Adaptive 3D Face Reconstruction from Unconstrained Photo Collections

Joseph Roth, Yiying Tong, and Xiaoming Liu
Department of Computer Science and Engineering, Michigan State University
{rothjos1, ytong, liuxm}@msu.edu

Abstract

Given a collection of "in-the-wild" face images captured under a variety of unknown pose, expression, and illumination conditions, this paper presents a method for reconstructing a 3D face surface model of an individual along with albedo information. Motivated by the success of recent face reconstruction techniques on large photo collections, we extend prior work to adapt to low quality photo collections with fewer images. We achieve this by fitting a 3D Morphable Model to form a personalized template and developing a novel photometric stereo formulation, under a coarse-to-fine scheme. Superior experimental results are reported on synthetic and real-world photo collections.

1. Introduction

Computer vision has had much interest in the long-standing problem of *3D surface reconstruction*, expanding from constrained desktop objects to in-the-wild images of large outdoor objects [1]. Face reconstruction [23, 28], the process of creating a detailed 3D model of a person's face, is important with applications in face recognition, video editing, avatar puppeteering, and more. For instance, accurate face models have been shown to significantly improve face recognition by allowing the rendering of a frontal-view face image with neutral expression [43], thereby suppressing intra-person variability. The face presents additional challenges than general surface reconstruction due to non-rigid deformations caused by expression variation.

For some, usually graphics, applications a highly detailed model may be reconstructed in a constrained scenario using depth scanners [26, 12], calibrated stereo images [6], stereo videos [7, 34] or even high-definition monocular videos [13, 9]. However, for other applications such as biometrics, it is important to work on unconstrained photos like those typical of online image searches or from surveillance cameras. These photo collections present additional challenges since no temporal information may be used, images are of low resolution and quality, and occlusions may exist.

Photometric stereo-based reconstruction methods have proven effective for unconstrained photo collections.

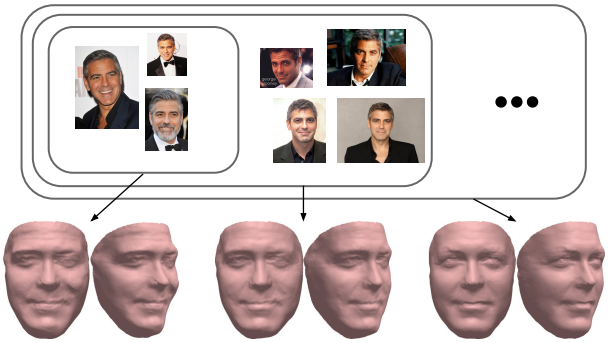


Figure 1. The proposed system reconstructs a detailed 3D face model of the individual, adapting to the number and quality of photos provided.

Beginning with Kemelmacher-Shlizerman and Seitz's work [23] which reconstructs a 2.5D depth map and extended by Roth *et al.* [28] to a full 3D mesh, photometric stereo-based approaches jointly estimate the surface normals, albedo, lighting conditions, and pose angles. Both techniques aim to identify a single representative face from the entire collection, which is challenging given the expression variation among images. By selecting a different consistent subset of images for each vertex on the face, the typical expression of the individual is used to drive the face reconstruction. However, there are still major limitations in photometric stereo-based reconstruction. One is that they require a sufficiently large collection of photos fo reconstruction. Theoretically, only four images are necessary if they are in perfect correspondence, but in practice the approaches use over one hundred images. Another is that the subset selection is binary and only makes use of $\sim$10% of the images for each vertex on the face.

Motivated by the success of the state of the art, we propose a novel adaptive photometric stereo-based reconstruction method from an unconstrained photo collection. Here, "adaptive" refers to the fact that our algorithm can handle a much wider range of photo collections, in terms of the number, resolution, and ethnicity of face images. Specifically, given a collection of unconstrained face images, we automatically detect faces and estimate 2D landmarks [37]. We then fit a 3D Morphable Model (3DMM) jointly to the

collection such that the projection of its annotated 3D landmarks are aligned with the 2D estimated landmarks [43] to create a personalized template. Each image has its pose estimated and is back-projected onto the personalized template to establish correspondence, and a dependability of each vertex is estimated to weight its influence in the reconstruction. The correspondence is used to jointly estimate the albedo, lighting conditions, and surface normals while the template is used to regularize the estimation. The template is then deformed to match the estimated surface normals and produce a reconstructed surface. A coarse-to-fine process is employed to first capture the generic shape and then fill in the details. To demonstrate the capabilities of the proposed approach, quantitative and qualitative experiments are performed on synthetic and in-the-wild photo collections, with comparison to the state of the art.

In summary, this paper makes three main contributions.

◇ A 3D Morphable Model is fit jointly to 2D landmarks for template personalization. Prior work used either a fixed template or landmark-based deformation that does not work well for small collections, with no prior face distribution.

◇ Photometric stereo is solved in a joint Lambertian image rendering formulation, with an adaptive template regularization that allows for graceful degradation to a small number of images. A dependability measure is proposed to weight the influence of images for face parts that are more confident to produce an accurate reconstruction.

◇ A coarse-to-fine reconstruction scheme is proposed to produce the similar quality reconstruction, with substantially lower computational cost.

2. Prior Work

We present a brief summary of relevant prior work on photometric stereo and face reconstruction.

Photometric stereo Classic photometric stereo estimates the surface normals of an object from a fixed camera orientation based on different light conditions. Photometric stereo was first proposed with knowledge of the light conditions [35] and even current methods still use this approach for cooperative subjects [16, 14]. Later it was discovered that even without knowledge of the light source photometric stereo can take advantage of the low rank nature of spherical harmonics [15, 40, 24, 4, 5, 36]. Most recent works can take multiple camera positions and put images into correspondence using Structure from Motion and even estimate arbitrary non-linear camera response maps [29]. Most photometric stereo techniques reconstruct from a common viewpoint and produce a 2.5D face surface which can only take advantage of frontal images. Photometric stereo usually uses SVD to find the low rank spherical harmonics, but then has to resolve an ambiguity using integrability or prior knowledge of the object. Such approaches require a sufficient number of images to obtain an accurate reconstruction, especially for non-rigid objects like the face where expression variation can disturb the low rank assumption. We propose using a personalized template to solve photometric stereo without using SVD, allowing the reconstruction to adapt to a small number of images.

Face reconstruction Face reconstruction creates a 3D face model from a set of input such as image(s), video, or depth data. It is a difficult problem with much recent interest and a variety of applications. In the biometrics community, pose, expression, and illumination are the main challenges of face recognition and all may be improved with accurate person-specific face models [43, 25, 41]. In graphics, high fidelity models with skeletal structures are useful for animations, puppeteering, and post processing videos. Face reconstruction began with cooperative subjects and expensive hardware where range scanners, multi-camera stereo [6, 7], or photometric stereo with known light arrays [16] can produce highly accurate models. There is recent interest from the graphics community in face reconstruction from videos [32, 13, 10, 30, 19, 9, 17] and even from RGB-D sequences [33]. But none of these techniques are directly comparable with ours since videos or special setups provide more information than unconstrained photo collections.

There are a series of recent works on reconstructing faces from photo collections [23, 28, 27]. The seminal work [23] creates a 2.5D model, locally consistent with the photo collection. It is extended in a few different directions, one in [38] where they use the surface normals from frontal faces to improve the fitting of a 3DMM, two in [22, 31] where the technique is used to generate a 3DMM, and three in [28] where the technique is expanded to handle pose variation and reconstructs a 3D model. Our work continues by improving the 3D reconstruction technique to adapt to lower-quality photo collections with fewer input images.

3. Algorithm

In this section, we present the details of the proposed approach and describe the motivational differences from prior art. We describe the basic preprocessing to obtain automatic landmark alignment. The main algorithm is broken down into three major steps. 1) Fit the 3DMM template to produce a coarse person-specific template mesh. 2) Estimate the surface normals of the individual using a photometric stereo (PS)-based approach. 3) Reconstruct a detailed surface using the estimated normals. Figure 2 provides an illustrated overview of the algorithm.

3.1. Photo Collection Preprocessing

A photo collection is a set of n images containing the face of an individual and may be obtained in a variety of ways, *e.g.*, a Google image search for a celebrity or a personal photo collection. The first step is to detect and crop faces from the images. We use the built-in face detection model from Bob [2] which was trained on various face

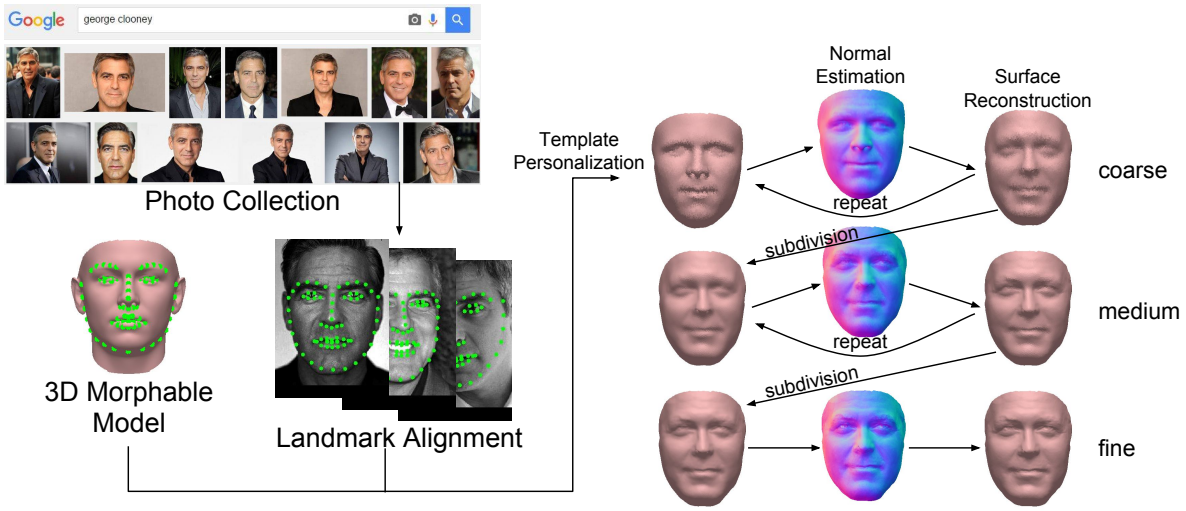


Figure 2. Overview of face reconstruction. Given a photo collection, we apply landmark alignment and use a 3DMM to create a personalized template. Then a coarse-to-fine process alternates between normal estimation and surface reconstruction.

datasets, such as CMU-PIE, that include profile view faces. The face detector is a cascade of Modified Census Transform (MCT) local binary patterns classifiers. Given the face bounding box, we convert the image to the intensity channel and crop outside of the face bounding box in order to ensure inclusion of the entire face. To estimate 2D landmarks, we employ the state-of-the-art cascade of regressors approach [37] to automatically fit 68 landmarks denoted as $\mathbf{W} \in \mathbb{R}^{2\times 68}$ onto each image.

3.2. Template Personalization

The initial template plays a vital role in the reconstruction process. Many aspects of the process depend upon the current template such as establishing correspondence across the photos, initial normal estimation during photometric recovery, and even Laplacian regularization during surface reconstruction. A good template should match the overall metric structure of the individual so that when it is projected onto photos of different poses, correspondence is established. Nevertheless, the template needs not contain fine facial details since those will be fleshed out by photometric normal estimation.

Prior work [28] used a single east Asian face mesh as a template, and employed landmark-based deformation to register the generic mesh to the person of interest. This technique was basically Structure from Motion (SFM) for the landmarks while the rest of the face was regularized by the curvature of the template mesh. The resultant template has two major limitations. One, the template has Asian influences that could potentially fit poorly to different ethnicities. Two, the SFM technique breaks down when fitting to a small number of photos with limited pose variations.

In light of these limitations, we propose to use a 3DMM instead of a single template mesh. The 3DMM is shown to accurately represent arbitrary face shapes based on a linear combination of scanned faces. Dense correspondence is established among the scans, and then [11] decomposes them into a set of bases for identity and another for expression.

$$\mathbf{X} = \bar{\mathbf{X}} + \sum_{k=1}^{199} \mathbf{X}_k^{\text{id}} \alpha_k^{\text{id}} + \sum_{k=1}^{29} \mathbf{X}_k^{\text{exp}} \alpha_k^{\text{exp}}, \qquad (1)$$

is the 3DMM composed of the mean shape $\bar{\mathbf{X}}$, a set of identity bases $\mathbf{X}^{\text{id}}$, and a set of expression bases $\mathbf{X}^{\text{exp}}$. $\mathbf{X} \in \mathbb{R}^{3\times p}$ is the 3D coordinates of p vertices in a triangulated mesh.

Typically, 3DMM fitting aims to minimize the difference between a rendered image and the observed photo [8], but recently, Zhu *et al*. propose an efficient fitting method based on landmark projection errors [43]. Our method extends [43] by jointly fitting the 3DMM to all n faces. To fit the 3DMM to a face image, we assume weak perspective projection $s\mathbf{R}\mathbf{X} + \mathbf{t}$, where s is the scale, $\mathbf{R}$ is the first two rows of a rotation matrix, and $\mathbf{t}$ is the translation on the image plane.

Given the 2D alignment results $\mathbf{W}$, the model parameters are estimated by minimizing the projection error of the landmarks that are labeled manually once onto the 3DMM,

$$\arg\min_{s,\mathbf{R},\mathbf{t},\alpha^{\text{id}},\alpha^{\text{exp}}} \|\mathbf{W} - (s\mathbf{R}[\mathbf{X}]_{\text{land}} + \mathbf{t})\|_F^2, \qquad (2)$$

where $[\mathbf{X}]_{\text{land}}$ selects the annotated landmarks from the entire model and $\|\cdot\|_F$ is the Frobenius norm. Furthermore, as the yaw angle increases, the 2D landmark alignment returns points along the contour or silhouette of the face, but the projected 3D landmarks would be obscured behind the cheek. [43] proposes a novel landmark marching technique where the 3D landmarks are moved along the surface to match the 2D silhouette under the current pose estimate.

We extend this process to jointly fit n faces of the same person by assuming a common set of identity coefficients

α^{id} but a unique set of expression α_i^{exp} and pose parameters per image. The error function then becomes,

$$\underset{s_i,\mathbf{R}_i,\mathbf{t}_i,\alpha^{\text{id}},\alpha_i^{\text{exp}}}{\arg\min} \sum_{i=1}^{n} \frac{1}{n} \|\mathbf{W}_i - (s_i\mathbf{R}_i[\bar{\mathbf{X}} + \sum_{k=1}^{199} \mathbf{X}_k^{\text{id}}\alpha_k^{\text{id}} + \sum_{k=1}^{29} \mathbf{X}_k^{\text{exp}}\alpha_{ki}^{\text{exp}}]_{\text{land}_i} + \mathbf{t}_i)\|_F^2, \quad (3)$$

where $[\cdot]_{\text{land}_i}$ is used because different poses of face images determine varying ranges of landmark marching, *i.e.*, different selections of vertices. This minimization is not jointly convex, but it can be solved by alternating estimation since it is linear with respect to each variable. Once the parameters are learned, we generate a personalized template $\mathbf{X}^0$ using the identity coefficients and the mean of the expression coefficients.

Model projection Correspondence between images in the collection is established based on the current template mesh $\mathbf{X}^0$. Given $\mathbf{X}^0$ and the projection parameters solved per image during model fitting, we sample the intensity of the projected location of vertex j in image i and place the intensity into a correspondence matrix $\mathbf{F} \in \mathbb{R}^{n\times p}$. That is, $f_{ij} = \mathbf{I}_i(u,v)$ where $\mathbf{I}_i$ is the ith image and $\langle u,v\rangle^\intercal = s_i\mathbf{R}_i\mathbf{x}_j + \mathbf{t}_i$ is the projected 2D location of vertex j in the image.

3.3. Photometric Normal Estimation

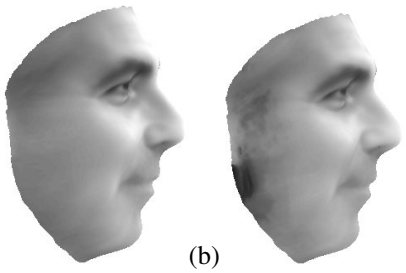
Fitting the 3DMM based on limited landmarks reconstructs a face with the overall shape of the individual, without the fine facial details, since it has few parameters. Even a traditional 3DMM is constrained by the span of the face bases and lacks the representational power to accurately reconstruct arbitrary, unseen faces. To recover these fine details, we use a photometric stereo-based approach to estimate the normals which in turn drives the reconstruction of surface details.

In computer graphics, a 3D model, projection model, texture map, and light sources are combined under a lighting model to render images. Computer vision aims to solve the inverse problem, *i.e.*, inferring the model parameters from one or multiple images. In either case, simplifying assumptions must be made. For graphics, the assumptions are because of either limited understanding about reflectance properties of different surfaces or computational efficiency. For vision, assumptions or prior knowledge are required to make the under-constrained inverse problem solvable.

We assume a Lambertian lighting model where the intensity at a projected point is defined by a linear combination of lighting parameters and the surface normal,

$$\mathbf{I}(u,v) = \rho_j \left(k_{\text{a}} + k_{\text{d}} \left(l^x n_j^x + l^y n_j^y + l^z n_j^z\right)\right), \quad (4)$$

where ρ_j is the surface albedo at vertex j, n_j^x, n_j^y, n_j^z is the unit surface normal at vertex j, k_{a} is the ambient coefficient, k_{d} is the diffuse coefficient, and l^x, l^y, l^z is the unit light source direction. For simplicity, we define $\mathbf{l} = \langle k_{\text{a}}, k_{\text{d}}l^x, k_{\text{d}}l^y, k_{\text{d}}l^z\rangle^\intercal$ for the lighting, $\mathbf{n}_j = \langle 1, n_j^x, n_j^y, n_j^z\rangle^\intercal$ for the normal, and $\mathbf{s}_j = \rho_j\mathbf{n}_j$ for the shape, so that $\mathbf{I}(u,v) = \mathbf{l}^\intercal\mathbf{s}_j$.



(a) (b)

Figure 3. Effect on albedo estimation with (a) and without (b) dependability. Skin should have a consistent albedo, but without dependability the cheek shows ghosting effects from misalignment.

To solve the Lambertian equation, prior work recognized that 95% of the variation in a face image set is explained by the first four principal components of $\mathbf{F}$ [5]. Thus, singular value decomposition (SVD) is used to factor $\mathbf{F}$ into a light matrix $\mathbf{L}^\intercal$, where each row is the light coefficients of image i, and $\mathbf{S}$, where each column is the shape coefficients of vertex j. Unfortunately, SVD alone cannot determine the true lighting and shape matrices since any invertible 4×4 matrix $\mathbf{A}$ forms a valid solution, $\mathbf{F} = \mathbf{L}^\intercal\mathbf{S} = \tilde{\mathbf{L}}^\intercal\mathbf{A}^{-1}\mathbf{A}\tilde{\mathbf{S}}$. To resolve this ambiguity the template face is typically used to constrain $\mathbf{A}$ to a numerically stable solution. In our study, we discover that this SVD approach fails to reconstruct for small image collections or when too much noise enters the rank-4 approximation from either extreme expression for people like Jim Carrey, or inconsistent occlusions such as long hair from women.

Instead we propose to solve the unknowns in an energy minimization approach with the following loss function,

$$\underset{\rho_j,\mathbf{l}_i,\mathbf{n}_j}{\text{argmin}} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} \|f_{ij} - \rho_j\mathbf{l}_i^\intercal\mathbf{n}_j\|^2 + \lambda_n\|\mathbf{n}_j - \mathbf{n}_j^t\|^2 \right), \quad (5)$$

where $\mathbf{n}_j^t$ is the current surface normal of the template at vertex j. This function may be solved by initializing $\mathbf{n}_j$ to $\mathbf{n}_j^t$ and ρ_j to 1 and then solving in an alternating manner for lighting, albedo, and normals.

3.3.1 Dependability

Not every part of each image is created equal. Clearly non-visible parts are not dependable, but even some visible parts may not help. For example, a low-resolution image will contribute less information than a higher-resolution one. Parts of faces changed by expression will have different surface normals. Faces with inaccurate landmark alignment will be out of correspondence. Many different factors play a role in the dependability of a projected point within an image. In the end, we found that simply using

$d_{ij} = \max(\cos(\mathbf{c}_i^\intercal \mathbf{n}_j), 0)$ where $\mathbf{c}_i$ is a unit camera vector perpendicular to the image plane is a good measure of dependability. This decreases the weight as a vertex approaches perpendicular to the camera since it is more susceptible to small changes in the pose estimation, whereas a vertex pointing towards the camera is more dependable. Fig. 3 shows the albedo estimation with and without dependability. We update Eqn. 5 to,

$$\underset{\rho_j, \mathbf{l}_i, \mathbf{n}_j}{\operatorname{argmin}} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} \|d_{ij}(f_{ij} - \rho_j \mathbf{l}_i^\intercal \mathbf{n}_j)\|^2 + \lambda_n \|\mathbf{n}_j - \mathbf{n}_j^t\|^2 \right). \quad (6)$$

3.3.2 Lighting and albedo estimation

We begin by initializing $\mathbf{n}_j$ to the template surface normal at vertex j and ρ_j to 1. While keeping the surface normals fixed, we alternate between solving the light coefficients and the surface albedo. We let this converge before estimating the surface normal, which allows the current surface normal to influence which local minimum solution is found. Solving for albedo is then an overconstrained least squares solution, *i.e.*, $\rho_j = (\mathbf{d}_j^\intercal \mathbf{L}^\intercal \mathbf{n}_j)/(\mathbf{d}_j^\intercal \mathbf{f}_j)$. Similarly, the lighting for an image has the closed form solution $\mathbf{l}_i^\intercal = (\mathbf{f}_i \circ \mathbf{d}_i)/(\mathbf{S} \circ \mathbf{d}_i)$, where $\circ$ is the Hadamard or entry-wise product.

3.3.3 Surface normal estimation

Once the lighting and surface reflectance properties are estimated, we finally estimate the surface normals. Similar to [23], we use a local subset of images to estimate the surface normal at each vertex. The goal of the local selection is to capture the dominant local expression among the collection, instead of a smoothed average of all expressions; it also serves to filter occlusions or areas with poorly fit templates. Given a subset of images $\mathcal{B} = \{i \mid \|\mathbf{l}_i^\intercal \mathbf{s}_j - f_{ij}\|^2 < \epsilon_n\}$, we minimize the following energy for each vertex:

$$\underset{\mathbf{n}_j}{\operatorname{argmin}} \sum_{i \in \mathcal{B}} \|d_{ij}(\rho_j \mathbf{l}_i^\intercal \mathbf{n}_j - f_{ij})\|^2 + \lambda_n \|\mathbf{n}_j - \mathbf{n}_j^t\|^2. \quad (7)$$

The regularization helps keep the face close to the initialization. But since the summation is not averaged, as more photos are added to the collection, the regularization has less weight and the estimated normals can deviate to match the observed photometric properties of the collection. In contrast, when the photo collection is small, the regularization term will play a relatively larger weight in determining the desired surface normal. Thus, this adaptive weighting handles a diverse photo collection size.

3.4. Surface Reconstruction

Given the surface normals $\mathbf{n}_j$ that specify the fine details of the face, we reconstruct a new surface $\mathbf{X}$ following the procedure outlined in [28]. We briefly summarize the procedure, and refer the reader to [28] for full details.

Algorithm 1: Adaptive 3D face reconstruction

Data: Photo collection
Result: 3D face mesh $\mathbf{X}$
`// Template personalization`
estimate landmarks $\mathbf{W}_i$ for each image
fit the 3DMM via Eq. 3 to generate template $\mathbf{X}^0$
remesh to the coarse resolution
for *resolution* $\in$ {*coarse, medium, fine*} **do**
  repeat
    estimate projection $s_i, \mathbf{R}_i, \mathbf{t}_i$ for each image
    establish correspondence $\mathbf{F}$ via backprojection
    estimate lighting $\mathbf{L}$ and albedo ρ via Eq. 6
    estimate surface normals $\mathbf{N}$ via Eq. 7
    reconstruct surface $\mathbf{X}^{k+1}$ via Eq. 8
  until $\frac{1}{p}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 < \tau$
  subdivide surface

The overall energy for surface reconstruction is composed of three parts,

$$\underset{\tilde{\mathbf{X}}}{\operatorname{argmin}}\ E_n + \lambda_b E_b + \lambda_l E_l. \quad (8)$$

We define $\tilde{\mathbf{X}}$ as a $3p$-dim reshaping of $\mathbf{X}$ collecting the x-coordinates followed by y and z, Δ is the Laplacian operator, $\mathcal{L}$ is its discretization up to a sign, H is the mean curvature, and H_j is the estimation based on the normals [28]. Then $E_n = \|\mathcal{L}\tilde{\mathbf{X}} - \mathbf{H}^k\|^2$ is the normal energy derived from the mean curvature formula $\Delta \mathbf{x} = -H\mathbf{n}$ and we collect and repeat $-H_j \mathbf{n}_j$ into a $3p$-dim vector $\mathbf{H}$. $E_b = \|\mathcal{L}_b \tilde{\mathbf{X}} - \mathcal{L}_b \tilde{\mathbf{X}}^k\|^2$ is the boundary energy, required since the mean curvature formula degenerates along the surface boundary into the geodesic curvature, which cannot be determined from the photometric normals. We therefore seek to maintain the same Laplacian along the boundary with $\mathcal{L}_{b,ij} = 1/e_{ij}$ where e_{ij} is the edge length connecting adjacent boundary vertices i and j. And $E_l = \sum_i \|s_i \mathbf{R}_i [\tilde{\mathbf{X}}]_{\text{land}} + \mathbf{t}_i - \mathbf{W}_i\|_F^2$, which uses the landmark projection error to provide a global constraint on the face, without which, the integration of the normals can have numeric drift across the surface of the face. Unlike [28] we do not include a shadow region smoothing since we use the template normal as a regularizer during normal estimation.

3.5. Adaptive Mesh Resolution

Algorithm 1 describes the order of steps as put together in the final face reconstruction system. When putting the steps together, we use a coarse-to-fine scheme to first fit the overall face shape and later adapt to the details present in the collection. To begin, we use ReMESH [3] to uniformly resample the personalized mesh $\mathbf{X}^0$ to a coarse $6,248$ $(= p)$ vertices. The resampling is done once offline on the mean

Figure 4. Synthetic data with expression, pose, lighting variation.

shape and is transferred to a personalized mesh by using the barycentric coordinates of the corresponding triangle. The algorithm is repeated within each detail level until it converges. After convergence, Loop subdivision [18] is performed to increase the resolution of the mesh, multiplying the number of vertices by 4. Moving from the coarse to fine level, we decrease ϵ_n and λ_n to increase selectivity of images used for surface normal estimation and lower template normal regularization. This helps the coarse reconstruction stay smooth and fit the generic structure while allowing the fine reconstruction to capture the details. We would like to stop the reconstruction automatically after the coarse or medium level if the photo collection does not contain enough information for detailed reconstruction, since the fine level may overfit to noise and lead to poor quality reconstruction. But we have yet to identify a good stopping criterion so we leave this for future work.

4. Experimental Results

To examine the effectiveness of the proposed approach, we experiment using synthetic data, personal photo collections with ground truth scans, and Internet images of celebrities and political figures. For baselines, we compare against prior photometric stereo-based approaches [28, 23]. Stereo imaging or video-based reconstruction techniques have access to additional information and are not compared. Furthermore, because the proposed approach uses a 3DMM to create the initial personal template, we do not compare against 3DMM either. Despite only using the landmarks for 3DMM fitting, the proposed approach can theoretically use any state-of-the-art 3DMM as initialization.

4.1. Experimental Setup

Data Collection We gather the three types of photo collections. Synthetic images are rendered from subject M001 of the BU-4DFE database [39] using the provided texture and selecting random frames from the 6 expression sequences (Fig. 4). A Lambertian lighting model re-illuminates the face with light sources randomly sampled from a uniform distribution in front of the face. Personal photos are used with ground truth models of the subjects created with a Minolta VIVID 910 range scanner at VGA resolution capturing 2.5D depth scans accurate to 220 microns. Given frontal and both 45° yaw scans, we stitch them together using Geomagic Studio to create a full 3D model. For Internet images, we query the Bing image search API with a person's full name. Face clustering is performed with Picasa to filter out spurious results and locate the subject of interest.

Table 1. Error comparison on synthetic data.

Method	Neutral	30° Yaw	Expression
Ours	**3.22%**	**3.82%**	**4.40%**
[28]	6.13%	7.48%	6.59%

Table 2. Error comparison of PC2 with different image numbers.

# Images	1	5	10	20	40
Ours	4.19%	4.07%	4.03%	3.46%	3.18%
[28]	-	8.77%	5.40%	4.73%	4.13%

Metrics To quantitatively evaluate the reconstruction performance we compute the average distance between the ground truth and reconstructed surfaces. The two surfaces are aligned by Procrustes superimposition of the 3D landmarks from the internal part of the face. The normalized vertex error is computed as the distance between a vertex in the ground truth mesh and the closest vertex in the reconstructed surface divided by the eye-to-eye distance. We report the average normalized vertex error.

Parameters The parameters for the algorithm are set as follows: $\tau = 0.005$, $\lambda_l = 0.01$, $\lambda_b = 10$, $\lambda_n = [1, 0.1, 0.01]$, and $\epsilon_n = [0.2, 0.08, 0.08]$ for coarse, medium, and fine resolution respectively.

4.2. Results, Comparisons, and Discussions

Synthetic The synthetic dataset allows us to test the algorithm's robustness to pose and expression *independently*. We generate three different sets of 50 images each: frontal faces with neutral expression, neutral expression faces with random yaw angles between ±30°, and frontal faces with random expressions. The ground truth model is taken as the neutral expression and reconstructions are aligned to the model using manually annotated 3D landmarks around the eyes, nose, and mouth. Table 1 shows that the proposed approach outperforms prior work in all scenarios. We see the proposed algorithm is more robust to pose than expression variation. Hopefully the improved capability of landmark alignment for large-pose faces [20, 21, 42] will further improve 3D reconstruction performance.

Personal photo collections To evaluate the reconstruction empirically on in-the-wild images, we capture two personal photo collections as well as ground truth 3D models of their neutral expression. Photo collection 1 (PC1) consists of 39 professional photos taken at a wedding. The proposed approach has 5.10% error while [28] has 8.31% on this set. Both results are relatively poor, which we hypothesize is due to the post processing usually done on professional photos of this nature, which invalidates the Lambertian assumption. Photo collection 2 (PC2) consists of 40 images captured on an iPhone by moving around to get different overhead lights and having the subject make random expressions and poses. This collection is similar to the popular selfies.

Figure 5. Qualitative comparison on celebrities. The proposed approach incorporates more of the sides of the face and neck than [23] while producing a better depth estimate than [28].

Figure 6(c) shows the resulting reconstruction with different numbers of images and photo resolution overlaid with the reconstruction error to demonstrate how different error amounts appear. This error measurement does a good job of capturing the global reconstruction error. We also compare with the prior work [28] for decreasing image numbers in Table 2. This shows our method has consistently lower errors, especially with lower numbers of images.

Internet collections A reconstruction may have a very good fit to the overall structure of the individual, but fail to capture some of the fine details that help define the person. For example, missing facial wrinkles will have a very minor impact on the surface-to-surface error, but can play a large role in convincing a human that the reconstruction is accurate. We strive not just for a metrically correct reconstruction, but also for a visually compelling reconstruction. After all, one major goal of using the photometric normals is to allow for reconstruction of the details outside of the span of a traditional 3DMM. We process the same set of celebrities used in [23] and [28], George Clooney (359 photos), Kevin Spacey (231), Bill Clinton (330), and Tom Hanks (264). The resolution of the images is scaled to 500 vertical pixels to match [28]. Figure 5 presents a side by side comparison between the various approaches. Our reconstruction is able to capture a larger surface area stretching to the neck and all the way back to the ears, while still capturing the fine details of the face. Fig. 7 presents more examples using 25-50 photos demonstrating the ability of our algorithm to generalize across races and genders. Note the ability to even reconstruct hairstyles for some people. The contrast between the personalized template and final reconstruction shows the limitation of landmark-based 3DMM fitting and the power of normal-based surface reconstruction.

Figure 7. Reconstruction results for Jinping Xi, Robin Williams and Sonya Sotomayor. From Left to right, personalized template, final reconstruction, and estimated albedo rendered on the surface.

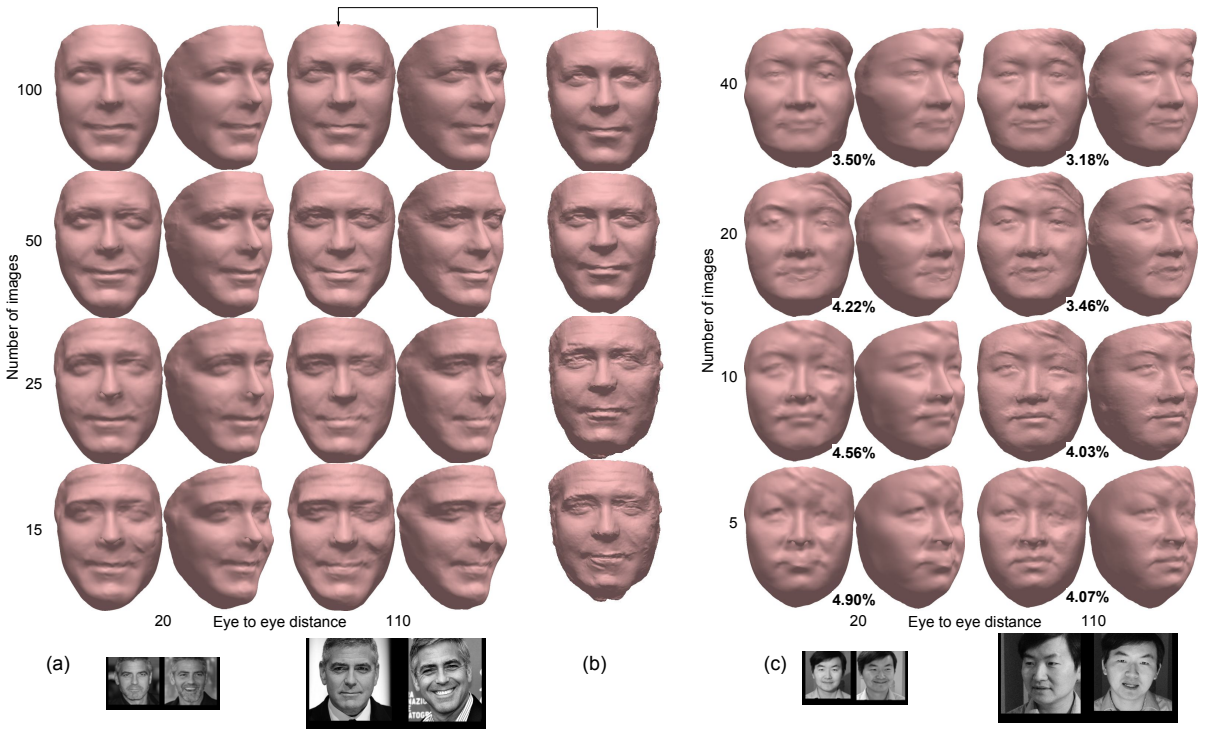


Figure 6. (a) George Clooney with different quality images. (b) Reconstruction without coarse-to-fine process. (b) Personal collection with different quality images. Reconstruction errors of our method are overlaid on each face pair.

Efficiency Written in a mixture of C++ and Matlab, the algorithm runs on a commodity PC with an AMD A10-5700 3.40 GHz CPU and 8 GB RAM. The processing time is $O(np + p^2)$ and we report times w.r.t. 100-image collections. Preprocessing, including face detection, cropping, and landmark alignment, takes 38 seconds. Template personalization takes 5 seconds. Photometric normal estimation and surface reconstruction take 6, 22, and 94 seconds for each iteration of the coarse, medium, and fine resolution, respectively. A typical reconstruction of George Clooney takes 5 coarse iterations, 2 medium, and 1 fine for a total time of 3.5 minutes.

Number of images One critique of photometric stereo-based reconstructions in the past is their dependence on a large number of images, typically several hundreds, which is too many for most applications. Figure 6(a) shows the reconstruction results for George Clooney with varied image numbers and resolutions. When only a few images exist, the algorithm relies more on the template face to regularize the photometric normals. This allows the reconstruction to gracefully degrade; as more images are available, the algorithm uses the additional data to create a more accurate and detailed reconstruction. Even with low resolution it is able to capture wrinkles on the forehead since the sampling across multiple images acts as super-resolution.

We also present the reconstruction errors for PC2 with different numbers of images in Figure 6(c). Note that the proposed approach can reconstruct a reasonable appearing face with only a few images and the error decreases as more images are used. The minimal number of images for PC2 is less than Fig. 6(a) since personal photo collections tend to be higher quality.

Coarse to Fine The coarse-to-fine scheme benefits both efficiency and quality. If the coarse-to-fine scheme is not used and instead the reconstruction starts at the fine resolution, it takes 4 iterations to converge for a total time of 7 minutes or double the time. Also, Fig. 6(b) shows the resultant reconstructions which are similar for large amounts of images, but noisy for small collections since the coarse step allows for more template regularization.

5. Conclusions

We presented a method for reconstructing a 3D face model from an unconstrained 2D photo collection which adapts to lower quality and fewer images. By using a 3DMM to create a personalized template which adaptively influences reconstruction in a coarse-to-fine scheme, we can efficiently create a more accurate model than prior work as demonstrated by experiments on synthetic and real-world data. There are numerous paths for future work, *e.g.*, fusing 3DMM and photometric stereo-based reconstructions so it can gracefully degrade down to a single image, and automatically identifying the detail level of reconstruction possible from an arbitrary photo collection.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications ACM*, 54(10):105–112, 2011. 1

[2] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel. Bob: a free signal processing and machine learning toolbox for researchers. In *ACMMM*, pages 1449–1452. ACM Press, 2012. 2

[3] M. Attene and B. Falcidieno. ReMESH: An interactive environment to edit and repair triangle meshes. In *SMI*, pages 271–276, 2006. 5

[4] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):218–233, 2003. 2

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

[6] T. Beeler, B. Bickel, P. Beardsley, R. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. *ACM Trans. Graph.*, 29(3), 2010. 1, 2

[7] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, and M. Gross. High-quality passive facial performance capture using anchor frames. *ACM Trans. Graph.*, 30(4):75:1–75:10, 2011. 1, 2

[8] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, 2003. 3

[9] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Trans. Graph.*, 34(4):46:1–46:9, 2015. 1, 2

[10] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: a 3D facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graphics*, 20(3):413–425, 2014. 2

[11] B. Chu, S. Romdhani, and L. Chen. 3D-aided face recognition robust to expression and pose variations. In *CVPR*, pages 1899–1906, 2014. 3

[12] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi. 3D scanning deformable objects with a single RGBD sensor. In *CVPR*, pages 493–501, 2015. 1

[13] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.*, 33(6):158, 2013. 1, 2

[14] M. F. Hansen, G. A. Atkinson, L. N. Smith, and M. L. Smith. 3D face reconstructions from photometric stereo using near infrared and visible light. *CVIU*, 114(8):942–951, 2010. 2

[15] H. Hayakawa. Photometric stereo under a light source with arbitrary motion. *J. Optical Soc. America A.*, 11(11):3079–3089, 1994. 2

[16] C. Hernández, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):548–554, 2008. 2

[17] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3D avatar creation from hand-held video input. *ACM Trans. Graph.*, 34(4):45, 2015. 2

[18] A. Jacobson et al. gptoolbox: Geometry processing toolbox, 2015. http://github.com/alecjacobson/gptoolbox. 6

[19] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3D face alignment from 2D videos in real-time. In *FG*, 2015. 2

[20] A. Jourabloo and X. Liu. Pose-invariant 3D face alignment. In *ICCV*, 2015. 6

[21] A. Jourabloo and X. Liu. Large-pose face alignment via CNN-based dense 3D model fitting. In *CVPR*, 2016. 6

[22] I. Kemelmacher-Shlizerman. Internet-based morphable model. In *ICCV*, 2013. 2

[23] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, 2011. 1, 2, 5, 6, 7

[24] K. Lee, J. Ho, and D. Kriegman. Nine points of light: Acquiring subspaces for face recognition under variable lighting. In *CVPR*, pages 129–139, 2001. 2

[25] X. Liu and T. Chen. Pose-robust face recognition using geometry assisted probabilistic modeling. In *CVPR*, volume 1, pages 502–509, 2005. 2

[26] R. Newcombe, D. Fox, and S. Seitz. Dynamicfusion: Reconstruction and tracking on non-rigid scenes in real-time. In *CVPR*, pages 343–352, 2015. 1

[27] W. Peng, C. Xu, and Z. Feng. 3D face modeling based on structure optimization and surface reconstruction with b-spline. *Neurocomputing*, 179:228–237, Feb. 2016. 2

[28] J. Roth, Y. Tong, and X. Liu. Unconstrained 3D face reconstruction. In *CVPR*, 2015. 1, 2, 3, 5, 6, 7

[29] B. Shi, K. Inose, Y. Matsushita, and P. Tan. Photometric stereo using internet images. In *3DV*, pages 361–368, 2014. 2

[30] F. Shi, H.-T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.*, 33(6), 2014. 2

[31] P. Snape, Y. Panagakis, and S. Zefeiriou. Automatic construction of robust spherical harmonic subspaces. In *CVPR*, 2015. 2

[32] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, pages 796–812. Springer, 2014. 2

[33] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.*, 34(6), 2015. 2

[34] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph.*, 31(6):187:1–187:11, Nov. 2012. 1

[35] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980. 2

[36] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *ACCV*, pages 703–717, 2010. 2

[37] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for accurate face alignment. In *ICCVW*, pages 392–396, 2013. 1, 3

[38] C. Yang, J. Chen, N. Su, and G. Su. Improving 3D face details based on normal map of hetero-source images. In *CVPRW*, pages 9–14. IEEE, 2014. 2

[39] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3D dynamic facial expression database. In *FG*, 2008. 6

[40] 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. *Int. J. Comput. Vision*, 35:203–222, 1999. 2

[41] D. Zeng, Q. Zhao, and J. Li. Examplar coherent 3D face reconstruction from forensic mugshot database. *J. Image Vision Computing*, 2016. 2

[42] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *CVPR*, 2016. 6

[43] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, pages 787–796, 2015. 1, 2, 3