Inverse Rendering of Faces with a 3D Morphable Model

Oswald Aldrian, *Student Member*, *IEEE*, and William A.P. Smith, *Member*, *IEEE*

Abstract—In this paper, we present a complete framework to inverse render faces with a 3D Morphable Model (3DMM). By decomposing the image formation process into geometric and photometric parts, 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 global solution is guaranteed. We start by recovering 3D shape using a novel algorithm which incorporates generalization error of the model obtained from empirical measurements. We then describe two methods to recover facial texture, diffuse lighting, specular reflectance, and camera properties from a single image. The methods make increasingly weak assumptions and can be solved in a linear fashion. We evaluate our findings on a publicly available database, where we are able to outperform an existing state-of-the-art algorithm. We demonstrate the usability of the recovered parameters in a recognition experiment conducted on the CMU-PIE database.

Index Terms—Inverse rendering, face shape, texture and illumination analysis

---

1 INTRODUCTION

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 [21]). 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 and Hanrahan [25] 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 paper, namely, by restricting our consideration to the class of human faces.

Whether in 2D (eye-center-aligned [37] or shape-free, warped images [9]) or 3D (depth maps [19], fields of surface normals [34], or meshes in dense correspondence [5]), human faces have been shown to be highly amenable to description using a linear statistical model. Two-dimensional 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 modeling stage [8]. On the other hand, 3D approaches use face shape and reflectance data collected using face capture devices, allowing face intrinsics to be modeled 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 [6].

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

1.1 Contributions

Both shape and reflectance properties contribute to appearance. Solving for both simultaneously leads to a nonconvex optimization problem [6] which is notoriously difficult to solve. In this paper, we instead propose estimating shape using geometric features alone in a manner which is independent of illumination and reflectance effects. We use the position of sparse 2D feature points, though silhouettes or edges would also be suitable features. 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 [24]. With a shape estimate to hand, we are then able to derive linear methods for reflectance and illumination analysis.

• *The authors are with the Department of Computer Science, University of York, Deramore Lane, York YO10 5GH, United Kingdom.*
E-mail: {oa525, william.smith}@york.ac.uk.

We present two efficient approaches which allow recovery of texture model parameters and specular reflectance properties. The first method assumes arbitrarily distributed but monochromatic illumination (of known color). We use a specular invariant color subspace and spherical harmonics to model the illumination environment. This leads to a bilinear system in the texture parameters and spherical harmonic coefficients. We subsequently estimate the specular reflectance function by fitting a higher order spherical harmonic to the specular difference image obtained by subtracting the estimated diffuse image from the input. The second method allows arbitrarily complex, unknown illumination and specular reflectance (the weakest assumptions of any available algorithm for fitting a morphable model). This model leads to a multilinear system in the texture parameters, spherical harmonic coefficients, and specular reflection parameters. We propose two ways to regularize the problem by encouraging the environment to be gray or minimizing the length of the illumination vector. Both methods allow the global optimum to be obtained.

In Section 1.2, we review the relevant literature. In Section 2, we introduce some preliminary concepts, namely, morphable models, spherical harmonic lighting, and the SUV color space. In Section 3, we describe our approach to fitting the shape model to sparse feature points. In Section 4, we describe two approaches which inverse render texture and illumination properties. In Section 5, we present experimental results on synthetic and real-world imagery, including face recognition results on the CMU PIE database. Finally, we offer some conclusions and directions for future work in Section 6.

1.2 Related Work

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

1.2.1 Inverse Rendering

In the context of arbitrary objects, inverse lighting for a Lambertian surface was considered by Marschner and Greenberg [21]. Ramamoorthi and Hanrahan [25] 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 [18] 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. [16] 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. Rather than a material library, Smith and Hancock [35] used a statistical shape model in conjunction with a parametric reflectance model to recover shape and reflectance properties from single color images.

In the context of faces, Marschner et al. [22] 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 [14] incorporated the Torrance and Sparrow [36] 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. [13] 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.

1.2.2 Statistical Face Modeling

Since the late 1980s, 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 [37] popularized 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 which can be used generatively to synthesize appearances. The 3D analog was proposed by Atick et al. [2]. 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 misregistrations as well as identity variation.

The correspondence problem was addressed in 2D by Craw and Cameron [9], who used manually labeled landmark points to warp images to the mean shape before applying PCA. This was developed further by Cootes et al. [7], who modeled 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 [31] and minimum description length [11] approaches. In 3D, Blanz and Vetter [6] used a version of optical flow applied to cylindrical parameterizations 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. [32] based on establishing a common embedding across all training samples.

1.2.3 Morphable Model Fitting

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. [2] fitted their 3D statistical face model to images using a gradient-descent-based optimization of the shape parameters. Blanz and Vetter [5], [6] 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 nonlinear optimization. The method required careful initialization and relied on a stochastic optimization 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. [28] 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 optimization of pose and illumination parameters and the overall objective is therefore nonlinear. Romdhani and Vetter [30] 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 optimize. Blanz et al. [4] showed how to fit a morphable model to a sparse sample of feature point positions. Their approach required careful selection of a global regularization parameter and required iterative reestimation of the perspective projection parameters. The most similar work in spirit to that presented here is due to Zhang and Samaras [38]. 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 modeled by a linear approach.

1.2.4 Pose/Illumination Insensitive Face Recognition

Face recognition under extreme variations of illumination and pose has presented a serious research challenge. Appearance-based approaches [3], [15], [20] 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 synthesize 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 [39].

1.3 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 [29]. 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 toward specific imaging environments. In particular, no existing method allows for both arbitrarily complex illumination and non-Lambertian reflectance. Our method is the most general to date and the first to guarantee globally optimal solutions in both the estimated shape and texture. Moreover, because every step of our algorithm is linear, it is highly efficient.

2 PRELIMINARIES

In this section, we revise some concepts that are used throughout the paper. We begin by describing 3D morphable models (3DMM), a type of statistical model which we use to constrain the inverse rendering problem. We then introduce spherical harmonic lighting, an efficient representation for reflectance under arbitrary illumination. Finally, we describe the SUV color space of Zickler et al. [40]. This is a source-dependent color space which we make use of in Section 4.1 for specular invariant model fitting.

2.1 3D Morphable Models

We use a 3D morphable model of shape and texture as a constraint to ensure that the results of our inverse rendering correspond to a plausible face appearance. 3DMMs were introduced by Blanz and Vetter [5] and have subsequently been applied successfully to applications in vision and graphics. A 3D morphable model is constructed from m face meshes which are in *dense correspondence*. That is to say, vertices with the same index in each mesh correspond to the same point on the face. 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:

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

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-normalized shape parameter vector as: $\mathbf{c}_s = [a_1/\sigma_{s,1} \, \ldots \, a_{m-1}/\sigma_{s,m-1}]^{\mathrm{T}}$.

2.2 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 [29]. We use the well-known spherical harmonic framework to efficiently represent reflectance under complex

illumination. Spherical harmonic basis functions are well suited to being 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. [26], [27] showed that lighting, 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:

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

$\mathcal{H}_{l,m}$ are orthonormal basis functions, where the subscript l denotes the degree and m the corresponding order. The degree l determines the polynomial degree used to compute the basis functions in terms of the surface normals. $\mathcal{L}_{l,m}$ are corresponding weightings which are termed the lighting coefficients. A symmetric BRDF as a function of the incident elevation angle can be expanded using the same set of basis functions. See [26] for a detailed description. Note that a symmetric BRDF is only a function of degree l and does not depend on 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 normalization constant.

In most real-world cases, we cannot 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 modeled by coefficients: $l_{l,m} = \Lambda_l \hat{\rho}_l \mathcal{L}_{l,m}$. We denote the concatenation of the diffuse coefficients as a parameter vector, l. Previous experimental results have shown that unconstrained complex illumination can be well approximated by a linear subspace. A second degree spherical harmonic approximation ($l = \{0, 1, 2\}$) accounts for at least 98 percent in the variability of the reflectance function. This result was independently derived by both Basri and Jacobs [3] and Ramamoorthi [27].

2.3 SUV Color Subspace

Recently, Zickler et al. [40] proposed a linear transformation in RGB color space, which is invariant to specularities and preserves diffuse shading. According to the dichromatic model, observations $\mathbf{I}_k$ are linear combinations of the diffuse color $\mathbf{D}$ and the specular color $\mathbf{S}$:

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

where the subscript c represents R, G, and B, respectively. The coefficients σ_d and σ_s are geometric scale 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 [40] proposes an efficient operation which transforms observations into specular invariant representations. The new representation is termed SUV *color 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 color of the light source $\mathbf{S}$ and therefore satisfies the condition $\mathbf{R}\mathbf{S} = (1, 0, 0)$. 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:

$$I_U = \sigma_d \mathbf{r}_2^{\mathrm{T}} \mathbf{D} = \mathbf{r}_2^{\mathrm{T}} \mathbf{I}_{RGB}, \quad (1)$$

$$I_V = \sigma_d \mathbf{r}_3^{\mathrm{T}} \mathbf{D} = \mathbf{r}_3^{\mathrm{T}} \mathbf{I}_{RGB}. \quad (2)$$

The vectors $\mathbf{r}_2^{\mathrm{T}}$ and $\mathbf{r}_3^{\mathrm{T}}$ correspond to the second and third row of the rotation matrix $\mathbf{R}$, which can be obtained using quaternions.

3 SHAPE MODEL FITTING

As part of our proposed framework, we present a novel algorithm for shape parameter estimation under unknown pose given the 2D coordinates of a sparse set of $N \ll p$ feature points. In order to obtain a linear solution, we decompose the problem into two steps which can be iterated and interleaved: 1) estimation of the camera projection matrix using known 3D-2D correspondences, and 2) estimation of 3D shape parameters using a known camera projection matrix. We initialize 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 reestimate the camera matrix, we can iterate the process, which typically converges in ≤ 5 iterations. Note that since the system is bilinear in the unknown shape and camera parameters, alternating least squares converges to a unique solution.

3.1 Estimating the Camera Projection Matrix

In order to obtain a linear solution, 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 normalized 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.

We assume an affine camera and compute the normalized projection matrix, $\tilde{\mathbf{C}} \in \mathbb{R}^{3\times4}$, using the *Gold Standard Algorithm* [17]. 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 minimizes: $\sum_i \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{C}}\tilde{\mathbf{X}}_i\|^2$, subject to

the affine constraint $\tilde{\mathbf{C}}_3 = [0\ 0\ 0\ 1]$. This system can be solved using least squares. The camera matrix is obtained by performing a denormalization step: $\mathbf{C} = \mathbf{T}^{-1}\tilde{\mathbf{C}}\mathbf{U}$.

3.2 Modeling Feature Point Variance

In order to explain the difference between observed and modeled 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 regularization weight parameter. The first source of variance is the generalization 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. Generalization error is spatially varying, i.e., some regions of the face are harder to generalize to than others, and it is this affect that is captured by having per-feature point variances. 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. The total variance of a feature point is the sum of the 3D variance projected to 2D and the 2D variance.

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: $\mathbf{v}'_i = \mathbf{V}\mathbf{V}^{\mathrm{T}}(\mathbf{v}_i - \bar{\mathbf{v}}) + \bar{\mathbf{v}}$. The vector of squared errors is given by: $\mathbf{e}_i = (\mathbf{v}_i - \mathbf{v}'_i)^2$. We define $\hat{\mathbf{e}}_i$ as the vector formed by subselecting 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 variance associated with the coordinates of the feature points: $\sigma^2_{3D,j} = \frac{1}{k}\sum_{i=1}^{k} \hat{\mathbf{e}}_{i,j}$, where $\sigma^2_{3D,j} \in \mathbb{R}^3$. This provides empirical means to predict how a feature point is likely to vary from its true position due to generalization errors. The units of $\sigma^2_{3D,j}$ are mm. The result can be used for 3D-3D reconstruction.

Defining a matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma^{-2}_{3D,j})$, the objective $\mathbb{E}$ becomes

$$\mathbb{E} = (\mathbf{V}\mathbf{a} + \bar{\mathbf{v}} - \mathbf{y})^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}(\mathbf{V}\mathbf{a} + \bar{\mathbf{v}} - \mathbf{y}). \quad (3)$$

This equation can be brought into a standard form, $(\mathbf{A}\mathbf{x} + \mathbf{b})^{\mathrm{T}}\boldsymbol{\Omega}(\mathbf{A}\mathbf{x} + \mathbf{b})$, and solved for the parameters very efficiently (see the supplementary material, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.206).

In order to predict how this results in variation in the image plane, we project the variances to 2D in units of pixels. The 3D variance of the jth feature point in homogeneous coordinates is given by $[\sigma^2_{3D,j,x}\ \sigma^2_{3D,j,y}\ \sigma^2_{3D,j,z}\ 1]^{\mathrm{T}}$. We define $\bar{\mathbf{C}} \in \mathbb{R}^{3\times 4}$ 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: $\sigma^2_{2D,j} = \bar{\mathbf{C}}\sigma^2_{3D,j} + \eta^2$. We use a value of $\eta = \sqrt{3}$ pixels in our experiments.

3.3 A Probabilistic Approach

The 3D shape parameters are obtained using a probabilistic approach which follows that of Blanz et al. [4]. 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 $p(\mathbf{c}_s) = \nu \cdot \exp(-)\ (\frac{1}{2}\|c_s\|^2)$, where ν is a normalization constant. The conditional likelihood of data $\mathbf{y}$ is given by

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

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. 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:

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

which can be maximized by minimizing the exponent:

$$\mathbb{E} = -2 \cdot \log P(\mathbf{c}_s|\mathbf{y}) = \sum_{i=1}^{3N} \frac{[y_{m2D,i} - y_i]^2}{\sigma^2_{2D,i}} + \|c_s\|^2. \quad (4)$$

To simplify, we bring (4) into standard form. The variances are rewritten as $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma^{-1}_{2D,i})$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{\Sigma}$. We set $\mathbf{A} = \mathbf{P}\hat{\mathbf{V}}_h$, $\mathbf{b} = \mathbf{P}\bar{\mathbf{v}} - \mathbf{y}$, and $\mathbf{x} = \mathbf{c}_s$.

4 TEXTURE MODEL FITTING

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. One very simple approach is based on the assumption that the surface is Lambertian and all illumination in the scene is of a single, known color (though its distribution may be arbitrary). In this case, taking ratios between color channels yields systems of linear equations with texture parameters as the only unknowns. In addition, the ratios cancel for occlusions (shadowing). However, in practice, the assumptions are too restrictive and the method performs poorly on real images. We present two more sophisticated approaches that account for specular reflection.

4.1 Method 1: Specular Invariant Model Fitting

The ratio method makes very limiting assumptions about reflectance, neglecting specularities entirely. In our first approach, we retain the flexibility of having an arbitrary distribution of illuminants but still require illumination of a fixed and known color (an extension of the method

would allow a relaxation to two source colors, as described in [40]). Under these assumptions, it is possible to transform to a specular-invariant space in which linear fitting of the texture model can take place.

4.1.1 Image Formation Process

We assume a dichromatic reflectance model comprised of additive Lambertian and specular terms. Unlike the previous work of Vetter et al. [4], [28], [30], 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 color. Implicit in the use of spherical harmonic illumination is the assumption that the object is globally convex (i.e., occlusions are neglected). This leads to the following image formation model:

$$I_{\{r,g,b\}} = S_{\{r,g,b\}} \int_{\Omega_{\mathbf{n}}} L(\omega)[\rho_{\{r,g,b\}}(\mathbf{n} \cdot \omega) + s(\mathbf{n}, \omega, \mathbf{v})] \, d\omega, \quad (5)$$

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

$$\mathbf{I}_{mod} = \mathcal{I}. * (\mathcal{C} \times_1 \mathbf{b} \times_2 \mathbf{l} + \mathcal{S}\mathbf{x}), \quad (6)$$

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 color 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 find an eighth order approximation to be sufficient for 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

$$\mathbf{I}_{mod} = \mathcal{I}. * [(\mathcal{H}\mathbf{l}). * (\mathbf{Tb} + \bar{\mathbf{t}}) + \mathcal{S}\mathbf{x}], \quad (7)$$

where $\mathcal{H} \in \mathbb{R}^{3p \times 9}$ are diffuse spherical harmonic basis functions obtained from the surface normals of the estimated face shape, and $\mathbf{Tb} + \bar{\mathbf{t}} \in \mathbb{R}^{3p}$ denotes diffuse albedo as approximated by the linear texture model. Note that in (6) 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 (7). For convenience, we define

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

The subscript k indicates the rows corresponding to the R, G, and B channels of the kth vertex. According to the specular invariant SUV color space defined in (1), applying

a rotation to the observed intensities allows us to relate diffuse intensity only to the observations

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

where $\mathbf{I}_{RGB,k} \in \mathbb{R}^3$ is a single observation corresponding to the kth vertex. As in (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.

4.1.2 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 is comprised of two terms:

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

For K intensity observations, the individual parts for the U and V channels 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. As the fitting takes place in SUV space, the result is invariant to specularities, with the unique solution determined by the number of iterations. From our experience, the system converges within three-five iterations.

4.1.3 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 optimization 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.

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

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

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 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 three free parameters $\hat{\tau}_l$, where $l \in \{0, 1, 2\}$, which can be obtained by solving the following linear system of equations:

$$\mathbf{I}_s = \hat{\tau}_0 \mathcal{S}_0 \mathcal{L}_0 + \hat{\tau}_1 (\mathcal{S}_{1,-1}\mathcal{L}_{1,-1} + \mathcal{S}_{1,0}\mathcal{L}_{1,0} + \mathcal{S}_{1,1}\mathcal{L}_{1,1})$$
$$+ \hat{\tau}_2 (\mathcal{S}_{2,-2}\mathcal{L}_{2,-2} + \mathcal{S}_{2,-1}\mathcal{L}_{2,-1} + \mathcal{S}_{2,0}\mathcal{L}_{2,0} + \mathcal{S}_{2,1}\mathcal{L}_{2,1} + \mathcal{S}_{2,2}\mathcal{L}_{2,2}).$$

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 $\mathbf{x}_l \in \mathbb{R}^9$.

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:

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

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: $\mathbf{x} = [\mathbf{x}_l | \mathbf{x}_h] \in \mathbb{R}^{81}$. In practice, there is very little visual difference in using orders higher than 5 or even 3 to explain the specular reflectance.

4.1.4 Inverse Rendering Pipeline

Fig. 1 summarizes our fitting procedure graphically. The steps in the pipeline are as follows:

1. The input image is transformed into U-V color space. This results in two specular-invariant observations per vertex color. We show $\sqrt{u^2 + v^2}$ for visualization purposes.
2. Texture and diffuse lighting coefficients are estimated in a bilinear fashion. We use spherical harmonic basis functions up to order 2, which are computed analytically from 3D shape information.
3. This results in a diffuse-only reconstruction.
4. Taking the difference of the input-image and our diffuse estimation allows us 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 spherical harmonic basis functions are computed using the surface normals of the mesh computed at the shape fitting stage.
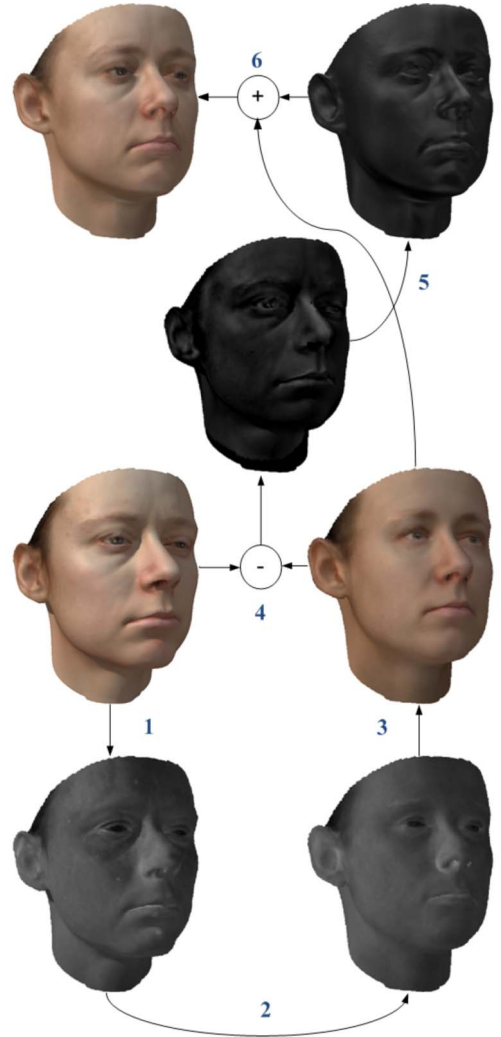

Fig. 1. Overview of the inverse rendering pipeline for Method 1. Each step in the pipeline requires only the solution of a system of linear equations.

4.2 Method 2: 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 color.

4.2.1 Image Formation Process

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

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

where the illumination function now has a wavelength dependence.

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, our modified basis functions are defined as follows:

$$\mathcal{U}(x,y,z)_k = \begin{bmatrix} \mathcal{H}(x,y,z)_k & \mathbf{0}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \mathcal{H}(x,y,z)_k & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} & \mathcal{H}(x,y,z)_k \end{bmatrix}. \quad (8)$$

We also take specular reflection of arbitrary unconstrained illumination into account. We substitute the basis functions $\mathcal{H}(x,y,z)_k$ in (8) 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, similarly to (6), $\mathcal{V}$ corresponds to a third order tensor spanning identity and diffuse illumination, however this time of arbitrary color. In matrix product notation, we state the image formation process as

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

4.2.2 Color Transformation
To make our fitting algorithm more flexible, we also allow for a linear color transformation, $\mathbf{M}(\cdot) \in \mathbb{R}^{3 \times 3}$, and offset, $\mathbf{o} \in \mathbb{R}^3$. This models the color response of the camera. A decomposition of $\mathbf{M}$ into individual contributions, $\mathbf{GC}$, is achieved as follows:

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

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

4.2.3 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 modeled as

$$\mathbf{I}_{mod,k} = \mathbf{M}[(\mathcal{U}_k\mathbf{l}) .* (\mathbf{T}_k\mathbf{b} + \bar{\mathbf{t}}_k) + \mathcal{S}_k\mathbf{x}] + \mathbf{o}. \quad (9)$$

4.2.4 Diffuse Inverse Rendering
We now show how the unknowns in (9) 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 color transformation parameters using the mean texture as initialization. We then correct for these transformations and denote the color corrected observations as $\bar{\mathbf{I}}$.

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 cannot 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 minimize the following objective function, which depends on the color transformation and observations:

$$\mathbb{E}_d = \|\bar{\mathbf{I}} - (\mathcal{U}\mathbf{l}) .* (\mathbf{Tb} + \bar{\mathbf{t}})\|^2. \quad (10)$$

To prevent overfitting, we introduce two sets of prior 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: $\mathbf{F}_r = [\mathbf{I}\ \mathbf{0}\ \mathbf{0}]$, $\mathbf{F}_g = [\mathbf{0}\ \mathbf{I}\ \mathbf{0}]$, and $\mathbf{F}_b = [\mathbf{0}\ \mathbf{0}\ \mathbf{I}]$, 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: $\mathbb{E}_3 = \|\mathbf{F}_r\mathbf{l} - \mathbf{F}_g\mathbf{l}\|^2 + \|\mathbf{F}_r\mathbf{l} - \mathbf{F}_b\mathbf{l}\|^2 + \|\mathbf{F}_g\mathbf{l} - \mathbf{F}_b\mathbf{l}\|^2$. All priors are added to (10) to form the overall cost function: $\mathbb{E}_a = \mathbb{E}_d + \lambda_1\mathbb{E}_1 + \lambda_2\mathbb{E}_2 + \lambda_3\mathbb{E}_3$.

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:

$$\frac{\partial \mathbb{E}_a}{\partial \mathbf{b}} = \frac{\partial \mathbb{E}_d}{\partial \mathbf{b}} + \frac{\partial \lambda_1 \mathbb{E}_1}{\partial \mathbf{b}} = 0, \quad (11)$$

$$\frac{\partial \mathbb{E}_a}{\partial \mathbf{l}} = \frac{\partial \mathbb{E}_d}{\partial \mathbf{l}} + \frac{\partial \lambda_2 \mathbb{E}_2}{\partial \mathbf{l}} + \frac{\partial \lambda_3 \mathbb{E}_3}{\partial \mathbf{l}} = 0, \quad (12)$$

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

4.2.5 Specular Inverse Rendering
As in the previous method (Section 4.1.3), 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 optimization 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.

5 EXPERIMENTS
In this section, we present a comprehensive experimental evaluation of our method. We begin by evaluating the accuracy of our shape reconstruction algorithm and show that it outperforms previously published methods. We then focus on photometric inverse rendering, where we explore simultaneous texture and illumination estimation. We use synthetic imagery to compare the accuracy of the three proposed methods to previously published state-of-the-art results [30]. In a recognition experiment on synthetic data, we show that both the shape and texture estimated by our method outperform state-of-the-art results [30], despite the geometric and photometric aspects of the fitting process being conducted in series. Finally, we show qualitative inverse rendering, illumination clustering, and recognition results for real-world imagery taken from the CMU PIE

TABLE 1
Shape Reconstruction Errors for 10 Out of Sample Faces

Face ID	Proposed	RLS	PPCA
001	0.1154	0.1422	0.1420
002	0.2136	0.3142	0.3918
006	1.9752	3.0120	5.9639
014	0.3392	0.2969	0.4150
017	0.2078	0.2861	0.2602
022	1.8462	1.7245	2.3073
052	0.2066	0.2217	0.1495
053	0.1740	0.2425	0.1692
293	0.0779	0.3942	0.1668
323	0.0917	0.1072	0.1219
Mean	0.5247	0.6741	1.0088

We compare our method to regularized least squares and probabilistic PCA. Errors are mean-squared euclidian error ($\times 10^{12}$) in units of μm^2.

database [33]. For feature points, we use a subset of 70 of the anthropometric landmarks suggested by Farkas [12]. Note that feature point saliency is an interesting topic in itself. Recent work has attempted to learn the most salient feature points from data [10].

5.1 Synthetic Data

The *Basel Face Model* is supplied with a database of synthetic images spanning 10 out-of-sample identities in nine pose angles and three 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 *Multifeature fitting algorithm* [30], for which shape and texture coefficients are provided with the model. We use the 99 most significant modes for shape reconstruction, and 60 most significant modes for texture.

5.1.1 3D-3D Shape Reconstruction

We begin by showing 3D-3D shape reconstruction results from a set of 70 feature points. We compare our proposed method against regularized least squares (RLS) with

isotropic variance and probabilistic PCA (PPCA). For RLS, we find the optimal regularization parameter which minimizes mean-squared error for all 10 faces via a line search. Comparing with PPCA is interesting since the variance for this method is also based on an optimality criterion. Table 1 shows quantitative results for the three methods in terms of mean-squared error. A qualitative comparison is shown in Fig. 2. Although quantitatively inferior, the PPCA result is perceptually better in some cases. The poor performance on faces 006 and 022 can be attributed to poor quality registration between the faces and the model.

5.1.2 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 [1]. Note also that it is straightforward to extend our method to silhouettes and edges. Pixels lying on an image edge are 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 sensitivity of our method to the number of feature points and noise in the feature point positions. In Fig. 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 five 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
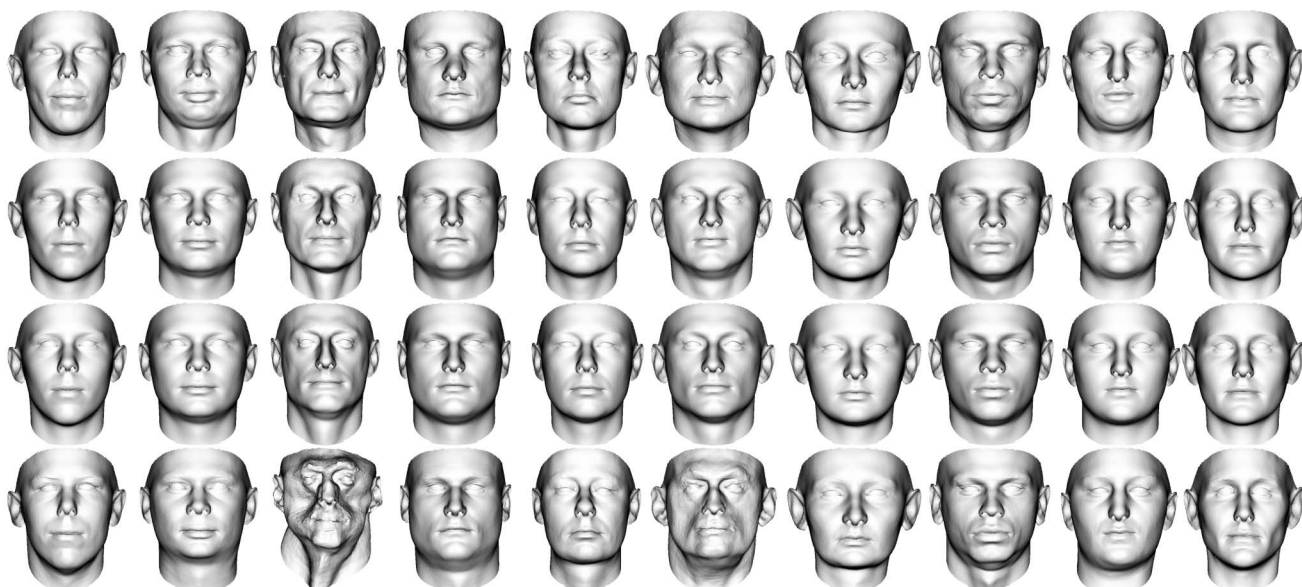

Fig. 2. Three-dimensional reconstructions from 70 feature points for 10 subjects. The top row shows ground-truth shape. The second row shows results for the proposed method. The third row shows reconstructions using regularized least squares. And the last row shows reconstructions using probabilistic PCA.
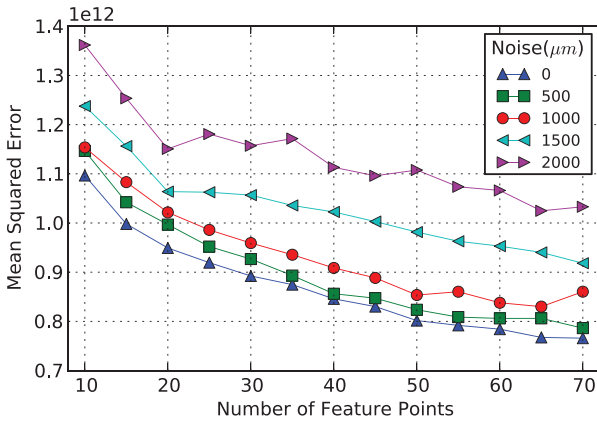
Fig. 3. Shape reconstruction error from 2D feature points, averaged over all subjects. For each number of feature points, the experiment is repeated 20 times with a random subset. The values shown in the figures are mean values.
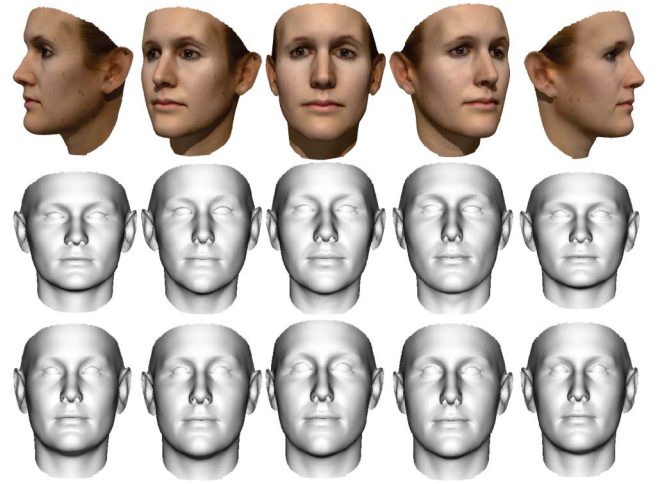

Fig. 5. Fitting results for subject 323 in five pose angles. The top row shows renderings. The second row shows fitting results for the comparison method. The last row shows fitting results for the proposed method.

effects: 1) It affects 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 Fig. 4 (bottom), we show performance as a function of pose angle. We also extract the estimated pose from the camera matrix and show the accuracy of our pose estimate in Fig. 4 (top). A qualitative comparison for one subject in nine pose angles can be seen in Fig. 5. Note that the BFM renderings exhibit significant perspective distortion which is not modeled by our affine camera. In our supplementary material, available online, 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 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 2 (left) shows shape recognition rates in comparison to a computationally more expensive reference method.

5.1.3 Texture

We evaluate texture reconstruction accuracy in terms of two error measures:

$$\mathbb{E}_g = \frac{1}{n} \left\| \mathbf{t}_g - \mathbf{t}_r \right\|^2 \quad \text{and} \quad \mathbb{E}_m = \frac{\left\| \mathbf{t}_g - \mathbf{t}_r \right\|^2}{\left\| \mathbf{t}_m - \mathbf{t}_r \right\|^2}, \qquad (13)$$

where $\mathbf{t}_r$ is the recovered texture, $\mathbf{t}_g$ the ground-truth texture, and $\mathbf{t}_m$ the ground-truth data projected into the 60 parameter model (i.e., the optimal model fit to the data). $\mathbb{E}_g$ is an absolute error measure on the texture values which lie in the interval $[0, 1]$. $\mathbb{E}_m$ is a relative error measure with respect to the best reconstruction possible by the model. A value of 1 would represent optimal performance. Table 3 shows performance as a function of pose angle, averaged over identity, and illumination. As for shape, we perform a recognition experiment using a single gallery texture parameter. Table 2 (right) shows recognition rates averaged over three illumination conditions and 10 subjects per pose.
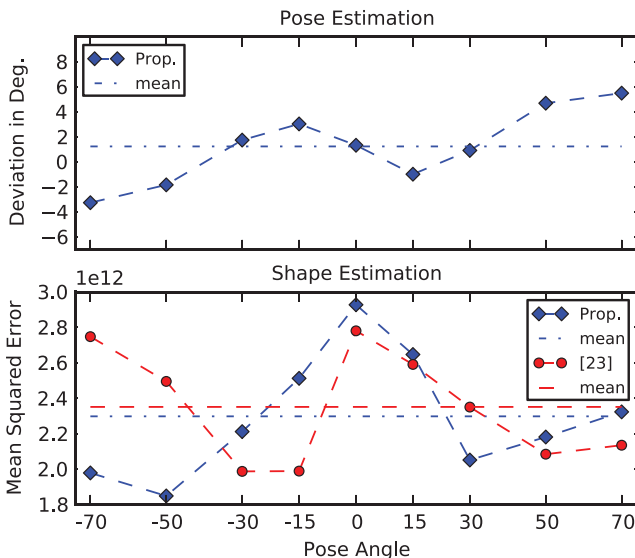

Fig. 4. The top figure shows recovered pose angles (minus ground truth) averaged over subjects. The bottom figure shows mean reconstruction errors for the proposed method and reference method [30], measured in μm^2.

TABLE 2
Mean Rank-1 Recognition Error Rates for All 270 Renderings Averaged over Three Illumination Conditions per Pose

	Shape		Texture	
Pose:	Prop.	[30]	M-2	[30]
$-70°$	96.7	87.8	92.0	81.0
$-50°$	100	93.6	94.8	92.0
$-30°$	100	94.4	94.9	89.9
$-15°$	100	91.6	98.4	91.7
$0°$	97.8	92.9	95.9	91.0
$15°$	98.9	90.7	94.9	88.1
$30°$	100	94.5	94.4	82.6
$50°$	100	96.3	96.0	84.7
$70°$	92.2	93.0	95.8	85.9
Mean	98.4	92.7	95.0	87.4

Results are shown for shape-only and texture-only recognition performance.

TABLE 3
Texture Reconstruction Errors as a Function of Pose Angle

Pose:	$\mathbb{E}_g \times 10^{-3}$			$\mathbb{E}_m$		
	M-2	M-1	[30]	M-2	M-1	[30]
$-70°$	5.6	8.3	5.9	5.2	7.9	5.4
$-50°$	5.7	6.9	6.8	5.3	6.5	6.3
$-30°$	5.8	6.9	11.8	5.4	6.6	11.1
$-15°$	5.9	5.4	9.4	5.4	5.2	8.8
$0°$	6.1	5.4	11.0	5.6	5.1	10.4
$15°$	6.7	7.1	10.9	6.1	6.8	10.5
$30°$	6.4	8.4	11.2	5.9	8.2	11.0
$50°$	6.6	7.5	8.2	6.0	7.2	8.2
$70°$	7.0	7.6	6.0	6.3	7.3	5.5
Mean	6.2	7.1	9.0	5.7	6.7	8.6

Errors are averaged over the 10 subjects and three illumination conditions.

The BFM renderings use a very simple illumination environment comprising a single directional and ambient source, both of white color. To demonstrate the performance of our method under complex illumination, we rendered images using environment lighting [23]. Results for two examples are shown in Fig. 6.

5.2 Real-World Images

Real-world face images are subject to outliers (e.g., background information, hair, or partial occlusions) and substantial variability of pose and illumination. To test the performance of our methods on such data, we use the CMU PIE database [33], which is comprised of 68 subjects in widely varying pose, illumination, and expression conditions. Moreover, in contrast to our synthetic data above, the images were obtained using a camera different from that used to obtain model data. We address this problem by estimating a linear color transformation between the model
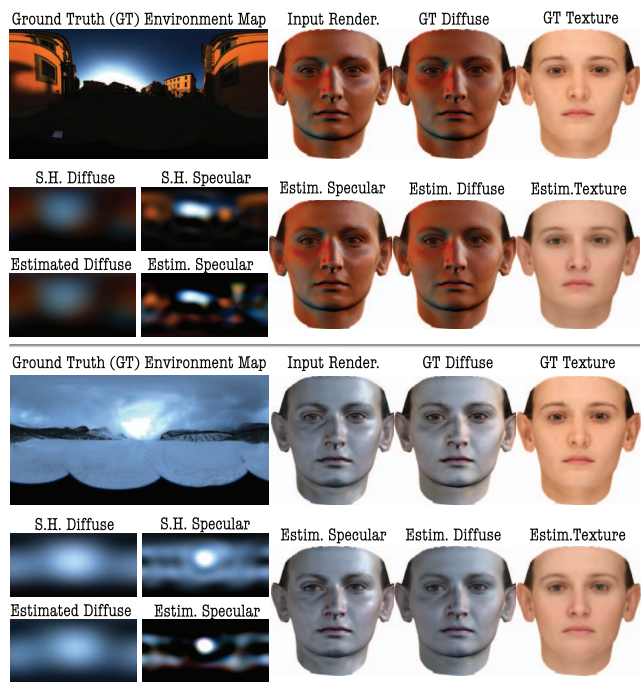

Fig. 6. Synthetic examples under environment illumination. "Input Rendering" serves as input to the algorithm. The second row shows reconstruction results using the proposed method. On the bottom left, estimated diffuse and specular environment maps are compared to ground truth.

and observations (see Section 4.2.2). The BFM does not model changes in expression. The experiments are therefore restricted to the expression neutral subset of the database, which nevertheless numbers 4,488 images.
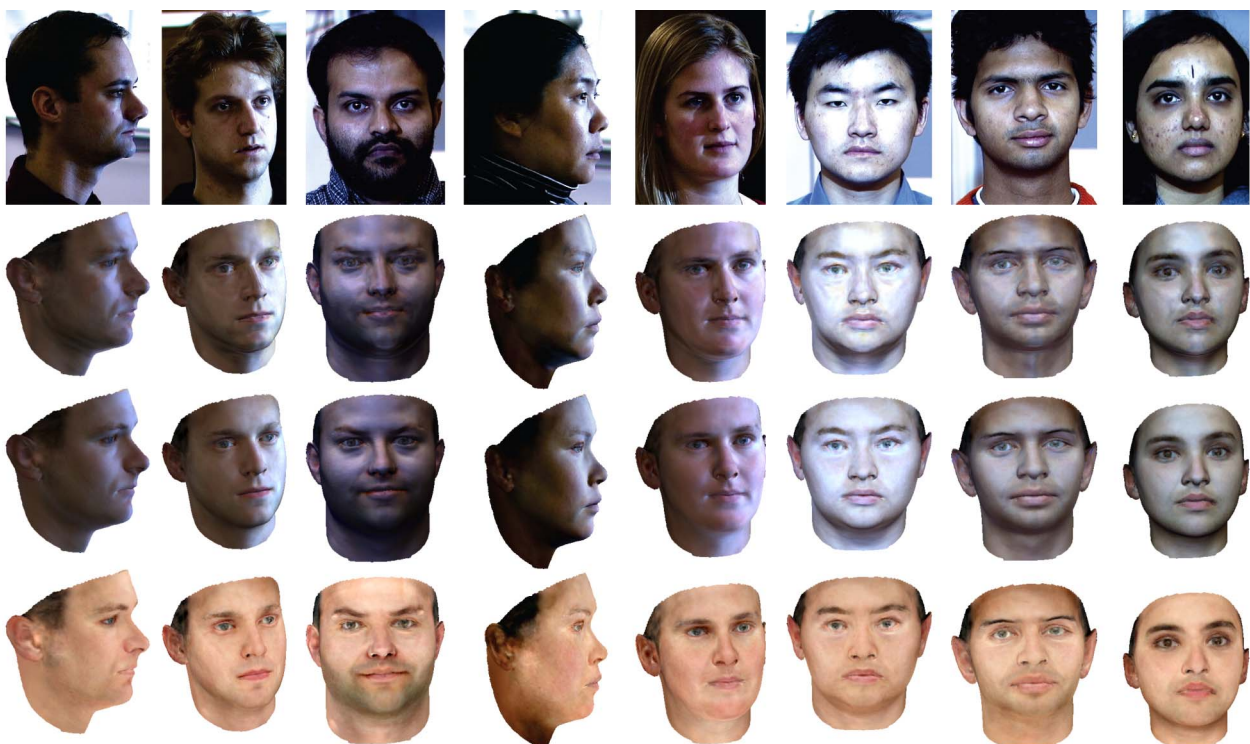

Fig. 7. Qualitative fitting results for eight randomly selected subjects, pose, and illumination conditions. Top row: Input images. Second row: Fitting result with all parameters estimated. Third row: Diffuse reconstruction. Last row: Texture only.
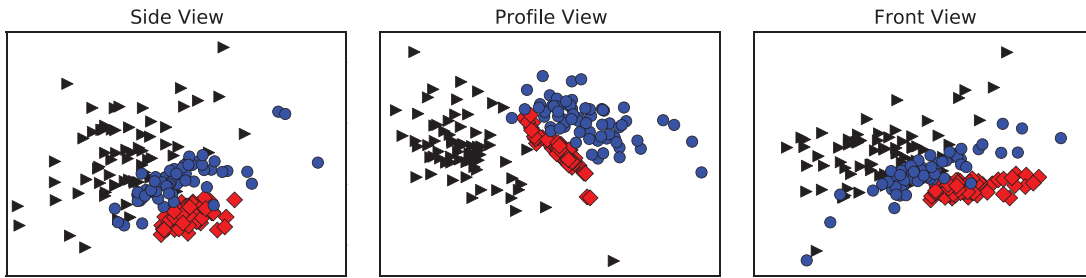
Fig. 8. Multidimensional scaling plots of irradiance parameters for three randomly selected illumination conditions. Results are shown for all 68 subjects of the database for side, profile, and front view, respectively.

5.2.1 Texture and Illumination Estimation

We begin by showing qualitative texture fitting results for a randomly selected set of images in Fig. 7. Note that specular reflections are successfully captured by the specular fit and that texture estimation is stable under a wide range of illumination conditions. The facial hair of the third subject is a failure case.

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. Fig. 8 shows the result of applying multidimensional scaling to the estimated illumination parameters. Results are shown for three randomly chosen lighting conditions for all 68 subjects in three pose angles. The three illumination conditions cluster well.

Fig. 9 shows two examples of cross illumination. For each subject, the last column shows a rendering using the illumination estimated from the input image in the same row and the texture estimated from the other input image. The fact that the images in the third and fourth columns are nearly identical shows that the relightings are stable even under dramatic changes in illumination.

5.2.2 Recognition

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 4.2. The BFM is a segmented model. In order to obtain high differentiability, we exploit this in our recognition experiment. We use the 99 most significant modes for the global model and each of the four segments, accounting for 495 texture coefficients in total. The same number of coefficients is used for the shape model. In order to handle outliers, we use a robust version of our texture fitting algorithm. We initially fit to the raw data, use a threshold to identify outlying observations, and then refit the model to inliers only.

Table 4 shows rank-1 recognition performance for our method and two reference methods. Overall, our method slightly outperforms that of Zhang and Samaras [38]. Performance is slightly worse than the Multifeature Fitting algorithm. However, our approach is computationally less expensive, requires very little parameter tuning, and converges to the globally optimal solution.

6 CONCLUSION AND FURTHER WORK

In this paper, we have proposed a complete framework for inverse rendering faces. By decomposing the image formation process into a geometric and photometric part, each can be solved as a multilinear system of equations. As opposed to previous methods, our approach is convex in each set of parameters and can be solved with a global optimum. Our shape fitting approach uses a sparse set of feature points and negates the need to empirically choose the weight between prior and data. We propose two texture fitting approaches which account for specular reflection and, the second of which allows for arbitrary, unknown illumination. This is the first morphable model

Fig. 9. Illumination transfer. First column: Input images of two subjects in same pose and different illumination condition. Second column: Estimated texture. Third column: Full model approximation. Last column: Estimated textures interchanged between illumination environments.

TABLE 4
Mean Recognition Rates for All 68 Subjects
of the CMU-PIE Database Averaged over 22 Illumination
Conditions per Pose Using a Single Gallery Image

Comparison of Recognition Results				
Gallery View	Probe View			
	front	side	profile	mean
Multi Feature Fitting algorithm [30] :				
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 [38] :				
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

fitting algorithm to account for both of these simultaneously. We obtain state-of-the-art results for both shape and texture using data with known ground truth. Our approach is efficient. Solving for the linear system for texture parameters is the asymptotically dominant step in a single iteration. This process has a complexity of $O(m^2 p)$, where m is the number of texture parameters and p the number of vertices. Hence, our algorithm is linear in the number of vertices.

The proposed methods make use of prior terms for texture and illumination. In this work, regularization weights were found empirically. In future work, we would like to examine ways to automatically tune these parameters. To reduce the influence of outliers when minimizing 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 refitting technique. The ultimate goal of inverse rendering is to model each contributing factor individually and independently. To further increase generalization ability, we would also like to include global illumination effects such as occlusions of the imaging environment.
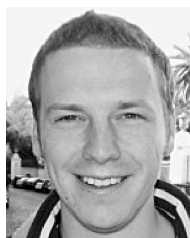
REFERENCES

[1] B. Amberg and T. Vetter, "Optimal Landmark Detection Using Shape Models and Branch and Bound," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 455-462, 2011.

[2] J.J. Atick, P.A. Griffin, and A.N. Redlich, "Statistical Approach to SFS: Reconstruction of 3D Face Surfaces from Single 2D Images," *Neural Computation,* vol. 8, no. 6, pp. 1321-1340, 1996.

[3] R. Basri and D.W. Jacobs, "Lambertian Reflectance and Linear Subspaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 2, pp. 218-233, Feb. 2003.

[4] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel, "A Statistical Method for Robust 3D Surface Reconstruction from Sparse Data," *Proc. Second Int'l Symp. 3D Data Processing, Visualization, and Transmission,* pp. 293-300, 2004.

[5] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis of 3D Faces," *Proc. ACM Siggraph,* pp. 187-194, 1999.

[6] V. Blanz and T. Vetter, "Face Recognition Based on Fitting a 3D Morphable Model," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 9, pp. 1063-1074, Sept. 2003.

[7] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," *Proc. European Conf. Computer Vision,* pp. 484-498, 1998.

[8] N.P. Costen, T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Automatic Extraction of the Face Identity-Subspace," *Image Vision Computing,* vol. 20, pp. 319-329, 2002.

[9] I. Craw and P. Cameron, "Parameterising Images for Recognition and Reconstruction," *Proc. British Machine Vision Conf.,* pp. 367-370, 1991.

[10] C. Creusot, N. Pears, and J. Austin, "3D Landmark Model Discovery from a Registered Set of Organic Shapes," *Proc. IEEE CVPR Workshop Point Cloud Processing,* 2012.

[11] 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 Trans. Medical Imaging,* vol. 21, no. 5, pp. 525-537, May 2001.

[12] L. Farkas, *Anthropometry of the Head and Face.* Raven Press, 1994.

[13] M. Fuchs, V. Blanz, H. Lensch, and H.-P. Seidel, "Reflectance from Images: A Model-Based Approach for Human Faces," *IEEE Trans. Visualization and Computer Graphics,* vol. 11, no. 3, pp. 296-305, May/June 2005.

[14] A. Georghiades, "Recovering 3-D Shape and Reflectance from a Small Number of Photographs," *Proc. Eurographics Symp. Rendering,* pp. 230-240, 2003.

[15] 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 Analysis and Machine Intelligence,* vol. 23, no. 6, pp. 643-660, June 2001.

[16] D.B. Goldman, B. Curless, A. Hertzmann, and S.M. Seitz, "Shape and Spatially-Varying BRDFs from Photometric Stereo," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[17] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision.* Cambridge Univ. Press, 2000.

[18] A. Hertzmann and S.M. Seitz, "Example-Based Photometric Stereo: Shape Reconstruction with General, Varying BRDFs," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1254-1264, Aug. 2005.

[19] T. Heseltine, N. Pears, and J. Austin, "Three-Dimensional Face Recognition Using Surface Space Combinations," *Proc. British Machine Vision Conf.,* 2004.

[20] K. Lee, J. Ho, and D. Kriegman, "Acquiring Linear Subspaces for Face Recognition under Variable Lighting," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 684-698, May 2005.

[21] S.R. Marschner and D.P. Greenberg, "Inverse Lighting for Photography," *Proc. Fifth Color Imaging Conf.,* pp. 262-265, 1997.

[22] S. Marschner, S. Westin, E. Lafortune, K. Torrance, and D. Greenberg, "Reflectance Measurements of Human Skin," Technical Report PCG-99-2, Cornell Univ., 1999.

[23] Univ. of Southern California "High-Resolution Light Probe Image Gallery," http://gl.ict.usc.edu/Data/HighResProbes, 2011.

[24] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D Face Model for Pose and Illumination Invariant Face Recognition," *Proc. IEEE Int'l Conf. Advanced Video and Signal Based Surveillance,* 2009.

[25] R. Ramamoorthi and P. Hanrahan, "A Signal-Processing Framework for Inverse Rendering," *Proc. ACM Siggraph,* pp. 117-128, 2001.

[26] R. Ramamoorthi, "Modeling Illumination Variation with Spherical Harmonics," *Face Processing: Advanced Modeling and Methods,* Academic Press, 2005.

[27] R. Ramamoorthi and P. Hanrahan, "A Signal-Processing Framework for Reflection," *ACM Trans. Graphics,* vol. 23, no. 4, pp. 1004-1042, 2004.

[28] S. Romdhani, V. Blanz, and T. Vetter, "Face Identification by Fitting a 3D Morphable Model Using Linear Shape and Texture Error Functions," *Proc. European Conf. Computer Vision,* pp. 3-19, 2002.

[29] S. Romdhani, J. Ho, T. Vetter, and D.J. Kriegman, "Face Recognition Using 3-D Models: Pose and Illumination," *Proc. IEEE,* vol. 94, no. 11, pp. 1977-1999, Nov. 2006.

[30] S. Romdhani and T. Vetter, "Estimating 3D Shape and Texture Using Pixel Intensity, Edges, Specular Highlights, Texture Constraints and a Prior," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 986-993, 2005.

[31] K. Sidorov, S. Richmond, and D. Marshall, "An Efficient Stochastic Approach to Groupwise Non-Rigid Image Registration," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2208-2213, June 2009.

[32] K.A. Sidorov, S. Richmond, and D. Marshall, "Efficient Groupwise Non-Rigid Registration of Textured Surfaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2401-2408, June 2011.

[33] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 12, pp. 1615-1618, Dec. 2003.

[34] W.A.P. Smith and E.R. Hancock, "Recovering Facial Shape Using a Statistical Model of Surface Normal Direction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 12, pp. 1914-1930, Dec. 2006.

[35] W.A.P. Smith and E.R. Hancock, "A New Framework for Grayscale and Colour Non-Lambertian Shape-from-Shading," *Proc. Asian Conf. Computer Vision,* pp. 869-880, 2007.

[36] K. Torrance and E. Sparrow, "Theory for Off-Specular Reflection from Roughened Surfaces," *J. Optical Soc. of Am.,* vol. 57, no. 9, pp. 1105-1114, 1967.

[37] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 586-591, 1991.

[38] L. Zhang and D. Samaras, "Face Recognition from a Single Training Image under Arbitrary Unknown Lighting Using Spherical Harmonics," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 3, pp. 351-363, Mar. 2006.

[39] S. Zhou and R. Chellappa, "Rank Constrained Recognition under Unknown Illuminations," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures,* pp. 11-18, 2003.

[40] T. Zickler, S. Mallick, D. Kriegman, and P. Belhumeur, "Color Subspaces as Photometric Invariants," *Int'l J. Computer Vision,* vol. 79, pp. 13-30, 2008.

Oswald Aldrian received the Dipl.-Ing (FH) degree in telecommunications from HTW Berlin in 2008 and the master's degree in electronic systems from Dublin City University in 2009. He is currently working toward the PhD degree at the University of York, United Kingdom, where he is a member of the Computer Vision and Pattern Recognition Group. His research interests include computer vision, statistical modeling, machine learning, and various fields of applied mathematics. He is a student member of the IEEE and BMVA.

William A.P. Smith received the BSc degree in computer science and the PhD degree in computer vision from the University of York, United Kingdom, in 2002 and 2007, respectively. He subsequently joined the Computer Vision and Pattern Recognition group there as a lecturer. His research interests are in face modeling, shape-from-shading, reflectance analysis, and the psychophysics of shape-from-X. He has published more than 70 papers in international conferences and journals, was awarded the Siemens best security paper prize at BMVC 2007, and was a finalist (UK nominee) for the ERCIM Cor Baayen award 2009. He is an associate editor of the IET journal *Computer Vision* and has served as cochair of the ACM International Symposium on Facial Analysis and Animation in 2010 and 2012 and the CVPR 2008 workshop on 3D Face Processing. He is a member of the IEEE and BMVA.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.