Automatic Construction Of Robust Spherical Harmonic Subspaces

Patrick Snape Yannis Panagakis Stefanos Zafeiriou
Imperial College London
{p.snape,i.panagakis,s.zafeiriou}@imperial.ac.uk

Abstract

In this paper we propose a method to automatically recover a class specific low dimensional spherical harmonic basis from a set of in-the-wild facial images. We combine existing techniques for uncalibrated photometric stereo and low rank matrix decompositions in order to robustly recover a combined model of shape and identity. We build this basis without aid from a 3D model and show how it can be combined with recent efficient sparse facial feature localisation techniques to recover dense 3D facial shape. Unlike previous works in the area, our method is very efficient and is an order of magnitude faster to train, taking only a few minutes to build a model with over 2000 *images. Furthermore, it can be used for real-time recovery of facial shape.*

1. Introduction

The recovery of 3D shape from images represents an ill-posed and challenging problem. In its most difficult form, this involves recovering a representation of shape for an object from a single image, under arbitrary illumination. However, for any given image, there are an infinite number of shape, illumination and reflectance inputs that can reproduce the image [1]. Therefore, shape recovery is commonly performed by relaxing the problem by introducing prior information or by adding constraints. The most impressive results have been achieved by restricting the problem space to a single class of objects such as faces. For example, Blanz and Vetter's 3D morphable model (3DMM) [7] is one of the most well-known shape recovery techniques and concentrates on the recovery of facial shape. 3DMMs constrain their reconstruction capabilities to lying within the span of a linear combination of faces. This allows for the synthesis of a large range of novel faces. However, the major drawback of 3DMMs is their complexity of construction. Morphable models require a set of high quality 3D meshes and associated textures. Currently, collecting these meshes is a time consuming and expensive process involving specialised hardware and manual guidance. Once the meshes have been collected, they must be placed into correspondence which is a complex research issue in its own right.

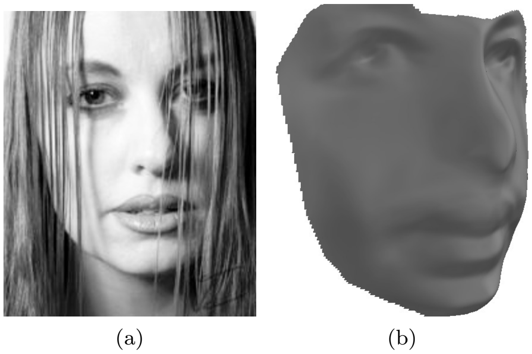

Figure 1: **An example reconstruction.** Given the input image (a) our algorithm can robustly recover dense 3D shape using only images.

In this paper, we look to borrow from ideas seen within the photometric stereo literature in order to recover shape from objects under unconstrained settings using *only a set of images*. Typically, these types of unconstrained photo collections are called "in-the-wild". We seek to construct our models in an automatic manner, without manual feature point placement or careful selection of the input images.

In particular, we seek to recover the shape of the object by exploiting the similarity within the object class. In the case of faces, there are millions of available images that can be utilised to build in-the-wild models. However, recovering shape from these images is incredibly challenging, as they have been captured in completely unconstrained conditions. No knowledge of the lighting conditions, the facial location or the camera geometric properties are provided with the images. To address these problems, we propose to recover a class specific spherical harmonic (SH) basis that exploits the low-rank structure of faces [5, 16]. Spherical harmonics are ideal for this purpose as they can be approximated by a low-dimensional linear subspace [5, 38]. By using the first order SH, 87.5% of the low-frequency component of the lighting is approximated. The first order SH can then be used to recover 3D shape as their discrete approximation directly incorporates the normals of the object. These normals can be integrated to provide a dense 3D surface [14].

Since we seek to recover a SH subspace, we require correspondence between our input images. This is achieved by locating a set of sparse features on the faces and then warping them into a single common reference frame. This method of achieving correspondence is powerful, as recent facial feature localisation techniques have incredibly low overhead [18, 39] and thus cause training to be efficient. The secondary benefit of this coarse alignment is that our basis can be coupled with existing facial alignment such as Active Appearance Models (AAM) [13, 34] in order to provide an appearance basis. We show that our recovered SH basis can be robustly learnt from automatically aligned, in-the-wild images. The basis can be used to recover both dense shape of generic faces and as a person specific appearance prior within AAM type algorithms.

Summarising, our contributions are:

1. We show the advantage of using a coarser alignment than optical flow for model construction. In particular, our training time for 2330 images from the HELEN dataset [25] is approximately **12 minutes**. We strongly believe that leveraging large numbers of images is important to build expressive models and thus training time is an important consideration.

2. A formal mathematical framework for performing efficient class specific uncalibrated photometric stereo using low-rank and sparsity constraints.

3. We show how our model can be coupled with existing facial alignment algorithms in order to provide low frequency dense shape for in-the-wild images.

2. Related Work

In the literature, there are many techniques that attempt to recover 3D facial shape from single images [7, 47, 30, 31, 28, 20, 44]. The most influential of these works was the 3D Morphable Model (3DMM) proposed in [7]. The 3DMM can produce very realistic reconstructions but has the disadvantage of having a complex model construction and fitting process. This reliance on accurate 3D meshes means that 3DMMs often suffer from an inability to recover complex facial attributes such as expression. Expression in dense 3D models has been addressed in the area of blendshapes [11, 10, 48], however these blendshapes are still complex to create as they require hundreds of meshes of individuals under varying expressions.

More general techniques for shape recovery such as the work of Barron *et al.* [3] do not perform well for inherently non-lambertian objects such as faces. However, shape-from-shading (SFS) has been shown to recover accurate facial shape by assuming a prior on the shape of faces [44, 20, 30, 29, 31, 17, 21]. In contrast to our proposal, SFS techniques rely on recovery of shape from a single image, whereas we consider large collections of images.

The most relevant techniques to this paper involve recovering shape from a collection of images under varying illumination. Typically, this involves solving some form of uncalibrated photometric stereo problem [4, 36, 35]. However, traditional uncalibrated photometric stereo techniques still assume that the images provided have been captured by a photometric stereo system under explicit directed lighting. The relaxation of the uncalibrated photometric stereo problem to a class of objects further increases the ambiguity inherent within the problem. Specifically, it is now necessary to separate the SH lighting from the identity of the individuals. This problem has been approached for both shape recovery and facial recognition purposes [27, 26, 30, 29, 51]. Lee *et al.* [27, 26] recover facial shape by separating illumination from identity in a manner that is similar to 3DMMs. Minsik *et al.* separate [30, 29] the appearance and identity via a low rank tensor decomposition that provides a very efficient reconstruction methodology. However, both Lee *et al.* and Minsik *et al.* still rely on previously built dense 3D models to perform their decomposition.

Recently, Kemelmacher-Shlizerman [19] proposed a method for building morphable models from images of faces downloaded from the Internet. This work shares similarities with ours in that it attempts to build a subspace that explicitly separates shape and appearance. However, in [19] they do not investigate a robust decomposition, but instead rely on a time consuming optical flow [22] based registration process to remove outliers from the images. Although this methodology allows for expression transfer, it does not allow the recovered shapes to be used within existing facial alignment techniques such as Active Appearance Models (AAMs). In contrast, our use of efficient facial alignment techniques to acquire correspondence substantially reduces our training time. It also allows our recovered basis to be coupled with the alignment techniques for simultaneous facial landmark localisation and dense surface recovery. However, the coarse geometric alignment we employ is more sensitive to corruptions such as occlusions and extreme facial pose. For this reason, we employ a low rank constraint [9, 37, 41, 12, 49, 32] to help remove these high frequency errors whilst maintaining the low frequency lighting variations. Although we share a similar optimisation framework to other robust principal component analysis problems such as [9, 37, 49, 32], we are the first to propose a low-rank decomposition that recovers a subspace of spherical harmonics.

3. Problem Formulation

In this section we describe how a spherical harmonic (SH) basis can be recovered using uncalibrated photometric stereo (PS) techniques. We then describe how this problem generalises to a multi-person dataset and how a representation of shape can be recovered per image. Finally, we discuss the importance of achieving correspondence between the images in an efficient and scalable manner.

3.1. Spherical Harmonic Bases

The lambertian reflectance model states that matte materials reflect light uniformly in all directions. This simple image formation model assumes that the intensity of light reflecting from a surface is a function of the shape of the surface and a linear combination of point light sources. More formally, given an image $I(x, y)$, the intensity at a given pixel (x, y) of a convex lambertian surface illuminated by a single light, can be expressed as

$$I(x, y) = \rho(x, y)\mathbf{l}^T\mathbf{n}(x, y), \tag{1}$$

where $\rho(x, y)$ is the albedo at the pixel and represents surface reflectivity, $\mathbf{l}$ is the vector denoting the single point light source illuminating the object and $\mathbf{n}(x, y)$ is the surface normal at the pixel.

If we now consider a collection of directional light sources placed at infinity, the lighting intensity at a given pixel can be expressed as a non-negative function of the unit sphere using a sum of spherical harmonics. Formally,

$$I(x, y) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \alpha_n \; \ell_{nm} \; \rho(x, y) \; Y_{nm}(\mathbf{n}(x, y)), \tag{2}$$

where $\alpha_n = \pi, 2\pi/3, \pi/4, \ldots,$ ℓ_{nm} are the coefficients of the harmonic expansion of the lighting and $Y_{nm}(\mathbf{n}(x, y))$ are the surface SH functions evaluated at the surface normal, $\mathbf{n}(x, y)$. As $n \to \infty$, the coefficients tend to zero, and thus the SH can be accurately represented by the lower order harmonics. In [15], it was shown that the first order SH function is guaranteed to represent at least 87.5% of the reflectance and experimentally verified to recover up to 95% in the case of faces. The first order SH expansion is also directly related to the objects surface normals:

$$\mathbf{Y}(\mathbf{n}(x, y)) = \rho(x, y)[1, \mathbf{n}_x(x, y), \mathbf{n}_y(x, y), \mathbf{n}_z(x, y)]^T, \tag{3}$$

where $\mathbf{n}_i(x, y)$ denotes the ith component of the normal vector. This is a particularly useful result as recovering the first order SH means directly recovering a representation of shape for an object.

3.2. Uncalibrated Photometric Stereo

Classical photometric stereo (PS) seeks to recover the normals of a convex object given a number of images under known different lighting with known directions. Traditionally, the following decomposition is performed

$$\mathbf{X} = \mathbf{N}\tilde{\mathbf{L}}, \tag{4}$$

where $\mathbf{X} \in \mathbb{R}^{d \times n}$ is the matrix of observations, and each of the n columns represents a vectorised image of the object with d total pixels, $\mathbf{N} \in \mathbb{R}^{d \times 3}$ contains the normal at every pixel and $\tilde{\mathbf{L}} \in \mathbb{R}^{3 \times n}$ is the matrix of lighting vectors per image. Assuming accurate light vectors and no shadowing artifacts, this problem is trivially solved as a linear least squares problem. Photometric stereo has been shown to provide accurate facial reconstructions despite faces not representing true lambertian objects. For example, there are many publicly available facial PS datasets such as the Photoface Database [50] and the Yale B [16] dataset.

If the lighting vectors are inaccurate or unknown, then PS is said to be uncalibrated. In [4], Basri *et al.* showed that $\mathbf{X}$ can be decomposed via a rank constrained singular value decomposition (SVD) to recover the SH bases and the lighting coefficients in the uncalibrated setting. First order SH are recovered by a rank 4 SVD and are accurate up to a 4×4 generalised Lorentz transformation. By enforcing constraints such as the integrability constraint [14], the first order SH, and thus the normals, can be recovered up to a generalised bas relief ambiguity (GBR). Formally, uncalibrated PS looks to recover

$$\mathbf{X} = \mathbf{B}\mathbf{L}, \tag{5}$$

where $\mathbf{X}$ is as before, $\mathbf{B} \in \mathbb{R}^{d \times 4}$ contains the first order SH basis images and $\mathbf{L} \in \mathbb{R}^{4 \times n}$ is the matrix of lighting coefficients. As previously mentioned, the solution to this problem is found by performing an SVD, $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, $\mathbf{B} = \mathbf{U}\sqrt{\mathbf{\Sigma}}$ and $\mathbf{L} = \sqrt{\mathbf{\Sigma}}\mathbf{V}^T$. Uncalibrated PS is useful as it is not always possible to recover accurate lighting estimations for every image.

3.3. Class Specific Uncalibrated Photometric Stereo

A generalisation of the uncalibrated PS problem for a specific class involves recovering a joint basis of appearance and illumination. In the case of SH for faces, this means attempting to separate the identity of the individual from their surface normals. This problem is a classic example of a bilinear decomposition problem and has been previously studied for use in 3D surface recovery [51, 30, 28, 27, 19]. In the case of SH, we seek to recover a low dimensional linear subspace that can recover normals for multiple individuals. This subspace implies that a face can be accurately reconstructed using a linear combination of basis shapes. This assumption is commonly employed in algorithms such as the 3DMM and AAMs. Assuming that we want to recover k such components for our shape subspace, and that we are using the first order SH, we will recover a $d \times 4k$ basis matrix that allows us to recover 3D facial shape for multiple individuals. Formally,

$$\mathbf{X} = \mathbf{B}(\mathbf{L} * \mathbf{C}), \tag{6}$$

where $\mathbf{B} \in \mathbb{R}^{d \times 4k}$ is the linear basis, $\mathbf{L} \in \mathbb{R}^{4 \times n}$ is the matrix of first order SH lighting coefficients, $\mathbf{C} \in \mathbb{R}^{k \times n}$ is the matrix of shape coefficients and $(\cdot * \cdot)$ denotes the Khatri-Rao product[23]. In fact, this is the exact decomposition problem solved by Kemelmacher-Shlizerman in [19] where they denote the combined coefficients matrix as $\mathbf{P} = \mathbf{L} * \mathbf{C}$. This was partially recognised by Zhou *et al.* [51], however they recover the lighting and shape coefficient separately by iteratively solving for each in an alternating fashion. Zhou *et al.* also do not provide any examples of the quality of the shape estimate that they recover.

Minsik *et al.* [30, 28] also attempt this decomposition by posing the problem in the form of a tensor. The decomposition can then be solved by applying a multilinear SVD. However, multilinear SVD requires a tensor representation and thus these techniques require prior data to recover results. A tensor representation is useful, however, for illustrating how to recover the $d \times 4$ first order SH for an individual, given their coefficients vector $\mathbf{c}_i \in \mathbb{R}^{k \times 1}$. We reshape the basis matrix $\mathbf{B}$ as a tensor which we denote $\mathbf{S} \in \mathbb{R}^{d \times k \times 4}$. The tensor product along the second mode, $\mathbf{S} \times_2 \mathbf{c}_i$, recovers the person specific shape of the ith column of $\mathbf{X}$. To recover $\mathbf{B}$ from $\mathbf{S}$, we perform matricisation of $\mathbf{S}$ along the first mode, denoted $\mathbf{S}_{(1)}$, to yield $\mathbf{S}_{(1)} = \mathbf{B} \in \mathbb{R}^{d \times 4k}$.

The problem given in (6) can now be solved within an optimisation framework, which we examine in detail in the next section.

3.4. Robust Construction Of Spherical Harmonic Bases

Inspired by recent advances in robust low-rank subspace recovery [9], we seek to modify Equation 6 to include new constraints that impose robustness. As mentioned previously, faces can be accurately reconstructed by a linear combination of faces taken from a low-dimensional basis. Therefore, we propose to decompose the image matrix into a low-rank part ($\mathbf{A}$) capturing the low frequency shape information and a sparse part ($\mathbf{E}$) accounting for gross but sparse noise such as partial occlusions and pixel corruptions. To promote low-rank and sparsity the nuclear norm (denote by $\|\cdot\|_*$) and the ℓ_1-norm (denote by $\|\cdot\|_1$) are employed, respectively. Formally we propose to solve the following non-convex optimisation problem:

$$\begin{aligned} &\underset{\mathbf{A},\mathbf{E},\mathbf{B},\mathbf{L},\mathbf{C}}{\operatorname{argmin}} && \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 + \frac{\mu}{2}\|\mathbf{A} - \mathbf{B}(\mathbf{L} * \mathbf{C})\|_F^2 \\ &\text{subject to} && \mathbf{X} = \mathbf{A} + \mathbf{E},\ \mathbf{B}^T\mathbf{B} = \mathbf{I}. \end{aligned} \tag{7}$$

Although the above problem is non-convex, an accurate solution can be obtained by employing the Alternating Directions Method (ADM) [6]. That is, to minimise the following augmented Lagrangian function:

$$\begin{aligned} &\mathcal{L}(\mathbf{A},\mathbf{E},\mathbf{B},\mathbf{C},\mathbf{L},\mathbf{Y}) = \\ &\quad \|\mathbf{A}\|_* + \lambda\|\mathbf{E}\|_1 + \frac{\mu}{2}\|\mathbf{A} - \mathbf{B}(\mathbf{L} * \mathbf{C})\|_F^2 + \\ &\quad \operatorname{tr}\left(\mathbf{Y}^T(\mathbf{X} - \mathbf{A} - \mathbf{E})\right) + \frac{\mu}{2}\|\mathbf{X} - \mathbf{A} - \mathbf{E}\|_F^2, \end{aligned} \tag{8}$$

with respect to $\mathbf{B}^T\mathbf{B} = \mathbf{I}$. Let t denote the iteration index. Given $\mathbf{A}_{[t]}$, $\mathbf{E}_{[t]}$, $\mathbf{B}_{[t]}$, $\mathbf{C}_{[t]}$, $\mathbf{L}_{[t]}$, $\mathbf{Y}_{[t]}$ and $\mu_{[t]}$, the iteration of ADM for Equation 7 reads:

$$\begin{aligned} \mathbf{A}_{[t+1]} &= \underset{\mathbf{A}_{[t]}}{\operatorname{argmin}} \mathcal{L}(\mathbf{A}_{[t]},\mathbf{E}_{[t]},\mathbf{B}_{[t]},\mathbf{C}_{[t]},\mathbf{L}_{[t]},\mathbf{Y}_{[t]}) \\ &= \|\mathbf{A}_{[t]}\|_* + \frac{\mu_{[t]}}{2}\Bigg(\|\mathbf{A}_{[t]} - \mathbf{B}_{[t]}(\mathbf{L}_{[t]} * \mathbf{C}_{[t]})\|_F^2 + \\ &\qquad \|\mathbf{X} - \mathbf{A}_{[t]} - \mathbf{E}_{[t]} + \frac{\mathbf{Y}_{[t]}}{\mu_{[t]}}\|_F^2\Bigg), \end{aligned} \tag{9}$$

$$\begin{aligned} \mathbf{E}_{[t+1]} &= \underset{\mathbf{E}_{[t]}}{\operatorname{argmin}} \lambda\|\mathbf{E}_{[t]}\|_1 + \\ &\frac{\mu_{[t]}}{2}\|\mathbf{X} - \mathbf{A}_{[t+1]} - \mathbf{E}_{[t]} + \frac{\mathbf{Y}_{[t]}}{\mu_{[t]}}\|_F^2, \end{aligned} \tag{10}$$

$$\mathbf{B}_{[t+1]} = \underset{\mathbf{B}_{[t]}^T\mathbf{B}_{[t]}=\mathbf{I}}{\operatorname{argmin}} \frac{\mu_{[t]}}{2}\|\mathbf{A}_{[t+1]} - \mathbf{B}_{[t]}(\mathbf{L}_{[t]} * \mathbf{C}_{[t]})\|_F^2, \tag{11}$$

$$\begin{aligned} &\left[\mathbf{L}_{[t+1]}, \mathbf{C}_{[t+1]}\right] = \\ &\quad \underset{\mathbf{L}_{[t]},\mathbf{C}_{[t]}}{\operatorname{argmin}} \frac{\mu_{[t]}}{2}\|\mathbf{A}_{[t+1]} - \mathbf{B}_{[t+1]}(\mathbf{L}_{[t]} * \mathbf{C}_{[t]})\|_F^2. \end{aligned} \tag{12}$$

Subproblem (9) admits a closed-form solution, given by the singular value thresholding (SVT)[8] operator as:

$$\mathbf{A}_{[t+1]} = D_{\mu_{[t]}^{-1}}\left[\mathbf{M}_{[t]} - \mathbf{A}_{[t]} + \mathbf{X} - \mathbf{E}_{[t]} + \frac{\mathbf{Y}_{[t]}}{\mu_{[t]}}\right], \quad (13)$$

where $\mathbf{M}_{[t]} = \mathbf{B}_{[t]}(\mathbf{L}_{[t]} * \mathbf{C}_{[t]})$ is introduced for brevity of the equation and the SVT is defined as $D_\tau(\mathbf{Q}) = \mathbf{U}\mathbf{S}_\tau\mathbf{V}^T$ for any matrix $\mathbf{Q}$ with SVD: $\mathbf{Q} = \mathbf{U}\mathbf{S}\mathbf{V}^T$. Subproblem (10) has a unique solution that is obtained via the elementwise shrinkage operator [9]. The shrinkage operator is defined as $\mathcal{S}_\tau[q] = \text{sgn}(q)\max(|q| - \tau, 0)$. Therefore, the solution of (10) is

$$\mathbf{E}_{[t+1]} = \mathcal{S}_{\lambda\mu_{[t]}^{-1}}\left[\mathbf{X} - \mathbf{A}_{[t+1]} + \frac{\mathbf{Y}_{[t]}}{\mu_{[t]}}\right]. \quad (14)$$

Subproblem (11) is a reduced rank Procrustes Rotation problem [52]. Its solution is given by $\mathbf{B}_{[t]} = \mathbf{U}\mathbf{V}^\top$ with

$$\mathbf{A}_{[t+1]}\left(\mathbf{L}_{[t]} * \mathbf{C}_{[t]}\right)^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \quad (15)$$

being the SVD of $\mathbf{A}_{[t+1]}\left(\mathbf{L}_{[t]} * \mathbf{C}_{[t]}\right)^\top$. However, due the unitary invariance of the Frobenius norm, Equation 12 becomes

$$\underset{\mathbf{L}_{[t+1]},\mathbf{C}_{[t+1]}}{\text{argmin}} \|\mathbf{B}^T_{[t+1]}\mathbf{A}_{[t+1]} - \mathbf{L}_{[t]} * \mathbf{C}_{[t]}\|_F^2. \quad (16)$$

Subproblem (12, 16) is a least squares factorisation of a Khatri-Rao product [40], which is solved as follows: Let $\mathbf{Q} = \mathbf{B}_{[t+1]}{}^\top\mathbf{A}_{[t+1]}, \mathbf{L} = \mathbf{L}_{[t]}$, and $\mathbf{C} = \mathbf{C}_{[t]}$. Furthermore, let $\mathbf{q}_i, l_i$, and c_i be the i_{th} columns of matrices $\mathbf{Q}, \mathbf{L}$, and $\mathbf{C}$, respectively. Clearly $\mathbf{q}_i = \mathbf{l}_i \oplus \mathbf{c}_i$, where $\oplus$ denotes the Kronecker product. For each column of $\mathbf{Q}$: Reshape $\mathbf{q}_i$ into a matrix $\widetilde{\mathbf{Q}}_i \in \mathbb{R}^{l\times k\times N}$ such that $\text{vec}\left(\widetilde{\mathbf{Q}}_i\right) = \mathbf{q}_i$. Obviously, $\widetilde{\mathbf{Q}}_i = \mathbf{c}_i \cdot \mathbf{l}_i^\top$ is a rank-one matrix. Compute the SVD of $\widetilde{\mathbf{Q}}_i$ as $\widetilde{\mathbf{Q}}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\top$. The best rank-one approximation of $\widetilde{\mathbf{Q}}_i$ is obtained by truncating the SVD as: $\mathbf{l}_i = \mathbf{u}_i\sqrt{\sigma_1}$ and $\mathbf{c}_i = \sqrt{\sigma_1}\mathbf{v}_i$, where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the first column vectors of $\mathbf{U}_i$ and $\mathbf{V}_i$, respectively, and σ_1 is the largest singular value. The ADM for solving (7) is outlined in Algorithm 1.

It is important to note that there are inherent ambiguities in this decomposition, both from the SVD to recover $\mathbf{B}$ and in the Khatri-Rao factorisation to recover $\mathbf{L}$ and $\mathbf{C}$. In particular, we are most concerned about how they may affect the recovered normals before we integrate them to recover depth. In order to resolve these ambiguities, we take the simplest possible approach, we recover the ambiguity matrix from a template set of normals provided by a known mean face.

Algorithm 1 Solving (7) by the ADM method.

Input: Data Matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and parameter λ.

Output: Matrices $\mathbf{A}$, $\mathbf{E}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{L}$.

```
Initialise: $\mathbf{A}_{[0]} = 0$, $\mathbf{E}_{[0]} = 0$, $\mathbf{B}_{[0]} = 0$, $\mathbf{C}_{[0]} = 0$, $\mathbf{L}_{[0]} = 0$, $\mathbf{Y}_{[0]} = 0$, $\mu_{[0]} = 10^{-6}$, $\rho = 1.1$, $\epsilon = 10^{-8}$
while not converged do do
    Fix $\mathbf{E}_{[t]}$, $\mathbf{B}_{[t]}$, $\mathbf{C}_{[t]}$, $\mathbf{L}_{[t]}$ and update $\mathbf{A}_{[t+1]}$ by
        $\mathbf{A}_{[t+1]} = D_{\mu_{[t]}^{-1}}\left[\mathbf{B}_{[t]}(\mathbf{L}_{[t]} * \mathbf{C}_{[t]}) - \mathbf{A}_{[t]} + \mathbf{X} - \mathbf{E}_{[t]} + \frac{\mathbf{Y}_{[t]}}{\mu_{[t]}}\right]$   (17)
    Fix $\mathbf{A}_{[t+1]}$, $\mathbf{L}_{[t]}$, $\mathbf{C}_{[t]}$, $\mathbf{L}_{[t]}$ and update $\mathbf{A}_{[t+1]}$ by
        $\mathbf{E}_{[t+1]} = \mathcal{S}_{\lambda\mu_{[t]}{}^{-1}}\left[\mathbf{X} - \mathbf{A}_{[t+1]} + \frac{\mathbf{Y}_{[t]}}{\mu_{[t]}}\right]$   (18)
    Update $\mathbf{B}_{[t+1]}$ by first performing the SVD on:
        $\mathbf{A}_{[t+1]}(\mathbf{L}_{[t]} * \mathbf{C}_{[t]})^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}$, $\mathbf{B}_{[t+1]} = \mathbf{U}\mathbf{V}^T$   (19)
    Update $[\mathbf{L}_{[t+1]}, \mathbf{C}_{[t+1]}]$ via a Least Squares Khatri-Rao factorization, as described in Section 3.4
    Update Lagrange multipliers by
        $\mathbf{Y}_{[t+1]} = \mathbf{Y}_{[t]} + \mu_{[t]}\left(\mathbf{X} - \mathbf{A}_{[t+1]} - \mathbf{E}_{[t+1]}\right)$   (20)
    Update $\mu_{[t+1]}$ by $\mu_{[t+1]} = \min(\rho\mu_{[t]}, 10^6)$
    Check convergence condition
        $\|\mathbf{X} - \mathbf{A}_{[t+1]} - \mathbf{E}_{[t+1]}\|_\infty < \epsilon$,
        $\|\mathbf{A}_{[t+1]} - \mathbf{B}_{[t+1]} - (\mathbf{L}_{[t+1]} * \mathbf{C}_{[t+1]})\|_\infty < \epsilon$   (21)
    $t \leftarrow t + 1$
end while
```

3.5. Efficient Pixelwise Correspondence

In contrast to the related work of Kemelmacher-Shlizerman [19], we achieved pixelwise correspondence between our images by using existing, efficient sparse facial alignment algorithms. This has two distinct advantages. Firstly, recent facial alignment algorithms such as those by Ren *et al.* [39] and Kazemi *et al.* [18] can produce a very accurate set of sparse facial features in the order of a single millisecond. In contrast, the optical flow method cited in [19] takes multiple seconds even for a small image. This means that our training time is drastically reduced in comparison to [19]. Ideally, our technique would be able to scale to the magnitude of thousands of images, whereas the alignment of [19] would quickly become infeasible as the number of images increases. In fact, the optical flow step is run multiple times as the collection flow algorithm is used [22] which involves an iterative algorithm of rank 4 decompositions and repeated optical flow. Secondly, the use of a direct alignment to a single reference frame

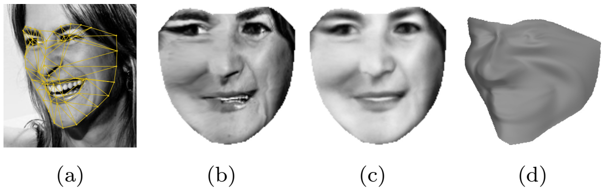

Figure 2: **Example of the low rank effect on warped pose**. (a) initial input image (b) input image after warping (c) warped image after the low rank constraint (d) recovered depth from (c).

enables the the usage of our basis in existing appearance based facial alignment algorithms such as AAMs. This means that our basis can be used to reconstruct dense 3D shape of faces directly from an existing AAM fitting provided the reference space of the AAM and our subspace is the same.

However, it is important to note that there are two potential drawbacks to our alignment technique. The alignment is based on a Piecewise Affine warping and is thus much coarser than the optical flow technique used in [19]. This is particularly amplified when larger poses are present in the input images. However, this is partly why the low rank component of our algorithm is so important. As Figure 2 shows, the robust decomposition of the basis helps correct these large global errors so that the shape subspace can be successfully recovered. Secondly, our technique does not contain a number of sub-clusters that can be used to warp expression onto our model. However, by using a large number of images that contain expression we directly include expression within our subspace. In [19], the recovered subspace will necessarily be devoid of expression as the global reference shape is neutral. This means that the subspace recovered by [19] will not be able to recover expressive 3D shape using efficient facial alignment algorithms.

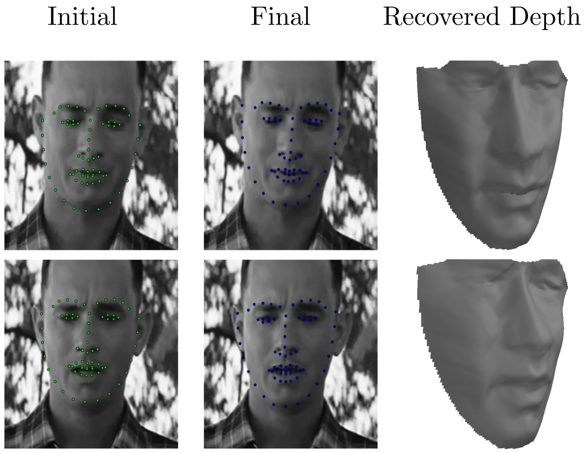

Figure 4: **Person specific model fitting for Tom Hanks**. Images of Tom Hanks coarsely aligned by a facial alignment method. Our algorithm improves the facial alignment and simultaneously recovers depth. Images shown are from a YouTube video of Tom Hanks.

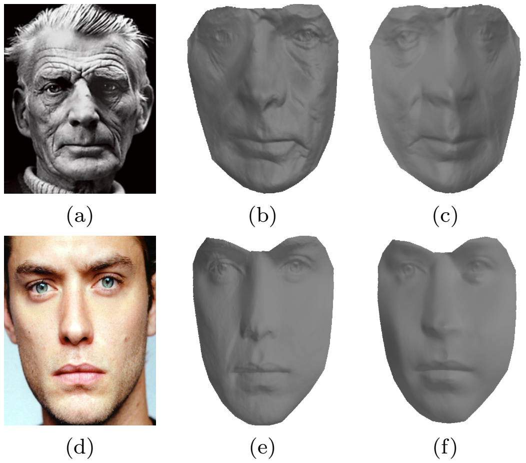

Figure 5: **Our subspace used for SFS**. Normals learnt automatically from the SH subspace of HELEN vs normals from the clean data of ICT-3DRFE. (b, e) the clean data (c, f) proposed subspace.

4. Experiments

In this section we provide a number of experiments that emphasise the increase in robustness of our reconstructions. We also show a new application to this type of model that involves improving the fitting results of an AAM using our constructed SH basis. Choosing the number of components, k, to recover is an important problem that was not properly addressed by Kemelmacher-Schlizerman in [19]. In these experiments we attempt to recover as many components as possible in order to strike a balance between cleanly reconstructed normals and identity. However, there is a trade-off when choosing the value of k. In particular, if the value of k is too large, then the decomposition is unable to separate the identity and shape and the subspace of shape no longer represents valid normals. This is one of the primary advantages of our robust decomposition, as it allows the value of k to be larger given the reduced rank of the images. However, a potential disadvantage of our proposed method is the sensitivity of the algorithm to the parameter λ, which must be tuned for every dataset. It is also important to stress that our main goal is to recover the low frequency shape information to provide plausible 3D facial surfaces under challenging conditions. However, in Section 4.3, we

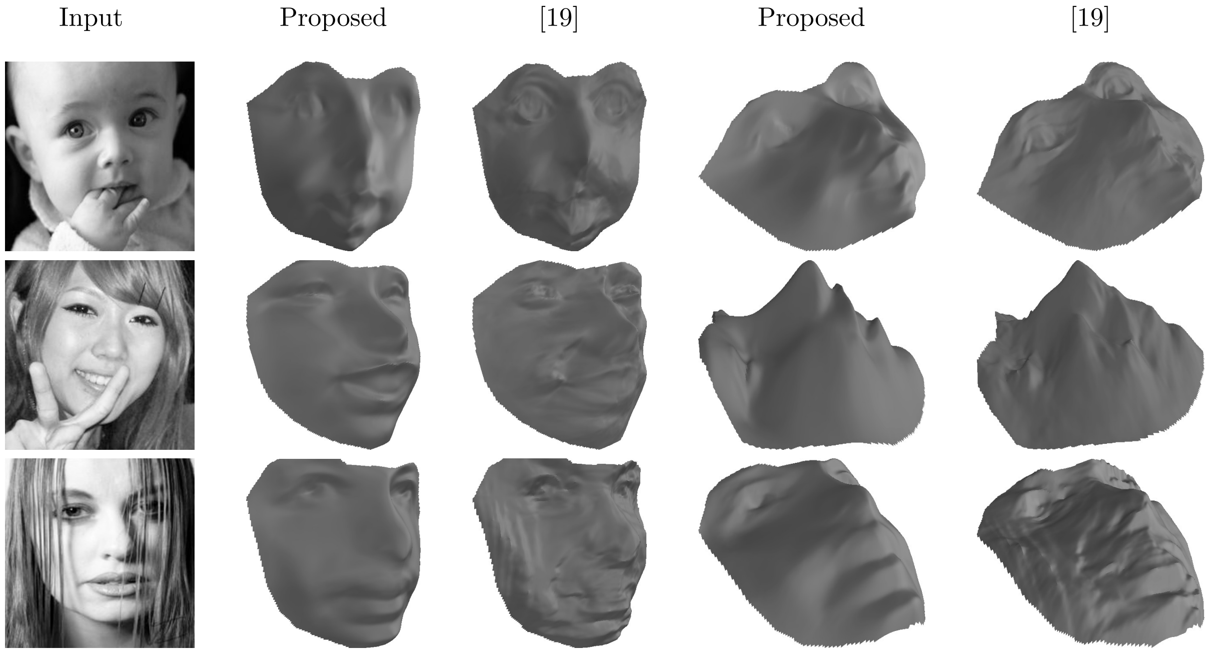

Figure 3: **Comparison with the blind decomposition of [19]**. Images from the HELEN[25] dataset.

show that our recovered subspace can be used in existing high frequency recovery algorithms such as SFS.

The area of 3D facial surface recovery is lacking any form of formal quantitative benchmark. The quantitative benchmark presented in [19] is performed on depth data recovered from photometric stereo. This is not ground truth depth data, as error is introduced during integration, and a more accurate evaluation would be the angular error of the recovered normals. However, in the presence of cast shadows, even the normals of photometric stereo are biased. For this reason, the lack of a standard and fair quantitative evaluation, we focus on qualitative results in this paper.

Specifically we performed the following experiments: (1) We built our subspace using the HELEN[25] dataset. We directly compare against the blind decomposition proposed in [19] and show particularly challenging images from the dataset. This experiment highlights the difficulty in constructing subspaces from large a set of in-the-wild images. (2) We show that the robust subspace learnt in (1) can be used within the shape-from-shading (SFS) framework of Smith *et al.* [44]. By recovering the normals from every image of HELEN, we can perform a secondary principal component analysis (PCA) on the normals in order to directly embed them within Smith's algorithm. In this experiment, we compare against a clean dataset of normals acquired from the ICT-3DRFE[46] database. (3) We show how our subspace can be combined with an existing facial alignment algorithm, namely project-out AAMs [34]. Our subspace can be used both as the appearance basis for the AAM and also as a methodology of recovering dense 3D shape.

In the following section we describe the construction of the bases and explain what processing was performed on each dataset.

4.1. Constructing The Robust Bases

The process of building the robust SH basis was the same for all datasets involved. Facial annotations consisting of 68 points were recovered through various methods for each dataset. In the case of the HELEN database, the manual annotations provided by the IBUG group were used [42, 43], in the case of the Yale B, Photoface and ICT-3DRFE databases, manual annotations were used and the in-the-wild images and video of Tom Hanks were automatically annotated by the one millisecond facial alignment method of [18] provided by the Dlib project [24].

These annotations were then warped via a Piecewise Affine transformation to a mean reference shape that was built from all the faces, training and testing, of the LFPW facial annotations provided by IBUG. This provided the dense correspondence required for performing matrix decompositions. To construct our SH bases, we performed the algorithm as described in Section 3.4 on the warped images. In order to provide the example reconstructions, the reconstructed images were warped back into their original shapes and then integrated using the method of Frankot and Chellappa

[14].

Table 1 gives examples of the training time taken for the in-the-wild Tom Hanks images and the HELEN dataset. It is important to note that part of the reason the training time is much lower for the Tom Hanks images is that they have an inherently lower rank than the HELEN images as they are all of the same individual. This greatly affects the convergence time and thus the timings do not scale linearly.

Data	W1	Train	W2	Tot
HELEN (2330)	8	730	25	763
T. Hanks (274)	1	21	4	26

Table 1: **Training Times.** Mean training times in seconds over 10 runs rounded to the nearest second. 'W1' denotes warping to the LFPW reference frame of (150 × 150) pixels, 'W2' denotes warping back to the original images and 'Train' denotes the total training time of our method described in Section 3.4. Original images were larger than the reference, hence the increase from 'W1' to 'W2'. Timings were recorded on an Intel Xeon E5-1650 3.20GHz with 32GB of RAM.

4.2. Comparison Using HELEN

In this set of experiments we wished to convey two results: (1) that we are capable of quickly constructing our basis on a large number of in-the-wild images, (2) that the our robust formulation of the problem gives superior performance to the blind decomposition used by [19]. In this experiment, $k = 200$ and the total number of components was thus $4k = 800$. Figure 3 shows the results from this experiment. As we can clearly see, on challenging images the blind decomposition is unable to separate the appearance from the illumination and thus the recovered normals are unable to recovery accurate shape.

4.3. Using The Subspace In SFS

The SFS technique of Smith *et al.* [44] relies on a PCA basis constructed from normals of a single class of object. It then seeks to recover the high frequency normal information directly from the texture. In order to create the PCA required by [44], we recovered spherical harmonics for every image in the dataset using the proposed algorithm. We then computed Kernel-PCA [45] on the normals recovered from the HELEN images and supplied them to [44]. The lighting vector is also an input to the algorithm and we recover it by solving a least squares problem with the known normals.

In order to provide a comparison for our reconstruction, we created a clean normal subspace using the data from the ICT-3DRFE [46] database. This database is primarily use for image relighting purposes, however, they provide a very accurate set of normals of faces under a wide range of expressions. The results of this experiment are shown in Figure 5. Although our subspace did not provide reconstructions that are as visually accurate as the subspace from ICT-3DRFE, they were still able to successfully recover a plausible representation of the high frequency shading information.

4.4. Automatic Alignment

In this experiment we used the Active Template Model (ATM) provided by the Menpo project [2] in order to perform a project-out type algorithm to align images of Tom Hanks. This model is similar to the Lucas-Kanade [33] method but uses a point distribution model (PDM) in order to perform non-rigid alignment between the images. In particular, the template image is fixed during optimisation of the PDM, and we use our subspace to provide a texture representing an approximation of the diffuse component of the image. This is essentially identical to the procedure performed within a project-out AAM.

We used a person specific SH subspace that was built on images of Tom Hanks that were downloaded automatically from the Internet. In this case, the images were automatically aligned using the DLib implementation of [18]. For this experiment, $k = 30$ and thus the total number of components $4k = 120$. We downloaded 200 frames from a Youtube video of Tom Hanks[1] and attempted to automatically align them using our subspace and the ATM. The ATM was initialised using another fitting of [18] which was then iteratively improved. At each global iteration, we recovered a new set of diffuse textures for each frame and then performed a refitting of every frame. This caused the images to align over a sequence of iterations. We performed 10 such iterations. Figure 4 shows two example frames where the alignment was improved and dense shape was also recovered.

5. Conclusion

We have proposed a robust method for automatically constructing generalised spherical harmonic subspaces. In particular, we have shown that by using a common reference frame as defined in algorithms such as AAMs, we can efficiently build models that have applications in shape recovery and facial alignment.

[1] https://www.youtube.com/watch?v=nFvASiMTDz0 from 3:43

Acknowledgements

Patrick Snape is funded by a DTA from Imperial College London and by a Qualcomm Innovation Fellowship. Yannis Panagakis is funded by the ERC under the FP7 Marie Curie Intra-European Fellowship. Stefanos Zafeiriou is partially supported by the EPSRC project EP/J017787/1 (4D-FAB).

References

[1] E. H. Adelson and A. P. Pentland. The perception of shading and reflectance. *Perception as Bayesian inference*, pages 409–423, 1996.

[2] J. Alabort-i Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou. Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 679–682, New York, NY, USA, 2014. ACM.

[3] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE T-PAMI*, 2015.

[4] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *IJCV*, 72(3):239–257, 2006.

[5] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE T-PAMI*, 25(2):218–233, 2003.

[6] D. P. Bertsekas. *Constrained optimization and lagrange multiplier methods*. 1982.

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[8] J.-F. Cai, E. J. Candes, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[9] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[10] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *TOG*, 32(4):1, 2013.

[11] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20(3):413–425, 2014.

[12] X. Cheng, S. Sridharan, J. Saragih, and S. Lucey. Rank minimization across appearance and shape for aam ensemble fitting. In *ICCV*, pages 577–584. IEEE, 2013.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE T-PAMI*, 23(6):681–685, 2001.

[14] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE T-PAMI*, 10(4):439–451, 1988.

[15] D. Frolova, D. Simakov, and R. Basri. Accuracy of spherical harmonic approximations for images of lambertian objects under far and near lighting. In T. Pajdla and J. Matas, editors, *ECCV*, volume 3021 of *Lecture Notes in Computer Science*, pages 574–587. Springer Berlin Heidelberg, 2004.

[16] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE T-PAMI*, 23(6):643–660, 2001.

[17] T. Hassner. Viewing real-world faces in 3d. In *ICCV*, pages 3607–3614. IEEE, 2013.

[18] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874. IEEE, 2014.

[19] I. Kemelmacher-Shlizerman. Internet based morphable model. In *ICCV*, pages 3256–3263. IEEE, 2013.

[20] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE T-PAMI*, 33(2):394–405, 2011.

[21] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *CVPR*, pages 1746–1753. IEEE, 2011.

[22] I. Kemelmacher-Shlizerman and S. M. Seitz. Collection flow. In *CVPR*, pages 1792–1799. IEEE, 2012.

[23] C. G. Khatri and C. R. Rao. Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhya: The Indian Journal of Statistics*, 1968.

[24] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[25] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Image Analysis and Processing*, pages 679–692. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[26] J. Lee, R. Machiraju, B. Moghaddam, and H. Pfister. Estimation of 3d faces and illumination from single photographs using a bilinear illumination model. In *Eurographics Symposium on Rendering*. Eurographics Association, 2005.

[27] J. Lee, B. Moghaddam, H. Pfister, and R. Machiraju. A bilinear illumination model for robust face recognition. In *ICCV*, pages 1177–1184. IEEE, 2005.

[28] M. Lee and C. H. Choi. Fast facial shape recovery from a single image with general, unknown lighting by using tensor representation. *Pattern Recognition*, 44(7):1487–1496, 2011.

[29] M. Lee and C.-H. Choi. A robust real-time algorithm for facial shape recovery from a single image containing cast shadow under general, unknown lighting. *Pattern Recognition*, 46(1):38–44, 2013.

[30] M. Lee and C.-H. Choi. Real-time facial shape recovery from a single image under general, unknown lighting by rank relaxation. *CVIU*, 120:59–69, 2014.

[31] Z. Lei, Q. Bai, R. He, and S. Z. Li. Face shape recovery from a single image using cca mapping between tensor spaces. In *CVPR*, pages 1–7, 2008.

[32] F. Lu, Y. Matsushita, I. Sato, T. Okabe, and Y. Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *CVPR*, pages 1490–1497. IEEE, 2013.

[33] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence*, 1981.

[34] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

[35] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *CVPR*, pages 1474–1481. IEEE, 2013.

[36] T. Papadhimitri and P. Favaro. A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima. *IJCV*, 107(2):139–154, 2014.

[37] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE T-PAMI*, 34(11):2233–2246, 2012.

[38] R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *JOSA*, 18(10):2448–2459, 2001.

[39] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, pages 1685–1692. IEEE, 2014.

[40] F. Roemer and M. Haardt. Tensor-based channel estimation and iterative refinements for two-way relaying with multiple antennas and spatial reuse. *IEEE Transactions On Signal Processing*, 58(11):5720–5735, 2010.

[41] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Raps: Robust and efficient automatic construction of person-specific deformable models. In *CVPR*, pages 1789–1796. IEEE, 2014.

[42] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshops*, pages 397–403. IEEE.

[43] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR Workshops*, pages 896–903. IEEE, 2013.

[44] W. A. P. Smith and E. R. Hancock. Recovering facial shape using a statistical model of surface normal direction. *IEEE T-PAMI*, 28(12):1914–1930, 2006.

[45] P. Snape and S. Zafeiriou. Kernel-pca analysis of surface normals for shape-from-shading. In *CVPR*, pages 1059–1066. IEEE, 2014.

[46] G. Stratou, A. Ghosh, P. Debevec, and L.-P. Morency. Exploring the effect of illumination on automatic expression recognition using the ict-3drfe database. *Image and Vision Computing*, 30(10):728–737, 2012.

[47] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE T-PAMI*, 31(11):1968–1984, 2009.

[48] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. In *SIGGRAPH*, SIGGRAPH, pages 77:1–77:10, New York, NY, USA, 2011. ACM.

[49] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In R. Kimmel, R. Klette, and A. Sugimoto, editors, *ACCV*, volume 6494 of *Lecture Notes in Computer Science*, pages 703–717. Springer Berlin Heidelberg, 2011.

[50] S. Zafeiriou, G. A. Atkinson, M. F. Hansen, W. A. P. Smith, V. Argyriou, M. Petrou, M. L. Smith, and L. N. Smith. Face recognition and verification using photometric stereo: The photoface database and a comprehensive evaluation. *IEEE Information Forensics and Security*, 8(1):121–135, 2013.

[51] S. Zhou, G. Aggarwal, R. Chellappa, and D. Jacobs. Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition. *IEEE T-PAMI*, 29(2):230–245, 2007.

[52] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.