Face Normals "in-the-wild" using Fully Convolutional Networks

George Trigeorgis
Imperial College London
g.trigeorgis@imperial.ac.uk

Patrick Snape
Imperial College London
p.snape@imperial.ac.uk

Iasonas Kokkinos
University College London
i.kokkinos@cs.ucl.ac.uk

Stefanos Zafeiriou
Imperial College London
s.zafeiriou@imperial.ac.uk

Abstract

In this work we pursue a data-driven approach to the problem of estimating surface normals from a single intensity image, focusing in particular on human faces. We introduce new methods to exploit the currently available facial databases for dataset construction and tailor a deep convolutional neural network to the task of estimating facial surface normals 'in-the-wild'. We train a fully convolutional network that can accurately recover facial normals from images including a challenging variety of expressions and facial poses. We compare against state-of-the-art face Shape-from-Shading and 3D reconstruction techniques and show that the proposed network can recover substantially more accurate and realistic normals. Furthermore, in contrast to other existing face-specific surface recovery methods, we do not require the solving of an explicit alignment step due to the fully convolutional nature of our network.

1. Introduction

Facial surface reconstruction from a single image is a problem that has attracted considerable attention over the past 25 years. This is in part due to both the multitude of applications related to face recognition and facial expression analysis, as well as its tractability due to the desirable properties of the physical structure of the human face. In contrast to the difficulty of the general case, the recovery of 3D facial shape has been highly successful. Human faces have a number of qualities that are desirable for shape recovery: they are extremely homogeneous in configuration (all healthy human faces have two eyes, a nose and mouth in the same approximate location), convex, exhibit approximately Lambertian reflectance [54, 17, 61, 43], are largely captured from a single direction (frontal) and are deformable and mostly not self occluding. Furthermore, there exists a large amount of publicly available imagery of faces and human faces are of significant interest to a number of fields including entertainment, medicine, and psychology. The two main lines of research consist of *(a)* Shape from Shading (SfS) methods, which can also potentially employ a statistical face prior [71, 3, 65, 12, 55, 59, 56], or *(b)* building and fitting a 3D Morphable Model (3DMM) [8, 7, 1]. A

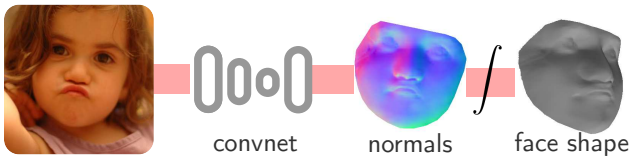

Figure 1: Depiction of our pipeline for 3D face shape estimation. Using a number of images of facial normals we train a fully convolutional network for normal estimation. Using the estimated normals we can retrieve the 3D face shape by classical normal integration techniques.

3DMM consists of a linear statistical model of the facial texture and surface which is learnt from a set of captured and well-aligned 3D facial scans. For many years, the only publicly available 3DMM was the Basel model [40], which was constructed from 200 Caucasian people displaying a neutral expression. Now, large-scale 3DMMs of neutral faces are available in LSFM [8] and expressive 3DMMs can be constructed by combining the statistical model of neutral faces with blendshapes [26, 9]. Nevertheless, fitting a 3DMM to single images requires solving a high-dimensional non-linear optimisation problem which is not only computationally demanding but also requires a near-optimal initialisation. Due to the difficulty of solving the original optimisation problem for 3DMMs, recent methods do not attempt to optimise the texture consistency term, but instead only fit the facial surface part of the 3DMM to a set of 2D facial landmarks [1, 26]. SfS [24], is the process of recovering surface by assuming that shading (*i.e.*, the intensity of a pixel in the image) is generated as a function of the

surface geometry and its interaction with light which is reflected/absorbed by the surface and captured by an imaging device. This function is generally modelled by the image irradiance equation:

$$I(x, y) \propto R(s_x(x, y), s_y(x, y)), \quad (1)$$

which states that the measured brightness of the image $I(x, y)$ is proportional to the radiance R at the corresponding point on the surface $s_x(x, y), s_y(x, y)$. The most commonly employed radiance function is the Lambertian function, which describes the measured brightness as being proportional to the cosine of the angle between the direction of the incident light and the surface normal. Explicitly, the Lambertian function describes the observed intensity at a single pixel as $I = \rho_d \boldsymbol{n}^\top \boldsymbol{s}$, where ρ_d is the albedo, $\boldsymbol{n}$ is the unit normal of the surface for the given pixel and $\boldsymbol{s}$ is a single unit point light placed at infinity. Although this is a relatively simple explanation for the potentially complex interaction between a surface and the light sources within an environment, it has been shown to describe up to 90% [68] of the low-frequency component of the lighting for images of a human face [5, 4, 68]. However, it is well known that shading alone is insufficient to disambiguate shape (*e.g.*, the well known bas-relief ambiguity [6]), hence generic SfS methods such as [3] are often suboptimal for more structured objects such as faces. Thus, statistical priors of facial surface normals have been utilised to constrain generic SfS methods in order to improve results. For example, generic methods such as that of Worthington et al. [65] have been extended by performing a linear projection of the recovered surface normals onto a constructed basis of facial normals [55, 56, 59]. Similarly, the work of Barron et al. [3] was extended to incorporate face specific priors by [35]. However, both of these methods required pre-built models in order to constrain their solutions. The current state-of-the-art SfS methods that do not require models [57, 28] combine ideas from uncalibrated photometric stereo [4] and low-rank tensor decompositions to robustly recover a combined model of shape and identity. Other methods have also explored the adaptation of fitted 3D templates with surface normals for more plausible surface recovery [45, 46, 30, 29]. However, the majority of these methods require an explicit alignment step in order to bring the facial model into correspondence with the facial image. Despite impressive advances in the area of facial alignment, this remains to be a challenging problem. Furthermore, dense alignment, as is required for the recovery of dense facial shape, is often achieved through highly expensive operations such as optical flow [28, 58]. Both 3DMMs and SfS are generative methods. In this paper, we take a different direction for estimation of the facial normals in unconstrained images and propose the first, to the best of our knowledge, discriminative deep learning methodology for the task of facial normal estimation. In particular, motivated by the success of deep learning to various tasks including object detection, dense semantic segmentation, and normal estimation of scenes [23, 2, 32, 62] etc., we propose to exploit the available large scale facial databases captured both in controlled, as well as in unconstrained conditions [8, 47] to train a fully convolutional deep network that maps image pixels to normals.

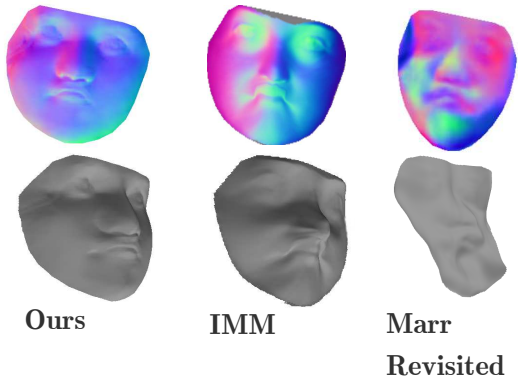

Figure 2: Facial surface normal estimation results from state-of-the-art techniques on the "in-the-wild" image of Fig. 1. Left to right: Proposed, IMM: state-of-the-art SfS technique [57], and generic state-of-the-art network [2].

More precisely, to acquire accurate ground truth of facial normals we synthesise images of faces created with the use of recently released Large-Scale 3D Facial Models (LSFM) [8] which contains facial shapes of individuals with diverse ethnicities and characteristics. To retrieve the 3D facial shape of the subject, we integrate the recovered normals using standard methods [14].

We provide experiments with multiple deep architectures using various loss functions appropriate for the task. We show that the proposed networks achieve state-of-the-art performance in estimation of facial normals in controlled conditions, as well as impressive reconstruction for very challenging "in-the-wild" facial images.

2. Prior work on Discriminative Surface Normal Estimation

Discriminative estimation of normals has recently received increased attention [2, 13, 62, 32, 44, 15]. One of the first methods was proposed in [70]. The training images were segmented using multiple unsupervised segmentation methods and then several dense features were extracted (*e.g.*, texton [38], SIFT [37], etc) and discriminative feature representations combining contextual and segment-based features were built. The ground truth normals were approximated by applying local feature coding by a weighted sum of representative normals and a discrim-

inative regressor (based on boosting) for these coefficients is trained. In the test phase, the likelihood of each representative normal was predicted by the classifier and the output normals were recovered as a weighted sum of representative normals. Richter and Roth [44] relax the requirement for external training data and instead use synthetic training data. The object silhouette is used to approximate an initial normal map, which is then used to approximate the object reflectance map in order to relight synthetic training data for the training the regressor.

One of the first methods that exploited the power of Deep Convolutional Neural Networks (DCNNs) for estimating the normals were proposed in [62]. The method in [62] used DCNNs to combine normal estimates from local and global scales, incorporating cues from room layout, edge labels and vanishing points. The method posed the surface normal regression problem as a classification one by applying the surface normal triangular coding technique from Ladicky et al. [70]. In particular, a codebook using *k*-means and a Delaunay triangulation was constructed over the words. Given this codebook and triangulation, a normal can be re-written as a weighted combination of the codewords in whose triangle it lies. At training-time, a softmax classifier is trained on the codewords. Recently, [15] used reliable surface normals reconstructed from multiview stereo as training data for a DCNN, which then predicts continuous normals from image intensity patches. This allows for object specific training and was shown to improve viewpoint specific reconstruction.

The first method that directly regresses to the surface normals was proposed in [13], which simultaneously trained a course-to-fine multi-scale DCNN for three tasks: depth prediction, surface normal estimation, and semantic labelling. The convolutional layers of the first scale (coarse level) were initialised by training on the object classification task over ImageNet [11]. The remaining network parameters for the mid- and fine levels were trained from scratch on the surface normal prediction task using NYU depth [50, 49]. The element-wise loss function used for surface estimation was the dot product between the ground-truth and the estimated surface normals.

Another regression based DCNN for normal estimation was proposed in [2]. Similar to [13], this method leverages the rich feature representation learnt by a DCNN trained on large-scale data tasks, such as object classification over ImageNet. The architecture combined a fully-convolutional architecture adapted from VGG-16 [52] with structures inspired by the hypercolumn representation [22]. The network was optimised using the ℓ_2-norm between ground-truth and estimated surface normals.

Most recently, another regression DCNN trained for surface estimation was proposed in [32]. This DCNN is a part of the so-called UberNet architecture which was proposed for jointly solving multiple image labelling tasks: such as detection of boundaries, saliency, semantic segmentation, human-parts prediction, surface normals recovery etc. The building block of Ubernet is VGG-16 [53]. For surface normal estimation, the ℓ_1-norm between ground-truth and the estimated surface normals was used.

All the above networks for surface normal estimation were trained on data samples displaying various indoor scenes [49, 50], hence, are likely sub-optimal for estimating the normals of human faces (please see Fig. 2). In this paper, we explore various DCNN architectures trained on facial databases for the task of facial surface normal estimation.

3. Databases of facial normals

Over the past two decades, the computer vision community has made considerable efforts to collect facial images for varying applications. Notable examples of early attempts include the FERET database [42] for face recognition and Cohn-Kadade database [27] for facial expression recognition. The interested reader may refer to [19] for a survey on face databases.

In this paper, we are interested in databases that can be used for training a DCNN for surface normal estimation. Ideally, we would use datasets that contain samples whose texture is captured in unconstrained conditions or whose texture is as close as possible to "in-the-wild" textures. Unfortunately, even with modern 3D capturing devices it is very difficult to acquire the 3D or 2.5D surface information from "in-the-wild" images. To mitigate this, we propose a learning strategy where we mix synthetic and real data for training the proposed network.

The databases appropriate for training our network are those that provide 3D surface scans, as well as databases captured under varying illuminations where the normals can be recovered using Photometric Stereo (PS) [64]. Currently, there are many databases that provide 3D facial scans, including FRGC [41], BU-3D [67], BU-4D [66] and BP4D-Spontaneous [72]. Nevertheless, collectively they do not contain more than 620 unique identities. Fortunately, a recent effort was made to collect a large database of faces and to build a large scale 3D Morphable Model (3DMM) [8]. In this paper, we use this database to generate a large amount of synthetic data.

The databases that contain samples captured under different illuminations include YALE-B [16], PIE [51] and MULTI-PIE [20], as well as the recently collected Photoface database [69]. The Photoface database [69] was collected using a custom-made four-source PS device designed to enable data capture with minimal interaction with people. The device was placed at the entrance to a busy workplace and captured many sessions from more than 450 people displaying various expressions. Each session comprises

four different images, under four different illuminants, from which the surface normal can be calculated using PS [64].

We also used the 3D Relightable Facial Expression database (ICT-3DRFE) [60] which contains 23 subjects and 15 expressions for a total of 345 images. The ICT-3DRFE dataset was acquired using a face scanning system that employs a spherical light stage with 156 white LED lights. This database can be used to synthesise high quality facial samples under different illuminations due to the separation of both specular and diffuse normals for each individual.

Finally, in order to incorporate the statistics of "in-the-wild" facial textures, we used the facial landmarks of the 300W data [48] to fit a 3DMM, following [26, 73]. We visually inspected the fittings and we kept those images for which the fitting was deemed acceptable.

In the remainder of this section we provide more details regarding how the data have been prepared with some visualisations of these data can be seen in Fig. 3.

3.1. Synthetic data generation from ICT-3DRFE

We generated synthetic data using the ICT-3DRFE database. The ICT-3DRFE dataset was captured using a high resolution face scanning system that employs a spherical light stage with 156 white LED lights. The lights are individually controllable in intensity and are used to light the face with a series of controlled spherical lighting conditions which reveal detailed shape and reflectance information. Linear polariser filters on the LED lights and an active polariser on the cameras allow specular and diffuse reflection to be recorded independently, yielding the diffuse and specular reflectance maps needed for photorealistic rendering under new lighting. We relit each sample under different random illuminations using the diffuse normals, as shown in Fig. 4.

3.2. Synthetic data generation using the LSFM 3DMM

As discussed above, the largest obstacle in solving the normal estimation problem for *in-the-wild* images is the lack of ground truth accurate normals in unconstrained scenarios. Although there are many databases suitable for normal recovery using Photometric Stereo (PS) [64], these lighting conditions are highly unrealistic. Also, the nature of PS capture set-ups is highly constrained and thus the variety in both identity and expression are low for these databases. For this reason, we constructed a large amount of synthetic data using rendered images. Specifically, we performed the following two steps *(1)* use a generative model of shape and texture to create a 3D instance of a face; *(2)* given this shape and a texture instance render it in a pseudo-photorealistic way on top of a randomly chosen scene.

The solution to *(1)* can be obtained by the use of three-dimensional statistical models of human facial shape and texture, known as 3D morphable models (3DMMs). A 3DMM is constructed by performing some form of dimensionality reduction, typically Principal Component Analysis (PCA), on a training set of 3D scans of faces that are in correspondence. Given this model, one can generate an infinite amount of realistic normals by synthesising a new instance $\boldsymbol{x}$ of the model. Specifically, choosing parameters from a normal distribution $\boldsymbol{c}_I \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and using the mean shape $\boldsymbol{\mu} \in \mathbb{R}^{3N}$ and weights of the model $\boldsymbol{W} \in \mathbb{R}^{3N \times k}$ we can synthesise a new instance $\boldsymbol{x} \in \mathbb{R}^{3N \times p}$,

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{W}_I \boldsymbol{c} \tag{2}$$

Booth *et al*. [8], provide a powerful 3DMM constructed with $9,663$ distinct subjects from a diverse set of demographics. Although this dataset is very diverse in terms of identity variation, it does not contain any diversity with respect to facial expression as all the subjects were captured in a neutral expression. To circumvent this, we use the expression bases created from the FaceWarehouse Database [9] to create a dual basis model of expression and identity, similar to [73],

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{W}_I \boldsymbol{c}_I + \boldsymbol{W}_E \boldsymbol{c}_E.$$

This process is further depicted in Fig. 5 where we detail the process that we used to generated the synthetic images of this dataset. As the true 3D facial structure is known, we obtain high quality ground truth normals for every synthesised image.

From the newly constructed mesh, we can retrieve the surface normals $\boldsymbol{n}$ at a vertex location $\boldsymbol{v} \in \mathbb{R}^3$ by the vector cross product of two edges of that the vertex's triangle,

$$\boldsymbol{n} = \frac{(\boldsymbol{v}^u - \boldsymbol{v}) \times (\boldsymbol{v}^v - \boldsymbol{v})}{\|(\boldsymbol{v}^u - \boldsymbol{v}) \times (\boldsymbol{v}^v - \boldsymbol{v})\|_2}$$

where $\boldsymbol{v}^u$ and $\boldsymbol{v}^v$ are vertices adjacent to $\boldsymbol{v}$ in the mesh structure along the positive horizontal and vertical directions.

A caveat with these generated samples is that a powerful regressor, as is the case with a large convolutional network, is that it can 'cheat' by taking into account various peculiarities of the synthesised samples such as the discontinuities between the face and the background or inaccurate lighting to learn to more easily recognise the pose and shape of the face. To account for this, we align these generated images to existing large-scale 2D datasets of "in-the-wild" images [47] in order to provide more realistic backgrounds. Each of these facial images contains a set of sparse annotations $\boldsymbol{s}_{2d} \in \mathbb{R}^{68 \times 2}$. Thus, we manually annotate the 3D mesh in the same manner in order to provide a set of 68 corresponding points $\boldsymbol{s}_{3d} \in \mathbb{R}^{68 \times 3}$ with the 2D images. Once we establish this correspondence, we can align the 3D shape to the image plane by employing a Perspective-n-Point (P-n-P) problem:

$$\boldsymbol{s}_{2d} = \boldsymbol{K}\boldsymbol{R}\boldsymbol{s}_{3d} + \boldsymbol{t} \tag{3}$$

Figure 3: From left to right: Photoface, ICT-3DRFE, 3D Morphable Models fitting, Synthesised image using a 3D Morphable model. Below are the associated ground truth normals for each dataset.

Figure 4: Relitting of the of the ICT-3DRFE dataset using the diffuse normals. On the left is the albedo texture, and on the right three examples of the relit texture.

where

$$\boldsymbol{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$$

is the matrix of intrinsic camera parameters, containing the focal length $\boldsymbol{f} \in \mathbb{R}^2$ and the principle point location $\boldsymbol{c} \in \mathbb{R}^2$. In this way we generate a supplementary 100 000 images of synthetic faces.

3.3. Synthetic data generation fitting a 3DMM

As has been previously mentioned, the constructed data using the 3DMM may not contain the desired facial texture of the "in-the-wild" images. To this end, we also fit the 3DMM to the "in-the-wild" images by employing the available sparse landmarks, similar to [1, 26, 73]. Specifically, to fit the 3DMM to the available images we employ the following optimisation problem

Figure 5: *1.* The generated shape and texture instance using the LSFM morphable model; *2.* Addition of expression using the FaceWarehouse expression basis; *3.* An image from the series Breaking Bad; *4.* The rendered aligned model.

$$\underset{\boldsymbol{c},\boldsymbol{R},\boldsymbol{t}}{\arg\min} \ \|\mathbb{P}(\boldsymbol{R}(\bar{\boldsymbol{s}} + \boldsymbol{U}\boldsymbol{c}) + \boldsymbol{t})) - \boldsymbol{s}_{2d}\|_F^2, \qquad (4)$$

where the goal is to recover the rotation $\boldsymbol{R}$, translation $\boldsymbol{t}$ and parameters $\boldsymbol{c}$ of the morphable model, under a weak-perspective projection $\mathbb{P}$. Beginning with just the mean 3D shape $\bar{\boldsymbol{s}}$, we optimise in an alternating manner first the pose parameters $\boldsymbol{R}, \boldsymbol{t}$ and then the shape model parameters $\boldsymbol{c}$.

Unfortunately, as shown in Fig. 3, the fittings do not accurately capture the identity of the person. However, these fittings can still be employed to regularise the optimisation problem. They ensure that the normals correctly capture the pose and expression of the subject.

3.4. Data from the Photoface Database

The last database we used was the Photoface database [69]. In the Photoface database, each ses-

sion contains four images captured under a different illumination. Examples of the Photoface are shown in Fig. 3. In order to estimate the normals from the images, we used the standard 4 source PS [64]. The standard PS assumes three or more grayscale images of a Lambertian object and constructs the following matrix equation:

$$\boldsymbol{I} = \boldsymbol{\rho} \odot \boldsymbol{N}\boldsymbol{L} \tag{5}$$

where $\boldsymbol{I} = [\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_N]$ is a $P \times N$ matrix containing irradiance values from all images, and P and N are the number of pixels and images respectively. Each row of $\boldsymbol{I}$ corresponds to a pixel position in an image, and each column corresponds to a different image. The albedo $\boldsymbol{\rho} \in \mathbb{R}^P$ combined with the normals matrix $\boldsymbol{N} \in \mathbb{R}^{P\times 3}$ represents the surface properties. The lighting matrix $\boldsymbol{L} = [\boldsymbol{l}_1, \ldots, \boldsymbol{l}_N] \in \mathbb{R}^{3\times F}$ represents the lighting directions and intensities, i.e., the j-th column of the matrix $\boldsymbol{L}$ corresponds to the lighting direction in the j-th image scaled by its intensity. Assuming that the light source vectors are known, we can solve a least squares version of the system in Eq. 5 for the albedo and the surface normal components at each pixel. Having available the albedo and normal information of a face, we can generate synthesised examples of the same subject by varying the light direction. We synthesised 3148 images by sampling random illuminations.

4. Model

As in [34, 36], we use a 'fully convolutional' network to extract an increasingly sophisticated hierarchy of features. Since the normal estimation task can benefit from both low and high level features, we use skip layers [21] that take intermediate layer activations as inputs and perform simple linear operations on them. In particular, we pool features from layers $\mathrm{conv1}, \mathrm{block2/unit_4}, \mathrm{block3/unit_6}, \mathrm{block4/unit_3}$ of the Resnet-50 [23] network. At each layer we learn linear mappings from the high-dimensional intermediate neuron activation space to the three-dimensional output space required for normal estimation.

We process these intermediate layers with batch normalisation [25] so as to bring the intermediate activations into a common scaling. As in [32] we keep the task-specific memory and computation budget low by applying linear operations within these skip layers, and fuse skip-layer results through additive fusion with learnt weights.

We appropriately place interpolation layers to ensure that results from different skip layers have commensurate dimensions, while, as in [39, 10], we use atrous convolution to increase the spatial resolution of high-level neurons. Finally, to account for the varying face sizes in the images we employ a 3-scale pyramid of our proposed network where at scales 2 & 3 we down-sample the image by half and a quarter times respectively by using a 2D average pooling operation, similar to [32]. The outputs of the different resolutions are combined through an additional fusion scheme that delivers the final normal estimates.

We consider two possible objective functions for the problem of surface normal regression. As our evaluation criterion is to minimise the angular distance between the network predictions $f(\boldsymbol{I})$ and the available ground truth normals $\boldsymbol{n}^*$ it is preferable to use the same loss function to train our fully convolutional network. To ensure that the resulting predictions are valid unit normal vectors we add a further ℓ_2 constraint, after of which we arrive at,

$$\begin{aligned} \mathcal{L}_{\text{cosine}} &= 1 - \sum_{i\in\mathcal{M}} f(\boldsymbol{I})_i^\top \boldsymbol{n}_i^* \\ \text{s.t. } & \|f(\boldsymbol{I})\|_2^2 = \|\boldsymbol{n}^*\|_2^2 = 1, \end{aligned}$$

where $\mathcal{M}$ is a mask containing the image indices corresponding to the visible face region. In addition to the cosine distance, we consider the smooth ℓ_1-loss [18] which was used in dense estimation tasks such as surface normal retrieval, segmentation [32] and object detection [18]. The ℓ_1-loss is regarded generally as a robust penaliser which helps to avoid the effect over over-smoothing the dense reconstructions [63]. To incorporate the smooth ℓ_1-loss we add again the ℓ_2 constrain for the network predictions,

$$\begin{aligned} \mathcal{L}_{\ell_1} &= \sum_{i\in\mathcal{M}} \text{smooth}_{L_1}(f(\boldsymbol{I})_i - \boldsymbol{n}_i^*) \\ \text{s.t. } & \|f(\boldsymbol{I})\|_2^2 = \|\boldsymbol{n}^*\|_2^2 = 1 \end{aligned}$$

5. Experiments

We conducted two sets of experiments. The first set is quantitative experiments on the Photoface database [69] where we consider as ground-truth the normals produced by the calibrated 4-source Photometric Stereo. For the purposes of this experiment we have withheld 100 subjects from the training set of all algorithms. Due to the lack of "in-the-wild" databases of normals, our second experiment is purely qualitative and includes images obtained from the Helen [33] and 300W [47] databases.

5.1. Experimental Setup

For learning the weights of the network we employ stochastic optimisation with Adam [31] with the default hyperparameters and one image per mini-batch. We use an initial learning rate of 0.001 with a polynomial decay rule, decreasing the learning rate by a factor of 10 every 10 000 iterations. To initialise the weights of the network we use the ImageNet-pretrained Resnet-50 model and initialise the weights of the new layers with random weights drawn from a Normal distribution.

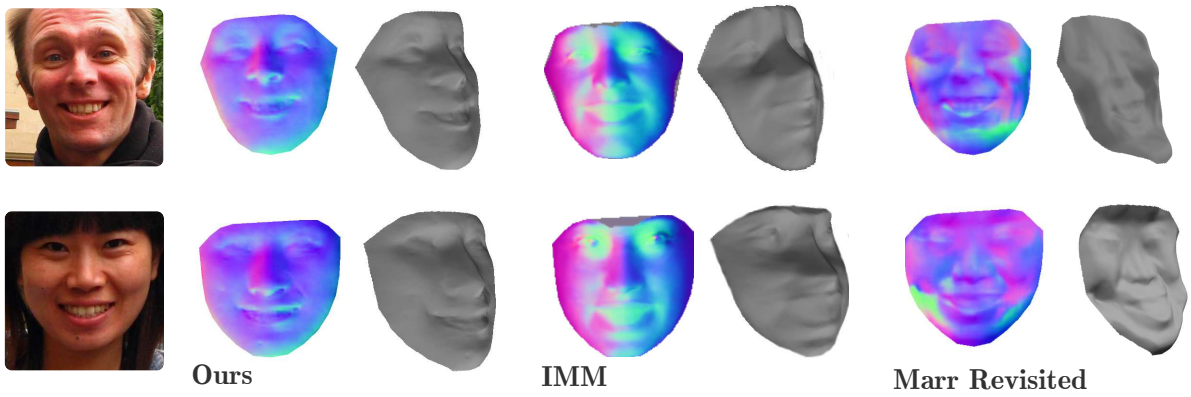

Figure 6: Example facial normal estimation and surface reconstruction from the Helen Dataset.

5.2. Experiments in Photoface

We compare with an array of state-of-the-art techniques for estimation of normals. Although we are concerned about the problem of surface normal estimation from a *single* image, we also provide experimental results for two well established techniques for SfS which require many images of the same subject under different illuminations as input [28, 57, 4]. First is Photometric Stereo with Unknown lighting (PS w/o Light), as porposed by [4] where we used the images from all four available illuinations to estimate the normals. Second, we applied the SfS method of [28, 57] (IMM). The method in [28, 57] reconstructs the facial normals from a collection of images of the same object. Hence, they have been applied on *all* the available data of Photoface to perform normal estimation. We applied the robust version of [28], proposed in [57], though the database does not contain occlusions and thus the results are expected to be very similar to [28]. We also compare against a landmark-driven fitting of the state-of-the art large scale 3DMM that we used for synthetic data generation in Section 3.2 (the model can describe both identity and expression variations). Finally, regarding state-of-the-art generic networks, we compare against the publicly available pre-trained networks [32, 2]. For all methods, we computed the angular error between the ground-truth and the estimated surface normals.

The results are summarised in Tab. 1. The proposed network has the best performance and achieves the lowest angular error. It is worth noting that the average performance of 3DMM fitting is good because it can capture general facial characteristics, but there are far fewer pixels with errors below 20° as 3DMMs lack the ability to capture high-frequency details of the facial surface. It is also worth noting that our method does not require an explicit alignment step, in comparison to both [28] and the landmark driven 3DMM estimation.

Table 1: Angular error for all the tested surface normal estimation methods. We show the results of the proposed network trained using the ℓ_1 loss.

Name	**Mean $\pm$ Std**	**$< 20°$**	**$< 25°$**	**$< 30°$**
PS w/o Light	42.9 ± 15.2	1.1%	13.1%	35.8%
IMM [28, 57]	24.2 ± 5.4	23.5	64.6%	88.3%
3DMM	26.3 ± 10.2	4.3%	56.05%	89.4%
Marr Rev. [2]	28.3 ± 10.1	31.8%	36.5%	44.4%
UberNet [32]	29.1 ± 11.5	30.8%	35.5%	55.2%
Proposed	$\mathbf{22.0} \pm 6.3$	36.63%	59.8%	79.6%

Loss	**Mean $\pm$ Std**	**$< 20°$**	**$< 25°$**	**$< 30°$**
Cosine Loss	$\mathbf{21.5} \pm 6.9$	29.9%	55.9%	81.5%
Smooth ℓ_1 Loss	22.0 ± 6.3	36.63%	59.8%	79.6%

Table 2: Angular error for the different loss functions.

Architecture	**Mean $\pm$ Std**	**$< 20°$**	**$< 25°$**	**$< 30°$**
Resnet + Cosine	$\mathbf{21.5} \pm 6.9$	29.9%	55.9%	81.5%
Pixelnet + Cosine	23.5 ± 6.3	35.17%	58.0%	78.2%

Table 3: Angular error for the different architectures.

Finally, we performed a series of experiments in order to evaluate the effect (a) of the loss function for the task (i.e., ℓ_1 vs cosine distance) and (b) of the network architecture (i.e., Resnet vs the PixelNet, which is based on VGG [2]). The experiments are summarised in Tables 2 and 3. As can be seen there is small difference between the performance of the two losses but the cosine distance is slightly better. Furthermore, the proposed architecture produces better results than PixelNet trained on exactly the same data and using the same loss function.

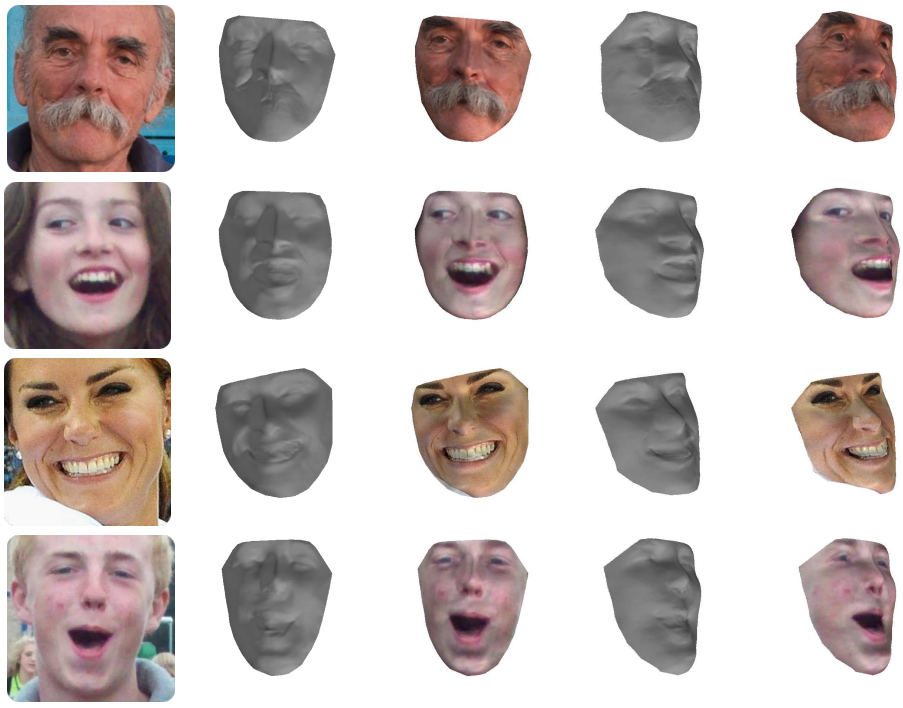

Figure 7: Representative surface reconstruction results from the challenging 300W dataset of "in-the-wild" facial images. The network generalises well to a diverse set of individuals and expressions. On the left is the original image from the 300W dataset. Next is the 3D shape reconstruction and the sampled texture from the image onto the shape.

5.3. Experiments "in-the-wild" databases

Since, there is no ground-truth for "in-the-wild" images, we can show only qualitative examples. For these experiments, we used the data provided by the 300W facial landmark localisation challenge[47, 48]. The methods we compare against are the robust version of the Internet Morphable Model (IMM) [28] proposed in [57] and, as before, a landmark-based fitting of the large scale 3DMM. The IMM reconstructs a collection of images, hence we have used 3000 "in-the-wild" facial images (the reconstruction process takes around 20 minutes). Fig. 6 shows some representative reconstruction cases of the proposed network versus IMM and the surface normal estimation network in [2]. For all surface reconstructions from normals we used the standard Frankot-Chellappa method [14].

It is evident that the proposed network provides very high-quality facial normals, even in images captured in very challenging recording conditions. Visual comparison versus the 3DMM are provided in the supplementary materials, since although the 3DMM can recover the pose and the expression, up to a certain extent, it cannot capture the fine-grained details. Finally, Fig. 7 shows more facial surfaces reconstructed by the proposed network.

6. Conclusions

We have presented the first, to the best of our knowledge, discriminative methodology tailored to facial surface estimation "in-the-wild". To this end, we capitalised on both the available facial database, as well as on the power of deep convolutional neural networks (DCNNs). We proposed methodologies for preparing training data for the task. We show that the proposed DCNN outperforms both the state-of-the-art facial surface normal estimation techniques, as well as the state-of-the-art pre-trained networks for normal estimation.

7. Acknowledgements

G. Trigeorgis was supported by EPSRC DTA award at Imperial College London. The work of P. Snape was partially funded by an EPSRC DTA and by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 688520 (TeSLA). S. Zafeiriou was partially funded by EPSRC Project EP/N007743/1 (FACER2VM). I. Kokkinos was supported by EU Horizon 2020 Project 643666 I-Support. We thank the NVIDIA Corporation for donating a Tesla K40 GPU used in this work.

References

[1] O. Aldrian and W. A. Smith. A linear approach of 3d face shape and texture recovery using a 3d morphable model. In *BMVC*, 2010. 1, 5

[2] A. Bansal, B. Russell, and A. Gupta. Marr Revisited: 2D-3D Alignment Via Surface Normal Prediction. *CVPR*, 2016. 2, 3, 7, 8

[3] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *T-PAMI*, 37(8):1670–1687, 2015. 1, 2

[4] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *IJCV*, 72(3):239–257, 2007. 2, 7

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

[6] P. N. Belhumeur, D. J. Kriegman, and A. L. Yuille. The bas-relief ambiguity. In *CVPR*, pages 1060–1066. IEEE, 1997. 2

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 1

[8] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *CVPR*, 2016. 1, 2, 3, 4

[9] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *T-VCG*, 20(3):413–425, 2014. 1, 4

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016. 6

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 3

[12] J.-D. Durou, M. Falcone, and M. Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *CVIU*, 109(1):22–43, 2008. 1

[13] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015. 2, 3

[14] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *T-PAMI*, 10(4):439–451, 1988. 2, 8

[15] S. Galliani and K. Schindler. Just look at the image: viewpoint-specific surface normal prediction for improved multi-view reconstruction. In *CVPR*, 2016. 2, 3

[16] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *T-PAMI*, 23(6):643–660, 2001. 3

[17] A. S. Georghiades, P. N. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *T-PAMI*, 23(6):643–660, 2001. 1

[18] R. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448, 2015. 6

[19] R. Gross. Face databases. In *Handbook of face recognition*, pages 301–327. Springer, 2005. 3

[20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 3

[21] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. *arXiv preprint arXiv:1411.5752*, 2014. 6

[22] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, pages 447–456, 2015. 3

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 2, 6

[24] B. K. Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. Technical report, MIT Artificial Intelligence Laboratory, 1970. 1

[25] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *JMLR*, 2015. 6

[26] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *CVPR*, 2016. 1, 4, 5

[27] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *FG*, pages 46–53. IEEE, 2000. 3

[28] I. Kemelmacher-Shlizerman. Internet-based morphable model. *ICCV*, 2013. 2, 7, 8

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

[30] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, pages 1746–1753. IEEE, 2011. 2

[31] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[32] I. Kokkinos. Ubernet : Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. *CoRR*, abs/1609.02132, 2016. 2, 3, 6, 7

[33] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, pages 679–692. Springer, 2012. 6

[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[35] C. Li, K. Zhou, and S. Lin. Intrinsic face image decomposition with human face priors. In *ECCV*, pages 218–233. Springer, 2014. 2

[36] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 6

[37] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[38] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 43(1):7–27, 2001. 2

[39] G. Papandreou, I. Kokkinos, and P. Savalle. Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection. In *CVPR*, pages 390–399, 2015. 6

[40] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *AVSS*, pages 296–301. IEEE, 2009. 1

[41] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005. 3

[42] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998. 3

[43] R. Ramamoorthi. Analytic pca construction for theoretical analysis of lighting variability in images of a lambertian object. *T-PAMI*, 24(10):1322–1333, 2002. 1

[44] S. R. Richter and S. Roth. Discriminative shape from shading in uncalibrated illumination. In *CVPR*, pages 1128–1136, 2015. 2, 3

[45] J. Roth, Y. Tong, and X. Liu. Unconstrained 3D face reconstruction. In *CVPR*, pages 2606–2615, 2015. 2

[46] J. Roth, Y. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. CVPR, 2016. 2

[47] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 2, 4, 6, 8

[48] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark Localization Challenge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 397–403, 2013. 4, 8

[49] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *ICCV-W*, pages 601–608. IEEE, 2011. 3

[50] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012. 3

[51] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *FG*, pages 46–51. IEEE, 2002. 3

[52] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 3

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[54] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *JOPT*, 4(3):519–524, Mar. 1987. 1

[55] W. A. Smith and E. R. Hancock. Recovering facial shape using a statistical model of surface normal direction. *T-PAMI*, 28(12):1914–1930, 2006. 1, 2

[56] W. A. Smith and E. R. Hancock. Facial shape-from-shading and recognition using principal geodesic analysis and robust statistics. *IJCV*, 76(1):71–91, 2008. 1, 2

[57] P. Snape, Y. Panagakis, and S. Zafeiriou. Automatic construction of robust spherical harmonic subspaces. In *CVPR*, pages 91–100, 2015. 2, 7, 8

[58] P. Snape, A. Roussos, Y. Panagakis, and S. Zafeiriou. Face flow. In *CVPR*, pages 2993–3001, 2015. 2

[59] P. Snape and S. Zafeiriou. Kernel-pca analysis of surface normals for shape from shading. In *CVPR*, pages 1059–1066, June 2014. 1, 2

[60] G. Stratou, A. Ghosh, P. Debevec, and L.-P. Morency. Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In *FG*, pages 611–618. IEEE, 2011. 4

[61] M. Turk and A. Pentland. Eigenfaces for recognition. *COGNEURO*, 3(1):71–86, 1991. 1

[62] X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *CVPR*, pages 539–547, 2015. 2, 3

[63] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic huber-l1 optical flow. 6

[64] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139–191139, 1980. 3, 4, 6

[65] P. L. Worthington and E. R. Hancock. New constraints on data-closeness and needle map consistency for shape-from-shading. *T-PAMI*, 21(12):1250–1267, 1999. 1, 2

[66] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *FG*, pages 1–6. IEEE, 2008. 3

[67] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *FG*, pages 211–216. IEEE, 2006. 3

[68] A. L. Yuille, D. Snow, R. Epstein, and P. N. Belhumeur. Determining generative models of objects under varying illumination: Shape and albedo from multiple images using svd and integrability. *IJCV*, 35(3):203–222, 1999. 2

[69] S. Zafeiriou, M. Hansen, G. Atkinson, V. Argyriou, M. Petrou, M. Smith, and L. Smith. The photoface database. In *CVPR*, pages 132–139, June 2011. 3, 5, 6

[70] B. Zeisl, M. Pollefeys, et al. Discriminatively trained dense surface normal estimation. In *ECCV*, pages 468–484. Springer, 2014. 2, 3

[71] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah. Shape-from-shading: a survey. *T-PAMI*, 21(8):690–706, 1999. 1

[72] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, and M. Reale. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *IMAVIS*, 32(10):692–706, 2014. 3

[73] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 4, 5