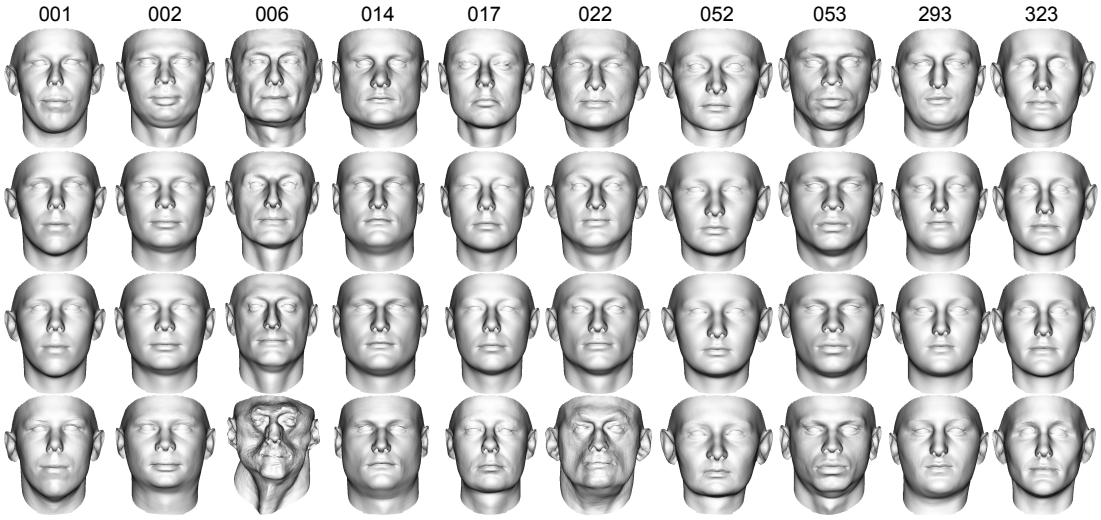

Figure 3.2: 3D reconstructions from 70 feature points for 10 subjects. Top row shows ground truth shape. Second row shows results the proposed method. Third row shows reconstructions using regularised least squares. And the last row shows reconstructions using probabilistic PCA.

3D–2D Shape reconstruction

We now demonstrate performance in reconstructing 3D shape from a sparse set of feature points projected to 2D. We assume that the 2D feature point positions are already known. Recent work has shown that facial feature points can be automatically located in a robust and efficient fashion using a local feature detector in conjunction with a shape model [110]. Note also that it is straightforward to extend our method to silhouettes and edges. Pixels lying on an image edge are

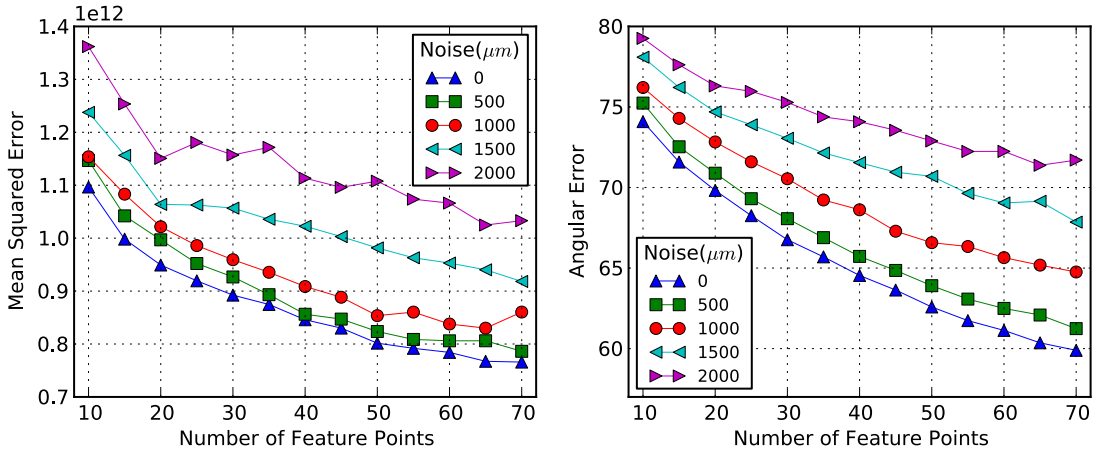

Figure 3.3: Shape reconstruction error from 2D feature points, averaged over all subjects. For each number of feature points, the experiment is repeated 20 times with a random subset. The values shown in the figures are mean values.

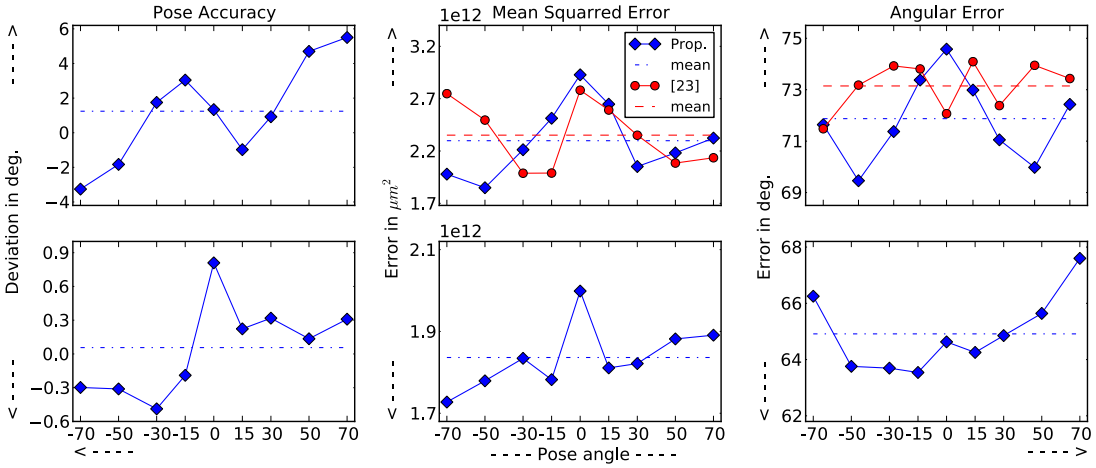

Figure 3.4: Top row, from left to right: First figure shows recovered pose angles (minus ground truth) averaged over subjects. Second and third plot show mean reconstruction errors measured in MSE and AE for the proposed method and reference method [55]. Bottom row shows results for the same setting rendered via an orthographic projection.The dashed line in all plots show mean over pose angles.

associated with the closest edge point in the model and simply become additional landmark points. The number of visible feature points depends on identity, pose, image resolution and level of noise present in the image. We begin by testing

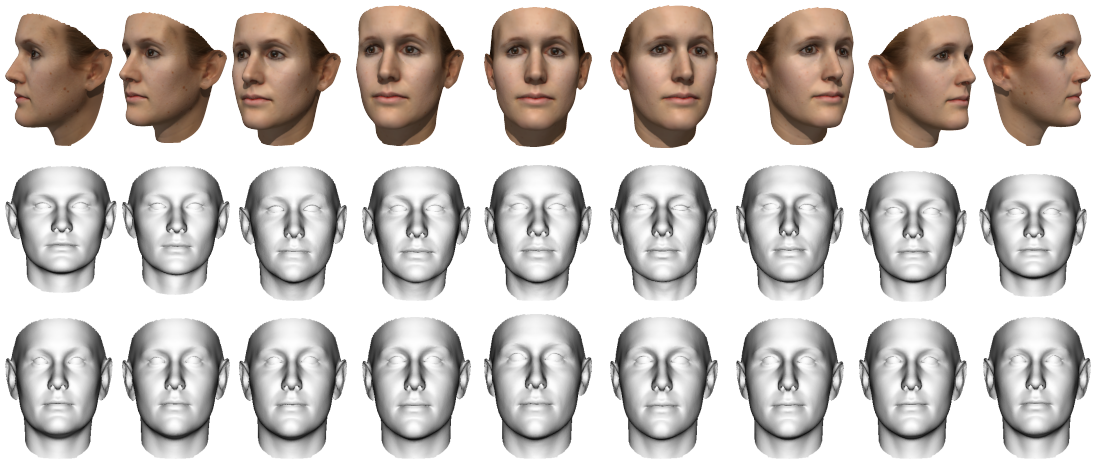

Figure 3.5: Fitting results for subject 323 in 9 pose angles. Top row shows renderings. Second row shows fitting results for the comparison method. Last row shows fitting results for the proposed method.

sensitivity of our method to the number of feature points and noise in the feature point positions. In Figure 3.3 we show the effect of varying the number of feature points, f, from 10 to 70. For $f < 70$, we select a random subset from the 70 feature points. We repeat this process 20 times and show averaged results. The experiment is repeated for 5 different noise levels. The results suggest that performance begins to sharply degrade for $f < 40$ feature points.

Pose variation in the BFM renderings consists of a rotation about the vertical axis. Changing pose has two effects: 1. it effects which feature points are visible, 2. it changes the information content of each feature point (e.g. in a frontal view, the location of the tip of the nose says little about nose length). In Figure 3.4 (top) we show performance as a function of pose angle in mean squared error and angular error. We also extract the estimated pose from the camera matrix and show the accuracy of our pose estimate. A qualitative comparison for one subject in 9 pose angles can be seen in Figure 3.5. Note that the BFM renderings exhibit significant perspective distortion which is not modelled by our affine camera. In Figure 3.4 (bottom) we show results for the same dataset under orthographic projection so that the effect of pose can be evaluated independently of perspective errors.

Finally, we use the estimated shape in a recognition experiment. We use one image per subject as the gallery image and associate each of the remaining probe

Figure 3.6: From left to right: subject in oil painting, reconstructed 3D shape in frontal view, projected 3D shape with texture mapped on it, cropped and rotated version, oil painting with adjusted pose.

Pose:	−70°	−50°	−30°	−15°	0°	15°	30°	50°	70°	**mean**
Prop.	96.7	100	100	100	97.8	98.9	100	100	92.2	98.4
[55]	87.8	93.6	94.4	91.6	92.9	90.7	94.5	96.3	93.0	92.7

Table 3.2: Mean rank-1 recognition error rates for all 270 renderings averaged over 3 illumination conditions per pose. Results are shown for shape-only.

images to the closest gallery image. We repeat, using every pose configuration as the probe image. Similarity is determined using angular distance on the shape parameter vectors. Table 3.2 shows shape recognition rates in comparison to a computationally more expensive reference method. Our method show a relative improvement of about 6%.

The recovered shape and camera matrix can be used to project the shape into the image. This provides means of obtaining correspondence between RGB values of image pixels and model vertices. As a simple application, we show in Figure 3.6, how this can be used to change the pose of a subject in an oil paining. The manipulation is not visible to an unaware human observer.

In the next two chapters, we use the RGB measurements and decompose them into texture, diffuse shading, specular highlights (Chapter 4) and global shading (Chapter 5).

3.3 Mixture distributions

The less feature points are available at inference stage, the more the solution relies on the prior, which is the mean face. Although the mean face has highest

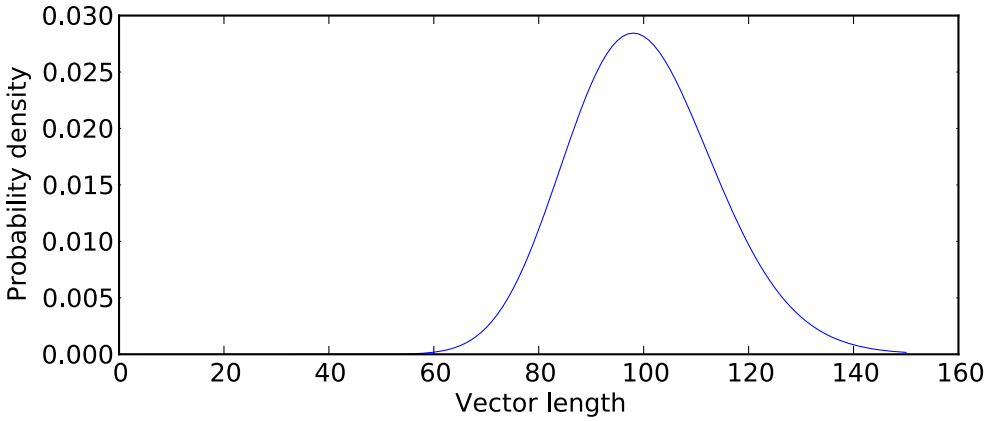

Figure 3.7: Empirical studies have shown, that length of identity vectors of a shape model follow a Chi-square distribution. The probability of observing a shape vector with length 0 (mean face) is negligible.

probability in terms of the PCA model, it is a most unlikely instance in the "real world". This has been confirmed in several studies [111]. For example, Patel and Smith [112] have shown, that the eigenvalue normalised shape vector of a model with n parameters follows a Chi-square distribution centred at n. More visually, faces lay close to a hyper-spherical manifold centred at the mean. Figure 3.7 shows the expected distribution on the length of parameter vectors for a model with $n = 100$ modes.

A Gaussian distribution of faces is practically motivated, and not based on empirical or theoretical facts. Assuming the elements of a vector $\mathbf{c}$ are distributed Gaussian, then the marginal and conditional distribution of a subset of $\mathbf{c}$ are also Gaussian distributed. And in both cases the solution can be found in closed-form. When dealing with faces, which consist of tens of thousands of vertices, this can be a significant advantage, because many applications in this domain require real-time or close to real-time performance. For their approach, Patel and Smith [112] require non-linear optimisation to solve for faces laying on a hyper-spherical manifold, which turns out to be costly in high dimensions.

Several studies indicate that facial features follow a bimodal distribution, with gender being the cause for bi-modality. For example Wu et al. [113] show, that facial needle maps (surface normals) cluster well into two distinct subspaces. They have used their framework to successfully classify gender.

A mixture of eigenfaces for face recognition has been proposed by Kim et

al. [114]. They show that a subspace mixture model achieves higher accuracy compared to single PCA model. Their approach is appearance based and operates in the domain of photographs. It indeed requires separate models to be built for identity-pose and identity-illumination. Also the number of training examples grows rapidly. We argue that a 2D model is intrinsically not the best choice. Effects of pose and illumination are naturally caused by the 3D nature of the problem.

Mixture models can be efficiently learnt with the EM algorithm [115]. Where in the E-step each training example is softly assigned to each cluster, and in the M-step, the mixture components are learnt independently using a weighted combination of the training examples. With each iteration of EM, the cost of the objective function reduces and the algorithm is guaranteed to converge. A globally optimal solution however is not guaranteed since the solution is dependent on initialisation. A different issue (which we experienced when implementing the model) is, that the parameters obtained after convergence do not represent the "ideal" solution. For a different application, this problem is also reported in [116]. Here, the authors state: *"... the standard practice of running EM until convergence to a local maximum in likelihood will not necessarily lead to optimal classification performance. In fact, we have often observed overfitting behavior, in which optimal classification is obtained after only two or three iterations of EM, but the likelihood continues to increase in subsequent iterations as observations are assigned to incorrect content classes"*. The authors address this problem by taking the number of iterations as an additional optimisation parameter.

The here mentioned issues with learning mixture models can be circumvented when training data is labelled and the unsupevervised learning problem turns into a supervised one. In this case we can build separate models for each class. Although this approach is more trivial, we avoid the local minima problem and only need to evaluate a single E-step at the inference stage. To summarise:

- Faces are not Gaussian distributed but lie on a spherical manifold.
- Solving this constrained problem directly requires non-linear optimisation.
- A simplification is to model the population with a mixture model, which favours solutions lying in distinct subspaces.

- Learning mixture model via EM is susceptible to local minima, and the most likely parameters obtained after convergence do not guarantee the "best" solution.
- Appearance based methods, are intrinsically not the best choice to model pose and illumination (especially not jointly).

Based on these observations, we build two PCA models (Female/Male). In a pre-trial, we examine how the model generalises to unseen data. We then extend our shape reconstruction algorithm so it can effectively deal with a mixture model. In both cases we compare the result with a single PCA model built from the same training data.

3.3.1 Motivation

To motivate our idea, we show how the features of out-of-sample faces are distributed on the example of the BFM. To do so we plot various features at different regions in input space (See Figure 3.8). The figure indicates, that the ten faces (of which five are female and five male) occupy two distinct regions. The plot also shows the BFM mean face for the same regions in input space. It should be clear that when reconstructing shape from sparse features (where dependency on prior is high) the mean face is neither representative for the set of female nor male faces, and having separate priors for each cluster is beneficial.

3.3.2 Statistical modelling

Unfortunately, we do not have access to the training data of the BFM. This would be ideal for our study and the results would be directly comparable to the ones in previous sections. The model release is however accompanied with attribute vectors for gender, height, weight and age. We can therefore sample from the BFM and translate the samples into a desired direction. We do this for 200 female and 200 male faces (5 examples for each class are shown in Figure 3.9) and build two separate PCA models. For the rest of this section, subscripts m

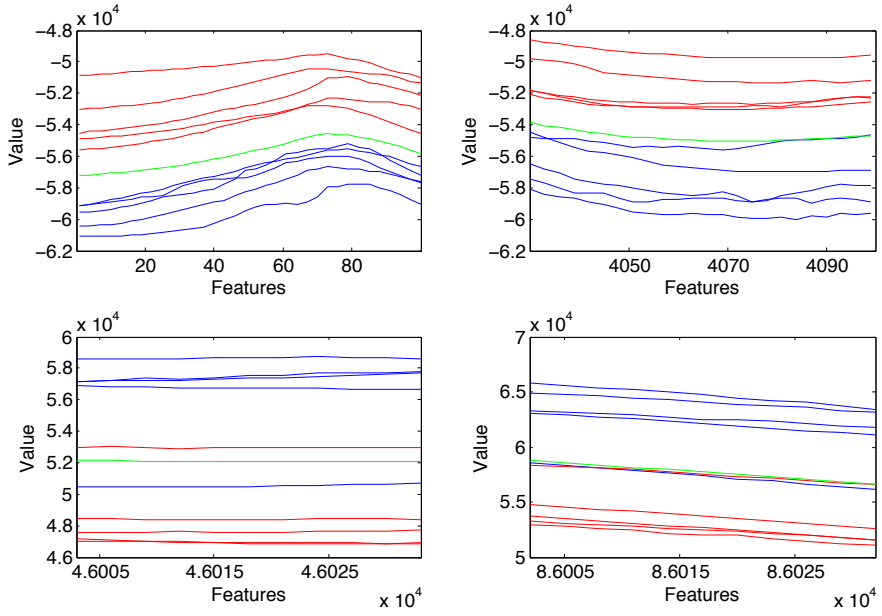

Figure 3.8: Distribution of features for various sections. Blue curves show male faces and red curves show female faces. The green curve shows BFM mean face.

and f indicate male and female, respectively. A new instance is now modelled as:

$$\mathbf{x} = w_m(\mathbf{V}_m\mathbf{a}_m + \bar{\mathbf{v}}_m) + w_f(\mathbf{V}_f\mathbf{a}_f + \bar{\mathbf{v}}_f),$$

subject to the constraints: $w_m, w_f \geq 0$ and $w_m + w_f = 1$. The same training data is used to build a single PCA model which we use for comparison. Figure 3.10 shows the energy captured by each of the 199 modes. In each case, the energy captured by modes > 100 can be regarded as noise and we discard them for our experiments. Figure 3.11 shows the mean face and the three most significant modes of variation for each model.

Figure 3.12 shows feature distribution for the same sections as in Figure 3.8. Dashed lines show class specific priors for female (red) and male (blue). The green line shows prior for the complete training data. As can be seen, class specific priors are more representative. There is however still a tendency towards the centre. One reasons therefore might be the small number of out-of-sample faces present in the plots. The fact that we have sampled from the BFM is also influential.

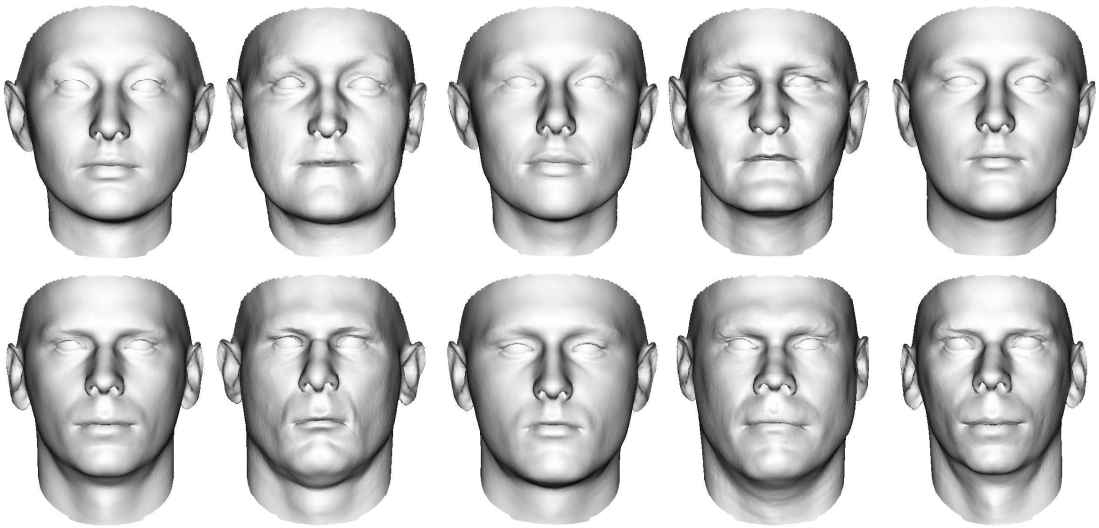

Figure 3.9: Top row shows five female examples, bottom row shows five male examples.

Modes	10		30		50		70	
49	1.19	1.43	1.02	1.20	0.97	1.06	0.88	1.01
99	1.11	1.20	1.01	1.09	0.82	0.85	0.76	0.81

Table 3.3: Reconstruction errors for 70 3D feature points. Left column shows bimodal and right column shows unimodal results. Values are $\times 10^{12} \mu m^2$.

3.3.3 3D–3D Shape reconstruction

In a first experiment we test the ability of the models to generalise to unseen data, where the data takes the form of 3D features. We use 70 feature points for our trial. Quantitative results are shown in Table 3.3. The gain in using a bi-modal model over a unimodal model is shown in Figure 3.13.

3.3.4 3D–2D Shape reconstruction

We extend our fitting algorithm proposed in Section 3.1 so it can effectively deal with class specific models. This is done by learning generalisation error of feature points in each cluster separately. The following shows one iteration of estimating camera matrix and shape alternatively. The loop is started with $w_m = w_f = 0.5$ and $\mathbf{a}_m = \mathbf{a}_f = \vec{\mathbf{0}}$. For each cluster we learn σ_{3D} in advance.

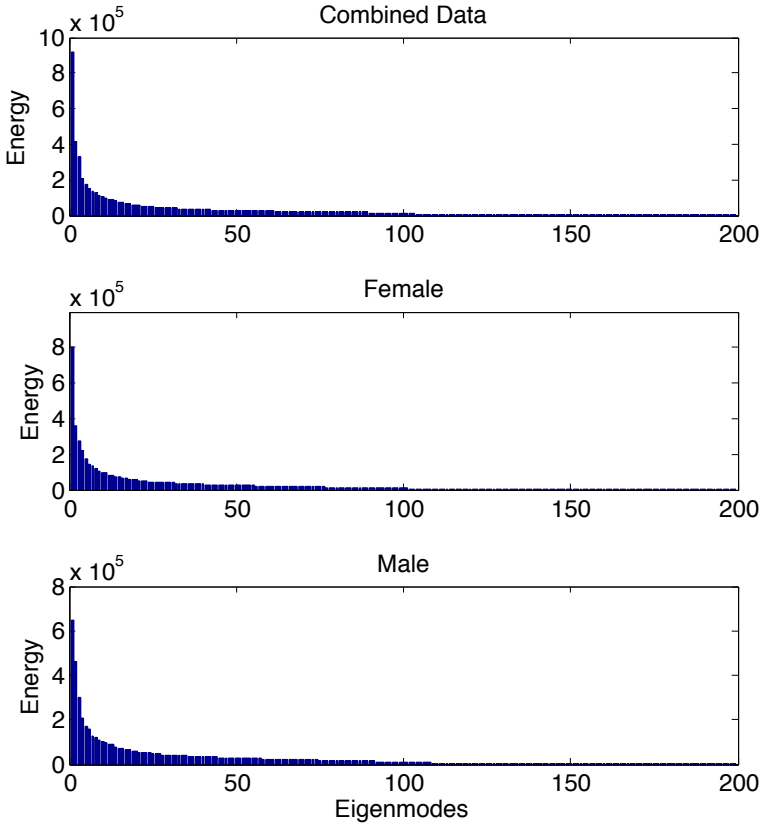

Figure 3.10: Eigenmode decay for model built from complete training-set (top), female samples only (middle) and male faces only (bottom).

1. Calculate camera matrix $\mathbf{C} = f(\mathbf{a}_m, \mathbf{a}_f, w_m, w_f)$
2. For both sets, project σ_{3D} onto image plane $\rightarrow \sigma_{2D}$
3. Calculate $\mathbf{a}_m$ and $\mathbf{a}_f$
4. Synthesise new shape $\mathbf{x} = w_m(\mathbf{V}_m\mathbf{a}_m + \bar{\mathbf{v}}_m) + w_f(\mathbf{V}_f\mathbf{a}_f + \bar{\mathbf{v}}_f)$
5. Measure squared inverse distances to cluster centroids $d_m = \frac{1}{\|\mathbf{x}-\bar{\mathbf{v}}_m\|}$ and $d_f = \frac{1}{\|\mathbf{x}-\bar{\mathbf{v}}_f\|}$
6. Evaluate weighting components $w_m = \frac{d_m}{d_m+d_f}$ and $w_f = \frac{d_f}{d_m+d_f}$

These steps are repeated until convergence. In practice, this is achieved within $5-10$ iterations. An alternative way to evaluate class probability is to measure reconstruction error on 2D feature points. In our experiments, we found

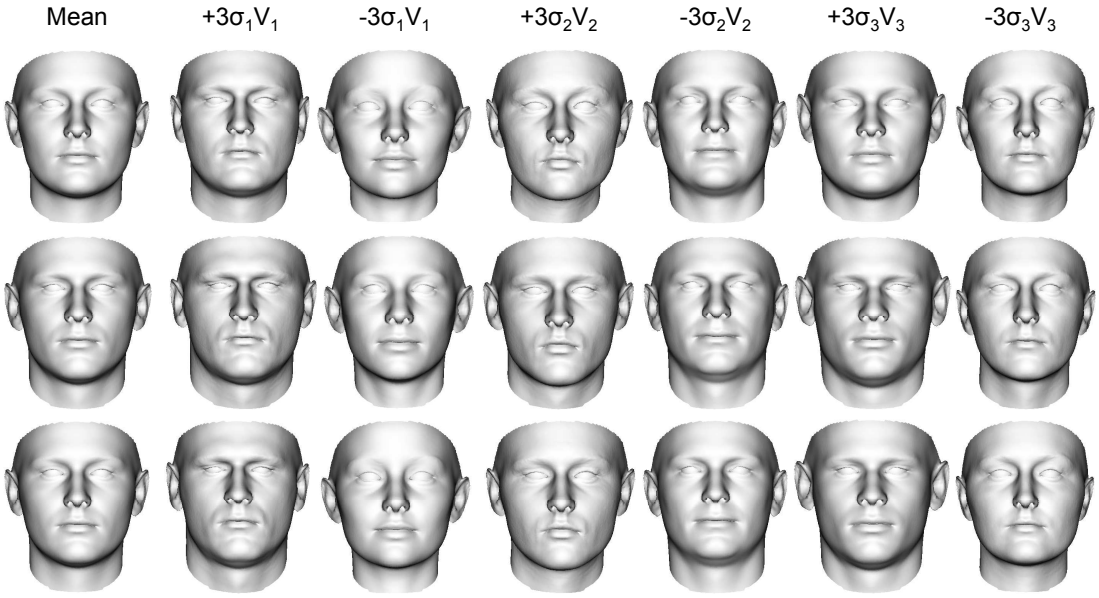

Figure 3.11: Top row shows mean face for female data and $\pm 3\sigma$ for the 3 most significant modes of variation "Model F". Second and third row shows the equivalent for male data "Model M" and combined data "Model C" respectively.

Modes	Bimodal	Unimodal
49	2.3606	2.3993
99	2.3895	2.4362

Table 3.4: Reconstruction error form 2D feature point (Basel renderings). Values are $\times 10^{12} \mu m^2$.

that distance to cluster centroids gives slightly better results. Table 3.4 shows reconstruction errors on the Basel renderings averaged over all subjects and pose angles.

3.3.5 Discussion

We have shown that a convex combination of class specific 3D morphable models is superior to a unimodal model. Our idea is motivated by examining feature distribution for 5 male and 5 female out-of-sample faces. Shape reconstruction error can be reduced by approximately $5-10$ % when fitting to both, 3D and 2D

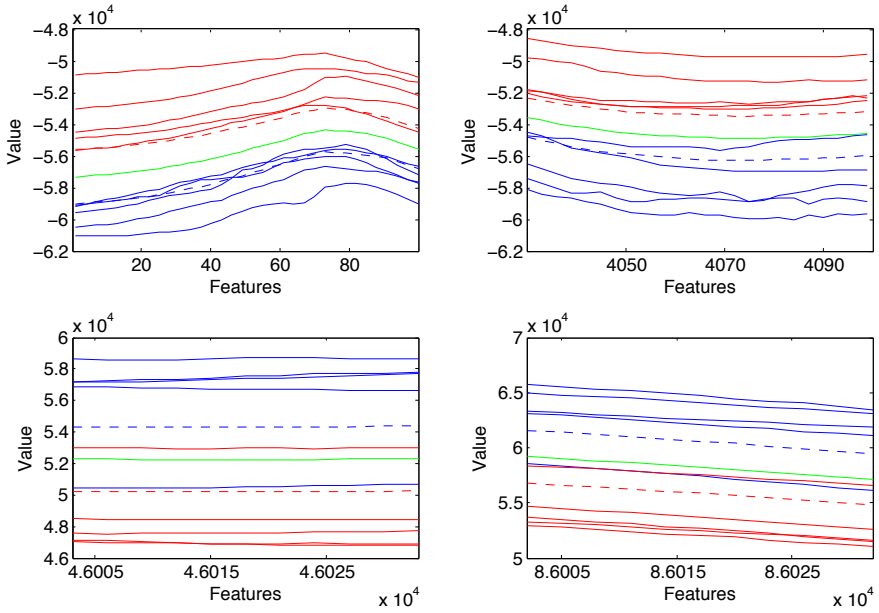

Figure 3.12: Distribution of features for various sections. Blue curves show male faces and red curves show female faces. Dashed curves show mean face for separate models. The green curve shows overall mean using the same training data.

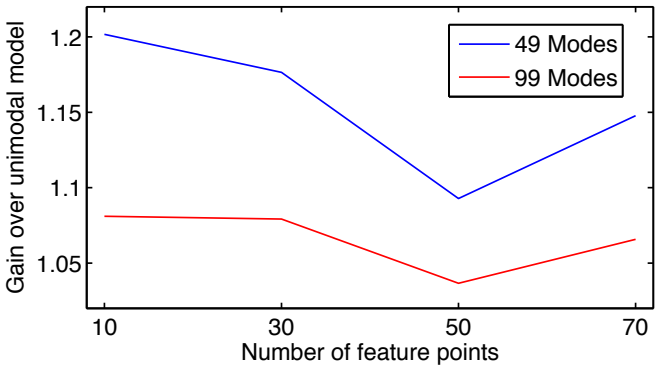

Figure 3.13: Gain in using a bimodal model.

feature points. In future work, we like to explore how the proposed method can be used in a gender classification task. The fact that we rely on sampling from an existing model (which is by construction unimodal) and we do not have access to the initial training-set, is a non negligible disadvantage. We use the BFM for subsequent experiments, since overall reconstruction errors are lower.

3.4 Conclusion

In this chapter, we proposed a novel shape fitting algorithm. Our approach uses a sparse set of feature points and negates the need to empirically choose the weight between prior and data. Our method is linear in the unknown shape and pose parameters can efficiently be solved in closed-form. The accuracy of our approach is comparable to a state-of-the-art analysis-by-synthesis algorithm, yet is orders of magnitude faster (less than a second using unoptimised Matlab code versus several minutes [9]). In addition, our empirical model of generalisation error was learnt using only 10 out-of-sample faces. Increasing this would likely improve results. We provide a comprehensive evaluation of the proposed method in terms of robustness to noise and pose accuracy. Our experiments indicate that the number of feature points alone is not the significant factor. On average, the shape reconstruction error is lower for close to profile views compared to front views, even though nearly half as many feature points are visible. This implies that the pose of a face effects the information content in a feature point observation. Examining the feature space shows, that out-of-sample faces occupy suggests 3D shapes to be bimodal distributed. We have therefore extended our algorithm to deal with mixture distributions. Experimental results indicate that using class specific priors outperform a standard PCA model, and that the level of gain increases with decreasing number of observations. More experiments with different training-sets are necessary to confirm the hypothesis.

Chapter 4

Texture and Illumination

In this chapter, we introduce three methods for modelling texture. The methods make increasingly weak assumptions about the illumination environment. The first method assumes Lambertian reflection only. The second and the third method explicitly take specular reflectance into account. The third method makes the least assumption about the illumination environment and models complex environment illumination of arbitrary colour. All three methods can be solved efficiently in a linear fashion.

4.1 Preliminaries

Before we propose our three methods for texture and illumination modelling, we revise some concepts that are used throughout this chapter. First we discuss spherical harmonic lighting, an efficient representation for reflectance under arbitrary illumination. Then, we describe the SUV colour space of Zickler et al. [117]. This is a source-dependent colour space which we make use of in Section 4.3 for specular invariant model fitting.

4.1.1 Spherical harmonic lighting

Illumination variation is responsible for large changes in 2D face appearance. In fact, changes in overall brightness due to lighting are almost always greater than changes due to identity [54]. We use the well known spherical harmonic

framework to efficiently represent reflectance under complex illumination. SH basis functions are well suited to be used in conjunction with 3DMMs, as the basis functions can be derived analytically from a given 3D face model.

Spherical harmonics are the natural extension of the Fourier representation to spherical functions. The seminal work of Ramamoorthi et al. [118; 119] showed that lighting, bidirectional reflectance distribution function (BRDF) and reflectance can be expressed using this series. Spherical harmonics form a set of orthonormal basis functions for the set of all square integrable functions defined on the unit sphere. In the frequency domain, the reflectance function is obtained by convolving the lighting function with the BRDF. The following shows the spherical expansion of a lighting function, L, as a function of the 3D surface normals x, y, z:

$$L(x, y, z) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \mathcal{L}_{l,m} \mathcal{H}_{l,m}(x, y, z),$$

where $\mathcal{H}_{l,m}$ are orthonormal SH basis functions and $\mathcal{L}_{l,m}$ are the corresponding weightings, which are termed the lighting coefficients. The subscript l denotes the degree and m its corresponding order. The basis functions are computed as follows:

$$\mathcal{H}_{l,m}(x, y, z) = \sqrt{\frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!}} P_{l,m}(z) e^{im\zeta},$$

where $\zeta = \arctan \frac{y}{x}$ and $P_{l,m}(z) = \frac{(1\text{-}z^2)^{m/2}}{2^l l!} \frac{d^{l+m}}{dz^{l+m}} (z^2\text{-}1)^l$ is the associated Legendre polynomial [120]. For $l = \{0, 1, 2\}$ the basis functions take the following form:

$$\mathcal{H}_{0,0} = \frac{1}{\sqrt{4\pi}}$$

$$\mathcal{H}_{1,\text{-}1} = \sqrt{\frac{3}{4\pi}} y \qquad \mathcal{H}_{1,0} = \sqrt{\frac{3}{4\pi}} z \qquad \mathcal{H}_{1,1} = \sqrt{\frac{3}{4\pi}} x$$

$$\mathcal{H}_{2,\text{-}2} = 3\sqrt{\frac{5}{12\pi}} xy \quad \mathcal{H}_{2,\text{-}1} = 3\sqrt{\frac{5}{12\pi}} yz \quad \mathcal{H}_{2,0} = \frac{1}{2}\sqrt{\frac{5}{4\pi}} (3z^2\text{-}1) \quad \mathcal{H}_{2,1} = 3\sqrt{\frac{5}{12\pi}} xz \quad \mathcal{H}_{2,2} = \frac{3}{2}\sqrt{\frac{5}{12\pi}} (x^2\text{-}y^2).$$

Figure 4.1 shows the above defined functions plotted on the surface of a sphere. A symmetric BRDF as a function of the incident elevation angle, can be expanded using the same set of basis functions. (See [119] for a detailed description. Note, that a symmetric BRDF is only a function of degree l and does not depend on

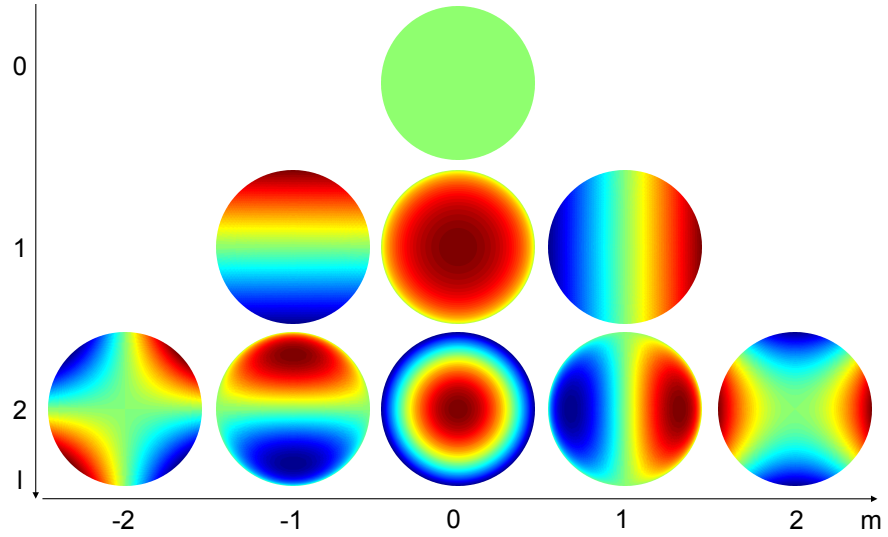

Figure 4.1: Graphical representation of spherical harmonic basis functions for $l = \{0, 1, 2\}$. The function values are symmetric about 0. Blue indicate negative and red positive values.

the harmonic order, m.) The BRDF parameters, $\hat{\rho}$, depend on surface properties. Any reflectance function, A, can be composed by multiplying corresponding frequency coefficients of lighting function and BRDF. In other words, the reflectance function is obtained by filtering the lighting function with the BRDF:

$$A(x, y, z) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \Lambda_l \hat{\rho}_l \mathcal{L}_{l,m} \mathcal{H}_{l,m}(x, y, z),$$

where $\Lambda_l = \sqrt{\frac{4\pi}{2l+1}}$ is a normalisation constant.

In most real world cases, we can not measure the coefficients of the lighting function directly. This would require the use of a light probe or panoramic camera placed in the scene. What we can measure is, in the Lambertian case, a low pass filtered version of the input signal, which is modelled by coefficients: $l_{l,m} = \Lambda_l \hat{\rho}_l \mathcal{L}_{l,m}$. We denote the concatenation of the diffuse coefficients as parameter vector, $\mathbf{l}$. Previous experimental results have shown that unconstrained complex illumination can well be approximated by a linear subspace. A second degree spherical harmonic approximation accounts for at least 98% in the variability of

the reflectance function:

$$B(x, y, z) = \sum_{l=0}^{2} \sum_{m=-l}^{l} l_{l,m} \mathcal{H}_{l,m}(x, y, z).$$

This result was independently derived by both Basri and Jacobs [96] and Ramamoorthi [118].

4.1.2 SUV Colour subspace

Recently, Zickler et al. [117] proposed a linear transformation in RGB colour space, which is invariant to specularities and preserves diffuse shading. Following the dichromatic model, observations $\mathbf{I}_k$ are linear combinations of the diffuse colour $\mathbf{D}$ and the specular colour $\mathbf{S}$:

$$I_c = \sigma_d D_c + \sigma_s S_c,$$

where the subscript c represents R,G and B respectively. The coefficients σ_d and σ_s are scaling factors dependent on material properties and shape. Separation of diffuse and specular components solely based on observations is an ill-posed problem. The method introduced in [117] proposes an efficient operation, which transforms observations into specular invariant representations. The new representation is termed SUV *colour space*. The transformation is defined as $\mathbf{I}_{SUV} = \mathbf{R}\mathbf{I}_{RGB}$, where the rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times3}$ aligns one of the axes (in this case the red axis) with the colour of the light source $\mathbf{S}$ and therefore satisfies the condition $\mathbf{RS} = (1, 0, 0)$. Due to this alignment, it can be shown that the intensities of the remaining channels (U and V) are functions of the diffuse part only and the following relation holds true:

$$I_U = \sigma_d \mathbf{r}_2^{\mathrm{T}} \mathbf{D} = \mathbf{r}_2^{\mathrm{T}} \mathbf{I}_{RGB}, \tag{4.1}$$

$$I_V = \sigma_d \mathbf{r}_3^{\mathrm{T}} \mathbf{D} = \mathbf{r}_3^{\mathrm{T}} \mathbf{I}_{RGB}. \tag{4.2}$$

The vectors $\mathbf{r}_2^{\mathrm{T}}$ and $\mathbf{r}_3^{\mathrm{T}}$ correspond the the 2nd and 3rd row of the rotation matrix $\mathbf{R}$, which can be obtained using quaternions. Assuming the source vector is the

light source vector $\mathbf{S}$ and the first coordinate axis $[1, 0, 0]$ is the destination vector $\mathbf{Z}$, we calculate quaternions $\mathbf{q} \in \mathbb{R}^4$ as follows:

$$\begin{aligned} \mathbf{S}_n &= \frac{\mathbf{S}}{\|\mathbf{S}\|} \\ \nu &= \frac{\mathbf{S}_n + \mathbf{Z}}{\|\mathbf{S}_n + \mathbf{Z}\|} \\ \mathbf{q}_1 &= \nu \times \mathbf{Z} \\ q_2 &= \nu \cdot \mathbf{Z} \\ \mathbf{q} &= [q_2, \mathbf{q}_1]. \end{aligned}$$

Assigning quaternions to elements $[a, b, c, d] = \mathbf{q}$, we construct $\mathbf{R}$:

$$\mathbf{R} = \begin{bmatrix} a^2 + b^2 - c^2 - d^2 & 2bc - 2ad & 2bd + 2ac \\ 2bc + 2ad & a^2 - b^2 + c^2 - d^2 & 2cd - 2ab \\ 2bd - 2ac & 2cd + 2ab & a^2 - b^2 - c^2 + d^2 \end{bmatrix},$$

which performs the desired alignment of the observations into a specular invariant representation.

4.2 Method 1: Colour channel ratios

Our statistical surface texture model captures variations in diffuse albedo. This forms one parameter of a number of possible parametric reflectance models which in turn determines the appearance of a face. By making assumptions about the surface reflectance and illumination, we are able to derive linear methods for fitting the texture model in an illumination-insensitive manner. Our first approach is the most restrictive, in that we neglect specular reflectance entirely and assume that all illumination in the scene is of the same, known colour.

4.2.1 Image formation process

The image formation model we use in our first texture fitting method assumes that surface reflectance is diffuse only and that illumination is provided by any combination of directional and ambient light sources of the same colour. Hence,

intensity is given by the following integral:

$$I_{\{r,g,b\}} = \rho_{\{r,g,b\}} S_{\{r,g,b\}} \int_{\Omega_{\mathbf{n}}} V_\omega L(\omega)(\mathbf{n} \cdot \omega) d\omega,$$

where $\Omega_{\mathbf{n}}$ is the upper hemisphere about the surface normal $\mathbf{n}$ and $\rho_{\{r,g,b\}}$ the diffuse albedo. We assume a distant lighting environment and hence, the incident radiance from direction ω is given by $L(\omega)$. This is globally scaled by the constant light source colour $S\{r, g, b\}$. V_ω is the visibility function which is equal to one if direction ω is unoccluded and zero otherwise. Note that every term in the integral is wavelength independent. The important observation is that taking ratios between pairs of colour channels cancels for all terms in the integral and is a function of the ratio of albedos only:

$$\frac{I_r S_g}{I_g S_r} = \frac{\rho_r}{\rho_g}. \tag{4.3}$$

Note that this relationship is independent of geometry entirely (both locally, via the surface normal, and globally, via occlusions).

4.2.2 Inverse rendering

We model diffuse albedo using our statistical texture model. Substituting the statistical model into Equation 4.3 (and assuming light source colour has been divided out of the image intensities) we obtain:

$$\frac{\mathbf{T}_{r(i)}\mathbf{b} + \bar{t}_{r(i)}}{\mathbf{T}_{b(i)}\mathbf{b} + \bar{t}_{b(i)}} = \frac{I_{r(i)}}{I_{b(i)}} \quad \text{and} \quad \frac{\mathbf{T}_{g(i)}\mathbf{b} + \bar{t}_{g(i)}}{\mathbf{T}_{b(i)}\mathbf{b} + \bar{t}_{b(i)}} = \frac{I_{g(i)}}{I_{b(i)}}, \tag{4.4}$$

where $\mathbf{T}_{r(i)}$ and $\bar{t}_{r(i)}$ represent the eigenvector and mean value for a corresponding observation $I_{r(i)}$ (in this case for the red channel). Image intensities are measured by sampling the image at the position of all visible (i.e. unoccluded) vertices in the face mesh. Equation 4.4 can be rewritten as follows:

$$(I_{b(i)}\mathbf{T}_{x(i)} - I_{x(i)}\mathbf{T}_{b(i)})\mathbf{b} = I_{x(i)}\bar{t}_{b(i)} - I_{b(i)}\bar{t}_{x(i)},$$

where the index x is substituted for r or g respectively. This gives us a linear system of equations of the following form:

$$\underbrace{\begin{pmatrix} I_{b(1)}\mathbf{T}_{r(1)} - I_{r(1)}\mathbf{T}_{b(1)} \\ I_{b(1)}\mathbf{T}_{g(1)} - I_{g(1)}\mathbf{T}_{b(1)} \\ \vdots \\ I_{b(k)}\mathbf{T}_{r(k)} - I_{r(k)}\mathbf{T}_{b(k)} \\ I_{b(k)}\mathbf{T}_{g(k)} - I_{g(k)}\mathbf{T}_{b(k)} \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} b_1 \\ \\ \vdots \\ \\ b_{m-1} \end{pmatrix}}_{\mathbf{b}} = \underbrace{\begin{pmatrix} I_{r(1)}\bar{t}_{b(1)} - I_{b(1)}\bar{t}_{r(1)} \\ I_{g(1)}\bar{t}_{b(1)} - I_{b(1)}\bar{t}_{g(1)} \\ \vdots \\ I_{r(k)}\bar{t}_{b(k)} - I_{b(k)}\bar{t}_{r(k)} \\ I_{g(k)}\bar{t}_{b(k)} - I_{b(k)}\bar{t}_{g(k)} \end{pmatrix}}_{\mathbf{h}},$$

with two equations per observed pixel value and can be solved using linear least squares: $\mathbf{b} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{h}$. A minimum of $k = m/2$ image-model correspondences are necessary to solve the system for m model parameters. In practice, many thousands of visible pixels are used.

4.3 Method 2: Specular invariant model fitting

The previous method makes very limiting assumptions about reflectance, neglecting specularities entirely. It also has no explicit model for illumination. Unfortunately, specularities cannot be accounted for using such a simple photometric invariant. In our next approach, we retain the flexibility of having an arbitrary distribution of illuminants but retain the requirement for fixed, known illumination colour (an extension of the method would allow a relaxation to two source colours as described in [117]). Under these assumptions, it is possible to transform to a specular-invariant space in which linear fitting of the texture model can take place.

4.3.1 Image formation process

We assume a dichromatic reflectance model comprising additive Lambertian and specular terms. Unlike the previous work of Blanz, Vetter and coworkers [55; 85; 86], we allow any combination of directed, ambient or extended light sources. However, to allow construction of the specular invariant space, we assume that all sources have the same colour. Implicit in the use of spherical harmonic illu-

mination is the assumption that the object is globally convex (i.e. occlusions are neglected). This leads to the following image formation model:

$$I_{\{r,g,b\}} = S_{\{r,g,b\}} \int_{\Omega_{\mathbf{n}}} L(\omega) \left[\rho_{\{r,g,b\}}(\mathbf{n} \cdot \omega) + s(\mathbf{n}, \omega, \nu) \right] d\omega,$$

where $s(\mathbf{n}, \omega, \nu)$ is an unknown specular reflectance function, which is assumed to be isotropic about the specular reflection direction $\mathbf{r} = 2(\mathbf{n} \cdot \nu)\mathbf{n} - \nu$, but otherwise unconstrained. ν is a unit vector in the viewing direction.

Substituting the statistical texture model for the diffuse albedo and using a spherical harmonic approximation to the diffuse and specular reflectances, the image formation process may be written in tensor notation as follows:

$$\mathbf{I}_{mod} = \mathcal{I}. * (\mathcal{C} \times_1 \mathbf{b} \times_2 \mathbf{l} + \mathcal{S}\mathbf{x}), \tag{4.5}$$

where $\mathcal{C}$ corresponds to a third order tensor spanning identity, expressed in terms of texture model coefficients $\mathbf{b}$, and diffuse illumination condition, expressed in terms of spherical harmonic illumination coefficients $\mathbf{l}$. $\mathcal{I} \in \mathbb{R}^{3p}$ denotes the colour of the light source $\mathbf{i} \in \mathbb{R}^3$, repeated according to number of vertices, p. We model specular contribution using an eighth order approximation, $\mathcal{S} \in \mathbb{R}^{3p \times 81}$, where the spherical harmonic basis is constructed by reflecting the viewing direction about surface normals. We found empirically, that an eighth order approximation is sufficient to model the specularities observed in typical skin reflectance. The specular spherical harmonic coefficients $\mathbf{x} \in \mathbb{R}^{81}$ capture information about the specular reflectance function and illumination environment. The symbol $.*$ denotes elementwise multiplication. In order to simplify the derivations, we may write the image formation process in matrix product notation:

$$\mathbf{I}_{mod} = \mathcal{I}. * [(\mathcal{H}\mathbf{l}). * (\mathbf{T}\mathbf{b} + \bar{\mathbf{t}}) + \mathcal{S}\mathbf{x}], \tag{4.6}$$

where $\mathcal{H} \in \mathbb{R}^{3p \times 9}$ are diffuse SH basis functions obtained from the surface normals of the estimated face shape, and $\mathbf{T}\mathbf{b} + \bar{\mathbf{t}} \in \mathbb{R}^{3p}$ denotes diffuse albedo as approximated by the linear texture model. Note that in Equation 4.5 the mean

texture $\bar{\mathbf{t}}$ is included in the tensor $\mathcal{C}$, the first element of $\mathbf{b}$ is fixed to 1 and the parameter vector is one dimension higher than in Equation 4.6. For convenience, we define:

$$\mathbf{t}_k = \mathbf{T}_k\mathbf{b} + \bar{\mathbf{t}}_k, \qquad \mathbf{d}_k = \mathcal{H}_k \cdot \mathbf{l}, \qquad \mathbf{s}_k = \mathcal{S}_k \cdot \mathbf{x}.$$

The subscript k indicates the rows corresponding to the R, G and B channels of the kth vertex. According to the specular invariant SUV colour space defined in Equation 4.1, applying a rotation to the observed intensities allows us to relate diffuse intensity only to the observations:

$$\begin{aligned} \mathbf{r}_2^{\mathrm{T}}\mathbf{I}_{RGB,k} &= \mathbf{r}_2^{\mathrm{T}}[\mathbf{i}. * (\mathbf{d}_k. * \mathbf{t}_k + \mathbf{s}_k)] \\ &= \mathbf{r}_2^{\mathrm{T}}(\mathbf{i}. * \mathbf{d}_k. * \mathbf{t}_k) \\ &= I_{U,k}. \end{aligned}$$

Where $\mathbf{I}_{RGB,k} \in \mathbb{R}^3$ is a single observation corresponding to the kth vertex. As in Equation 4.2, the same applies for the V channel, $I_{V,k}$. Thus, each observation, k, results in two specular invariant equations which relate texture model parameters and observed intensities.

4.3.2 Diffuse inverse rendering

The unknowns in our specular invariant representation are the texture parameters $\mathbf{b}$ and diffuse lighting coefficients $\mathbf{l}$. The result is a bilinear system of equations relating the unknowns to the observations via the specular invariant space. The objective function comprises two terms:

$$\mathbb{E} = \mathbb{E}_U + \mathbb{E}_V.$$

For K intensity observations, the individual parts for U and V channel are defined as follows:

$$\mathbb{E}_U = \sum_{k=1}^{K} \left[\mathbf{r}_2^{\mathrm{T}}\mathbf{I}_k - \mathbf{r}_2^{\mathrm{T}}(\mathbf{i}. * \mathbf{d}_k. * \mathbf{t}_k)\right]^2,$$
$$\mathbb{E}_V = \sum_{k=1}^{K} \left[\mathbf{r}_3^{\mathrm{T}}\mathbf{I}_k - \mathbf{r}_3^{\mathrm{T}}(\mathbf{i}. * \mathbf{d}_k. * \mathbf{t}_k)\right]^2.$$

As the error function $\mathbb{E}$ is quadratic in terms of the parameters $\mathbf{b}$ and $\mathbf{l}$, we calculate the partial derivatives of $\mathbb{E}$ with respect to each parameter by keeping the remaining parameter constant. This leads to a bilinear solution:

$$\frac{\partial \mathbb{E}}{\partial \mathbf{b}} = \frac{\partial \mathbb{E}_U}{\partial \mathbf{b}} + \frac{\partial \mathbb{E}_V}{\partial \mathbf{b}} \quad \text{and} \quad \frac{\partial \mathbb{E}}{\partial \mathbf{l}} = \frac{\partial \mathbb{E}_U}{\partial \mathbf{l}} + \frac{\partial \mathbb{E}_V}{\partial \mathbf{l}}$$

We set to zero and obtain closed-form solutions for $\mathbf{b}$ and $\mathbf{l}$, respectively. Both sets of parameters are obtained using alternating least squares. This allows us to recover albedo and diffuse lighting parameters in a specular invariant manner. As the objective function is convex in both parameters, a global solution is guaranteed with the accuracy of the solution determined by the number of iterations of alternating least squares. In practice this converges within 3-5 iterations. Because the problem is convex, the solution is independent of initialisation. However, an initialisation which seems to yield swift convergence is to set the texture parameter to zero (i.e. the mean texture) and solve for illumination first.

4.3.3 Specular inverse rendering

With diffuse reflectance factored into albedo and illumination estimates, we now proceed to model specular reflectance. This is solved in two steps. For low frequency, $l \leq 2$, we use the illumination environment estimated in the diffuse fitting stage. For higher frequencies, $3 \leq l \leq 8$, we use an unconstrained optimisation procedure and, hence, the contribution of higher frequency illumination to specular reflectance is free to vary independently. The problem can be stated as: $\bar{\mathbf{I}}_s = \bar{\mathbf{I}}_{s,l} + \bar{\mathbf{I}}_{s,h}$. We also assume that specular reflectance is symmetric about the reflection vector.

Low order specular reflectance Using the estimated parameters: $\mathbf{b}$ and $\mathbf{l}$, we synthesise a diffuse-only image and subtract from the input image to obtain the specular-only image, $\bar{\mathbf{I}}_s$:

$$\bar{\mathbf{I}}_s = \mathbf{I} - \mathcal{I}. * [(\mathcal{H}\mathbf{l}). * (\mathbf{T}\mathbf{b} + \bar{\mathbf{t}})].$$

We clamp negative values to zero (these are caused by cast shadows or errors in the diffuse estimate). The lighting coefficients $\mathcal{L}_{lm}$ are obtained by dividing diffuse coefficients, l_{lm} by the Lambertian BRDF parameters, which are constant for a given order.

Specular reflection requires an alternate basis set constructed with respect to the reflection vector. Hence, we reflect the the viewing direction about the normals and define new specular basis functions $\mathcal{S}(x', y', z')$. Since the illumination environment is already known, the isotropic specular reflectance function has only 3 free parameters $\hat{\tau}_l$, where $l \in \{0, 1, 2\}$ which can be obtained by solving the following linear system of equations:

$$\begin{aligned}\bar{\mathbf{I}}_{s,l} = {} & \hat{\tau}_0 \mathcal{S}_0 \mathcal{L}_0 + \hat{\tau}_1 (\mathcal{S}_{1,-1}\mathcal{L}_{1,-1} + \mathcal{S}_{1,0}\mathcal{L}_{1,0} + \mathcal{S}_{1,1}\mathcal{L}_{1,1}) \\ & + \hat{\tau}_2 (\mathcal{S}_{2,-2}\mathcal{L}_{2,-2} + \mathcal{S}_{2,-1}\mathcal{L}_{2,-1} + \mathcal{S}_{2,0}\mathcal{L}_{2,0} + \mathcal{S}_{2,1}\mathcal{L}_{2,1} + \mathcal{S}_{2,2}\mathcal{L}_{2,2}). \end{aligned} \tag{4.7}$$

Multiplying specular BRDF parameters $\hat{\tau}_l$ with the corresponding lighting coefficients, $\mathcal{L}_{l,m}$, results in the specular coefficients $x_{l,m}$, which are concatenated to form vector $\mathbf{x}_l \in \mathbb{R}^9$.

Higher order specular reflectance For orders $l \in \{3, \ldots, 8\}$ both lighting and BRDF are unknown. A unique solution does not exist for this problem. However, it is possible to solve for higher order coefficients which capture both illumination and reflectance properties. These could be factored into separate contributions by making additional assumptions. Obtaining these combined coefficients again requires solution of a linear system:

$$\bar{\mathbf{I}}_{s,h} = [\mathcal{S}_{8,-8}\mathcal{S}_{8,-7} \ldots \mathcal{S}_{3,-3} \ldots \mathcal{S}_{3,3} \ldots \mathcal{S}_{8,7}\mathcal{S}_{8,8}]\mathbf{x}_h. \tag{4.8}$$

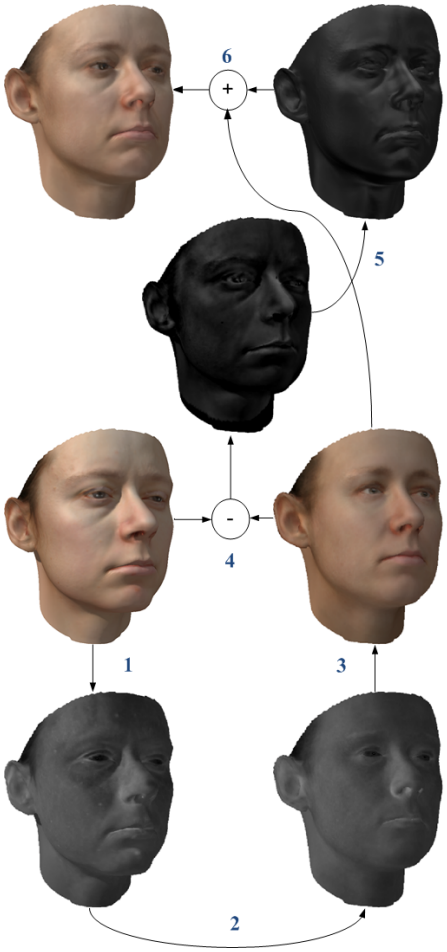

Figure 4.2: Overview of the inverse rendering pipeline (see text). Each step in the pipeline requires only the solution of a system of linear equations.

We solve for $\mathbf{x}_h \in \mathbb{R}^{72}$ using least squares, and construct the final specular coefficient vector by concatenating the low order and high order solutions: $\mathbf{x} = \begin{bmatrix} \mathbf{x}_l & \mathbf{x}_h \end{bmatrix} \in \mathbb{R}^{81}$.

4.3.4 Inverse rendering pipeline

Figure 4.2 summarises our fitting procedure graphically. The steps in the pipeline are as follows:

1. The input image is transformed into U-V colour space by aligning one of the RGB axis with the colour of the light source. We use quaternions for this

transformation. This results in two specular observations per vertex-colour.

2. We estimate coefficients of a linear texture model in conjunction with diffuse lighting parameters in a bilinear fashion. We use SH basis functions up to order 2, which are computed analytically from 3D shape information. Here we show $\sqrt{u^2 + v^2}$ for visualisation purposes.
3. This results in a diffuse-only reconstruction.
4. Taking the difference of the input image and our diffuse estimation allows to calculate a specular-only image.
5. Using an alternative set of SH basis functions (normals reflected around the viewing direction) enables estimation of the specular contribution of the input image.
6. Adding specular estimation to the diffuse-only reconstruction leads to the final result.

The SH basis functions are computed using the surface normals of the mesh computed at the shape fitting stage.

4.4 Method 3: Unconstrained illumination

In this section, we relax our assumptions further. We retain the assumption of additive Lambertian and specular terms, where the specular function is assumed to be isotropic about the specular reflection direction but is otherwise unconstrained. However, we allow any combination of directed, ambient or extended light sources of arbitrary and varying colour.

4.4.1 Image formation process

Our final image formation model allows for arbitrarily coloured environment illumination:

$$I_{\{r,g,b\}} = \int_{\Omega_{\mathbf{n}}} L_{\{r,g,b\}}(\omega) \left[\rho_{\{r,g,b\}}(\mathbf{n} \cdot \omega) + s(\mathbf{n}, \omega, \nu)\right] d\omega,$$

where the illumination function now has a wavelength dependence.

Texture and illumination

We now proceed to express this model in terms of a multilinear system of equations. We construct a set of basis functions, $\mathcal{U}(x, y, z)$, derived from $\mathcal{H}(x, y, z)$. For a single vertex k, the modified basis functions are defined as follows:

$$\mathcal{U}(x, y, z)_k = \begin{bmatrix} \mathcal{H}(x, y, z)_k & \mathbf{0}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \mathcal{H}(x, y, z)_k & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} & \mathcal{H}(x, y, z)_k \end{bmatrix}. \tag{4.9}$$

We also take specular reflection of arbitrary unconstrained illumination into account. We substitute the basis functions $\mathcal{H}(x, y, z)_k$ in Equation 4.9 with ones constructed using the reflected view vectors and construct the specular set: $\mathcal{S}(x, y, z)_k$. We denote a column of this matrix as $\mathcal{S}(x, y, z)_k^c$, where $c \in \{r, g, b\}$. This notation will be used later when fitting the specular part. In tensor notation, the full image formation process can be stated as:

$$\mathbf{I}_{mod} = \mathcal{V} \times_1 \mathbf{b} \times_2 \mathbf{l} + \mathcal{S}\mathbf{x},$$

where similar to Equation 4.5, $\mathcal{V}$ corresponds to a third order tensor spanning identity and diffuse illumination, however this time of arbitrary colour. In matrix product notation, we state the image formation process as:

$$\mathbf{I}_{mod} = (\mathcal{U}\mathbf{l}). * (\mathbf{T}\mathbf{b} + \bar{\mathbf{t}}) + \mathcal{S}\mathbf{x}.$$

Linear colour transformation

To make the fitting algorithm more flexible and allows fitting to different imaging properties, we estimate a linear colour transformation $\mathbf{M}(\cdot) \in \mathbb{R}^{3\times3}$ and add offset $\mathbf{o} \in \mathbb{R}^3$, firstly proposed for morphable model fitting in [9], to the illuminated estimations $(\cdot)$. Decomposition of $\mathbf{M}$ into individual contributions is achieved as follows:

$$\mathbf{M} = \mathbf{G}\mathbf{C} = \begin{pmatrix} g_r & 0 & 0 \\ 0 & g_g & 0 \\ 0 & 0 & g_b \end{pmatrix} \cdot \left[c\mathbf{I} + (1 - c) \begin{pmatrix} 0.3 & 0.59 & 0.11 \\ 0.3 & 0.59 & 0.11 \\ 0.3 & 0.59 & 0.11 \end{pmatrix} \right],$$

where the entries g_r, g_g and g_b are gains for RGB respectively, and c corresponds to the contrast value. The gains are lower bound to 0 and contrast is constrained to lie between 0 and 1.

The complete model

The entire image formation process is a multilinear system which consists of two nested bi-affine parts. For a single vertex, k, the image formation is modelled as:

$$\mathbf{I}_{mod,k} = \mathbf{M}[(\mathcal{U}_k\mathbf{l}).*(\mathbf{T}_k\mathbf{b} + \bar{\mathbf{t}}_k) + \mathcal{S}_k\mathbf{x}] + \mathbf{o}. \tag{4.10}$$

4.4.2 Inverse rendering

We now show how the unknowns in Equation 4.10 can be recovered. This amounts to a series of linear least squares problems. The entire system can be iterated to convergence (which will correspond to the global minimum) but we have found that one pass, as described, is sufficient for good results. We begin by estimating the colour transformation parameters using the mean texture as initialisation. We then correct for these transformations and denote the colour corrected observations as $\bar{\mathbf{I}}$.

Diffuse component

The diffuse and specular shading coefficients, $\mathbf{l}$ and $\mathbf{x}$, both depend on a single lighting function: $\mathcal{L} = [\mathcal{L}_r^{\mathrm{T}}\mathcal{L}_g^{\mathrm{T}}\mathcal{L}_b^{\mathrm{T}}]$. As the lighting function can not be estimated directly from a single 2D image, we start by estimating $\mathbf{l}$ and $\mathbf{b}$ in a bilinear fashion. Ignoring the specular part at this stage, we minimise the following objective function, which depends on the colour transformation and observations:

$$\mathbb{E}_d = \|\bar{\mathbf{I}} - (\mathcal{U}\mathbf{l}).*(\mathbf{T}\mathbf{b} + \bar{\mathbf{t}})\|^2. \tag{4.11}$$

Regularisation

To prevent overfitting, we introduce two priors on the parameters which encourage simplicity: $\mathbb{E}_1 = \|\mathbf{b}\|^2$ and $\mathbb{E}_2 = \|\mathbf{l}\|^2$. We also define a "grayworld" prior, $\mathbb{E}_3$, which prefers white illumination. We implement this constraint by encouraging the difference between $\mathcal{L}_r^{\mathrm{T}}$, $\mathcal{L}_g^{\mathrm{T}}$ and $\mathcal{L}_b^{\mathrm{T}}$ to be small. To do so we define three filter matrices: $\mathtt{F}_r = \begin{bmatrix}\mathbf{I} & \mathbf{0} & \mathbf{0}\end{bmatrix}$, $\mathtt{F}_g = \begin{bmatrix}\mathbf{0} & \mathbf{I} & \mathbf{0}\end{bmatrix}$, and $\mathtt{F}_b = \begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{I}\end{bmatrix}$, where $\mathbf{I}, \mathbf{0} \in \mathbb{R}^{9\times 9}$ correspond to the identity and null matrix respectively. The constraint takes the following form: $\mathbb{E}_3 = \|\mathtt{F}_r\mathbf{l} - \mathtt{F}_g\mathbf{l}\|^2 + \|\mathtt{F}_r\mathbf{l} - \mathtt{F}_b\mathbf{l}\|^2 + \|\mathtt{F}_g\mathbf{l} - \mathtt{F}_b\mathbf{l}\|^2$. All priors are added to Equation 4.11 to form the overall cost function: $\mathbb{E}_a = \mathbb{E}_d + \lambda_1\mathbb{E}_1 + \lambda_2\mathbb{E}_2 + \lambda_3\mathbb{E}_3$.

Diffuse bi-affine system

The objective function $\mathbb{E}_a$ is convex in $\mathbf{b}$ and $\mathbf{l}$. We treat both sets as independent contributions, find the partial derivatives, set to zero:

$$\frac{\partial \mathbb{E}_a}{\partial \mathbf{b}} = \frac{\partial \mathbb{E}_d}{\partial \mathbf{b}} + \frac{\partial \lambda_1 \mathbb{E}_1}{\partial \mathbf{b}} = 0, \tag{4.12}$$

$$\frac{\partial \mathbb{E}_a}{\partial \mathbf{l}} = \frac{\partial \mathbb{E}_d}{\partial \mathbf{l}} + \frac{\partial \lambda_2 \mathbb{E}_2}{\partial \mathbf{l}} + \frac{\partial \lambda_3 \mathbb{E}_3}{\partial \mathbf{l}} = 0, \tag{4.13}$$

and solve for both sets using alternating least squares. The solution is independent of initialisation, although using the mean texture results in swift convergence, typically within ≤ 5 iterations.

Specular component

Similar to the previous method (Section 4.3.3), specular reflectance is solved in two steps. For low frequency, $l \leq 2$, we use the illumination environment estimated in the diffuse fitting stage. For higher frequencies, $3 \leq l \leq 8$, we use an unconstrained optimisation procedure. Since in this section we allow the illumination environment to be of arbitrary colour, Equation 4.7 is modified:

$$\begin{aligned}\bar{\mathbf{I}}_{s,l} = &\hat{\tau}_0 \sum_c (\mathcal{S}_0^c \mathcal{L}_0^c) + \hat{\tau}_1 \sum_c (\mathcal{S}_{1,-1}^c \mathcal{L}_{1,-1}^c + \mathcal{S}_{1,0}^c \mathcal{L}_{1,0}^c + \mathcal{S}_{1,1}^c \mathcal{L}_{1,1}^c) + \\ &\hat{\tau}_2 \sum_c (\mathcal{S}_{2,-2}^c \mathcal{L}_{2,-2}^c + \mathcal{S}_{2,-1}^c \mathcal{L}_{2,-1}^c + \mathcal{S}_{2,0}^c \mathcal{L}_{2,0}^c + \mathcal{S}_{2,1}^c \mathcal{L}_{2,1}^c + \mathcal{S}_{2,2}^c \mathcal{L}_{2,2}^c).\end{aligned}$$

Multiplying specular BRDF parameters, $\hat{\tau}_l$, with the corresponding lighting coefficients, $\mathcal{L}^c_{l,m}$, results in the specular coefficients, $x^c_{l,m}$. These are concatenated to form vector $\mathbf{x}_l \in \mathbb{R}^{27}$. For higher orders $l \in \{3, \ldots, 8\}$, we change Equation 4.8 to :

$$\bar{\mathbf{I}}^c_{s,h} = \begin{bmatrix} \mathcal{S}^c_{8,-8} & \mathcal{S}^c_{8,-7} & \ldots & \mathcal{S}^c_{3,-3} & \ldots & \mathcal{S}^c_{3,3} & \ldots & \mathcal{S}^c_{8,7} & \mathcal{S}^c_{8,8} \end{bmatrix} \mathbf{x}^c_h,$$

solve for $\mathbf{x}^c_h \in \mathbb{R}^{72}$ and construct $\mathbf{x}_h = \begin{bmatrix} \mathbf{x}^r_h & \mathbf{x}^g_h & \mathbf{x}^b_h \end{bmatrix} \in \mathbb{R}^{216}$.

Colour transformation parameters

Having an estimate for $\mathbf{b}, \mathbf{l}$ and $\mathbf{x}$, we can synthesise vertex k and solve for the unknown colour transformation. In the previous paragraphs $\bar{\mathbf{I}}$ referred to the model without colour transformation applied to. We stick with this notation and denote $\bar{\mathbf{I}}_o$ for the model without offset applied to.

$$\mathbf{I}_{mod,k} = \mathbf{M}\bar{\mathbf{I}}_{o,k} + \mathbf{o} = \mathbf{GC}\bar{\mathbf{I}}_{o,k} + \mathbf{o}.$$

This leads to the second bi-affine system:

$$\mathbb{E}_c = \|\bar{\mathbf{I}} - (\mathbf{GC}\bar{\mathbf{I}}_{o,k} + \mathbf{o})\|^2.$$

Again, we set the partial derivatives with respect to the parameters $\mathbf{g}$ and c to zero and solve using alternating least squares.

4.5 Experiments

In this section, we present thorough evaluation for the proposed methods. We show results on synthetic data, real world imagery (captured by ourselves) and the CMU-PIE database [121].

4.5.1 Texture

The out-of-sample faces are accompanied with ground truth data. This allows to compare the fittings in terms of an absolute reconstruction error. Our evaluation

	$\mathbb{E}_g \times 10^{-3}$				$\mathbb{E}_m$			
Subject:	M3	M2	M1	[55]	M3	M2	M1	[55]
001	6.0	7.9	37.3	8.8	5.0	6.6	31.2	7.4
002	4.1	6.6	24.7	7.5	3.1	5.0	18.7	5.7
006	8.2	6.4	16.4	7.5	6.3	4.9	12.6	5.7
014	3.7	6.2	25.0	4.4	4.3	7.2	28.9	5.1
027	17.0	9.0	67.8	14.2	12.9	6.8	51.5	10.8
022	3.5	6.0	28.2	11.0	3.2	5.3	25.2	9.8
052	5.0	8.9	39.8	12.4	5.3	9.5	42.5	13.3
053	4.6	6.9	31.1	12.0	4.7	7.0	31.9	12.2
293	5.3	5.9	35.3	7.4	8.1	9.0	53.8	11.4
323	4.7	6.9	35.4	5.2	4.3	6.2	31.6	4.7
Mean	6.2	7.1	34.1	9.0	5.7	6.7	32.8	8.6

Table 4.1: Texture reconstruction errors over all 9 pose angles and 3 illumination conditions for different subjects. The 60 most significant modes are used for reconstructions.

is based on the following two error measures:

$$\mathbb{E}_g = \frac{1}{n}\|\mathbf{t}_g - \mathbf{t}_r\|^2 \quad \text{and} \quad \mathbb{E}_m = \frac{\|\mathbf{t}_g - \mathbf{t}_r\|^2}{\|\mathbf{t}_m - \mathbf{t}_r\|^2}, \tag{4.14}$$

where, $\mathbf{t}_g$ is the ground truth texture data provided with the out-of-sample faces and $\mathbf{t}_m$ is the ground truth data projected into the 60 parameter model (i.e. the optimal model fit to the data). $\mathbf{t}_r$ is the texture recovered using the proposed methods and the reference method. Individual values within each texture vector are within $\mathbb{R} \in [0, 1]$. $\mathbb{E}_m$ is a relative error measure which relates our result to the best possible the model can achieve for a particular out-of-sample face (a value of 1 denotes optimal performance). Table 4.1 shows fitting results averaged over subjects, and Table 4.2 compares the reconstructions averaged over pose angles. The methods M2 and M3 show improved results over [55]. Qualitative results for method M3 are shown in Figure 4.3. However, the results can still be improved by a factor of more than five (the best value we obtained for $\mathbb{E}_m = 5.7$). In a final experiment, we performed a recognition experiment using the obtained parameters for method M3. In a similar fashion as for the experiment in Section 3.2.1, each of the 27 renderings per subject serve as gallery image. The remaining fitting results are associated with the closest gallery image in terms of angular distance. Table 4.3 shows recognition error rates in percentage averaged over 3 illumination conditions per pose.

	$\mathbb{E}_g \times 10^{-3}$				$\mathbb{E}_m$			
Pose:	M3	M2	M1	[55]	M3	M2	M1	[55]
$-70°$	5.6	8.3	33.7	5.9	5.2	7.9	32.3	5.4
$-50°$	5.7	6.9	38.1	6.8	5.3	6.5	36.8	6.3
$-30°$	5.8	6.9	33.9	11.8	5.4	6.6	32.6	11.1
$-15°$	5.9	5.4	27.5	9.4	5.4	5.2	26.3	8.8
$0°$	6.1	5.4	25.3	11.0	5.6	5.1	24.3	10.4
$15°$	6.7	7.1	27.9	10.9	6.1	6.8	26.8	10.5
$30°$	6.4	8.4	36.1	11.2	5.9	8.2	35.0	11.0
$50°$	6.6	7.5	45.1	8.2	6.0	7.2	43.4	8.2
$70°$	7.0	7.6	39.3	6.0	6.3	7.3	37.6	5.5
Mean	6.2	7.1	34.1	9.0	5.7	6.7	32.8	8.6

Table 4.2: Texture reconstruction errors over all 10 subjects and 3 illumination conditions for different pose angles. The 60 most significant modes are used for reconstructions.

Pose:	$-70°$	$-50°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$50°$	$70°$	**mean**
M3:	92.0	94.8	94.9	98.4	95.9	94.9	94.4	96.0	95.8	95.0
[55]	81.0	92.0	89.9	91.7	91.0	88.1	82.6	84.7	85.9	87.4

Table 4.3: Mean rank-1 recognition error rates for all 270 renderings averaged over 3 illumination conditions per pose. Results are shown for texture-only.

4.5.2 Environment map rendering

The rendering provided with the BFM use a very simple illumination environment comprising a single directional and ambient source, both of a white colour. To simulate complex illumination, we rendered images using environment maps. We provide qualitative results for texture fitting method M3. To do so we used environment maps provided by the University of Southern California [122]. Results for two examples are shown in Figure 4.4. The image on the top row termed "Input Rendering", serves as input to our algorithm. The other images on the top row are shown for comparison purposes only.

4.5.3 Grayscale synthetic data

Proposed method M3 can be used to fit the colour model to grayscale imagery. We convert the rendering for one subject to grayscale and use the colour transformation described in Section 4.4.1 to fit. Figure 4.5 shows the results for subject one of the test-set in two poses, 0 and -70, and 3 illumination conditions. For

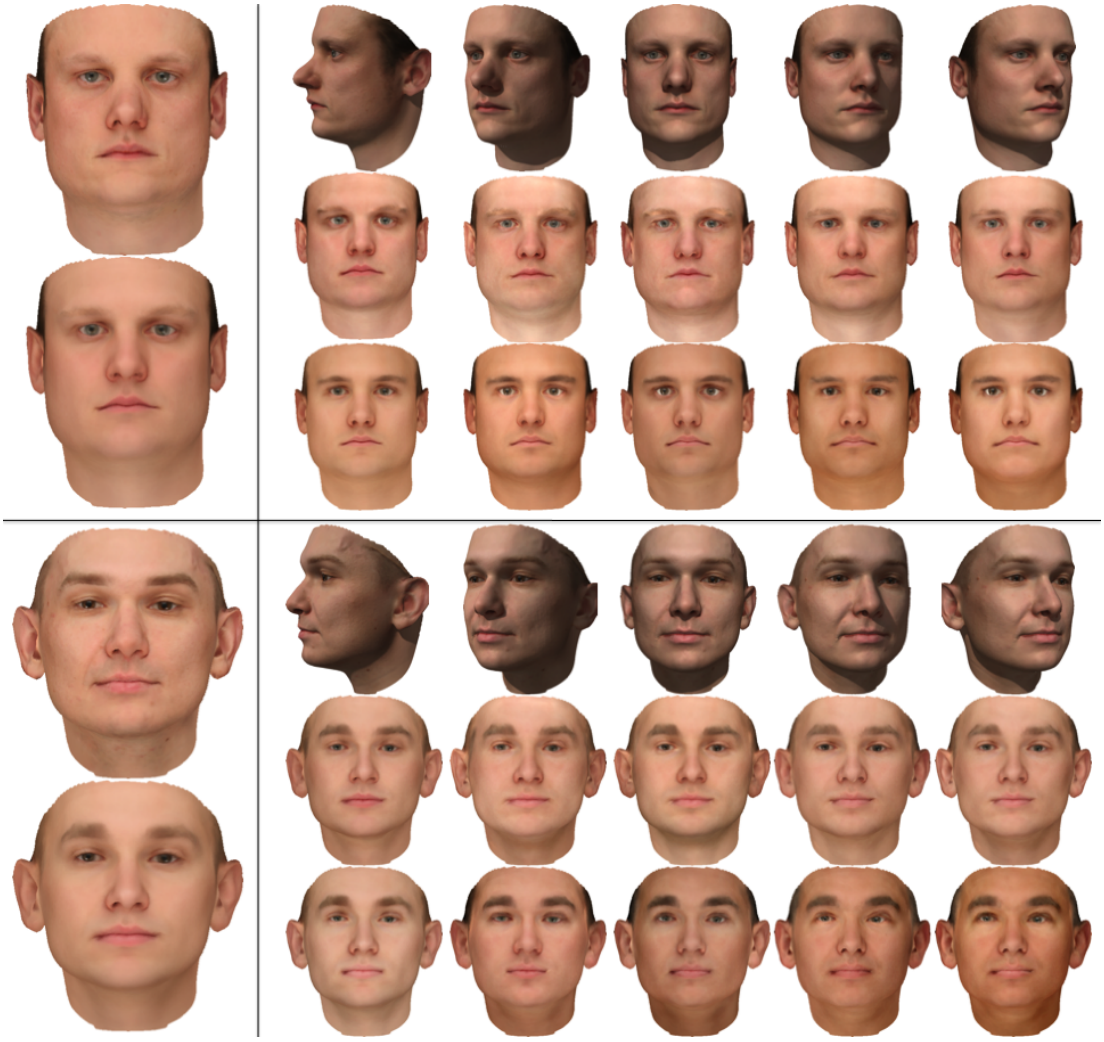

Figure 4.3: Fitting results to out-of-sample renderings for two subjects. Left column shows ground truth (top) projected into the model using the 60 most significant modes (below). Second column: Top row shows out-of-sample renderings in different pose and illumination conditions. Second row shows texture reconstruction using method M3. Last row shows texture fitting results using the reference method [55].

comparison, we also show fitting results for the original colour images. Using the error measure introduced in Equation 4.14, and averaging over all 27 renderings we obtain the following values for face one: $\mathbb{E}_g = 6.3$ and $\mathbb{E}_m = 5.27$.

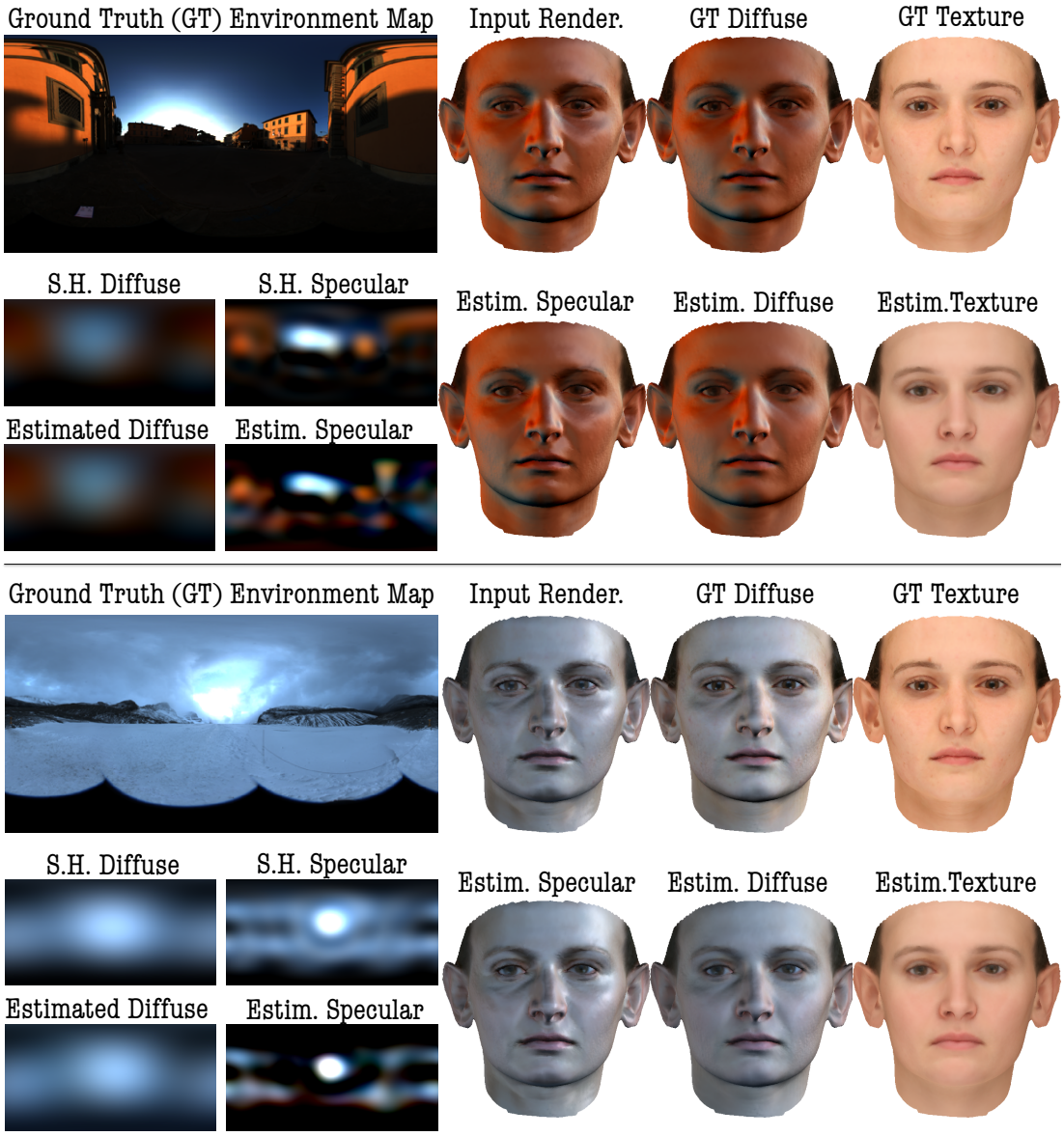

Figure 4.4: Two examples of out-of-sample face convolved with spherical harmonic approximation of additive specular and diffuse environment maps. "Input Rendering" serves as input to the algorithm. The second row shows reconstruction results for method M3. Estimated diffuse and specular environment maps are shown compared to ground truth on the bottom left hand side.

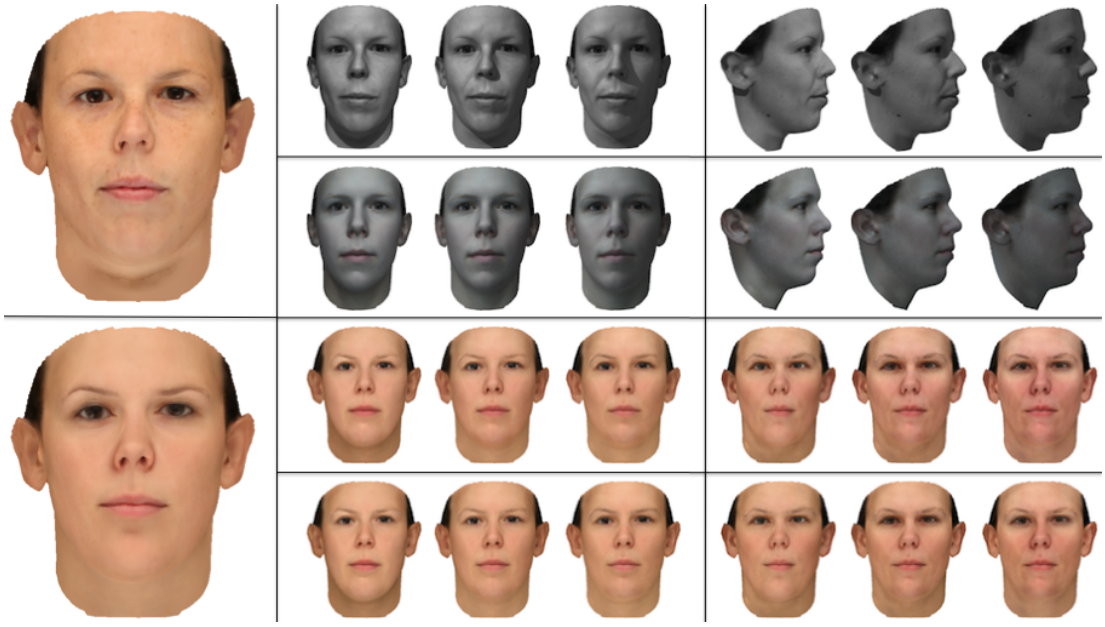

Figure 4.5: Fittitng results to grayscale images. First column shows ground truth shape and texture (top) and texture projected into the model (60 modes). Column 2 and 3: Top row shows input images 0 and −70 in 3 illumination conditions respectively. Second row shows fitting results with all parameters estimated, including shape. Third row shows the fitting results for texture only. Last row shows fitting results to the equivalent RGB images using the same method. They demonstrate how fitting accuracy to grayscale images degrades for varying illumination in more extreme pose angles.

4.5.4 Real world images

We captured photographs of 4 subjects disjoint from the training data using a Nikon D200 camera. This part of the experimental section demonstrates the full inverse rendering process, including shape and camera properties. Visual results are shown in Figure 4.6.

4.5.5 Discussion

Experimental results on synthetic data show that our methods show promise. In some cases, for instance proposed method M3, it even allows us to outperform more computationally expensive anaysis-by-synthesis approaches. Synthetic data however is not representative of real world conditions where data can be severely

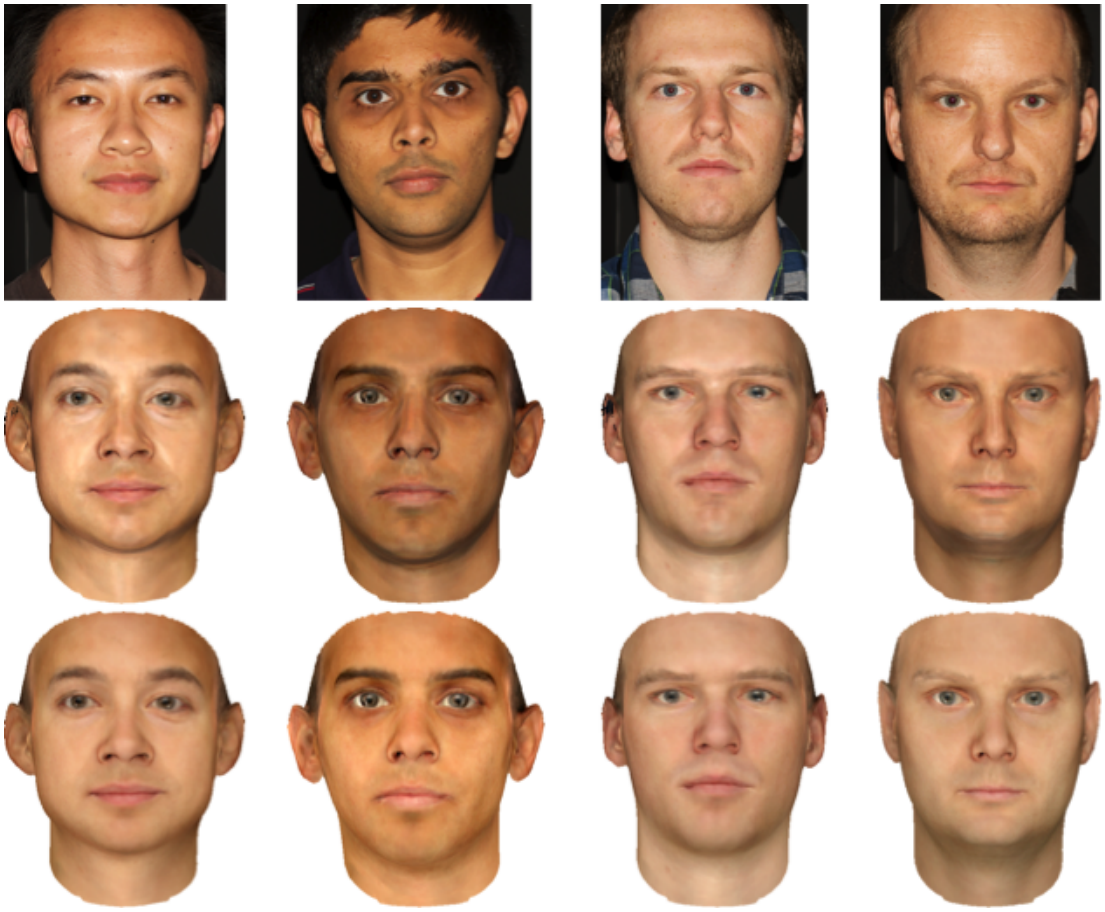

Figure 4.6: Face shape and appearance modelling from photographs. Top row shown photographs of 4 subjects, not part of the training data. Second row shows a full model of the photograph (Shape,camera properites, albedo, colour transformation, diffuse and specular lighting). Last row shows diffuse albedo only.

afflicted by outliers, partial occlusions and substantial variability of pose and illumination. In the following section, we address these issues and show qualitative and quantitative results on the CMU-PIE database.

4.5.6 CMU-PIE Database

The CMU-PIE database is a de-facto standard for face recognition experiments [121]. The database consists of 68 subjects in different pose, illumination and expression conditions. Moreover, in contrast to our synthetic data above, the

images were obtained using a camera different to that used to obtain model data. We address this problem by estimating a linear colour transformation between the model and observations (see Section 4.4.1).

Recognition

The statistical model used throughout this work does not account for changes in expression. The experiments are therefore restricted to the expression neutral subset of the database, which nevertheless numbers 4488 images in total. In order to obtain correspondence between shape and texture, we use the shape and pose coefficients published with the BFM. This allows for a fair comparison of photometric accuracy, as the 3D shapes for both algorithms are the same. Since illumination is complex, we only show fitting results obtained by the method proposed in Section 4.4. The BFM is a segmented model. In order to obtain high differentiability, we make use of this feature in this experimental section. We use 99 most significant modes for the global model and each of the 4 segments, accounting for 495 texture coefficients in total. The same number of coefficients is used for the shape model.

As opposed to synthetic data, real world imagery is likely to contain outliers (e.g. background information, hair or partial occlusions). The proposed algorithms minimise the L_2 norm in order to derive a closed-form solution. Unfortunately the L_2 norm is sensitive to outliers and the obtained minimum can be heavily influenced by them. In order to deal with outliers in a principled way, a two pass approach is implemented as follows:

- In a pre-fitting stage, texture, diffuse lighting and specular coefficients together with camera parameters are estimated to synthesise a test model. Each vertex-colour of the test model is compared to the input image. Vertices which deviate further from the input image than a threshold value are classified as outliers and discarded in the main-fitting step.
- In the second step, all inverse operations (specular, diffuse and camera parameters) are applied to the test model to obtain a texture only model. The texture only model is compared to the estimated texture values and vertices

Comparison of Fitting Results				
Gallery View	Probe View			
	front	side	profile	mean
Multi Feature Fitting algorithm [55] :				
front	98.9	96.1	75.7	90.2
side	96.9	99.9	87.8	94.9
profile	79.0	89.0	98.3	88.8
mean	91.6	95.0	87.3	91.3
Zhang and Samaras [88] :				
front	96.5	94.6	78.7	89.9
side	93.9	96.7	78.6	89.7
profile	81.8	81.5	89.8	84.3
mean	90.7	90.9	82.3	87.9
Proposed Method :				
front	99.5	95.1	70.4	88.3
side	92.0	99.5	83.7	91.8
profile	73.7	84.0	98.5	85.4
mean	88.4	92.9	84.2	88.5

Table 4.4: Mean recognition rates for all 68 subjects of the CMU-PIE database averaged over 22 illumination conditions per pose. For all experiments, a single image (front, side or profile) is chosen a gallery image and compared against the remaining images (probe set).

which deviate further than a pre-defined threshold value are excluded in the main-fitting step as well.

Table 4.4 shows recognition rates for proposed method 3. The table also shows recognition performance for two reference methods. Overall, our method slightly outperforms that of Zhang and Samaras. Performance is slightly worse than the Multi Feature Fitting algorithm. However, our approach is computationally less expensive, requires very little parameter tuning and converges to the globally optimal solution. Qualitative results for randomly selected images are shown in Figure 4.7.

Relighting and illumination clustering

To examine the robustness of the estimated illumination environments, we show two demonstrations. First, we apply illumination parameters to textures obtained by a different image of same subject in the same pose. This allows for direct qualitative comparison. Figure 4.8 shows three examples of "cross-illumination". For each subject, the last column shows the texture estimated under the other

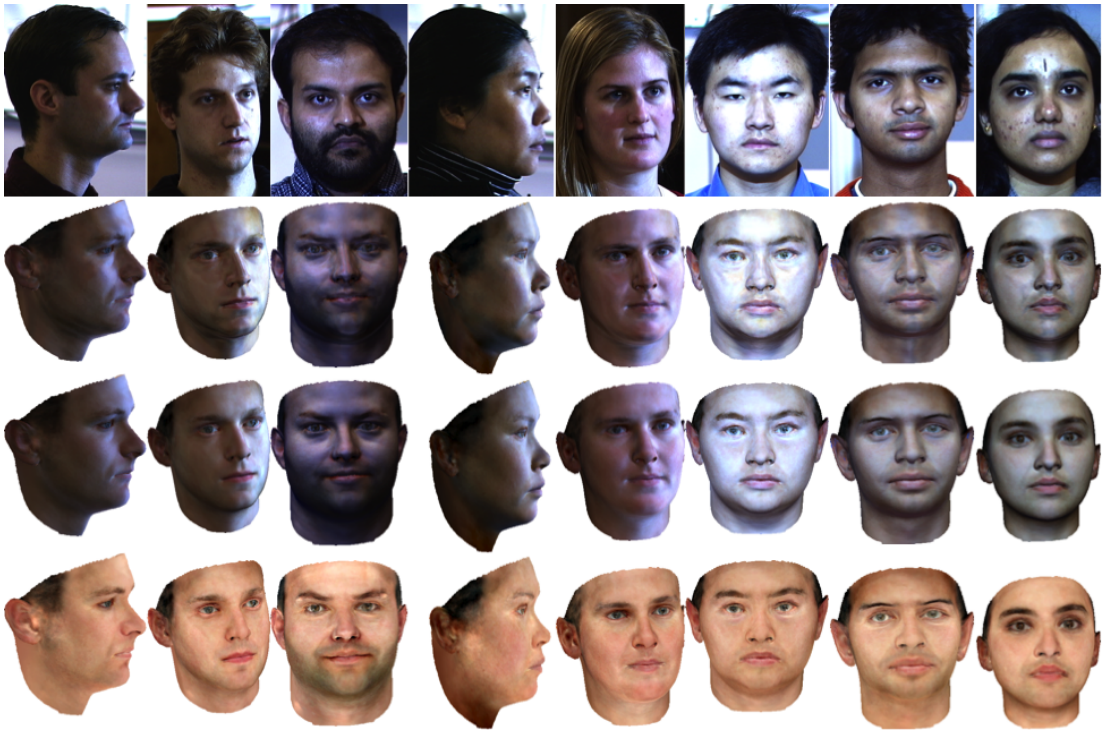

Figure 4.7: Qualitative fitting results for 8 randomly selected subjects, pose and illumination conditions. Top row: Input images. Second row: Fitting result with all parameters estimated. Third row: Diffuse reconstruction. Last row: Texture only.

illumination condition rendered using the estimated illumination environment. The relightings are stable even under dramatic changes in illumination.

Second, we test how stable the estimated illumination environments are across different subjects. Ideally, for fixed pose and illumination all 68 subjects should yield the same irradiance map, since diffuse BRDF parameters are constant in the Lambertian case. Figure 4.9 shows 2 dimensions of multidimensional scaling (MDS) plots for 3 randomly chosen lighting conditions for all 68 subjects in 3 pose angles. As can be clearly seen, the three illumination conditions cluster well and are being distinguished by our algorithm.

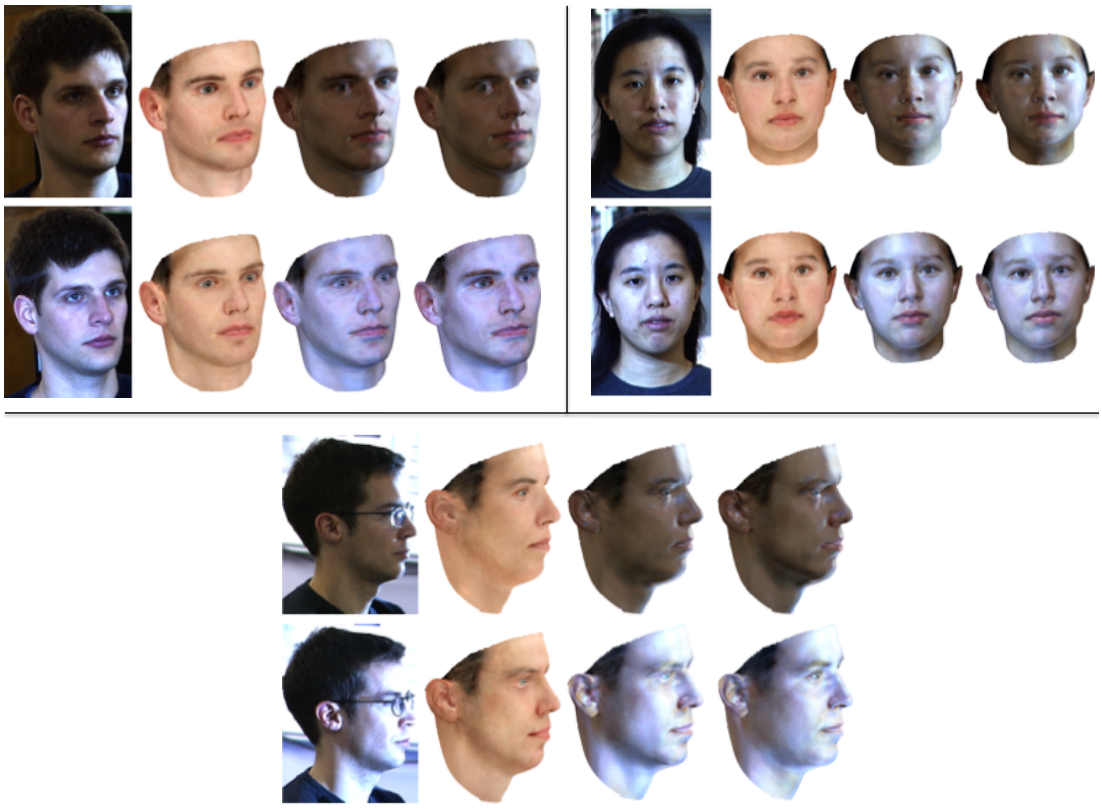

Figure 4.8: Illumination transfer for 3 subjects in same pose and different illumination condition. First column: Input image . Second column: Estimated texture. Third column: Full model approximation. Last column: Estimated textures interchanged between illumination environments.

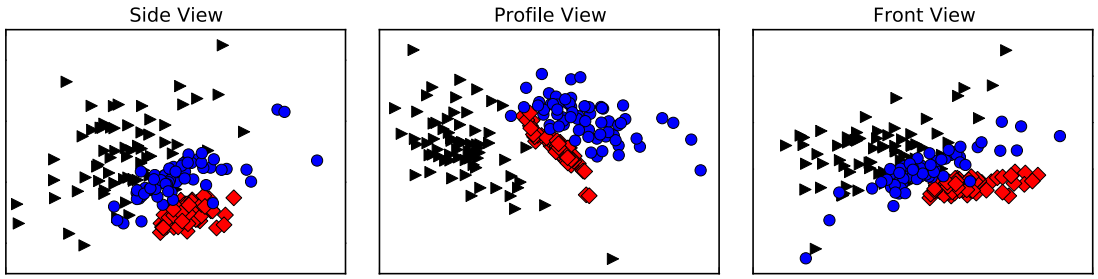

Figure 4.9: Multidimensional scaling plots of irradiance parameters for 3 randomly selected illumination conditions. Results are shown for all 68 subjects of the database for side, profile and front view, respectively. The x and y axis represent the first and the second MDS dimension.

Rerendering

In Figure 4.10, we show how the estimated coefficients for shape, pose, texture and illumination are used to re-render the result into the input image. We compare our result with the one obtained by the Multi feature fitting algorithm [55]. For all subjects a novel pose (30 degrees) is shown with the corresponding illumination estimate.

Illumination transfer

To demonstrate the stability of the estimated parameters, we show an example of illumination transfer. We estimate illumination and identity in three images. Illumination is then applied across all subjects. For comparison, we show the ground truth images present in the dataset. The result is shown in Figure 4.11.

4.6 Conclusion

We introduced 3 methods for fitting a linear texture model with increased generalisation ability. Proposed algorithm M3 makes the least restrictive assumptions for fitting a 3D morphable model, allowing for unconstrained arbitrary complex illumination. We validate our theoretical assumptions with experiments on different datasets, where we obtain state-of-the-art results. The proposed methods make use of prior terms for texture and illumination. Until now, the influence of regularisation terms is found empirically. In future work, we would like to examine ways to automatically tune these parameters. To reduce the influence of outliers when minimising the L_2 norm, we would also like to investigate how a RANSAC implementation can increase performance. Because our fitting process is linear, it would be ideally suited to such an iterative re-fitting technique.

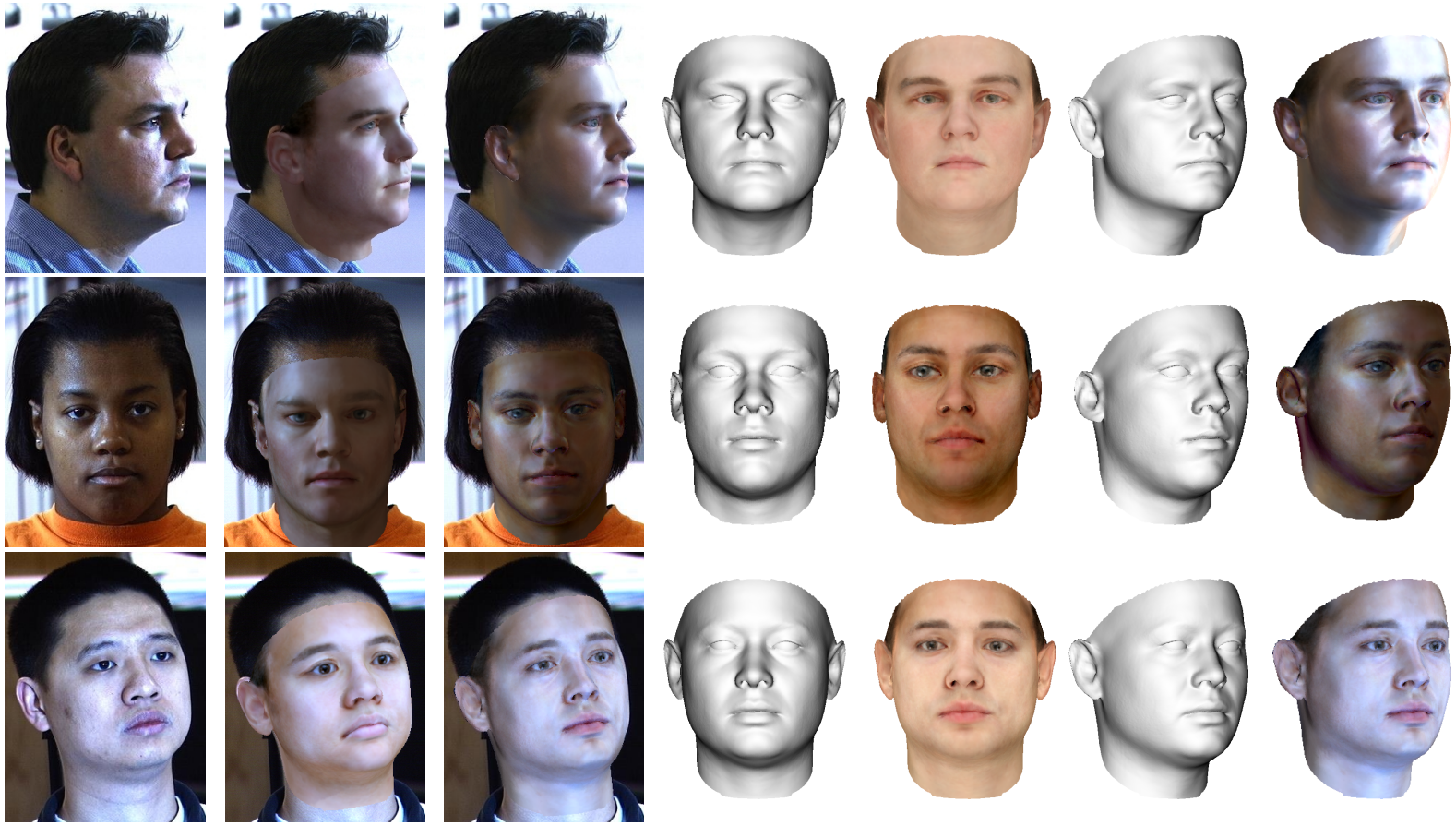

Figure 4.10: From left to right: Input Image. Reconstruction rendered into input image [55]. Reconstruction rendered into input image for the proposed method. Shape reconstruction. Shape reconstruction with recovered texture mapped on it. Novel pose (30 deg.). Novel pose with estimated illumination. The second example demonstrates an unavoidable effect when the texture model is put to its limits. In that case, illumination is wrongly estimated in order to reduce overall pixel error.

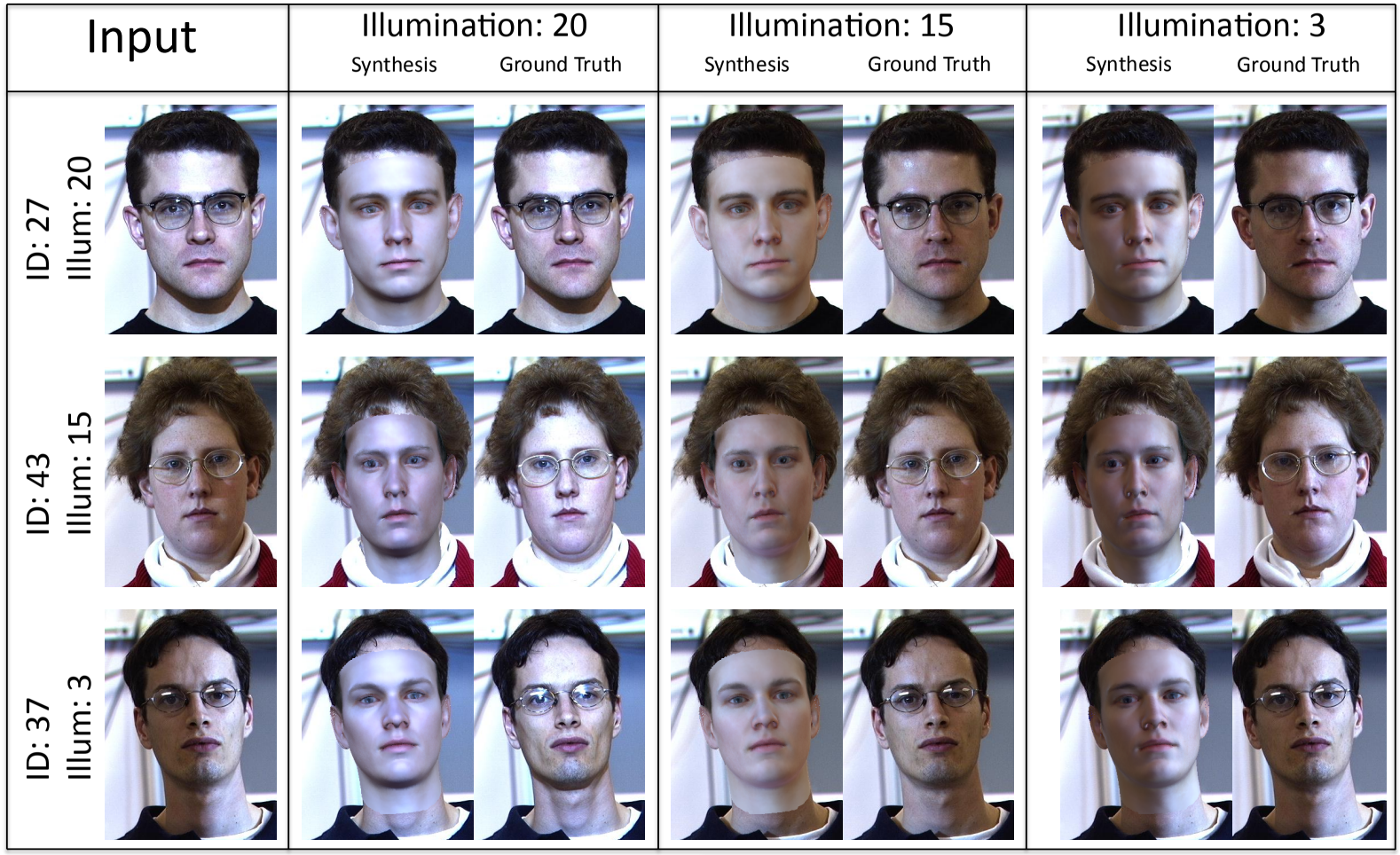

Figure 4.11: Illumination transfer example. We estimate identity and illumination in three input images. The illumination is then transferred to all identities and the result rendered into the image. Ground truth is shown for comparison next to the renderings.

Chapter 5

Global Shading

To date, no fitting algorithm for 3D morphable models takes full effects of global shading into account. Global shading is a phenomenon that takes inter-reflections and occlusions between all objects into account and is a function of the entire scene. Global shading depends on the complexity of the scene and its calculation is computationally very demanding. An in depth coverage of the topic and algorithms for computing global shading can by found in [123] by Pharr and Humphreys. In this thesis, we consider global shading effects which are a function of the object of interest and the illumination environment only. The approximation we consider is known as ambient shading, which we describe in greater detail in the next section. In this chapter, we show how ambient shading can be efficiently incorporated into the image formation process. Another useful approximation for capturing global illumination effects is to use modified surface normals that account for the dominant unoccluded direction from which light arrives. These normals are termed bent normals (further described in Section 5.2). We demonstrate how substituting surface normals with bent normals results in further improvements.

5.1 Ambient occlusion

Ambient occlusion (AO) is a phenomenon which can be observed when non-convex objects are illuminated by ambient light. It is a function of global shape. Convex objects like spheroids or the Johnson solids do not exhibit this phe-

nomenon. When a surface is illuminated by perfectly ambient illumination, shading arises because of partial occlusions of the incident hemisphere by other parts of the surface. This means that the pattern of shading under ambient illumination is determined by global geometry. In contrast, shading under point source illumination is a function solely of the local normal direction. Ambient shading arises in the real world when an object is observed under a cloudy sky. Another recent example of its utility is when a lightstage (spherical illumination rig) is used to produce a close approximation to ambient illumination [68]. The occlusion value at a particular point p of an object $\mathcal{O}$ is defined by how "much" p is occluded by $\mathcal{O}$. We therefore use the expression self-occlusion synonymously. Formally, the normalised intensity $a_p \in [0, 1]$ under unit-strength ambient illumination is given by:

$$a_p = \frac{1}{\pi} \int_{\Omega(\mathbf{n}_p)} (\mathbf{n}_p \cdot \omega) V_{p,\omega} d\omega, \tag{5.1}$$

where $\mathbf{n}_p$ is the local normal direction, $\Omega(\mathbf{n}_p)$ the upper hemisphere about $\mathbf{n}_p$ and the visibility function $V_{p,\omega}$ is equal to zero if point p is occluded in direction ω and one otherwise.

In Section 4.3 and 4.4 we use spherical harmonics to approximate illumination. We decided for this set of basis functions, as they can accurately resemble complex environment illumination. A major drawback however is, that precisely speaking, only convex shaped objects can be approximated with this technique. Faces on the other hand are non-convex at certain regions and an approximation error is unavoidable.

As a motivating example, consider the face shown in Figure 5.1 (left). The face is illuminated by white ambient light only. Note, that on non-convex parts of the face, like the nostrils, eyes, ears or lips, shading is visible. The image next shows a fitting result to the input rendering for algorithm described in Section 4.4. The algorithm fails to approximate the shadows present in the input image. In order to recreate the shaded regions, either the texture or the illumination must be darkened. This causes larger modelling errors. The result of incorporating ground truth AO into the image formation process is shown in the third image of Figure 5.1. The algorithm successfully models the shadows present in the input image.

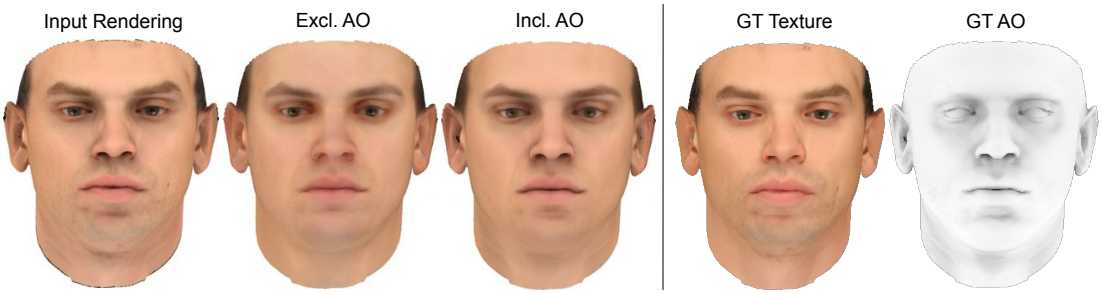

Figure 5.1: Global shading example. Left image shows input rendering. Second image shows model approximation using fitting algorithm described in Section 4.4. Third image shows result for the same fitting algorithm with SH basis functions pre-multiplied with ground truth AO (obtained via Meshlab). Fourth and fifth images show ground truth texture and ambient occlusion.

To further justify our findings, we conducted a trial with eight subjects in three pose angles and three illumination conditions. We calculate ground truth AO using Meshlab [124]. The subjects are not part of the training data of which the morphable model is constructed. The test-set consists of 72 samples. Pose angles are chosen 0, 45 and -60 degrees. The samples are rendered in an increasingly complex environment, starting with uniform-white, "glacier" and "pisa". The environment maps are courtesy of Debevec et al. [122] and are widely used for forward and inverse rendering purposes. The 2D images are obtained using the rendering toolkit "PBRT-v2" [123]. We subsequently run our fitting algorithm on the 2D renderings. Note, that in this pre-trial, we incorporate ground truth AO (as calculated by Meshlab) into the image formation process. All reconstructions are compared against ground truth values using the L_2 norm as distance measure:

$$\mathbb{E}_{\mathbf{g}} = \frac{1}{n}\|\mathbf{t}_g - \mathbf{t}_r\|^2,$$

where $\mathbf{t}_g$ is the ground truth texture and $\mathbf{t}_r$ is the recovered texture. The small letter $n = 160470$ corresponds to the total number of observations. Figure 5.2 shows the outcome of the trial for all 72 renderings, also averaged over pose angles and illumination. The figure shows, that incorporating AO reduces texture reconstruction error; overall and for each individual setting (pose and illumination).

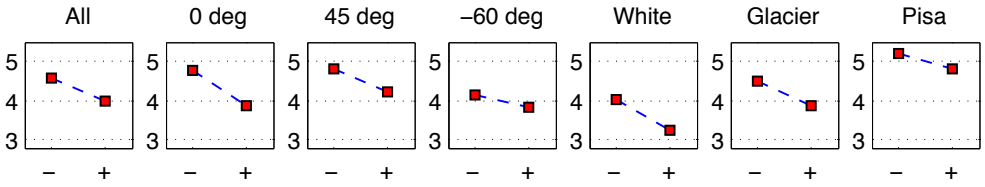

Figure 5.2: Texture reconstruction errors for all 8 subjects, also averaged over pose angles and illumination environment. The sign"-" indicates reconstructions without ambient occlusion and "+" reconstruction errors with ground truth ambient occlusion. Errors on the vertical axis are $\times 10^{-3}$.

5.2 Bent normals

We further refine our algorithm by incorporating bent normals [106] into the image formation process. Combining AO and bent normals is popular in graphics, because they can be precomputed and subsequently used in real-time rendering applications. Using surface normals to model illumination with spherical harmonics leads to an error in estimating the direction of the light source. An example thereof is shown in Figure 5.3.

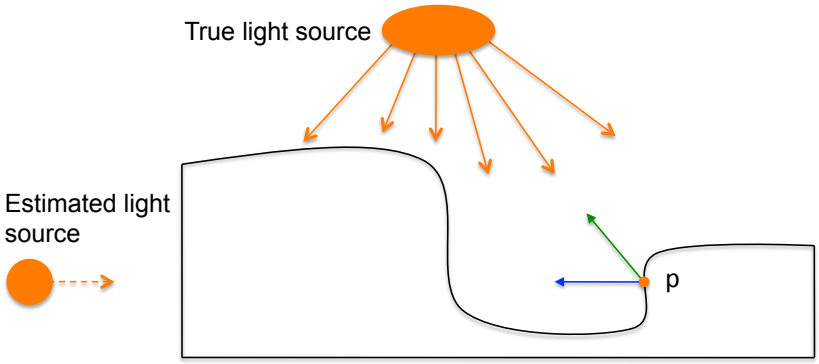

Figure 5.3: SH basis functions constructed from surface normals (blue arrow) would associate a non-existing light source to the reflectance at point p. Bent normals (green arrow) provide a better approximation of the true underlying illumination for non-convex objects.

This means, that the more non-convex an object is, the less accurate the illumination environment can be estimated. In order to compensate for this systematic error, we propose using bent normals to model illumination in this

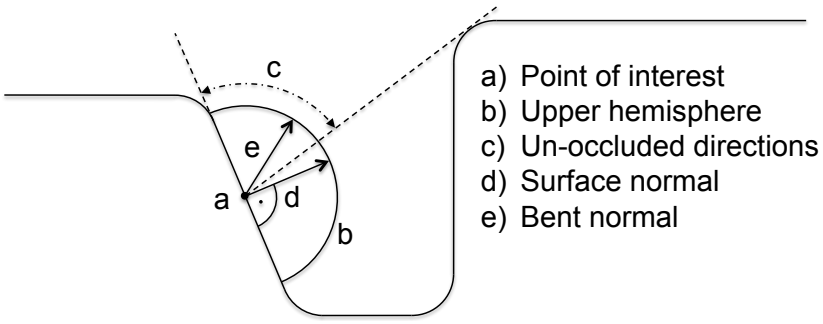

Figure 5.4: Environment lighting modelled via surface normals results in a systematic error for non-convex parts of a shape. Bent normals (the direction of least occlusion) provide a more accurate approximation to the true underlying process.

thesis. Figure 5.4 illustrates the concept of bent normals further.

Accurately calculating AO and bent normals for a given shape is computationally expensive. We therefore propose to build a statistical model of AO and bent normals and learn the relationship between shape parameters and AO/bent normal parameters. We subsequently relax AO and infer the parameter as part of the fitting process, i.e. not relying on initial shape estimate and model inference. The result is that the texture map is not corrupted by dark pixels in occluded regions and the accuracy of the estimated texture map and illumination environment is increased. We present results on synthetic images with corresponding ground truth.

5.3 Image formation process

As in the previous chapter, we allow complex environment illumination of arbitrary colour. Our image formation process models additive Lambertian and specular terms. The Lambertian term further distinguishes between an ambient term, which is pre-multiplied with an occlusion model and an unconstrained part. This captures our approximation to global illumination.

5.4 The physical image formation process

Consider a surface with additive diffuse (Lambertian) and specular reflectance illuminated by a distant spherical environment. The image irradiance at a point p with local surface normal $\mathbf{n}_p$ is given by an integral over the upper hemisphere $\Omega(\mathbf{n}_p)$:

$$i_p = \int_{\Omega(\mathbf{n}_p)} L(\omega) V_{p,\omega} \left[\rho_p(\mathbf{n}_p \cdot \omega) + s(\mathbf{n}_p, \omega, \nu)\right] d\omega, \tag{5.2}$$

where $L(\omega)$ is the illumination function (i.e. the incident radiance from direction ω). $V_{p,\omega}$ is the visibility function, defined to be zero if p is occluded in the direction ω and one otherwise. ρ_p is the spatially varying diffuse albedo and we assume specular reflectance properties are constant over the surface.

A common assumption in computer vision is that the object under study is convex, i.e.: $\forall p, \omega \in \Omega(\mathbf{n}_p) \Rightarrow V_{p,\omega} = 1$. The advantage of this assumption is that the image irradiance reduces to a function of local normal direction which can be efficiently characterised by a low order spherical harmonic.

However, under point source illumination this corresponds to an assumption of no cast shadows and under environment illumination it neglects occlusion of regions of the illumination environment. In both cases, this can lead to a large discrepancy between the modelled and observed intensity and, in the context of inverse rendering, distortion of the estimated texture. Heavily occluded regions are interpreted as regions with dark texture.

Many approximations to Equation 5.2 have been proposed in the graphics literature and several could potentially be incorporated into an inverse rendering formulation. However, in this thesis we demonstrate that even the simplest approximation yields an improvement in inverse rendering accuracy. Specifically, we use the AO and bent normal model proposed by Zhukov et al. [105] and Landis [106]. AO is based on the simplification that the visibility term can be moved outside the integral:

$$i_p = a_p \int_{\Omega(\mathbf{n}_p)} L(\omega) \left(\rho_p(\mathbf{n}_p \cdot \omega) + s(\mathbf{n}_p, \omega, \nu)\right) d\omega,$$

where the AO $a_p \in [0, 1]$ at a point p is given by Equation 5.1. Under this model,

light from all directions is equally attenuated, i.e. the directional dependence of illumination and visibility are treated separately. For a perfectly ambient environment (i.e. $\forall \omega \in S^2, L(\omega) = k$), the approximation is exact. Otherwise, the quality of the approximation depends on the nature of the illumination environment and reflectance properties of the surface. An extension to AO is the so-called bent normal. This is the average unoccluded direction. It attempts to capture the dominant direction from which light arrives at a point and can be used in place of the surface normal for rendering.

5.4.1 Model approximation of the image formation process

The image formation model stated in Equation 5.2 is approximated using the following multilinear system:

$$\mathbf{I}_{mod} = (\mathbf{Tb} + \bar{\mathbf{t}}). * [(\mathcal{U}_a \mathbf{l}_a) . * (\mathbf{Oc} + \bar{\mathbf{o}}) + \mathcal{U}_b \mathbf{l}_b] + \mathcal{S}\mathbf{x}. \tag{5.3}$$

The factors and terms are defined as follows:

$\mathbf{Tb} + \bar{\mathbf{t}} \rightarrow$ Diffuse Albedo
$\mathcal{U}_a \mathbf{l}_a \rightarrow$ DC lighting component
$\mathbf{Oc} + \bar{\mathbf{o}} \rightarrow$ Ambient occlusion
$\mathcal{U}_b \mathbf{l}_b \rightarrow$ Diffuse lighting
$\mathcal{S}\mathbf{x} \rightarrow$ Specular contribution

Given a 3D shape and the camera projection matrix, the unknown photometric coefficients are: $\mathbf{b}, \mathbf{l}_a, \mathbf{c}, \mathbf{l}_b$ and $\mathbf{x}$.

5.4.2 Inverse rendering

We estimate unknown coefficients given a single 2D image. To get correspondence between a subset of the model vertices and the image, we a fit a statistical shape model using the method described in Chapter 3. Assuming an affine mapping, the method alternates between estimating rigid and non-rigid transformations. Given a 3D shape, we have access to surface normals, which can be used to construct SH basis functions.

As part of this thesis, we propose to construct a spherical harmonic basis from bent normals rather than surface normals. Surface normals are still of interest to us for two reasons. Firstly, they serve as reference for the proposed modification, and secondly, they are incorporated as part of a joint model which infers bent normals from 3D shape very efficiently. The image formation model (Equation 5.3), and its algebraic solution with respect to the unknowns is not affected by this substitution.

The basis set $\mathcal{U}_a \in \mathbb{R}^{N\times 3}$ contains ambient constants per colour channel and is invariant of the normals. $\mathcal{U}_b \in \mathbb{R}^{N\times 24}$ contains linear and quadratic terms with respect to the normals. Similarly, the set $\mathcal{S} \in \mathbb{R}^{N\times s}$ contains higher order approximations (up to polynomial degree s), which are used to model specular contributions. However, in the specular case, the normals are rotated about the viewing direction, ν. The three sets are sparse and can directly be inferred given a shape estimate. The parameters $\mathbf{l}_a, \mathbf{l}_b$ and $\mathbf{x}$ depend on a single lighting function $\mathbf{l}$, and are coupled via Lambertian and specular BRDF parameters. We use the method described in Section 4.4 to obtain the lighting function from the reflectance parameters. In order to prevent overfitting, we add prior terms for texture and AO to the objective function: $\mathbb{E}(\mathbf{b}, \mathbf{l}_a, \mathbf{l}_b, \mathbf{x}, \mathbf{c})$. Both priors penalise complexity and the overall objective takes the following form:

$$\mathbb{E} = \|\mathbf{I}_{mod} - \mathbf{I}_{obs}\|^2 + \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2, \tag{5.4}$$

where $\mathbf{I}_{obs}$ are RGB measurements mapped onto the visible vertices of the projected shape model. Under the assumption that each aspect contributes independently to the objective, the system can be solved with a global optimum for each parameter-set. As opposed to conventional multilinear systems, which can only be solved up to global scale, our system factors the mean for texture and occlusion; this property makes the solution unique. In the same way as described in Sections 4.3 and 4.4 of the previous chapter, we equate the partial derivatives of Equation 5.4 to zero and obtain closed-form solutions for each parameter-set.

5.5 Statistical modelling

Our proposed framework requires five statistical models. A PCA model for 3D shape, diffuse albedo and AO, respectively. And principal geodesic analysis (PGA) models for surface normals and bent normals. Coefficients for global shape, texture and AO are obtained directly in the fitting pipeline. Coefficients for bent normals on the other hand cannot be obtained in the same way, due to higher order dependencies. We therefore propose a generative method to infer bent normals from 3D shape. A supplemental surface normal model is used to reduce generalisation error. Comparative results to using vertex data only are presented in the experimental section.

5.5.1 Shape model

As in previous chapters, we use a 3D morphable model to describe shape. For convenience, we use a slightly different notation for the parameter vector in this chapter. A 3D shape is approximated as follows:

$$\mathbf{v} = \bar{\mathbf{v}} + \sum_{i=1}^{m-1} a_i \mathbf{V}_i,$$

where $\mathbf{a} = [a_1 \ \dots \ a_{m-1}]^{\mathrm{T}}$ corresponds to a vectors of parameters. In this chapter, we define the variance-normalised vector as: $\mathbf{e}_a = [a_1/\sigma_{a,1} \ \dots \ a_{m-1}/\sigma_{a,m-1}]^{\mathrm{T}}$. Otherwise, notation is equivalent to other parts of the thesis.

5.5.2 Surface normal model

In addition to 3D vertices, we make use of surface normals to capture local shape variation. Different to data laying on a Euclidian manifold ($\mathbb{R}^n$), surface normals are part of a Riemannian manifold and cannot be simply modelled by applying PCA to the samples. Fletcher et al. [125] introduced the concept of PGA, which can be seen as a generalisation of PCA to the manifold setting. Smith and Hancock showed how this framework can used to build a statistical model of surface normals [17]. We briefly revise PGA for data laying on S^2. Our description closely resembles the seminal work of Fletcher et al. [125].

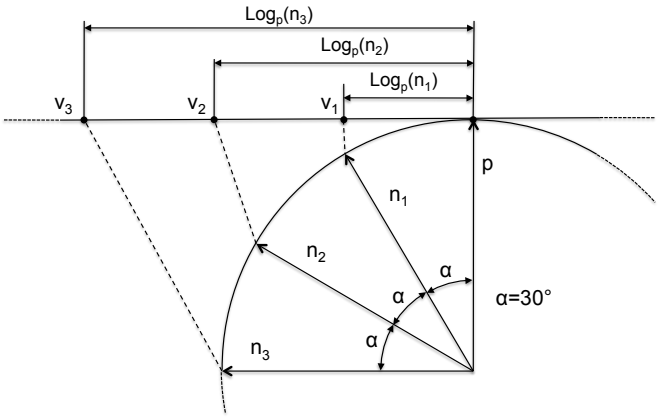

Figure 5.5: Distance preserving mapping of three normals $\mathbf{n}_1$, $\mathbf{n}_2$ and $\mathbf{n}_3$ onto the tangent plane. The mapping is defined with respect to north pole defined by: $\mathbf{p}$.

A normal $\mathbf{n} = (n_x, n_y, n_z)$ projects onto a point $\mathbf{v} = (v_x, v_y)$ on the tangent plane, with respect to the plane-centre defined by point $\mathbf{p} = (0, 0, 1)$, using the Log map:

$$\mathbf{Log_p}(\mathbf{n}) = \left(n_x \cdot \frac{\Theta}{\sin\Theta}, n_y \cdot \frac{\Theta}{\sin\Theta} \right),$$

where the angle $\Theta = \arccos(n_z)$. Points on the tangent plane can be back projected onto S^2 using the Exponential map:

$$\mathbf{Exp_p}(\mathbf{v}) = \left(v_x \cdot \frac{\sin\|\mathbf{v}\|}{\|\mathbf{v}\|}, v_y \cdot \frac{\sin\|\mathbf{v}\|}{\|\mathbf{v}\|}, \cos\|\mathbf{v}\| \right).$$

The two mappings are isometric and bijective for metric spaces: $\mathbb{R}^3$ and S^2.The concept is further illustrated in Figure 5.5. Having defined the projections, a mean vector $\Delta\mu$ intrinsic to the manifold can be computed using an iterative procedure. In the first step, an arbitrary surface normal is set being the mean, μ_j. Each normal is now projected to the tangent plane with respect to μ_j and the intrinsic mean updated: $\Delta\mu_{j+1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{Log}_{\mu_j}(\mathbf{n}_i)$. The extrinsic mean is obtained by applying the exponential map: $\mu_j = \mathbf{Exp}_{\mu_j}(\Delta\mu)$. The algorithm has converged when $\|\Delta\mu\| < \epsilon$. For reasonably small values of ϵ, this is usually achieved in ≤ 10 iterations.

Having found a mean vector μ, the principal geodesics can be found by mapping all normals to the tangent plane using the Log map with respect to μ, and

applying PCA to the corresponding design matrix. We construct the following model for surface normals:

$$\mathbf{n} = \mathbf{Exp}_{\mu}\left(\sum_{i=1}^{m-1} b_i \mathbf{N}_i\right),$$

where $\mathbf{b} = [b_1 \ \ldots \ b_{m-1}]^{\mathrm{T}}$ are parameter vectors. Variance-normalised they are defined as: $\mathbf{e}_b = [b_1/\sigma_{b,1} \ \ldots \ b_{m-1}/\sigma_{b,m-1}]^{\mathrm{T}}$.

5.5.3 Ambient occlusion model

For each of the m shape samples, we compute ground truth AO using Meshlab [124]. Each vertex is assigned a single integer $o_i \in [0, 1]$, which corresponds to the occlusion value. A value of 1 indicates a completely unoccluded vertex. As for shape, we construct a PCA model for AO:

$$\mathbf{o} = \bar{\mathbf{o}} + \sum_{i=1}^{m-1} c_i \mathbf{O}_i,$$

where $\mathbf{c} = [c_1 \ \ldots \ c_{m-1}]^{\mathrm{T}}$ is a parameter vector. The computed mean value is defined as: $\bar{\mathbf{o}}$, and the $\mathbf{O}_i$'s are modes of variation capturing decreasing energy $\sigma^2_{c,i}$. We define $\mathbf{e}_c = [c_1/\sigma_{c,1} \ \ldots \ c_{m-1}/\sigma_{c,m-1}]^{\mathrm{T}}$.

5.5.4 Bent normal model

From a modelling perspective, bent normals are equivalent to surface normals (samples on a spherical manifold). The model is constructed in the same way as the one described in Section 5.5.2:

$$\mathbf{b} = \mathbf{Exp}_{\mu}\left(\sum_{i=1}^{m-1} d_i \mathbf{B}_i\right).$$

Note that here $\Delta\mu$ refers to the intrinsic mean of the bent normals. As in previous defined models, $\mathbf{d} = [d_1 \ \ldots \ d_{m-1}]^{\mathrm{T}}$ is a vector of parameters. Variance-normalised they are defined as: $\mathbf{e}_d = [d_1/\sigma_{d,1} \ \ldots \ d_{m-1}/\sigma_{d,m-1}]^{\mathrm{T}}$.

5.5.5 Ambient occlusion and bent normal inference

We infer bent normals (and for one setting AO) from shape data using a generative non-parametric model. We decided to use a probabilistic linear Gaussian model with class specific prior functions. In a discrete setting, a joint instance comprising the knowns and unknowns is represented as a feature vector $\mathbf{f} = [\mathbf{f}_k^{\mathrm{T}}\mathbf{f}_c^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{f}_k^{\mathrm{T}} = [\mathbf{e}_a^{\mathrm{T}}\mathbf{e}_b^{\mathrm{T}}]^{\mathrm{T}}$ or $\mathbf{f}_k^{\mathrm{T}} = \mathbf{e}_a^{\mathrm{T}}$ (depending on whether only vertices or vertices and surface normals are used) and $\mathbf{f}_c^{\mathrm{T}} = \mathbf{e}_c^{\mathrm{T}}$ or $\mathbf{e}_d^{\mathrm{T}}$ (AO or bent normal inference). The training-set consists of m instances re-projected into the corresponding models. To ensure the scales of both models are commensurate, we use variance normalised parameter vectors. We construct the design matrix $\mathbf{D} \in \mathbb{R}^{(m_k+m_c)\times m}$ by stacking the parameter vectors. From a conceptual perspective our model is equivalent to a probabilistic PCA model [109]. The non-probabilistic term is described with a noise term, ϵ. Noise is assumed normally distributed, and its distribution is assumed stationary for all features. A joint occurrence is described as follows:

$$\mathbf{f} = \mathbf{W}\alpha + \mu + \epsilon.$$

The parameter α and the noise term are assumed to be Gaussian distributed:

$$p(\alpha) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad p(\epsilon) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}). \tag{5.5}$$

In probabilistic terms, a joint instance is written as: $p(\mathbf{f}|\alpha) \sim \mathcal{N}(\mathbf{W}\alpha + \mu, \sigma^2\mathbf{I})$. The characteristic model parameters are: $\mathbf{W}$, μ and σ^2 ($\mathbf{I}$ is the identity matrix of appropriate dimension). The most likely values can be obtained by applying PCA to the mean-free design matrix: $\bar{\mathbf{D}} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{f}_i - \mu)(\mathbf{f}_i - \mu)^{\mathrm{T}} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{V}^{\mathrm{T}}$, and setting:

$$\mu_{ml} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{f}_i, \qquad \sigma_{ml}^2 = \frac{1}{m-1-u}\sum_{i=u+1}^{m-1}\mathbf{\Sigma}_{i,i}^2, \qquad \mathbf{W}_{ml} = \mathbf{U}_u(\mathbf{\Sigma}_u^2 - \sigma_{ml}^2\mathbf{I})^{\frac{1}{2}}.$$

The small letter u corresponds to the number of used modes. Our aim is to estimate $\mathbf{f}_c$ given $\mathbf{f}_k$. For data which is jointly Gaussian distributed, the following

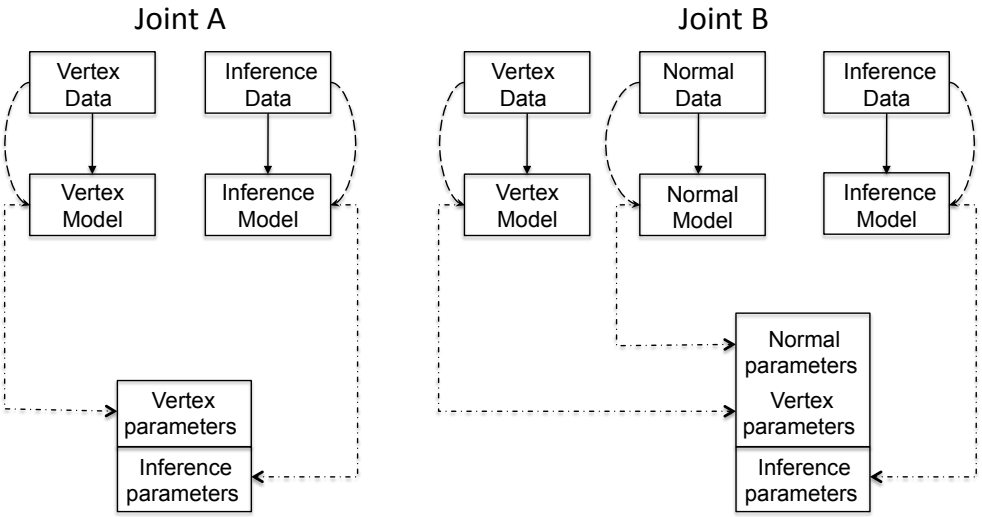

Figure 5.6: Building process for both models starts with high dimensional training data, which is used to build subspace models (PCA for vertices and ambient occlusion, and PGA for surface and bent normals). The training data is reprojected into the models, which yields a low dimensional parameter for each instance. The parameters are jointly modelled via PPCA.

margninalisation property holds true:

$$p(\mathbf{f}) = p(\mathbf{f}_k, \mathbf{f}_c) \sim \mathcal{N}\left(\begin{bmatrix} \mu_k \\ \mu_c \end{bmatrix}, \begin{bmatrix} \mathbf{W}_k\mathbf{W}_k^{\mathrm{T}} & \mathbf{W}_k\mathbf{W}_c^{\mathrm{T}} \\ \mathbf{W}_c\mathbf{W}_k^{\mathrm{T}} & \mathbf{W}_c\mathbf{W}_c^{\mathrm{T}} \end{bmatrix} \right).$$

Therefore we can write the probability $p(\mathbf{f}_k|\alpha) \sim \mathcal{N}(\mathbf{W}_k\alpha+\mu_k, \sigma^2\mathbf{I})$. According to the specifications, we also have knowledge of how α is distributed (see relations 5.5). Applying Bayes' rule, we can infer the posterior for alpha as follows:

$$p(\alpha|\mathbf{f}_k) \sim \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}_k^{\mathrm{T}}(\mathbf{f}_k - \mu_k), \sigma^2\mathbf{M}^{-1}), \tag{5.6}$$

where $\mathbf{M} = \mathbf{W}_k\mathbf{W}_k^{\mathrm{T}} + \sigma^2\mathbf{I}$ (see for example [126] for an in-depth explanation of linear Gaussian models). Given an estimate for α, we can obtain the posterior distribution for the missing part $p(\mathbf{f}_c|\alpha)$, with the mode centre $(\mathbf{W}_c\alpha + \mu_c)$ corresponding to the MAP estimate.

5.6 Experiments

As in other chapters, we use the BFM [27] to represent shape and diffuse albedo. Since we do not have access to the initial training data, we sample from the model to construct a representative set. This is only required for the shape model, as neither AO nor bent normals exhibit a dependency on texture. In order to capture the span of the model, we sample ± 3 standard deviations for each of the $k = 199$ principal components plus the mean shape. Because of the non-linear relationship between shape and AO/bent normals, we additionally sample 200 random faces from the model. This accounts for a total of $m = 599$ training examples. For each sample, we calculate ground truth AO and bent normals using Meshlab [124]. We retain $m_{a,b,c,d} = 199$ most significant modes for each model.

Using a physically based rendering toolkit (PBRT v2), we render 3D faces of eight subjects in three pose angles and three illumination conditions. The subjects are not part of the training-set. To cover a wide range, we chose pose angles: $-60°$, $0°$ and $45°$. The faces are rendered in the following illumination environments: 'White', 'Glacier' and 'Pisa', where the latter two are obtained from [122]. Skin reflectance is composed of additive Lambertian and specular terms with a ratio of 10/1. The test-set consists of $8 \times 3 \times 3 = 72$ images in total.

For each of the samples, we first recover 3D shape and pose from a sparse set of feature points using algorithm described in Chapter 3. We project shape into the images and obtain RGB values for a subset of the model vertices, $\tilde{n} = 3\tilde{p}$. Using the proposed image formation process and objective function, we decompose the observations into its contributions: texture, ambient shading, diffuse shading and specular reflection. We compare five settings: In a reference method (algorithm described in Section 4.4), we use SH basis functions constructed from surface normals and do not account for AO. We refer to this setting as: 'V'. The second method incorporates AO predicted from the statistical model and uses surface normals to construct SH basis functions: 'A'. In a third setting, AO and bent normals are predicted from the model; we refer to this method as: 'B'. In the fourth setting, we estimate AO as part of the fitting pipeline: 'Fit A'. And finally, we derive the SH basis functions from predicted bent normals and fit AO, termed: 'Fit B'. An overview of the five methods is shown in Table 5.1. Each of the settings

Method	Attribute	
	Ambient Occlusion	Bent Normals
V	–	–
A	inferred	–
B	inferred	inferred
Fit A	fitted	–
Fit B	fitted	inferred

Table 5.1: Summary of the 5 fitting method we compare with each other. The term "inferred" relates to attributes predicted from the statistical model. The term "fitted" means that the attribute is estimated as a free parameter within the objective function. A blank '–' indicates that the attribute is not used.

is evaluated according to three quantitative measures:

1. Texture reconstruction error
2. Fully synthesised model error
3. Illumination estimation accuracy

We also investigate how accurately bent normals are predicted from the joint model by comparing against ground truth. The last part of this section shows qualitative reconstructions for texture and full model synthesis and an application of illumination transfer. Out-of-sample faces are labelled with three digit numbers $(001 - 323)$.

5.6.1 Ambient occlusion and bent normal generalisation error

In this section, we investigate generalisation error of the AO and bent normal model. In a first trial, we examine how well the model generalises to unseen data by projecting out-of-sample data into the model. This setting is referred to as 'Model'. In a second and third trial, we measure the error induced by predicting AO and bent normals from shape. The first method uses only vertex data, which we term 'Joint A'. The second method additionally incorporates surface normals, termed 'Joint B'. Reconstruction error is measured in mean squared error for the AO model and mean angular distance for the bent normal model. They are

Face :	001	002	014	017	052	053	293	323	mean
Ambient Occlusion Generalisation Error: $\mathbb{E}_a$									
Model:	37.3	13.2	13.4	19.7	13.8	24.7	10.3	11.4	17.9
Joint A:	183.7	40.0	42.1	53.2	65.5	119.2	31.9	34.1	71.2
Joint B:	160.6	33.1	39.2	49.6	46.6	84.9	26.6	28.5	58.6
Bent Normal Generalisation Error: $\mathbb{E}_b$									
Model:	3.32	2.68	2.79	2.67	2.56	3.10	2.19	2.41	2.72
Joint A:	4.53	4.04	4.82	4.04	4.13	5.49	3.83	3.73	4.33
Joint B:	3.94	3.57	4.42	3.66	3.88	4.75	3.15	3.34	3.84

Table 5.2: Ambient occlusion and bent normal approximation error for eight subjects. Errors are measured in mean angular distance (Bent normals) and mean squared error (Ambient occlusion).

defined as follows:

$$\mathbb{E}_a = \|\mathbf{o}_g - \mathbf{o}_r\|^2, \qquad \mathbb{E}_b = \frac{1}{p}\sum_{i=1}^{p} \arccos\left(\frac{\|\mathbf{b}_g^i \cdot \mathbf{b}_r^i\|}{\|\mathbf{b}_g^i\|\|\mathbf{b}_r^i\|}\right),$$

where $\mathbf{o}_g$ and $\mathbf{o}_r$ are ground truth and reconstructed AO, respectively. The symbol $\mathbf{b}_g^i$ corresponds to the i'th ground truth bent normal and $\mathbf{b}_r^i$ to the reconstruction. Table 5.2 shows reconstruction errors for eight out-of-sample faces.

As can be seen in the table, adding surface normals to the statistical model reduces inference error by about 41% for the AO and 17% for the bent normal model. Therefore, we use the data inferred from model 'Joint B' for the previously introduced fitting methods: 'A', 'B' and 'Fit B'.

5.6.2 Texture reconstruction error

We use the 60 most significant principal components to model texture. Our evaluation is based on squared Euclidian distance:

$$\mathbb{E}_t = \frac{1}{n}\|\mathbf{t}_g - \mathbf{t}_r\|^2,$$

between ground truth texture $\mathbf{t}_g$ and reconstruction $\mathbf{t}_r$. Individual values within each texture vector are within $\mathbb{R} \in [0, 1]$. Table 5.3 shows texture reconstruction errors for all subjects averaged over illumination and over pose angles using ground truth shape. Table 5.4 shows the same results for shape obtained by fitting

$\mathtt{E_t}, \forall:$	001	002	014	017	052	053	293	323	0°	45°	−60°	mean
V	3.742	3.175	2.892	13.55	4.555	3.027	3.794	3.895	5.054	5.025	4.405	4.829
A	3.731	3.005	2.244	12.31	3.799	2.663	2.948	3.303	4.135	4.551	4.064	4.250
B	3.113	3.155	2.329	10.21	3.232	2.789	2.704	3.538	3.900	3.915	3.837	3.884
Fit A	3.693	3.004	2.038	12.52	3.735	2.681	3.075	3.227	4.127	4.409	4.206	4.247
Fit B	3.104	3.220	2.168	10.68	3.011	2.498	2.803	3.431	3.863	3.843	3.886	3.864

Table 5.3: Texture reconstruction errors averaged over subjects and pose angles (GT Shape all methods). Individual entries are $\times 10^{-3}$.

$\mathtt{E_t}, \forall:$	001	002	014	017	052	053	293	323	0°	45°	−60°	mean
V	3.394	3.866	3.417	13.13	5.108	3.454	3.710	3.856	5.277	5.165	4.534	4.991
A	3.328	3.945	3.032	11.89	4.515	3.331	2.870	3.019	4.504	4.717	4.251	4.491
B	2.974	4.113	3.537	9.469	3.973	3.234	2.792	3.534	4.461	4.143	4.005	4.203
Fit A	3.213	3.603	2.843	12.63	4.387	3.511	3.254	3.109	4.596	4.704	4.410	4.569
Fit B	2.803	3.716	3.111	10.09	3.509	2.978	2.992	3.352	4.224	3.973	4.028	4.069

Table 5.4: Texture reconstruction errors averaged over subjects and pose angles (FP Shape all methods). Individual entries are $\times 10^{-3}$.

to sparse feature points using method proposed in Chapter 3. Using bent normals reduces the error by about 9%. Also, fitting AO does only give an advantage over predictions when using shape reconstructions from feature points.

5.6.3 Full model composition error

The difference between the fully synthesised model and the images is examined in this part. As for texture, we measure the difference in squared Euclidian distance:

$$\mathbb{E}_f = \frac{1}{\tilde{n}} \|\mathbf{f}_g - \mathbf{f}_r\|^2,$$

where entries in $\mathbf{f}_g$ and $\mathbf{f}_r$ are within range $[0, 1]$. Error is normalised over the number of observations $\tilde{n}$, and differs for subjects and pose. But is constant for the five methods. Results for this measurements are shown in Table 5.5 (ground truth shape) and Table 5.6 (reconstructed shape). Fitting AO results in about 11% lower errors over predicting it from the statistical model and the use of bent normals does not give further improvements.

$\mathtt{E_f}, \forall:$	001	002	014	017	052	053	293	323	0°	45°	−60°	mean
V	2.739	2.651	2.361	3.610	2.602	2.564	2.794	2.768	2.094	3.257	2.932	2.761
A	2.774	2.534	2.255	3.590	2.553	2.566	2.672	2.759	1.992	3.280	2.866	2.713
B	3.091	2.526	2.181	3.727	2.604	2.646	2.639	2.797	1.903	3.381	3.045	2.776
Fit A	2.447	2.356	2.086	3.321	2.335	2.341	2.463	2.508	1.809	2.997	2.640	2.482
Fit B	2.467	2.309	1.998	3.304	2.261	2.304	2.370	2.476	1.684	2.949	2.676	2.436

Table 5.5: Full reconstruction errors averaged over subjects and pose angles (GT Shape all methods). Individual entries are $\times 10^{-3}$.

$\mathtt{E_t}, \forall:$	001	002	014	017	052	053	293	323	0°	45°	−60°	mean
V	2.312	2.933	3.007	3.740	2.659	2.603	2.419	2.782	2.404	3.219	2.797	2.807
A	2.364	2.901	2.934	3.937	2.591	2.655	2.313	2.782	2.391	3.314	2.724	2.810
B	2.472	3.004	2.861	4.133	2.663	2.658	2.395	2.735	2.350	3.369	2.875	2.865
Fit A	2.006	2.628	2.655	3.462	2.304	2.267	2.040	2.469	2.106	2.892	2.439	2.479
Fit B	2.007	2.641	2.585	3.441	2.319	2.231	2.074	2.394	2.068	2.831	2.486	2.461

Table 5.6: Full reconstruction errors averaged over subjects and pose angles (FP Shape all methods). Individual entries are $\times 10^{-3}$.

5.6.4 Environment map approximation error

In this section, we compare lighting approximation error for the five methods. We obtain ground truth lighting coefficients by rendering a sphere in the same illumination conditions than the faces. The material properties are also set to be equal. As the normals and the texture of the sphere are known, we deconvolve the image formation and extract lighting coefficients. This also makes sense for white illumination, as with this procedure we obtain the overall magnitude of light source intensity. We divide reflectance vectors by the corresponding BRDF parameters and use them as ground truth: $\mathbf{l}_g$. For each of the images, we compute angular distance between the recovered lighting coefficients $\mathbf{l}_r$ and the ground truth:

$$\mathbb{E}_l = \arccos\left(\frac{\|\mathbf{l}_g \cdot \mathbf{l}_r\|}{\|\mathbf{l}_g\|\|\mathbf{l}_r\|}\right).$$

Results for the experiments are shown in Table 5.8, where we have averaged over all subjects and pose angles.

5.6.5 Qualitative results

For visual comparison, we show qualitative results for three methods ('V', 'Fit A' and 'Fit B'). Figure 5.7 displays fitting results for various subjects, pose angles

$E_1, \forall:$	White	Glacier	Pisa	mean
V	23.87	13.83	17.43	18.38
A	33.67	23.33	24.99	27.33
B	23.15	15.12	13.28	17.18
Fit A	19.51	14.25	15.93	16.56
Fit B	19.15	13.27	12.90	15.11

Table 5.7: Light source approximation error (GT Shape) for for the three methods. Entries represent angular error averaged over all subjects and pose angles.

$E_1, \forall:$	White	Glacier	Pisa	mean
V	10.14	16.61	19.05	15.25
A	14.94	24.65	26.38	21.99
B	7.93	14.77	14.87	12.52
Fit A	9.12	15.31	16.94	13.79
Fit B	6.44	12.70	14.03	11.06

Table 5.8: Light source approximation error (FP Shape) for for the three methods. Entries represent angular error averaged over all subjects and pose angles.

and illumination conditions. The Figure shows full model synthesis and texture reconstructions for methods 'V' and 'Fit B'.

As can be seen in Table 5.4 and 5.6, quantitative differences between method 'Fit A' and 'Fit B' are less pronounced. This also applies for perceptual differences. Methods 'Fit B' notably obtains more accurate reconstructions for regions which are severely occluded. Figure 5.8 shows magnifications of full model synthesis of the nose region for two subjects.

A most important feature to be extracted is diffuse albedo. As an identity specific parameter it should be consistently estimated across pose and illumination. Figure 5.9 shows model synthesis and texture reconstructions for method 'Fit B' for one subject in all pose and illumination combinations, including shape and pose estimates.

5.6.6 Illumination transfer

To demonstrate the stability of the estimated parameters, we combined lighting coefficients (ambient, diffuse and specular) estimated from a set of subjects with identity parameters (shape, texture and AO) and pose from the same set. The results are shown in Figure 5.10. Diagonal entries show fitting results to the actual images, and off-diagonal entries show cross illumination/identity results.

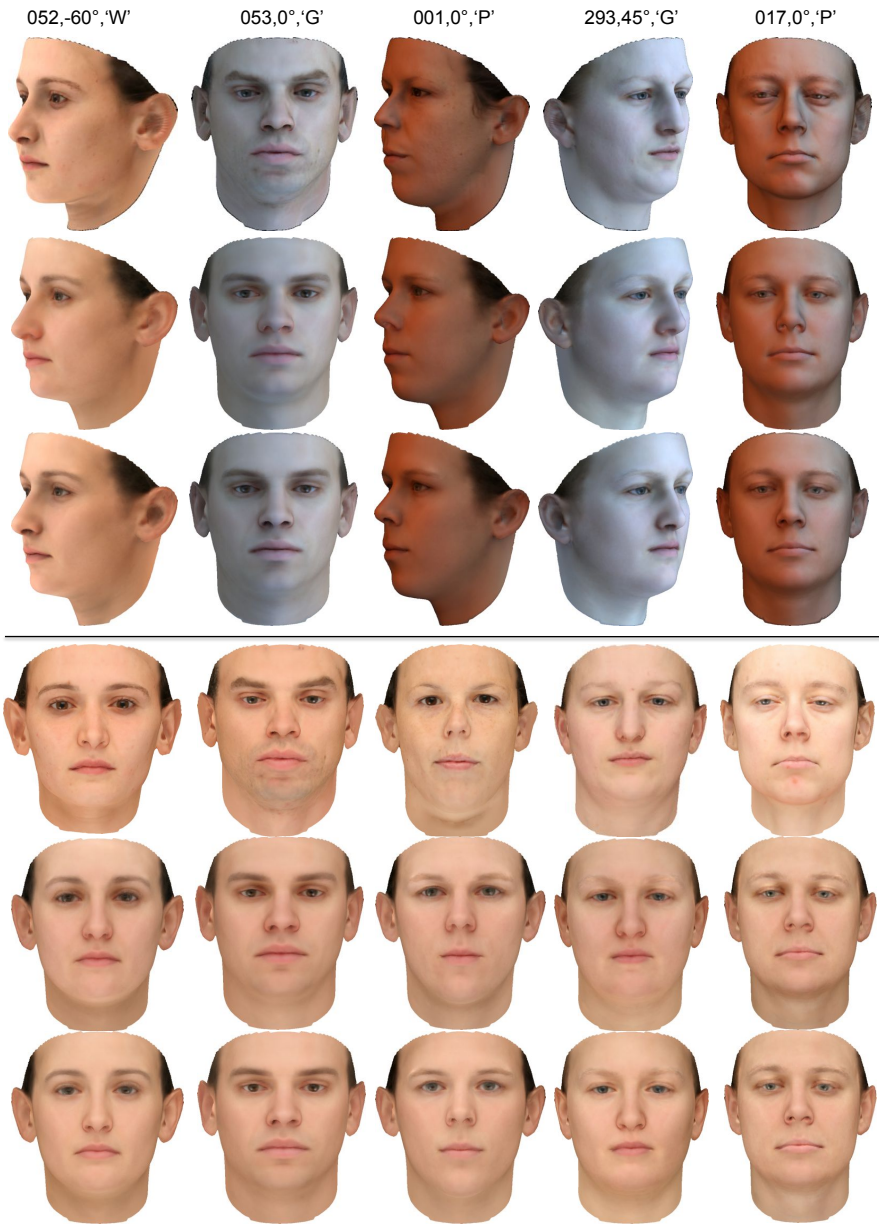

Figure 5.7: Comparison of method 'V' and 'Fit B' for five subjects in different pose angles and illumination condition. Top row shows input images. Second row shows full model synthesis for method 'V'. The third row shows full model synthesis for method 'Fit B'. And the fourth and fifth row show ground truth texture (and shape) and texture reconstruction using method 'V'. The last row shows texture reconstructions using method 'Fit B'. Labels on top of images are: Face ID, Pose, Illumination, where 'W','G' and 'P' corresponds to 'White', 'Glacier' and 'Pisa'.

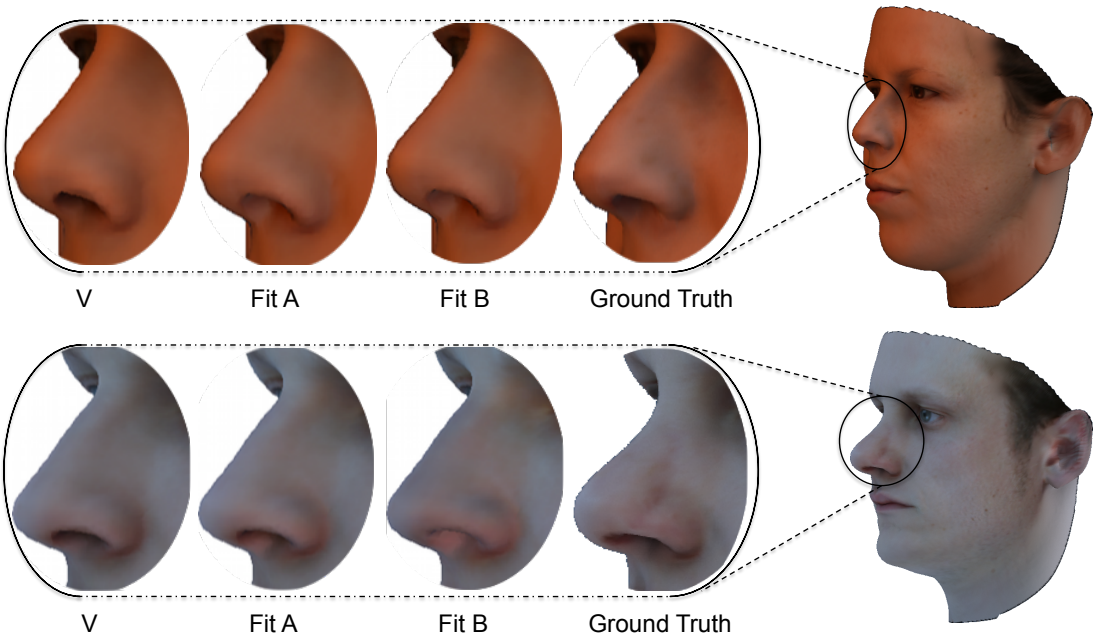

Figure 5.8: Comparison of the three methods: 'V', 'Fit A' and 'Fit B' for two subjects. Close up nose region face ID: 001 (top) and ID: 014 (bottom).

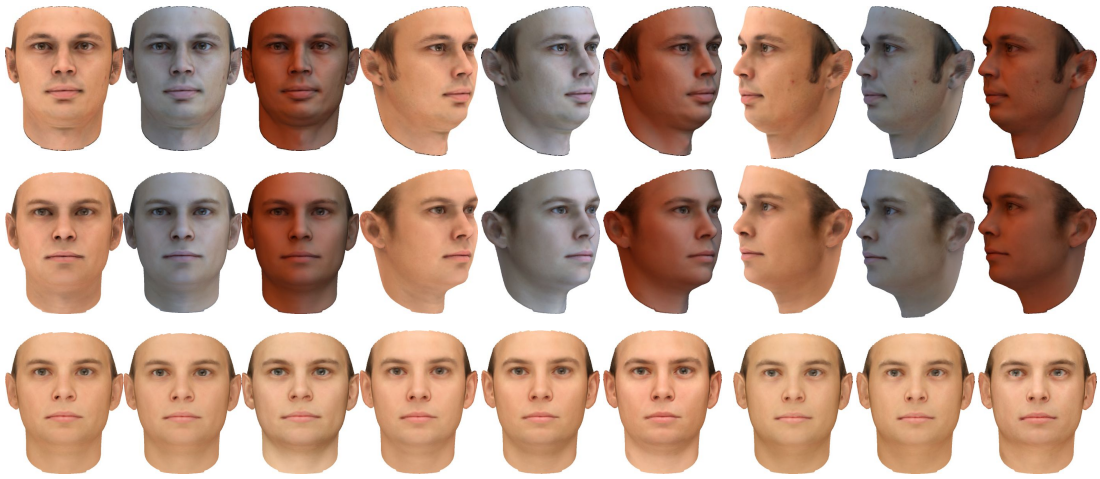

Figure 5.9: Fitting results for one subject (ID: 002) in all pose angles and illumination condition. Top row shows input images. Second row shows full model synthesis using method 'Fit B'. Bottom row shows texture reconstructions. Note, that the face shown, does not posses lower texture reconstruction error than method 'Fit A'.

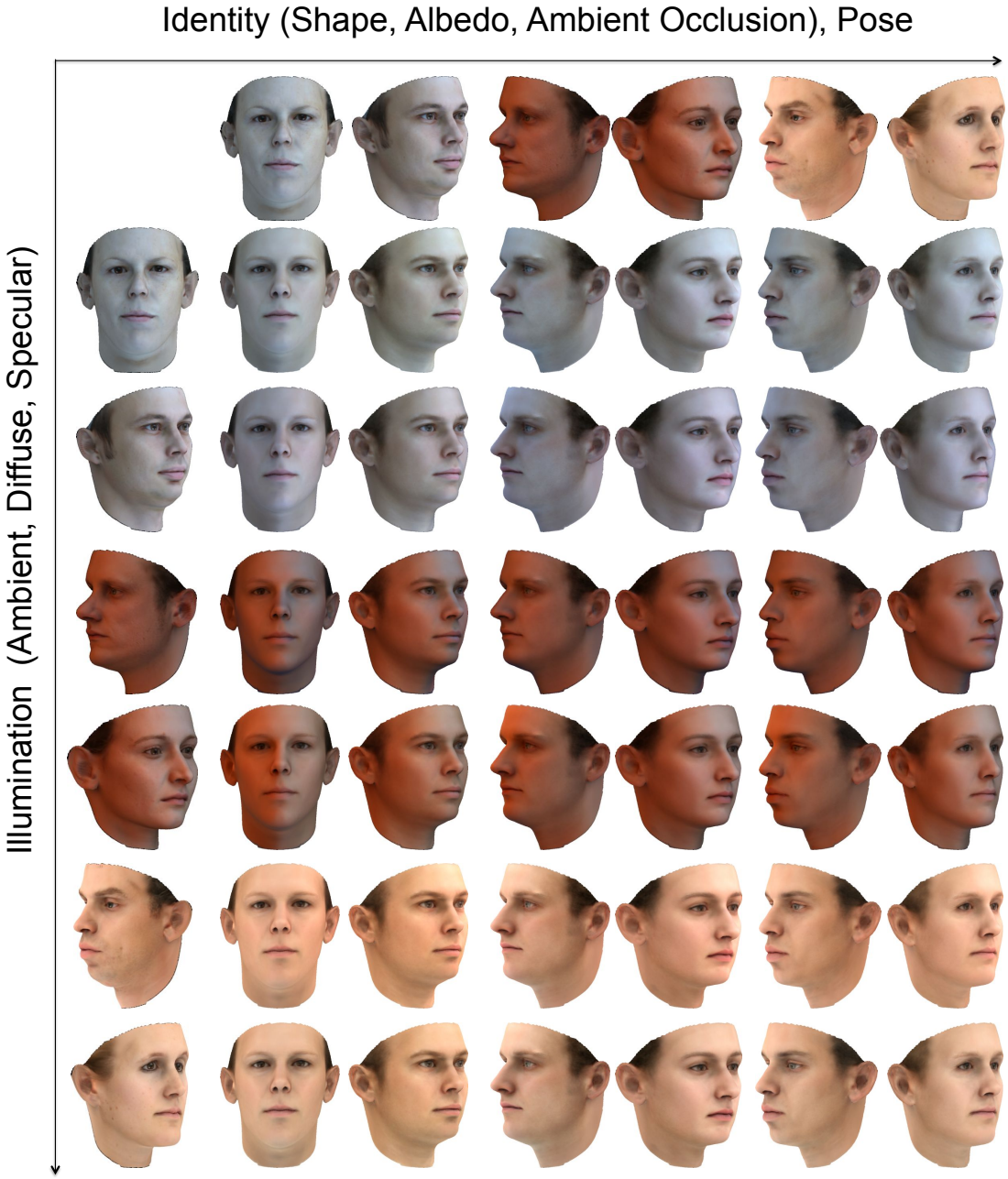

Figure 5.10: Illumination transfer example. Top row and first column show input images of six subjects. Estimated parameters for shape, pose, diffuse albedo and ambient occlusion from the columns are combined with lighting estimates obtained by the rows.

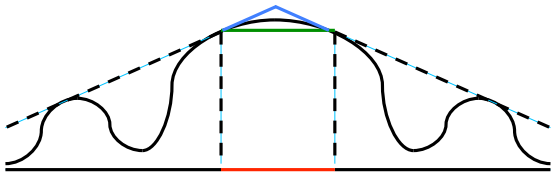

Figure 5.11: Ambient shading ambiguity: for this 1D surface, points between the vertical dotted lines would all have maximal intensity (as they are completely unoccluded). However, any convex curve between the green and blue curves would also generate the same image.

5.7 Shape from ambient occlusion

Shading arises under ambient illumination due to partial occlusion of the incident hemisphere. The general formulation of the shape-from-ambient shading (SFAS) problem is ill-posed. In this section we present a model-based approach which can solve the problem for the object class human faces. We show that a linear statistical model is sufficient to capture the relationship between 3D shape and ambient shading.

Estimation of 3D shape from ambient shading has received very little attention in comparison to the classical point source shape from shading problem. Langer and Bülthoff [103] showed the importance of ambient shading in human perception of shape. Langer and Zucker [70] were the first to study the problem from a computational perspective. More recently, Prados et al. [104] showed that the problem is ill-posed under certain conditions (see Figure 5.11) and amounts to the solution of a strongly non-local and non-linear Integro-Partial Differential Equation.

5.7.1 Experiments

We use the same data ($m = 599$ training instances) and model introduced in Section 5.5 for the experiments. The only difference is, that the knowns and unknowns are reversed (we assume known AO and unknown shape). A graphical chart of how the joint model is built from training data can be seen in Figure 5.12. The joint data consists of eigenvalue normalised parameters and is modelled via PPCA.

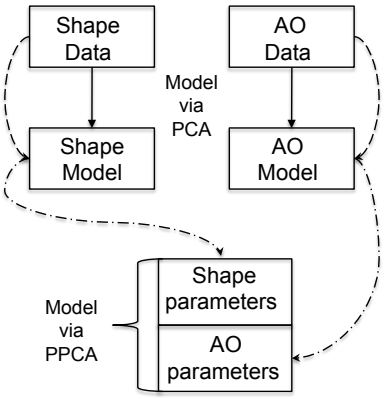

Figure 5.12: Graphical representation of the model building pipeline. The dashed arrows can be read as "project into", and the dot-dashed arrows "forwards" the eigenvalue normalised parameters.

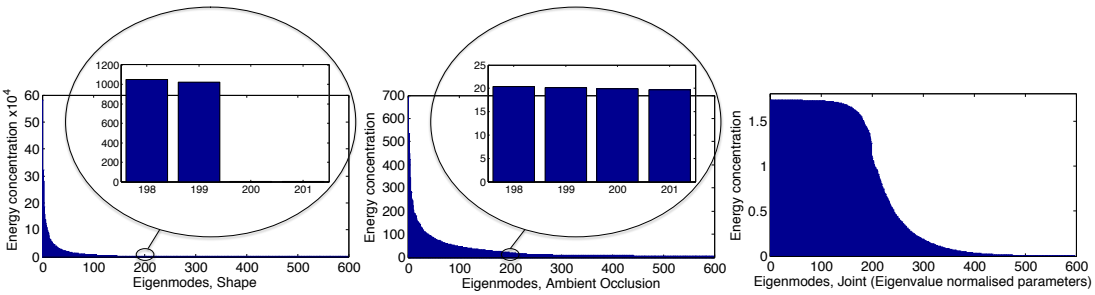

Figure 5.13: Eigenvalue decay for the three models. The magnification around the 200'th eigenvalue suggests a non-linear relationship between shape and ambient occlusion.

Energy concentration for shape, AO and joint model can be seen in Figure 5.13. For the joint model, we retain 350 most significant modes. We assume perfect correspondence between ambient shading values and the model, since our aim is to evaluate whether a linear model is sufficient for the SFAS problem in principle.

5.7.2 Results

Figure 5.14 shows the MAP estimates for the 10 faces using inference algorithm 5.6. In order to quantitatively evaluate our findings, we calculate distances between all 10 reconstructions to all 10 ground truth shapes. We define distance

	Ground Truth Face									
Rec. Face	001	002	006	014	017	022	052	053	293	323
001	**76.9**	108.3	100.2	98.2	83.3	96.8	77.0	85.2	99.5	84.5
002	91.6	**59.7**	94.2	82.3	98.0	91.9	93.0	84.1	104.2	111.1
006	92.7	81.6	**66.9**	77.3	83.7	76.0	112.0	100.2	82.5	87.2
014	100.7	89.4	80.7	**67.7**	92.1	89.6	91.3	93.8	84.9	89.1
017	87.0	90.7	85.6	91.6	**52.3**	82.0	90.3	96.3	87.9	74.4
022	89.6	91.4	83.8	82.7	83.4	**66.1**	103.5	102.0	98.5	93.4
052	77.8	100.8	98.9	106.0	80.5	97.7	**47.6**	90.5	93.7	80.5
053	87.6	92.4	84.4	85.4	104.9	76.8	99.3	**59.3**	99.4	106.8
293	94.5	75.3	95.2	78.6	98.8	92.5	109.3	99.5	**64.8**	104.8
323	77.2	105.8	92.0	109.8	86.3	96.6	80.9	96.2	64.9	**49.9**

Table 5.9: Angular error (in degrees) between reconstructed faces and ground truth faces projected into the model. 199 modes are used for all cases. Bold numbers highlight lowest error per row and per column. This demonstrates that not only the reconstructions are closest to the corresponding ground truth (row), but also the ground truth is closest to the corresponding reconstruction (column).

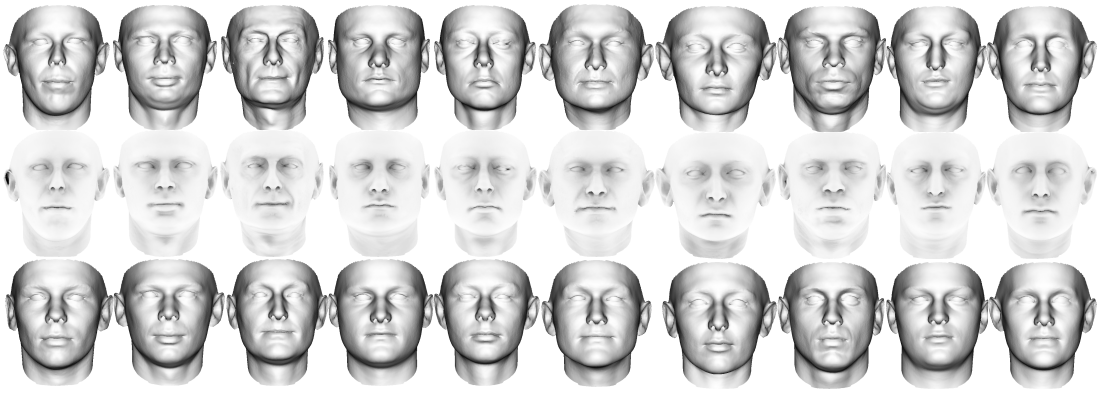

Figure 5.14: Qualitative reconstruction results for 10 subjects, disjoint from the training data. Top row shows ground truth shape. Second row shows ambient occlusion for ground truth shape and the third row shows reconstructions predicted by the joint model.

$\mathbb{D}(\mathbf{a}, \mathbf{b})$ between 2 shape instances: $\mathbf{a}$ and $\mathbf{b} \in \mathbb{R}^{199}$using angular distance:

$$\mathbb{D}(\mathbf{a}, \mathbf{b}) = \arccos\left(\frac{\|\mathbf{a} \cdot \mathbf{b}\|}{\|\mathbf{a}\|\|\mathbf{b}\|}\right),$$

where vectors representing ground truth are obtained by projecting the shapes into the shape model. Table 5.9 shows the obtained distances.

5.8 Conclusions

In this chapter, we have presented a generative method to estimate global illumination in an inverse rendering pipeline. To do so, we learn and incorporate a statistical model of ambient occlusion and bent normals into the image formation process. The resulting objective function is convex in each paramter-set and can be solved accurately and efficiently using alternating least squares. In addition to qualitative improvements, empirical results show that reconstruction accuracy for texture, lighting and full model synthesis increases by about $10 - 18\%$. In future work, we would like to explore the performance of the proposed fitting algorithm in a recognition experiment. Further, we presented a practical method to infer 3D face shape under ambient illumination. Our experiments show that the method achieves high accuracy. The current framework assumes the ground truth ambient occlusion to be known. In future experiments, we would like to render 3D shapes with a physically based ray tracing system and infer the 3D shape from ambient occlusion present in the 2D renderings.

Chapter 6

Conclusions and future work

In this final chapter, we summarise the achievements we have made, outline potential extensions and improvements and give some final remarks.

6.1 Summary

We have proposed a complete framework to inverse render faces. By decomposing the image formation process into a geometric and photometric part, we can solve each subset as a multilinear system of equations. As opposed to previously introduced methods for fitting a 3D morphable model to images, the proposed methods are convex in each set of parameters and can be solved with a global optimum. Besides a novel way to recover shape from a sparse set of feature points, which negates the need to empirically find a parameters that trades-off model dominance versus data, we introduced 3 methods for fitting a linear texture model with increased generalisation ability. Proposed method 3 makes the least restrictive assumptions for fitting a 3D morphable model, allowing for unconstrained arbitrary complex illumination. We demonstrate the performance of our method on challenging data, where we are able to obtain state-of-the-art results. We have further modified the image formation process to account for global shading effects. Our approximation thereof is known as ambient occlusion. The method is refined by using bent normals to construct spherical harmonic basis functions. By doing so, we show how a systematic error present for non-convex objects can be significantly reduced.

6.2 Future work

During the course of this thesis, we have identified several shortcomings, ways for improvements and directions for future work.

- The proposed methods make use of prior terms for texture and illumination. In this work, regularisation weights were found empirically. In future work, it would be interesting to examine ways to automatically tune these parameters. Either in a similar way as we proposed for shape: By learning how the texture model generalises to unseen data and incorporating this knowledge into the objective function. Or using an approach based on semidefinite programming [127].
- To reduce the influence of outliers when minimising the L_2 norm, one can investigate how a RANSAC implementation can increase performance. Because our fitting process is linear, it would be ideally suited to such an iterative re-fitting technique. It would also be interesting to examine how performance changes when minimising the L_1 norm.
- Future work might look at how automatic feature detectors can be incorporated into our method. Recently Amberg et al. [110] proposed a reliable method for getting correspondence of image and model features. A further way of getting correspondence are contours and silhouettes. After an initial shape estimate, model edges can be directed towards the closest edge in the image. This process is known as data assignment can be solved via an iterative closest point algorithm [128].
- A further point worth to investigate is, how a shape estimate can be refined from estimated ambient occlusion. The foundations thereof have already been proposed in Section 5.7 of this thesis. In future work we like to rerun recognition experiments on the CMU-PIE database and and include ambient occlusion estimation and shape refinement.
- In Appendix A we propose a method to non-linearly synthesise novel shapes from a small number of sample shapes in order to obtain a larger population. The same approach can be used to build a texture model. In a similar fashion as Mohammed et al. [129], the texture PCA model would only act

as a global constraint. We can divide overlapping patches in (u, v) space and synthesise non-linear patches. When fitting, we undo illumination (diffuse specular and global) and force agreement with (small slice) of neighbouring pixels and the global model.

6.3 Remarks

In this final section, we would like to give some remarks on our proposed framework and face recognition / modelling in general. In order for face recognition systems to be widely deployed, they must be reliable and work under unconstrained conditions. For applications where cooperative subjects can be assumed, one can argue that current state-of-the-art computer vision algorithms can solve the problem. Here the question arises of how the algorithms can best be implemented in hardware and work in real-time, or how the interface to the system can be optimised.

A different picture arises for face recognition under unconstrained conditions and where subjects can not be assumed cooperative. The primary problem might not be the lack of available models or algorithms to solve the individual parts, but rather in how the right techniques can be selected and combined in order to achieve sufficient recognition rates. Even if each component can be designed and optimised for maximum efficiency, merely putting these parts together does not guarantee magnificent performance. It is the overall setup which is equally important and which can make the significant difference. Often trade-offs have to be made and an apparently better component has to be sacrificed to maximise the overall performance. Along these lines we see the greatest challenge and potential for improving the performance of face recognition systems. Nevertheless, the individual components are important. Our framework is proposed such that each contributing factor is incorporated separately and independently. Our image formation model has a modular design which allows components to be exchanged and new ones to be added. We have demonstrated this by incorporating ambient occlusion and substituting surface normals with bent normals. Other possible examples include a change of skin reflectance model, or adding models for facial hair or glasses.

Appendix A

Non-linear shape composition

Models based on PCA assume a linear relationship amongst the training examples. When the number of training examples is large, and more importantly representative of the target population, the model is able to generalise well to unseen data. In this final section, we explore ways how to generalise from a small number of faces (for instance $5-10$) to a larger population. In the case of human faces, a PCA model would lack the ability to generalise well to novel faces. The possible variability of the population is simply to high. Segmenting the model [27] is one way to address the problem. Segmenting the model however has two disadvantages: The boundaries where the segments meet show discontinuities, and have to be blended via a smoothing function. The second disadvantage of segmenting the model is related to the inference part. For instance, reconstructing faces from feature points requires that sufficient observations (number of feature points) are available in each segment. This can be problematic, when observations are sparse.

What we propose in this appendix is to synthesise a larger, possibly infinite, set of training examples from a small number of basis shapes in a non-linear fashion. We further address the problem of moving outside of the space spanned by the training examples, while preserving local consistency. One way is working on the meshes (represented as graphs) directly. This however is delicate due to the large number of inter-dependencies. A coarse simplification is to treat faces as chain models and compose new samples within such a framework. This however results in faces which are not globally smooth (depending on how the

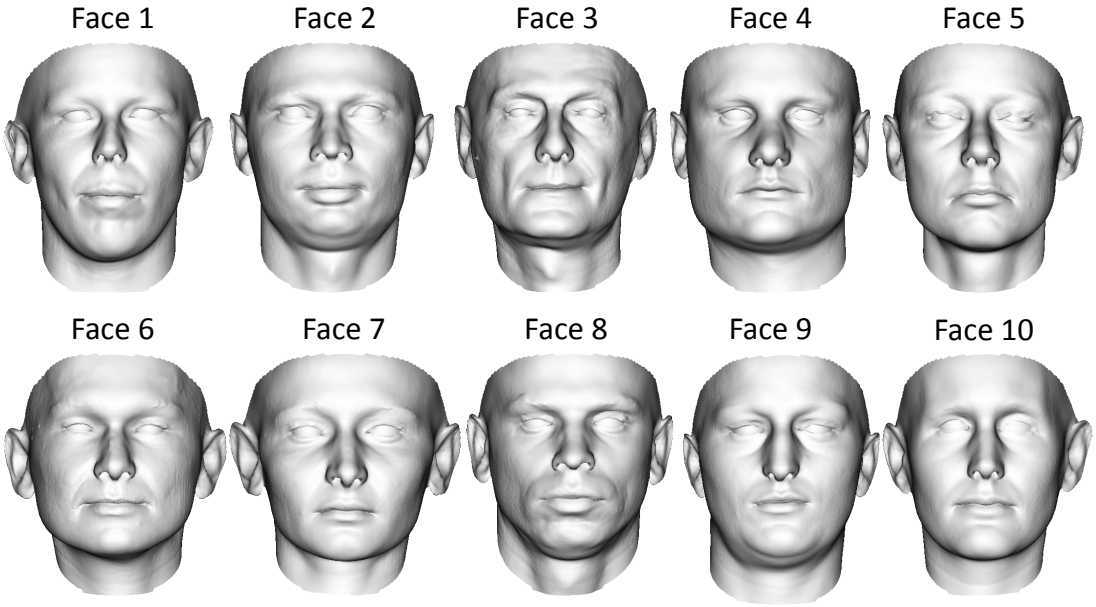

Figure A.1: Training examples used as basis shapes for non-linear shape composition.

chain traverses the mesh). Our approach lays in between these extremes. We compose new examples from existing ones in two dimensional (u, v) reference frame (see Section 2.2.5).

A.1 Approach

In order to exploit local correlation, we compose novel faces in 500×500 (u, v) reference frame. A shape $\mathbf{v}_i \in \mathbb{R}^{160470}$ and its embedding $\mathbf{D}_i \in \mathbb{R}^{500\times500}$ are different representations of the same instance. We create random fields via *"Random Field Simulation"* which is a collection of Matlab functions courtesy of P. Constantine. Creating exponential fields of that size is computationally demanding. To increase efficiency, we create 100×100 random maps and interpolate them to the desired size (500×500) using Matlabs built in interpolation toolkit. As input to our algorithm, we consider ten out-of-sample faces from the BFM, which are shown in Figure A.1. We randomly select $s \in \{2, 3, 4\}$ faces from the training-set $\mathcal{V} = \{1, \ldots, n\}$, where $n \in \{5, 10\}$, depending on the experimental setup. The first setup considers all 10 out-of-sample faces. The second case considers only

the 5 male faces (Faces: $2, 3, 4, 6$ and 8). As a preprocessing step, we subtract the mean $\bar{\mathbf{v}} = \sum_{i=1}^{s} \mathbf{v}_i$ of the selected basis shapes $\mathbf{v}_i$ from each instance and denote the result $\bar{\mathbf{v}}_i$. For each of the selected faces we create a random exponential field $\mathbf{E}_i$ (This accounts for non-linearity). We normalise the fields such that at each input point the random fields sum to a value greater than one:

$$\bar{\mathbf{E}}_i = \frac{\mathbf{E}_i}{\sum_{i=1}^{s} \mathbf{E}_i} \times (1 + 0.25s).$$

This allows to move outside of the space spanned by the samples. Finding a good normalising constant is a trade-off between how far we like to move outside of the sample space and face plausibility. We found empirically that a value of 0.25 per face introduces novelty while minimising unnatural shape deformations. The embeddings $\bar{\mathbf{D}}_i$ for each face $\bar{\mathbf{v}}_i$ are element-wise multiplied with the corresponding normalised random field: $\bar{\mathbf{D}}_{i,e} = \bar{\mathbf{D}}_i . * \bar{\mathbf{E}}_i$, and the novel face is the sum thereof $\bar{\mathbf{N}} = \sum_{i=1}^{s} \bar{\mathbf{D}}_{i,e}$. We map $\bar{\mathbf{N}}$ form embedding space to vector space $\bar{\mathbf{n}}$ and add the mean shape: $\mathbf{n} = \bar{\mathbf{n}} + \bar{\mathbf{v}}$. Figure A.2 shows a non-linear composition of faces 2, 4 and 9 and their corresponding weightings.

For the rest of this chapter, we use the term *linear model* for a PCA model built from basis shapes, and the term *non-linear model* for PCA model built from non-linear shape compositions (strictly speaking both models are linear with respect to the training samples).

A.2 Experiments

We consider two scenarios. The first setup uses $n = 10$ basis shapes and is referred to as *Case 1*. The second scenario uses only the male shapes ($n = 5$) and is referred to as *Case 2*. For both cases we synthesise a larger population and use the data to build PCA models. We examine eigenspace, model flexibility and generalisation error from feature points.

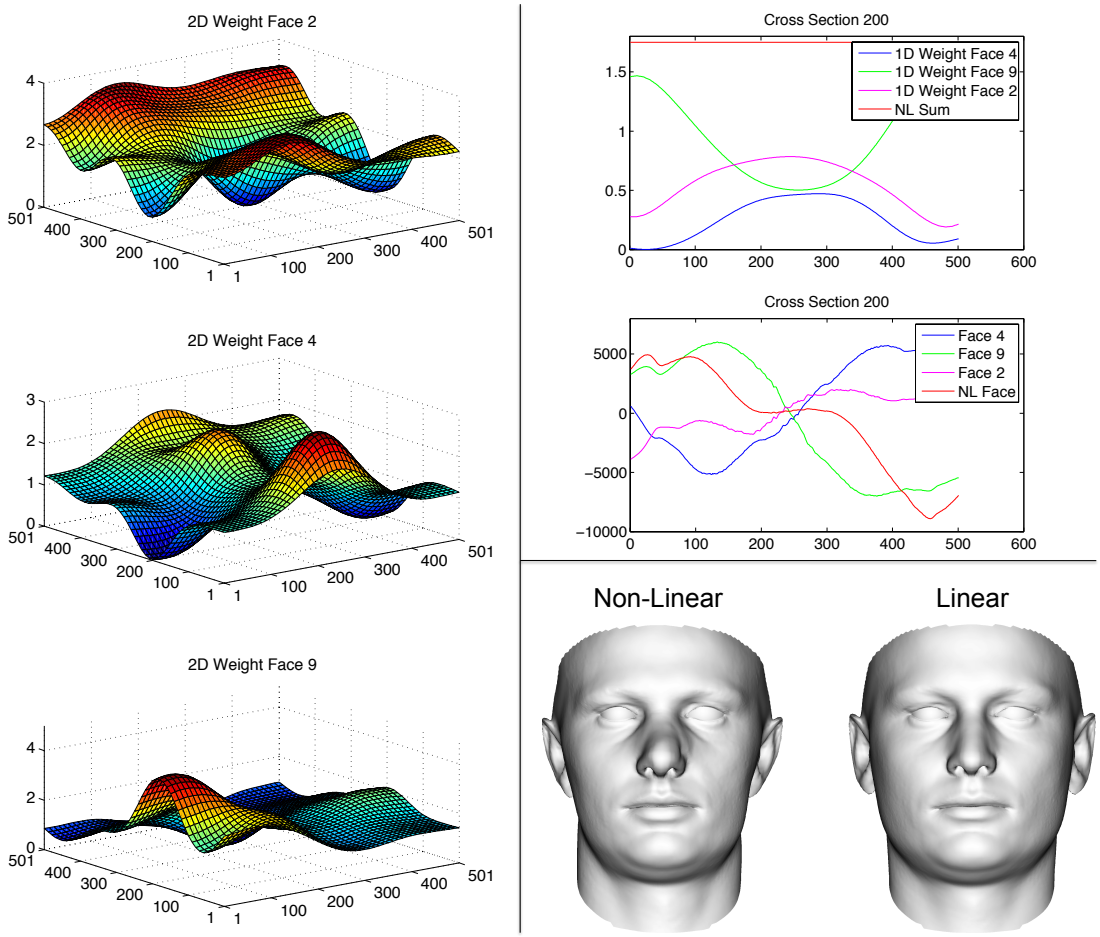

Figure A.2: Non-linear composition of a new training example from 3 existing faces. Left column shows random exponential maps created for each face (not yet normalisd). The plots on the right hand side show normalised local contributions for each face (cross-section 200 horizontally). The plot below shows absolute values for the resulting shape. Note, that because we normalise to 1.75 we are able to move outside the span of the training examples while preserving local consistency. The faces on the bottom show the resulting face (left) and the arithmetic mean of the samples for comparison (right).

A.2.1 Statistical modelling

For *Case 1* we compose 1000 novel faces using the proposed method. Figure A.3 (top) shows randomly selected training examples. Note, the third face on the top row shows an example of unnatural shape deformation. For *Case 2* we

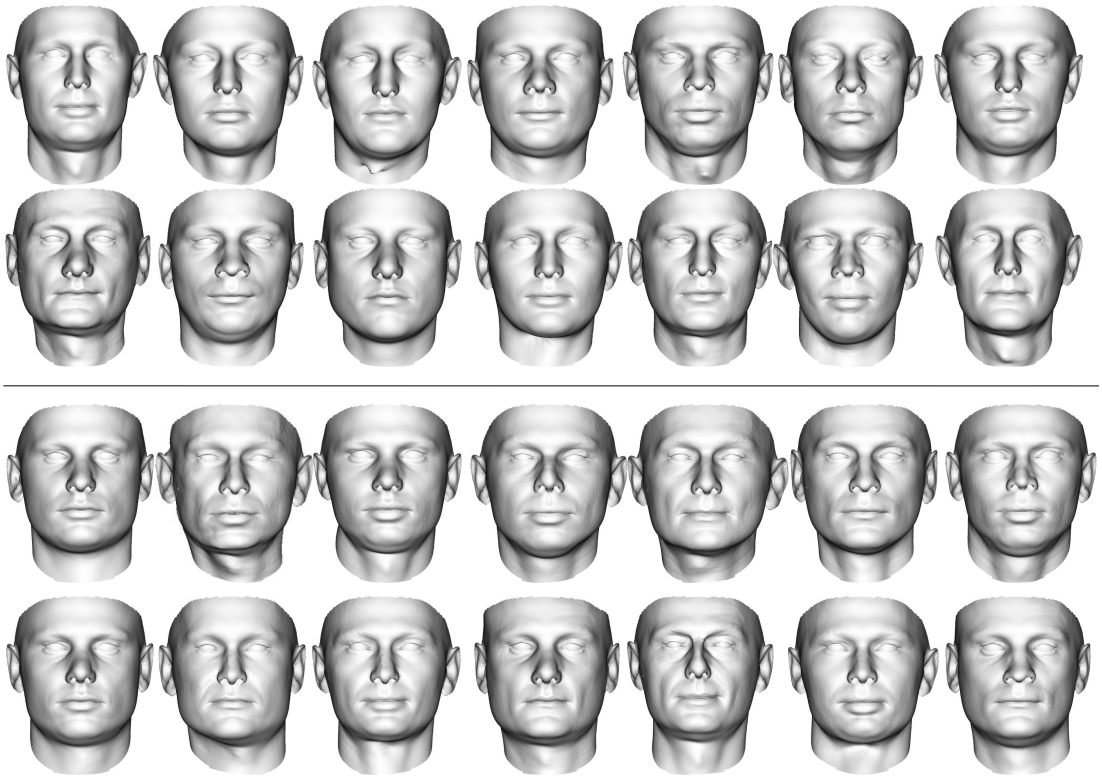

Figure A.3: Randomly selected training examples generated by the proposed method for $n = 10$ (top) and $n = 5$ (bottom). In general, the faces appear plausible. Unnatural deformations however are not entirely excluded (see for instance third face on tow row).

synthesise 500 novel faces. A randomly selected subset thereof is shown in Figure A.3 (bottom). We use the basis shapes to build a linear PCA model and the compositions to build a non-linear PCA model. For the non-linear models, the energy captured by each mode is shown in Figure A.4 (top) for *Case 1* and Figure A.4 (bottom) for *Case 2*. To make results comparable, we retain 199 most significant modes for both models.

A.2.2 Model generalisation

To test model flexibility, we project a test-face (not part of the training data) into the models. Depending on the case, the maximum number of modes is $n = \{5, 10\}$ for the linear models (including the mean shape). For the non-linear models, we

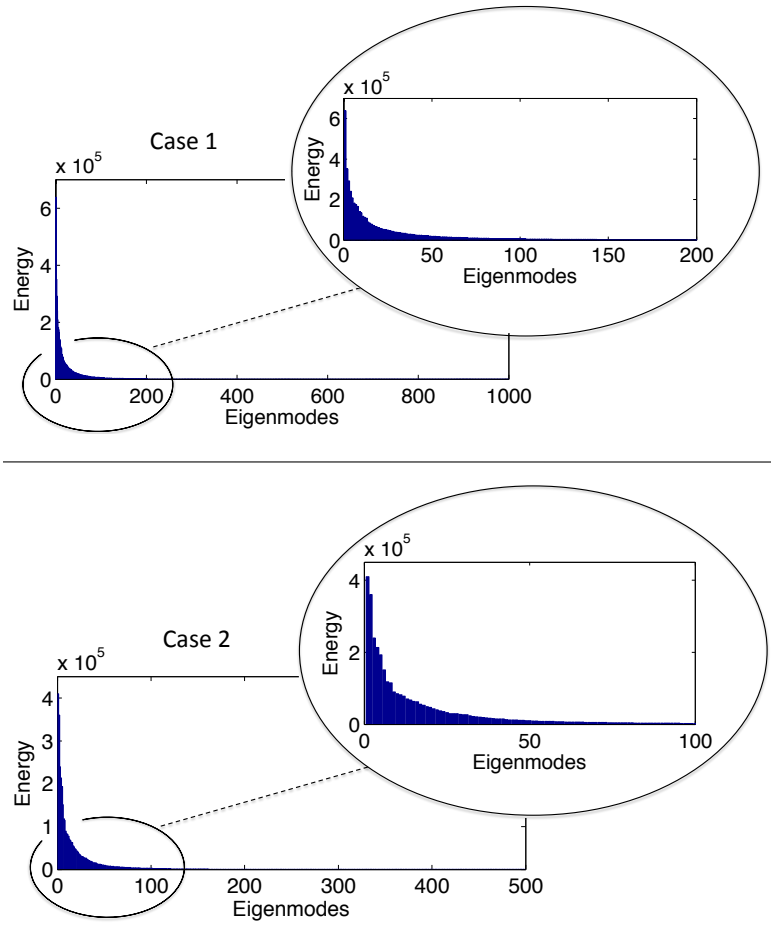

Figure A.4: Eigenvalue decay for models built from 1000 and 500 samples for *Case 1* and *Case 2*, respectively.

chose 200, 100, 50, 20 and n modes. Reconstruction errors are shown in Table A.1, and graphically in Figure A.5. Qualitative results are shown in Figure A.6.

A.2.3 3D–3D Shape reconstruction

In this section, we examine model generalisation from 3D feature points. As in the previous section, we use $m_l = n$ modes for the linear model. For the non-linear model, we allow the number of modes m_n to scale linearly with the number of feature points f. We chose $m_n = \frac{f}{2}$ for our experiments. The number of feature points are selected as: $\mathcal{F} = \{400, 300, 200, 100, 70, 50, 30, 20, 10\}$. Each $f \in \mathcal{F}$ is

Case 1	Non-Lin.					Lin.
Modes:	200	100	50	20	10	10
Error: $\times 10^{11}$	0.93	1.80	3.34	6.13	8.04	7.65
Case 2	Non-Lin.					Lin.
Modes:	200	100	50	20	5	5
Error: $\times 10^{11}$	2.46	3.73	5.44	9.06	12.96	13.03

Table A.1: Reconstruction error for out-of-sample face projected into model build from non-linear data using different number eigenmodes. For comparison we show best possible reconstruction error for a linear combination of training samples.

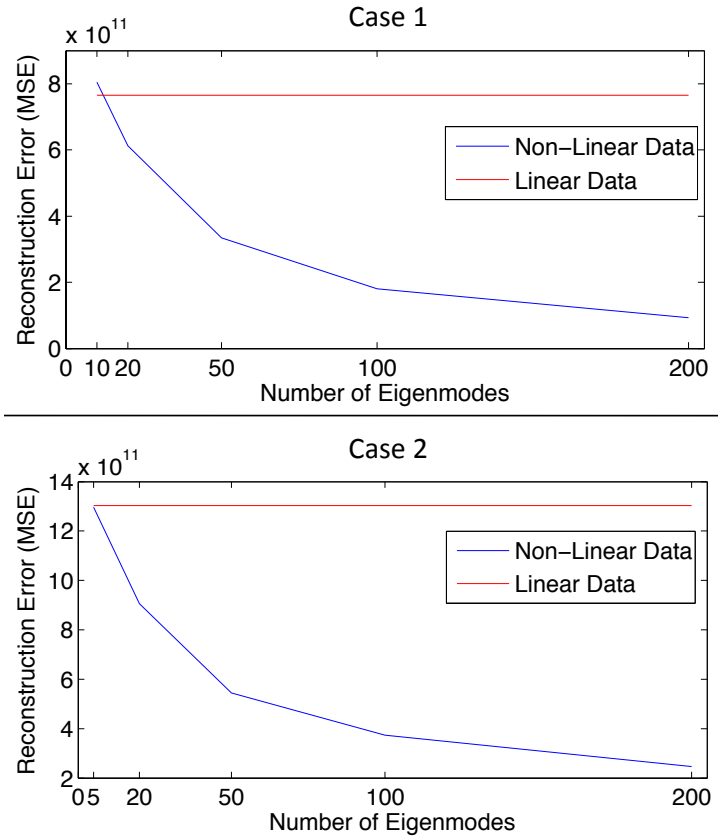

Figure A.5: Generalisation Error (MSE). Ground truth projected into the corresponding models.

selected randomly from the total number of vertices $T = 53490$. We repeat each case 100 times and report the mean value thereof. The results are shown in Table A.2 and plotted in Figure A.7.

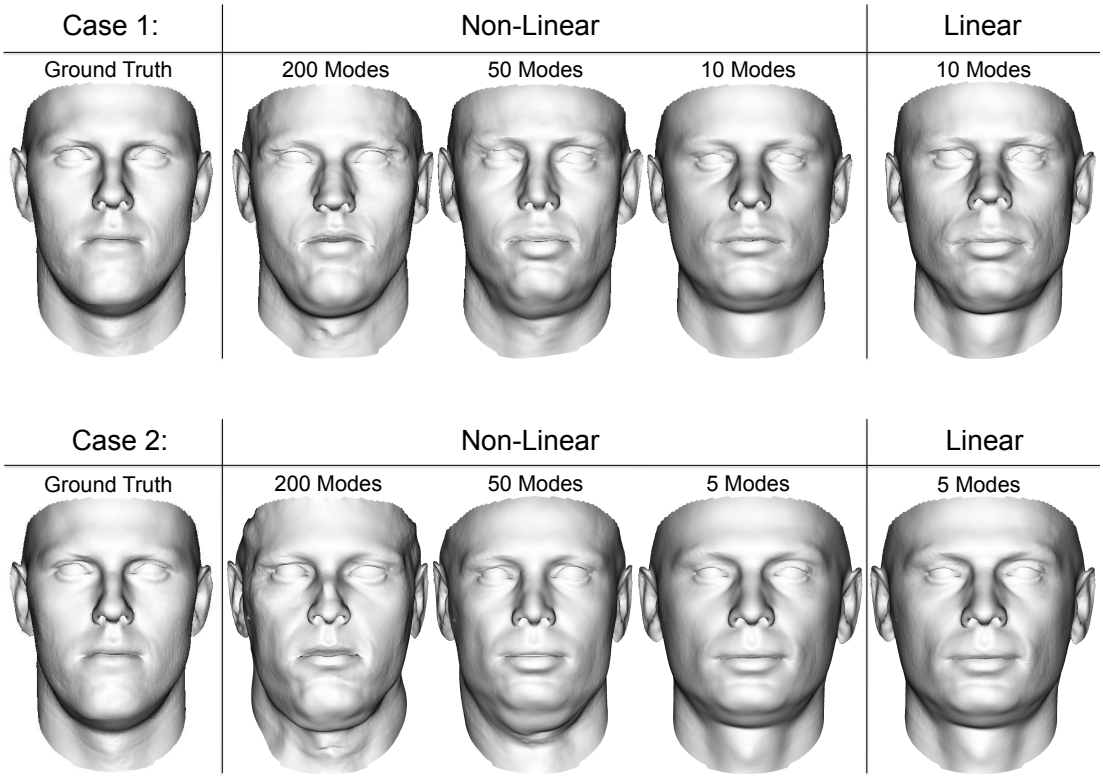

Figure A.6: Generalisation Results. Ground truth projected into the corresponding models.

Case 1									
Feat. No.:	400	300	200	100	70	50	30	20	10
Non. Lin.	1.85	2.23	2.85	4.81	5.27	6.79	9.21	10.2	14.4
Lin.	7.73	7.76	7.81	8.01	8.11	8.35	8.92	9.56	11.9
Case 2									
Feat. No.:	400	300	200	100	70	50	30	20	10
Non. Lin.	3.99	4.53	5.59	7.35	7.78	9.93	12.0	13.3	14.4
Lin.	13.0	13.0	13.1	13.2	13.3	13.4	13.8	14.1	15.6

Table A.2: Reconstruction error for out-of-sample face projected into model build from non-linear data using different number eigenmodes. For comparison we show best possible reconstruction error for a linear combination of training samples.

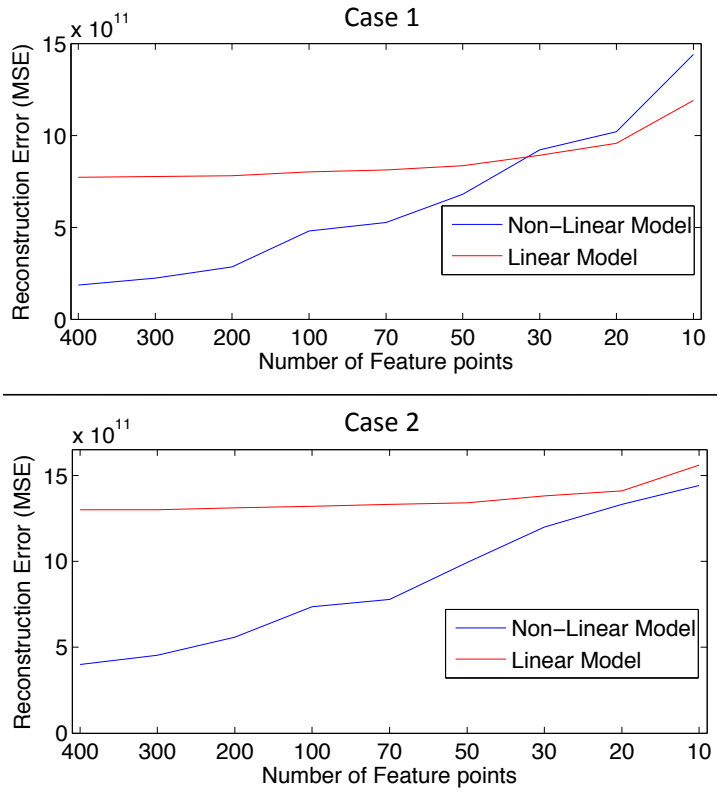

Figure A.7: Generalisation Error (MSE). Feat. points projected into the corresponding models.

A.3 Discussion

The experimental results have attested, that the proposed method of synthesising novel faces allows to build more flexible morphable models with improved generalisation abilities. In order to benefit from the improvements, the required number of feature points is $f > 35$ for *Case 1*. For *Case 2* each selected f showed improvements. These are preliminary results and can not be understood as a thorough validation. In both cases, but in particular in *Case 2*, the number of basis shapes is extremely low. It would be elusive to assume, that a flexible morphable model with strong inference capabilities can be learnt from only 5 basis shapes. Unfortunately, the lack of available basis shapes prohibited are comprehensive examination. Besides the number of basis shapes, the following explores further suggestions for improvements.

Global consistency We have not enforced any form of global consistency for the training-set. A simple way to do so is exploiting symmetry. A more sophisticated approach is to fix the linear weights at certain feature points (for instance the Farkas points) to the same (or similar) value. The random fields are then conditioned on that points.

Unnatural deformations Some training examples show unnatural deformations at certain regions. This skews the statistics of the model. The problem can be circumvented by segmenting the training data and discard distorted segments. This has been proposed for skull data by Luethi et al. [19]. The downside is, that the model parameters can not be learnt in closed-form.

Random fields We have used exponential random fields as weighting functions. In general any smooth 2D function can be used. Examples include Gaussian random fields, random harmonic fields or affine. Combinations thereof are possible.

A.4 Conclusion

Preliminary results have shown, that it is possible to build flexible morphable models from limited training data. Instead of using the faces itself, we compose non-linear combinations thereof using exponential random fields. By normalising the weighting functions to value greater than one, we allow to move outside of the sample space by preserving local consistency.

In future work, we like to combine the ideas of class specific priors and non-linear shape composition and conduct a comprehensive study on a larger training and test-set. We would also like to explore how the proposed method can be used to construct a patch-based texture model. An appearance based version thereof was proposed by Mohammed et al. [129]. The advantage of synthesising texture in (u, v) reference frame is pose and illumination invariance.

References

[1] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face recognition by humans: 20 results all computer vision researchers should know about," *Department of Brain and Cognitive Sciences Massachusetts Institute of Technology Cambridge, MA*, 2005. 1

[2] K. Grill-Spector and R. Malach, "The human visual cortex," *Annual Review of Neuroscience*, vol. 27, pp. 649–677, 2004. 1

[3] E. Adelson, "Perceptual organization and the judgment of brightness," *Science*, vol. 262, no. 5142, pp. 2042–2044, 1993. 1

[4] B. W, P. Isola, I. Blank, and A. Oliva, "Establishing a database for studying human face photograph memory," in *Proceedings of the Cognitive Science Society*. Cognitive Science Society, 2012. 2

[5] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: A review," *Medical Image Analysis*, vol. 13, no. 4, pp. 543 – 563, 2009. 2

[6] C. Lorenz and N. Krahnstöver, "3D statistical shape models for medical image segmentation," in *Proceedings of the 2nd international conference on 3D digital imaging and modeling*. IEEE Computer Society, 1999, pp. 414–423. 2

[7] J. M. Ferryman, A. D. Worrall, G. D. Sullivan, and K. D. Baker, "A generic deformable model for vehicle recognition," in *Proc. BMVC*, 1995, pp. 127–136. 2

[8] V. Petrovic and T. F. Cootes, "Analysis of features for rigid structure vehicle type recognition," in *Proc. BMVC*, 2004, pp. 587–596. 2

[9] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *ACM Trans. Graphic. (Proceedings of SIGGRAPH)*, 1999, pp. 187–194. 2, 13, 16, 30, 32, 57, 71

[10] G. Vogiatzis, C. Hernandez, P. Torr, and R. Cipolla, "Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2241 –2246, 2007. 2

[11] R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *Int. J. Comput. Vis.*, vol. 72, no. 3, pp. 239–257, 2007. 2

[12] W. Burger and M. J. Burge, *Principles of Digital Image Processing: Core Algorithms*, 1st ed. Springer Publishing Company, Incorporated, 2009. 3

[13] S. Prince, *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012. 3, 9

[14] T. Kanade, "Picture processing system by computer complex and recognition of human faces," Ph.D. dissertation, Kyoto University, 1973. 3

[15] R. Jafri and H. R. Arabnia, "A survey of face recognition techniques," *Journal of Information Processing Systems*, pp. 41–68, 2009. 3

[16] W. A. P. Smith and E. R. Hancock, "Facial shape–from–shading and recognition using principal geodesic analysis and robust statistics," *Int. J. Comput. Vision*, vol. 76, no. 1, pp. 71–91, 2008. 3

[17] ——, "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. 3, 30, 34, 96

[18] B. Moghaddam, J. Lee, H. Pfister, and R. Machiraju, "Model–based 3D face capture with shape–from–silhouettes," in *Proc. IEEE Work. Analysis and Modeling of Faces and Gestures*, 2003, pp. 20–27. 3, 33

[19] M. Lüthi, T. Albrecht, and T. Vetter, "Building shape models from lousy data," in *Proc. of the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part II*, ser. MICCAI '09, 2009, pp. 1–8. 3, 126

[20] P. Paysan, M. Lüthi, T. Albrecht, A. Lerch, B. Amberg, F. Santini, and T. Vetter, "Face reconstruction from skull shapes and physical attributes," in *Proceedings of the 31st DAGM Symposium on Pattern Recognition*, 2009, pp. 232–241. 3

[21] P. Ekman, "Facial expressions of emotion: New findings, new questions," *Psychological Science*, vol. 3, no. 1, pp. 34–38, 1992. [Online]. Available: http://pss.sagepub.com/lookup/doi/10.1111/j.1467-9280.1992.tb00253.x 4

[22] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. Prince, "Probabilistic models for inference about identity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 144 –157, jan. 2012. 4, 11

[23] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proc. ECCV*, 2002, pp. 447–460. 4

[24] D. Vlasic, M. Brand, H. P. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM Trans. Graphic. (Proceedings of SIGGRAPH)*, vol. 24, no. 3, pp. 426–433, 2005. 4, 25

[25] 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. 5, 30, 32

[26] S. Romdhani, J. Ho, T. Vetter, and D. J. Kriegman, "Face recognition using 3D models: Pose and illumination," *Proc. of the IEEE*, vol. 94, no. 11, pp. 1977–1999, 2006. 7, 36

[27] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. IEEE Intl. Conf. on Advanced Video and Signal based Surveillance*, 2009. 7, 12, 16, 101, 117

[28] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proc. CVPR*, 1991, pp. 586–591. 9, 10, 11, 32

[29] B. Amberg, "Editing faces in videos," Ph.D. dissertation, Universitaet Basel, 2011. 9

[30] M. Breidt, H. H. Bülthoff, and C. Curio, "Towards building a 4D morphable face model," in *Proceedings of the ACM/SSPNET 2nd International Symposium on Facial Analysis and Animation*. New York, NY, USA: ACM, 2010. 9

[31] J. Modersitzki, *Fair: Flexible Algorithms for Image Registration*, ser. Fundamentals of Algorithms. Society for Industrial and Applied Mathematics, 2009. 9, 14

[32] L. Sirovich and M. Kirby, "Low–dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am.*, vol. 4, no. 3, pp. 519–524, 1987. 10, 30

[33] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 7, pp. 711–720, 1997. 11

[34] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, "Active shape models – their training and application," *Comput. Vis. Image Underst.*, vol. 61, pp. 38–59, 1995. 11

[35] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. ECCV*, 1998, pp. 484–498. 11, 32

[36] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *Proc. ECCV*, 1998, pp. 581–695. 11

[37] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The extended M2VTS database," in *In Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999, pp. 72–77. 11

[38] R. W. Frischholz and U. Dieckmann, "BioID: A multimodal biometric identification system," *Computer*, vol. 33, no. 2, 2000. 11

[39] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295 – 306, 1998. 11

[40] R. Gross, "Face databases," in *Handbook of Face Recognition*, A. S.Li, Ed. New York: Springer, February 2005. 12

[41] M. Hamouz, J. Tena, J. Kittler, A. Hilton, and J. Illingworth, "3D assisted face recognition: A survey," in *3D Imaging for Safety and Security*. Springer Netherlands, 2007, vol. 35, pp. 3–23. 12

[42] V. Blanz, "Automatische rekonstruktion der dreidimensionalen form von gesichtern aus einem einzelbild," Ph.D. dissertation, Universitaet Tuebingen, 2000. 12, 13

[43] S. Romdhani, "Face image analysis using a multiple features fitting strategy," Ph.D. dissertation, Universitaet Basel, 2005. 12, 13, 18, 19

[44] K. Zhou, J. Snyder, B. Guo, and H.-Y. Shum, "Iso-charts: Stretch-driven mesh parameterization using spectral analysis," in *Proc. Eurographics Symposium on Geometry Processing*, 2004, pp. 47–56. 12

[45] S. Huq, B. Abidi, S. G. Kong, and M. Abidi, "A survey on 3D modeling of human faces for face recognition," in *3D Imaging for Safety and Security*. Springer Netherlands, 2007, vol. 35, pp. 25–67. 13

[46] K. A. Sidorov, S. Richmond, and D. Marshall, "Efficient groupwise non-rigid registration of textured surfaces," in *Proc. CVPR*, 2011, pp. 2401 –2408. 14, 32

[47] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step non–rigid ICP algorithms for surface registration." in *Proc. CVPR*, 2007. 14

[48] A. Dedner, M. Lüthi, T. Albrecht, and T. Vetter, "Curvature guided level set registration using adaptive finite elements," in *Proceedings of the 29th*

DAGM conference on Pattern recognition. Berlin, Heidelberg: Springer-Verlag, 2007. 14

[49] A. Patel and W. A. P. Smith, "3D morphable face models revisited," in *Proc. CVPR*, 2009, pp. 1327–1334. 16

[50] USF HumanID 3D Face Database, Courtesy of Sudeep. Sarkar, University of South Florida, Tampa, FL. 16

[51] L. Farkas, *Anthropometry of the Head and Face*. New York: Raven Press, 1994. 17, 43, 44

[52] S. Romdhani and T. Vetter, "Efficient, robust and accurate fitting of a 3D morphable model," in *Proc. ICCV*, 2003. 19

[53] P. Haeuschen, *Ambient Occlusion: Grundlagen, Konzepte, und Beispielimplementation*. Vdm Verlag Dr. Mueller, 2008. 19

[54] J. Ho and D. Kriegman, *Face Processing: Advanced Modeling and Methods, Chap. On the Effect of Illumination and Face Recognition*, W. Zhao and R. Chellappa, Eds. U.S.A.: Academic Press, 2005. 19, 58

[55] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *Proc. CVPR*, vol. 2, 2005, pp. 986–993. 22, 33, 44, 46, 48, 64, 75, 76, 77, 82, 85, 86

[56] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," Dec. 2010. 22

[57] P. Ekman and W. V. Friesen, "A new pan-cultural facial expression of emotion," *Motivation and Emotion*, vol. 10, pp. 159–168, 1986. 23

[58] D. Cosker, R. Borkett, D. Marshall, and P. L. Rosin, "Towards automatic performance driven animation between multiple types of facial model," *Computer Vision, IET*, vol. 2, pp. 129–141, 2008. 23

[59] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978. 24

[60] A. Bronstein, M. Bronstein, and R. Kimmel, "Three–dimensional face recognition," *Int. J. Comput. Vis.*, vol. 64, no. 1, pp. 5–30, 2005. 24

[61] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," in *Proc. ECCV*, 2002, pp. 447–460. 24

[62] M. Alex, M. A. O. Vasilescu, and D. Terzopoulos, "Multilinear image analysis for facial recognition," in *Proc. ICPR*, 2002, pp. 511–514. 24

[63] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear subspace analysis of image ensembles," in *Proc. CVPR*, 2003, pp. 2–93. 24

[64] ——, "TensorTextures: multilinear image-based rendering," *ACM Trans. Graphic. (Proceedings of SIGGRAPH)*, vol. 23, no. 3, pp. 336–342, 2004. 25

[65] B. Amberg, R. Knothe, and T. Vetter, "Expression invariant 3D face recognition with a morphable model," in *Proc. IEEE Intl. Workshop on Analysis and Modeling of Faces and Gestures*, 2008. 25, 26

[66] P. Paysan, M. Lüthi, T. Albrecht, A. Lerch, B. Amberg, F. Santini, and T. Vetter, "Face reconstruction from skull shapes and physical attributes," in *Proceedings of the 31st DAGM Symposium on Pattern Recognition.* Berlin, Heidelberg: Springer-Verlag, 2009, pp. 232–241. 27

[67] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, and W. Sarokin, "Acquiring the reflectance field of a human face," in *ACM Trans. Graphic. (Proceedings of SIGGRAPH)*, 2000. 27

[68] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec, "Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination," in *Proc. Eurographics Symposium on Rendering*, 2007. 27, 89

[69] O. Alexander, M. Rogers, W. Lambeth, M. Chiang, and P. Debevec, "Creating a photoreal digital actor: The digital emily project," *Conference for Visual Media Production*, vol. 0, pp. 176–187, 2009. 28, 29

[70] M. S. Langer and S. W. Zucker, "Shape-from-shading on a cloudy day," *J. Opt. Soc. Am. A*, vol. 11, pp. 467–478, Feb 1994. 28, 110

[71] S. R. Marschner and D. P. Greenberg, "Inverse lighting for photography," in *Proc. Fifth Color Imaging Conference*, 1997, pp. 262–265. 30, 31

[72] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *ACM Trans. Graphic. (Proceedings of SIGGRAPH)*, 2001, pp. 117–128. 30, 31

[73] I. Craw and P. Cameron, "Parameterising images for recognition and reconstruction," in *Proc. BMVC*, 1991, pp. 367–370. 30, 32

[74] T. Heseltine, N. Pears, and J. Austin, "Three–dimensional face recognition using surface space combinations," in *Proc. BMVC*, 2004. 30

[75] N. P. Costen, T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Automatic extraction of the face identity–subspace," *Image Vis. Comput.*, vol. 20, pp. 319–329, 2002. 30

[76] A. Hertzmann and S. M. Seitz, "Example–based photometric stereo: Shape reconstruction with general, varying BRDFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1254–1264, 2005. 31

[77] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz, "Shape and spatially-varying BRDFs from photometric stereo," in *Proc. ICCV*, 2005. 31

[78] S. Marschner, S. Westin, E. Lafortune, K. Torrance, and D. Greenberg, "Reflectance measurements of human skin," Cornell University, Tech. Rep. PCG–99–2, 1999. 31

[79] A. Georghiades, "Recovering 3D shape and reflectance from a small number of photographs," in *Eurographics Symposium on Rendering*, 2003, pp. 230–240. 31

[80] K. Torrance and E. Sparrow, "Theory for off–specular reflection from roughened surfaces," *J. Opt. Soc. Am.*, vol. 57, no. 9, pp. 1105–1114, 1967. 31

[81] M. Fuchs, V. Blanz, H. Lensch, and H.-P. Seidel, "Reflectance from images: A model–based approach for human faces," *IEEE T. Vis. Comput. Graph.*, vol. 11, no. 3, pp. 296–305, 2005. 31

[82] J. J. Atick, P. A. Griffin, and A. N. Redlich, "Statistical approach to SFS: Reconstruction of 3D face surfaces from single 2D images," *Neural Comp.*, vol. 8, no. 6, pp. 1321–1340, 1996. 32

[83] K. Sidorov, S. Richmond, and D. Marshall, "An efficient stochastic approach to groupwise non-rigid image registration," in *Proc. CVPR*, 2009, pp. 2208 –2213. 32

[84] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor, "A minimum description length approach to statistical shape modelling," *IEEE Transactions on Medical Imaging*, vol. 21, pp. 525–537, 2001. 32

[85] S. Romdhani, V. Blanz, and T. Vetter, "Face identification by fitting a 3D morphable model using linear shape and texture error functions," in *Proc. ECCV*, 2002, pp. 3–19. 33, 64

[86] 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. 33, 42, 64

[87] R. Knothe, S. Romdhani, and T. Vetter, "Combining PCA and LFA for surface reconstruction from a sparse set of control points," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, 2006, pp. 637–644. 33

[88] L. Zhang and D. Samaras, "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 351–363, 2006. 33, 82

[89] H. Shim, I. Ha, T. Rhee, J. D. Kim, and C. Kim, "A probabilistic approach to realistic face synthesis," in *Proc. ICIP*, 2010. 33

[90] W. Y. Zhao and R. Chellappa, "Symmetric shape–from–shading using self–ratio image," *Int. J. Comput. Vision*, vol. 45, pp. 55–75, 2001. 34

[91] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3D structure recovery of faces," in *Proc. ECCV*, vol. 2, 2004, pp. 99–113. 34

[92] W. A. P. Smith and E. R. Hancock, "A new framework for grayscale and colour non–lambertian shape–from–shading," in *Proc. ACCV*, 2007, pp. 869–880. 34

[93] E. Prados and O. Faugeras, "A generic and provably convergent shape-from-shading method for orthographic and pinhole cameras," *Int. J. Comput. Vision*, vol. 65, no. 1-2, pp. 97–125, 2005. 34

[94] I. Kemelmacher and R. Basri, "Molding face shapes by example," in *Proc. ECCV*, 2006, pp. 277–288. 34

[95] 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. 34, 35

[96] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, 2003. 34, 36, 61

[97] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 1–15, 2005. 34

[98] S. Zhou and R. Chellappa, "Rank constrained recognition under unknown illuminations," in *Proc. IEEE Intl. Workshop on Analysis and Modeling of Faces and Gestures*, 2003, pp. 11–18. 35

[99] K. Nishino, P. N. Belhumeur, and S. K. Nayar, "Using eye reflections for face recognition under varying illumination," in *Proc. ICCV*, vol. 1, 2005, pp. 519–526. 35

[100] A. Pentland, B. Moghaddam, and T. Starner, "View–based and modular eigenspaces for face recognition," in *Proc. CVPR*, 1994. 35

[101] T. F. Cootes, K. N. Walker, and C. J. Taylor, "View–based active appearance models," in *Proc. Int. Conf. on Face and Gesture Recognition*, 2000, pp. 227–232. 35

[102] V. Blanz, P. Grother, J. Phillips, and T. Vetter, "Face recognition based on frontal views generated from non–frontal images," in *Proc. CVPR*, vol. 2, 2005, pp. 454–461. 35

[103] M. S. Langer and H. H. Büthoff, "Depth discrimination from shading under diffuse lighting," 2000. 36, 110

[104] E. Prados, N. Jindal, and S. Soatto, "A non-local approach to shape from ambient shading," in *Proc. IEEE Intl. Conf. on Scale Space and Variational Methods in Computer Vision*. Springer-Verlag, 2009, pp. 696–708. 36, 110

[105] S. Zhukov, A. Iones, and F. Kronin, "An ambient light illumination model," in *Rendering Techniques, Proceedings of the Eurographics Workshop*. Springer, 1998. 36, 93

[106] H. Landis, "Production-ready global illumination," *Siggraph Course Notes*, vol. 16, no. 3, 2002. 36, 91, 93

[107] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000. 39

[108] C. Creusot, N. Pears, and J. Austin, "3D landmark model discovery from a registered set of organic shapes," in *Proc. CVPR workshop on Point Cloud Processing*, 2012. 44

[109] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, pp. 611–622, 1999. 44, 99

[110] B. Amberg and T. Vetter, "Optimal landmark detection using shape models and branch and bound," in *Proc. ICCV*, 2011, pp. 455 –462. 45, 115

[111] A. M. Burton and J. R. Vokey, "The face-space typicality paradox: Understanding the face-space metaphor," *The Quarterly Journal of Experimental Psychology Section A*, vol. 51, no. 3, pp. 475–483, 1998. 49

[112] A. Patel and W. A. P. Smith, "Exploring the identity manifold: constrained operations in face space," in *Proc. ECCV*. Springer-Verlag, 2010, pp. 112–125. 49

[113] J. Wu, W. A. Smith, and E. R. Hancock, "Gender discriminating models from facial surface normals," *Pattern Recognition*, vol. 44, no. 12, pp. 2871 – 2886, 2011. 49

[114] H.-C. Kim, D. Kim, and S. Y. Bang, "Face recognition using the mixture-of-eigenfaces method," *Pattern Recognition Letters*, vol. 23, no. 13, pp. 1549 – 1558, 2002. 50

[115] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977. 50

[116] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, Jun. 2000. 50

[117] T. Zickler, S. Mallick, D. Kriegman, and P. Belhumeur, "Color subspaces as photometric invariants," *Int. J. Comput. Vision*, 2008. 58, 61, 64

[118] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for reflection," *ACM Trans. Graphic. (Proceedings of SIGGRAPH)*, vol. 23, no. 4, pp. 1004–1042, 2004. 59, 61

[119] R. Ramamoorthi, "Modeling illumination variation with spherical harmonics," in *Face Processing: Advanced Modeling and Methods*. Academic Press, 2005. 59

[120] E. W. Weisstein, "Associated Legendre Polynomial," http://mathworld.wolfram.com/AssociatedLegendrePolynomial.html. 59

[121] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, 2003. 74, 80

[122] University of Southern California, "High-resolution light probe image gallery," 2011, http://gl.ict.usc.edu/Data/HighResProbes. 76, 90, 101

[123] M. Pharr and G. Humphreys, *Physically Based Rendering: From Theory to Implementation*, ser. Morgan Kaufmann. Elsevier Science, 2010. 88, 90

[124] Visual Computing Laboratory, Institute of the National Research Council of Italy, "Meshlab," http://meshlab.sourceforge.net/. 90, 98, 101

[125] P. T. Fletcher, S. Joshi, C. Lu, and S. M. Pizer, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans. Med. Imaging*, vol. 23, no. 8, pp. 995–1005, 2004. 96

[126] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, ser. Adaptive computation and machine learning. MIT Press, 2006. 100

[127] G. Lanckriet, N. Cristianini, P. Bartlett, and L. E. Ghaoui, "Learning the kernel matrix with semi-definite programming," *Journal of Machine Learning Research*, vol. 5, p. 2004, 2002. 115

[128] P. Besl and N. McKay, "A method for registration of 3D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, pp. 239–256, 1992. 115

[129] U. Mohammed, S. Prince, and J. Kautz, "Visio-lization: Generating novel facial images," *ACM Trans. Graphic. (Proceedings of SIGGRAPH)*, vol. 28, no. 3, 2009. 115, 126