Adaptive 3D Face Reconstruction from Unconstrained Photo Collections

Joseph Roth, Yiying Tong, *Member, IEEE,* and Xiaoming Liu, *Member, IEEE*

Abstract—Given a photo collection of "unconstrained" face images of one individual captured under a variety of unknown pose, expression, and illumination conditions, this paper presents a method for reconstructing a 3D face surface model of the individual along with albedo information. Unlike prior work on face reconstruction that requires large photo collections, we formulate an approach to adapt to photo collections with a high diversity in both the number of images and the image quality. To achieve this, we incorporate prior knowledge about face shape by fitting a 3D morphable model to form a personalized template, following by using a novel photometric stereo formulation to complete the fine details, under a coarse-to-fine scheme. Our scheme incorporates a structural similarity-based local selection step to help identify a common expression for reconstruction while discarding occluded portions of faces. The evaluation of reconstruction performance is through a novel quality measure, in the absence of ground truth 3D scans. Superior large-scale experimental results are reported on synthetic, Internet, and personal photo collections.

Index Terms—Face reconstruction, photometric stereo, unconstrained.

1 Introduction

3D reconstruction, the process of inferring a 3D model from 2D imagery, is a long-standing problem in computer vision. Beginning with simple and constrained desktop objects and expanding to large, complex, or unconstrained objects including outdoor scenes [1], many different approaches have been proposed. One specific object, the face, has seen a recent growth of research for creating detailed 3D models, often called *face reconstruction*. Not only are faces one of the most commonly photographed objects, but having an understanding of the 3D face shape enables many applications in face recognition [2], 3D expression recognition [3], facial animations [4], avatar puppeteering [5], and more. Face reconstruction is important for biometrics since the estimation of pose, expression, and illumination, the confounding factors of face recognition, may all be improved with accurate person-specific models [2], [6], [7]. Faces are particularly challenging for multi-image reconstructions since they are non-rigid with deformations caused by expression variation, aging, weight changes, etc.

Face reconstruction itself is a broad topic spanning many different situations depending on the input type and desired level of detail. Table 1 provides an overview of the most common scenarios and approaches. It is important to note the large variety of different scenarios and methods for face reconstruction. Oftentimes, commercial systems will combine multiple methods in order to overcome the shortcomings of one single approach.

This work specifically looks at the scenario of using unconstrained photo collections. *Unconstrained* means there is no knowledge about the cameras, the illumination of the scene, or the pose and expression of the subject. A *photo collection* refers to a set of images of the same subject taken with no knowledge of the time of capture; this is in contrast to a video, which can be viewed as a consecutive set of images with a consistent and small time intervals between images. For example, Google Photos clusters your personal photos to create collections for each individual. Compared to other unconstrained settings, a photo collection possesses more information than a single image, but has fewer constraints than a video which satisfies temporal assumptions.

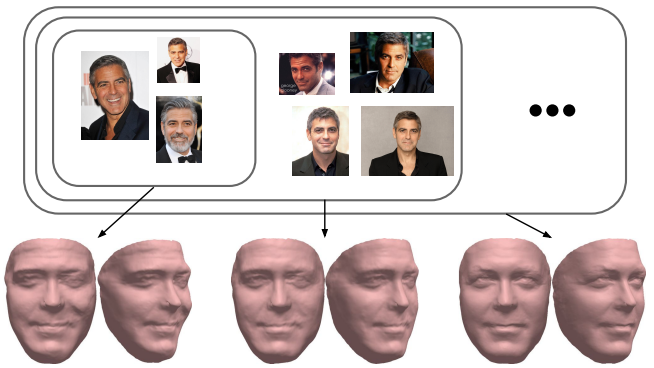


Fig. 1. The proposed system reconstructs a detailed 3D face model of the individual, adapting to the number and quality of photos provided.

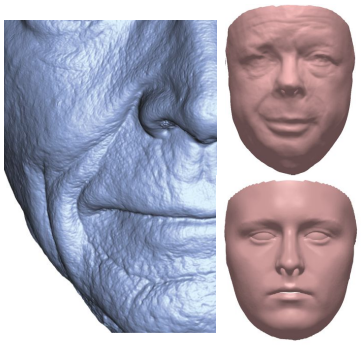
Dense correspondence among images is the basis for multi-view stereo and photometric stereo reconstruction approaches. Unlike videos where optical flow [15] can take advantage of consistent lighting over time, photo collections are hard to establish dense correspondence and particularly challenging for reconstruction. The recent work of [11] demonstrates the promise of enhanced dense correspondence for arbitrary face images, but it is still not accurate enough for direct multi-view stereo reconstruction. In the meantime, photometric stereo-based reconstruction methods have proven most effective for unconstrained photo collections. Starting with the reconstruction of 2.5D depth maps [14] and extending to full 3D meshes [16], photometric stereo-based methods not only perform face reconstruction but also explicitly estimate the albedo and per image light-

• *J. Roth, Y. Tong, and X. Liu are with the Department of Computer Science and Engineering, Michigan State University. E-mail: liuxm@cse.msu.edu*

Manuscript received Oct 10, 2016.

TABLE 1
Overview of face reconstruction approaches. Example of levels of detail pore (left), wrinkle (top), and smooth (bottom).

	Input	Method	Detail
Constr.	Point cloud	Range scanner	± 0.03mm max
	Synchronized images	Multi-view stereo [8]	0.088mm mean, pore
	Time-multiplexed	Photometric stereo [9]	wrinkle
Uncon.	Single image	3DMM [10]	smooth
		CNN [11]	smooth
	Video	Optical flow tracking [12]	wrinkle
	RGB-D Video	Dynamic fusion [13]	wrinkle, deformable
	Photo collection	Photometric stereo [14]	wrinkle



ing and pose conditions. All prior photo collection methods reconstruct a single representative face from the collection, which is challenging given the expression variation.

While prior photo collection-based reconstruction methods are compelling, they have many limitations. Frontal images are required for [14], and even though [16] can use non-frontal images, we demonstrate its significant performance drop with large pose variation. Another limitation of [16] is the requirement of a sufficiently large photo collection. Theoretically, only four images are necessary for a photometric stereo-based approach, but in practice prior approaches report results on collections with over one hundred images each, for two primary reasons. One, their singular value decomposition solution to photometric stereo is susceptible to noise with small collections. Two, prior approaches perform a local selection step where only $\sim 10\%$ of images are used for each vertex of the model.

In our face reconstruction framework, given a collection of unconstrained face images, we first align 2D landmarks [17] to all detected faces. We then create a personalized face model by fitting a 3D morphable model (3DMM) jointly to the collection such that the estimated 2D landmarks align with the projection of the associated 3D landmarks on the model. Dense correspondence is established across the collection by estimating the pose for each image and back-projecting the image onto the personalized template. The albedo, lighting, and surface normals are estimated globally with an energy minimization approach using an adaptive template regularization to allow reconstruction of small photo collections down to a single face. The surface normals are further refined locally using a novel structural similarity feedback, increasing the amount of images used for local selection to $\sim 50\%$. Reconstruction of the face model deforms the mesh to match the estimated surface normals. A coarse-to-fine process is employed to first capture the generic shape and then fill in the details. We perform extensive experimental evaluations to show qualitatively and quantitatively the performance of the proposed face reconstruction method.

In short, this paper makes the following contributions.

◇ A 3D morphable model is fit jointly to 2D landmarks of multiple images for model personalization. Prior work used either a fixed template or landmark-based deformation that does not work well for small collections.

◇ A joint Lambertian image rendering formulation with an adaptive template regularization solves the photometric stereo problem, allowing for graceful degradation to a small number of images.

◇ A pose-based dependability measure is proposed to weight the influence of face regions based on confidence.

◇ Structural similarity, a measure correlated with human perception, drives the local selection of images for estimating surface normals.

◇ A new reconstruction quality measure based on structural similarity enables evaluation without ground truth.

2 Prior Work

We now review face reconstruction techniques in constrained and unconstrained scenarios, photometric stereo estimation and how we differ from traditional approaches, and reconstruction quality assessment.

Constrained Face Reconstruction The most accurate scenario for face reconstruction is *constrained*. With a constrained subject, *i.e.*, when the person is captured under known conditions (lighting, pose, and expression) with known equipment, highly detailed models may be reconstructed. Commercial *range scanners* capture fine details ranging from ± 0.03mm accuracy for the $\$75,000$ Konica Minolta Vivid 9i [18] to ± 0.5mm with the reasonably priced $\$1,441$ IIIDScan [19]. These detailed scans are often considered as ground truth for evaluating other reconstruction methods, but only capture depth maps and need to fuse multiple scans to form a 3D reconstruction.

Synchronized *multi-camera stereo* setups [8], [20] can capture pore-level detail with 0.088mm accuracy demonstrated on a physical mask. While real faces with their high specularity and non-uniform albedo will produce higher quantitative errors, it can still recover realistic pores and wrinkles for real faces. The most economical constrained option is *photometric stereo* where a single camera takes multiple images with different known lighting conditions [21]. Accuracy is not reported in world distances, but based on visual inspection it achieves wrinkle and not quite the pore details of prior work. These constrained approaches capture detailed, metrically accurate rigid models. Since these approaches require subject cooperation, they are routinely used to capture actor models for video games and films.

Unconstrained Face Reconstruction In *unconstrained* situations, where the exact camera, lighting conditions, pose, and expression are unknown, the accuracy depends on the type of input. The simplest form is a *single image*, where limited information about the surface is present and approaches

must rely on prior knowledge. The most common approach uses a statistical model of face shape distributions as a 3DMM to produce a smooth reconstruction. The first 3DMM fitting approach uses an image rendering loss function to estimate model parameters that produce a similar synthetic image to the observed real image [10]. However, it takes several minutes to converge for a single image. There are also improved 3DMM fitting algorithms, including [22], [23], [24]. Recently, CNN-based approaches to fitting the 3DMM have shown efficient means of performing sparse [25] and dense [11], [26] reconstructions.

A *video* provides sufficient information for unconstrained reconstruction [27], [28], [29], [30], [31], [32], [33], [34] where each frame comes from the same camera and temporal coherence may be assumed. Video-based reconstructions produce smooth details in real-time and are great for expression transfer, avatar puppeteering, and other consumer entertainment applications. With more processing time, video-based reconstructions produce accurate wrinkle details and can also be used for avatar construction. Recently, approaches using a depth or RGB-D video [13], [35] can produce a dynamic model deforming to every video frame.

Harder than videos are *photo collections* [14], [16], [36]. For detailed reconstructions using photometric stereo, [14] reconstructs a 2.5D height field that is extended in [16] to a true 3D surface. A smooth surface can also be reconstructed using B-splines in [36]. Photo collections are used for a variety of non-reconstruction purposes such as improved fitting of a 3DMM [37] and even creating a 3DMM without range scans [12], [38]. Our work focuses on photometric stereo-based reconstruction for its ability in producing wrinkle details, but we address the limitations of current approaches such as the restriction to large collections with mainly frontal images as input.

Photometric Stereo Classic photometric stereo estimates the surface normals of a *rigid* object from a *fixed* camera orientation by observing the reflectance under different lighting conditions. Photometric stereo was first proposed with knowledge of the lighting conditions [39] and even current methods still use this approach for cooperative subjects [9], [21]. Later it was discovered that even without detailed knowledge of the light source, photometric stereo can take advantage of dominant contribution of the low-rank spherical harmonics in lighting due to the mainly diffuse reflectance of face [40], [41], [42], [43], [44], [45]. Most recent work can take multiple camera positions and put images into correspondence using Structure from Motion and even estimate arbitrary non-linear camera response maps [46]. Most photometric stereo techniques reconstruct from a common viewpoint and produce a 2.5D face surface which can only make use of frontal images. Solutions usually use singular value decomposition (SVD) to find a low rank approximation of the spherical harmonics, using an integrability constraint or prior knowledge of the object to resolve ambiguity. SVD approaches require sufficient 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 a novel, adaptive, template regularized, image rendering solution to photometric stereo, where we solve the same loss function as traditional 3DMM fitting for the lighting, albedo, and surface normals instead of the shape. This enables the solution to work using substantially fewer images.

TABLE 2
Notations.

Symbol	Dim.	Description
$\mathbf{I}$	matrix	image
n	scalar	number of images
q	scalar	number of landmarks (68)
$\mathbf{W}$	2xq	2D landmark matrix
p	scalar	number of mesh vertices
$\mathbf{X}$	3xp	3D shape model
$\mathbf{N}$	4xp	surface normal matrix
$\mathbf{L}$	4xn	lighting matrix
$\mathbf{D}$	nxp	dependability matrix
$\mathcal{L}$	pxp	sparse Laplacian
s	scalar	scale
$\mathbf{R}$	2x3	rotation matrix
$\mathbf{t}$	2x1	translation vector
ρ_j	scalar	albedo at vertex j
$\mathbf{F}$	nxp	image correspondence
H_j	scalar	mean curvature

Quality Assessment Evaluating the quality of a reconstruction technique is a challenging problem. Many face reconstruction works avoid a quantitative assessment due to the lack of ground truth or the inability of measurements to distinguish between fine details in the reconstruction. Some works label a sparse set of landmarks on the images and measure the projection error at these points [11], [26]. With a ground truth model, a surface-to-surface distance can be evaluated such as the L^2 distance or the Hausdorff distance, but the L^2 distance focuses only on the low frequency details, while the Hausdorff distance is susceptible to the error at outliers. A recent work [47] uses a weighted normal distance between the reconstructed face and the mean face in order to assign quality to different reconstructions. We propose a structural similarity based metric to quantitatively evaluate reconstruction performance in the absence of ground truth scans, and demonstrate how it aligns with human perception of the reconstruction quality.

3 Algorithm

We now present the details of the proposed approach, describing the motivational differences from prior works. Figure 2 provides an overview of the different steps to face reconstruction. The algorithm assumes the existence of a photo collection with automatically annotated landmarks and a 3DMM. Notations used throughout this paper are provided in Table 2. The main algorithm is composed of three steps. Step 1: Fit the 3DMM template to produce a coarse person-specific template mesh. Step 2: Estimate the surface normals of the individual using a photometric stereo (PS)-based approach. Step 3: Reconstruct a detailed surface matching the estimated normals.

3.1 Inputs and Preprocessing

3.1.1 Photo collection

A photo collection is a set of n images containing the face of an individual and may be obtained in a variety of ways,

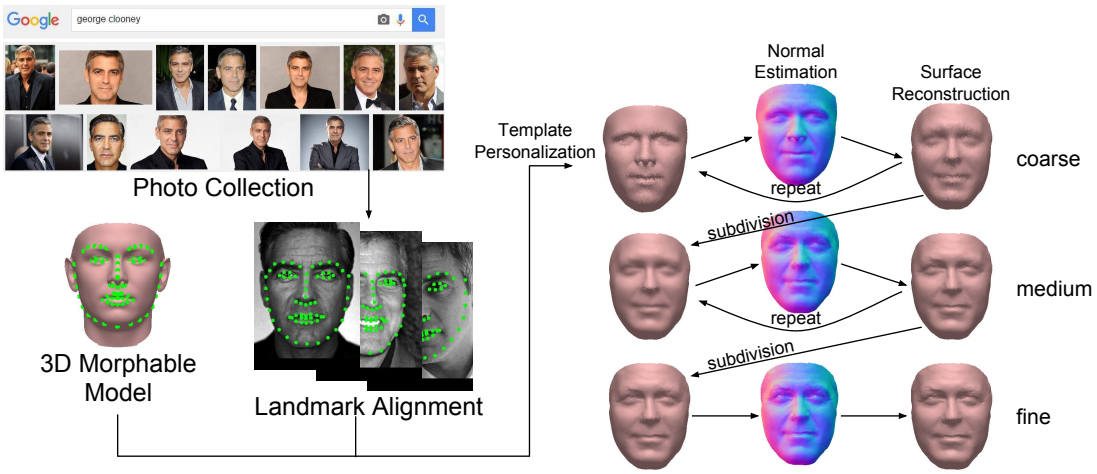


Fig. 2. Overview of face reconstruction. Given a photo collection, we apply landmark alignment and use a 3DMM to create a personalized template. Then a coarse-to-fine process alternates between normal estimation and surface reconstruction.

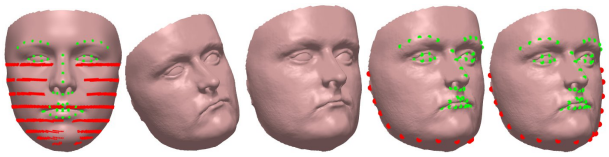


Fig. 3. The landmark marching process. (a) internal (green) landmarks and external (red) defined paths; (b) estimated face and pose; (c) face with roll rotation removed; (d) landmarks without marching; and (e) landmarks after marching corresponding to 2D image alignment.

e.g., a Google image search for a celebrity or a personal photo collection. We assume that the only face in each image belongs to the person of interest. To normalize the images, we automatically detect the face using the built-in face detection model from Bob [48] which was trained on various face 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. We filter out faces with a quality score < 25 to remove extremely poor quality faces or images without a face. Given the face bounding box from the detector, we scale the image to 110 pixels inter-eye distance and crop it to a total size of 437×437 to ensure that the entire face region is present in the image.

A Lambertian lighting assumption uses a linear encoding of the intensity of the lighted object. However, most cameras (and displays) use a non-linear gamma encoding of images in order to provide a subjectively equal step in brightness for humans. Since the exact cameras and image encoding are unknown for unconstrained collections, we apply a single derendering correction [49] to convert each image into the linear intensity scale.

3.1.2 Landmarks

Landmarks are the locations of common fiducial points such as the eyes, nose, or mouth on a face. In recent years, the automatic detection of landmarks [11], [50], [51] has seen rapid improvement due to large labeled datasets such as LFPW [52] and 300-W [53]. To estimate 2D landmarks, we employ the state-of-the-art cascade of regressors approach [17] to automatically fit $q = 68$ landmarks denoted as $\mathbf{W} \in \mathbb{R}^{2\times q}$ onto each image. Figure 3 shows the 68 landmarks used in this work. The landmarks can be separated into two groups. One, the internal landmarks on the eyebrows, eyes, nose, and mouth. These correspond to physical parts of the face and are consistent on all faces regardless of pose. Two, the external landmarks for the cheek / jaw along the silhouette of the face. These landmarks do not have a single correspondence to a point on the 3D face. As the face turns to non-frontal views, face alignment algorithms typically detect external landmarks on the facial silhouette. As a result, the external landmarks of two different poses correspond to different 3D model vertices.

Landmark marching It is therefore desirable to estimate pose specific vertices to maintain 3D-to-2D correspondences between the landmarks. In literature, there have been a few proposed approaches [4], [6], [54]. In this work, we follow the proposed *landmark marching* method from [6]. Specifically, for the external landmarks a set of horizontal *paths*, each containing a set of vertex indices, are defined to match the contour of the face as it turns. Given a non-frontal face image along with an estimated pose, we rotate the 3D model using the estimated yaw and pitch while ignoring the roll, and determine the corresponding vertex along each predefined path based on the maximum (minimum) x-coordinate for the right (left) side of the face. See Fig 3 for an illustration of the process.

3.2 Step 1: Model Personalization

The face model plays a vital role in the reconstruction process. The current face model directly establishes correspondence between photos, provides an initialization for surface normal estimation, and regularization during surface reconstruction. Therefore, it is important to begin with a good personalized model of the face. We desire the model to match the overall metric structure of the individual to provide accurate correspondence when projected onto photos of different poses. However, the model need not contain fine facial details since those will be determined by the photometric normal estimation.

Prior work used either a single face mesh [14] or a Structure from Motion-based (SfM) deformation of a single face mesh [16]. These models have two main limitations. One,

the model has fixed specific ethnicity and gender and may not generalize its fit across a diverse set of subjects. Two, the SfM technique requires multiple images with sufficient pose variation and may not work for small collections. Therefore, we propose supplementing additional prior information to help form a personalized template for a wide range of subjects with few images.

3.2.1 3D Morphable Model

In light of these limitations, we propose to use a 3DMM instead of a single template mesh. A 3DMM can approximate arbitrary face shapes and is one of the most successful models for describing the face. Represented as a statistical distribution of linear combinations of scanned face shapes, the 3DMM compactly represents wide variations due to identity and expression and is independent of lighting and pose. We use the 3DMM of the following form,

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

where $\mathbf{X} \in \mathbb{R}^{3\times p}$ denotes the shape (vertex locations) of a 3D face mesh composed of the mean shape $\bar{\mathbf{X}}$, a set of identity bases $\mathbf{X}^{\text{id}}$, and a set of expression bases $\mathbf{X}^{\text{exp}}$, with coefficients $\vec{\alpha}^{\text{id}}$ and $\vec{\alpha}^{\text{exp}}$. We use the 3DMM from [6] where the identity comes from the Basel Face Model [55] and the expression comes from Face Warehouse [29]. The separation of the bases into expression and identity is based on the method from [56].

Fitting a 3DMM entails finding the model coefficients and projection parameters which best match a face in a given image. Typically, 3DMM fitting aims to minimize the difference between a rendered image and the observed photo [22] using manually annotated landmarks for pose initialization. As automatic face alignment has improved, Zhu *et al.* recently propose an efficient fitting method based only on landmark projection errors [6]. To fit the 3DMM to a face image, they 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 $\mathbf{W}$ for one image, the model parameters are estimated by minimizing the projection error of the 3DMM to the landmarks,

$$\underset{s,\mathbf{R},\mathbf{t},\vec{\alpha}^{\text{id}},\vec{\alpha}^{\text{exp}}}{\arg\min} \|\mathbf{W} - (s\mathbf{R}[\mathbf{X}]_{\text{land}} + \mathbf{t})\|_F^2 + E_{\text{reg}}, \tag{2}$$

where $[\mathbf{X}]_{\text{land}}$ selects the annotated landmarks from the entire model and $\|\cdot\|_F$ is the Frobenius norm and E_{reg} is a regularizer (see Eq. 3) for the 3DMM coefficients. However, as discussed in Sec. 3.1.2, the pose must be known to march the external 3D landmarks along their paths to establish correspondence with $\mathbf{W}$, *i.e.*, the 3D landmark selection depends on the pose and the 3DMM coefficients.

Our single-image joint optimization of Eq. 2 follows [6], which is performed in an alternating manner for the pose parameters and the 3DMM coefficients. Initializing with the mean face, $\vec{\alpha}^{\text{id}} = \vec{\alpha}^{\text{exp}} = \vec{\mathbf{0}}$, first we solve for the pose (s, $\mathbf{R}$, and $\mathbf{t}$) [57], then update the landmarks through marching, and finally solve for the shape ($\vec{\alpha}^{\text{id}}$ and $\vec{\alpha}^{\text{exp}}$). All steps are over-constrained linear least squares solutions. In this work we perform four total iterations since it converges quickly.

We extend this process to jointly fit n faces of the same person by assuming a common set of identity coefficients $\vec{\alpha}^{\text{id}}$ but a unique set of expression $\vec{\alpha_i}^{\text{exp}}$ and pose parameters per image. The modified full error function is,

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

where σ_k is the variance of the kth shape coefficient, typically used in Tikhonov regularization, and $[\cdot]_{\text{land}_i}$ is used because different poses of face images have different selections of corresponding vertices. This function may be solved as before since it is linear with respect to each variable. Once the parameters are learned, we generate a personalized model $\mathbf{X}^0$ using the identity coefficients and the mean of the expression coefficients, indicating the typical expression of the individual in the collection.

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 3D vertex j on image i.

At the conclusion of Step 1, we have a personalized model for the subject matching their overall shape, as well as projection parameters for each image. The model at this stage is a smooth reconstruction for two reasons. One, the 3DMM only captures low-frequency shape details. Two, the model is fit based on a limited set of sparse landmarks so it requires a strong regularization, which favors a smooth result. Despite being smooth, the model allows for a set of dense correspondence to be established across the photo collection. These dense correspondences will be leveraged to produce the fine details of the face.

3.3 Step 2: Photometric Normal Estimation

To add in the wrinkle details to the personalized model, we use the dense correspondence along with a photometric stereo-based normal estimation. Intuitively, the differences in shading observed across the photo collection provide clues to the true surface normal, which may differ from the smooth version offered by the 3DMM estimate. Practically speaking, we will need to estimate the lighting conditions for each image and the surface albedo or reflectance of the face in order to estimate the surface normals.

3.3.1 Lighting Model

In computer graphics, lighting models are used to render a realistic synthetic image from geometric models with reflectance information. Whereas, computer vision uses them to solve the inverse problem, *i.e.*, inferring the model parameters from a real image. In either case, assumptions about

how to model a scene must be made. The assumptions may be due to limited understanding of the real world environment such as reflectance properties of surfaces, or they may be for computational efficiency or tractability. For example, we use a weak perspective camera projection model to tractably solve the pose and projection, and we use the 3DMM model prior knowledge of face shapes to personalize our initial shape model.

For lighting, we assume a Lambertian model, which allows accumulation of many far away light sources into a single vector, where the perceived intensity at a projected point is defined by a linear combination of lighting parameters and the surface normal,

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

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, I_a is the ambient light intensity, I_d is the directional light intensity, and l^x, l^y, l^z is the unit light source direction of the image. For simplicity, we combine the lighting coefficients and direction into a vector $\mathbf{l} = \langle I_\mathrm{a}, I_\mathrm{d} l^x, I_\mathrm{d} l^y, I_\mathrm{d} l^z \rangle^\intercal$, and define $\mathbf{n}_j = \langle 1, n_j^x, n_j^y, n_j^z \rangle^\intercal$ for the normal. Using the notation from the model projection we see that $f_{ij} = \mathbf{I}_i(u,v) = \rho_j \mathbf{l}_i^\intercal \mathbf{n}_j$. This lighting model is also called the first-order spherical harmonics lighting.

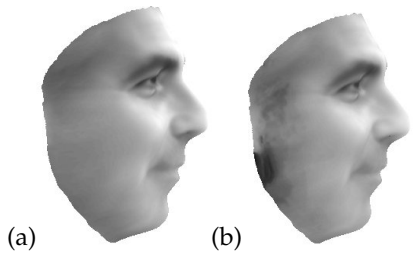
Ref [58] shows that theoretically 1st order spherical harmonics models a minimum of 87.5% of the lighting energy while a non-linear 2nd order will model 99.2%, but in practice they found 1st and 2nd order model 94-98% and 99.5% respectively. Furthermore, [43] demonstrates that shape reconstruction accuracy using 1st order is 95-98% while 2nd order is 97-99%. So, while a more complex lighting assumption may potentially increase the accuracy by a single percentage, it introduces non-linearity into the solution process. Therefore, we use the 1st order assumption in this work. In the future, if we allow other nonlinearities in the model a 2nd order assumption could be made.

Prior work jointly solved for the Lambertian formulation using SVD by factoring $\mathbf{F}$ into a light matrix $\mathbf{L}^\intercal$ and a shape matrix $\tilde{\mathbf{N}}$ which includes the albedo and surface normals [14], [16]. The SVD approach assumes the first four principal components of $\mathbf{F}$ encode the lighting variation while suppressing differences in expression, facial appearance, and correspondence errors. These assumptions hold for large collections of nearly frontal images because SVD can accurately recover the ground truth in the presence of sparse amounts of errors. However, we will show that small collections are susceptible to any correspondence errors from misalignment or expression variations. Furthermore, subjects with long hair that obscures the face and changes styles within the collection will express as an albedo change and affect the first principal component.

In light of the limitations of the SVD approach, we propose an energy minimization approach to jointly solve for albedo, lighting, and normals with,

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

where $\mathbf{n}_j^t$ is the current surface normal of the face mesh at vertex j, and λ_n is the regularization weight. The template regularization helps keep the face close to the initialization.



(a) (b)

Fig. 4. 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.

However, large deviation from the initialization is possible for large collections due to our weighting scheme. Since the summation is not averaged, as more photos are added to the collection, the regularization has less overall weight with λ_n independent of collection size. Thus, the estimated normals may deviate further to match the observed photometric properties of the collection. In contrast, when the photo collection is small, the regularization term will play an important role in determining the estimated surface normal. Thus, this adaptive weighting handles a diverse photo collection size. However, the outliers, which are mitigated by the SVD approach, can have a larger impact with the square error minimization. Therefore, to alleviate the problem of outliers, it is important to properly determine which images to use for each part of the face.

3.3.2 Dependability

While we have claimed to put the photo collection into correspondence $\mathbf{F}$, we certainly do not assume it to be perfect. We use a dependability measurement to weight the influence of different images for each vertex. What makes a part of the projected mesh on an image dependable? Clearly, the part must be visible for the given pose and not occluded by something in front of the face. Does the resolution of an image contribute to its dependability? If the face has a different expression, it may 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. We use $d_{ij} = \max(\cos(\mathbf{c}_i^\intercal \mathbf{n}_j), 0)$, where $\mathbf{c}_i$ is the projection direction (the normal of image plane), as the measure of dependability to handle self-occlusion and sampling artifacts. Other problems such as expression and and external occlusion are left to local selection (Sec. 3.3.4).

What does this dependability measure accomplish? First, backward facing self-occluded parts of the face are given a weight of 0. Second, regions of the image more susceptible to pose estimation errors are given lower weights. As the normal approaches perpendicular to the camera, slight perturbations of the pose can project a faraway surface point to the same pixel on the image. Whereas, a vertex pointing towards the camera is more stable and should be more dependable. Fig. 4 shows the albedo estimation with and without dependability. We update Eqn. 5 to,

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

What is not modeled by this dependability choice? First, regions in cast shadow by any external occlusion, such as cheek occluded by nose or sunglasses; second, landmark alignment errors; and third, expression differences. We will address these issues with the local selection step in Sec. 3.3.4.

3.3.3 Global Estimation

Now that we have a good idea of how to approach the normal estimation, we discuss how to minimize the energy in Eq. 6. While not jointly convex, it is convex for each step when solved in an iterative approach, it has a closed form solution for $\vec{\rho}$, $\mathbf{L}$, and $\mathbf{N}$ independently. We begin by initializing $\mathbf{n}_j$ to the template surface normal at vertex j and ρ_j to 1. We then alternate solving for the lighting coefficients, albedo, and the surface normals until convergence. Solving lighting through least squares has the solution,

$$\mathbf{l}_i = (\tilde{\rho} \circ \mathbf{N} \circ \mathbf{d}_i) \backslash (\mathbf{f}_i \circ \mathbf{d}_i), \tag{7}$$

where $\circ$ is the entrywise product, $\tilde{\rho}$ is $\vec{\rho}$ repeated four times to become the same size as $\mathbf{N}$, and $A \backslash b$ denotes the least squares solution of $Ax = b$ as the backslash operator in Matlab. Similarly, albedo has a closed form solution,

$$\rho_j = (\mathbf{d}_j^\intercal \mathbf{L}^\intercal \mathbf{n}_j) \backslash (\mathbf{d}_j^\intercal \mathbf{f}_j). \tag{8}$$

Finally, the normals are solved via least squares with damping λ_n,

$$\mathbf{n}_j = (\mathbf{B}^\intercal \mathbf{B} + \lambda_n \mathbf{I})^{-1} (\mathbf{B}^\intercal (\mathbf{f}_j \circ \mathbf{d}_j) + \lambda_n \mathbf{n}_j^t), \tag{9}$$

where $\mathbf{B} = \tilde{\rho} \circ \mathbf{D} \circ \mathbf{L}$.

3.3.4 Local Selection

As mentioned in Sec. 3.3.2, the dependability measure only handles small landmark alignment error, but does not consider expression changes, cast shadow, or other potential correspondence errors. To handle these other forms of errors, we use a local selection process as proposed in [14] to refine the photometric estimates. The goal of local selection is to find a collection of images for each vertex that are in local agreement, and re-estimate the surface normal using only those images. This prevents smoothing across all expressions, and can filter the occlusions. The basic approach of local selection is to identify a subset of images $\mathcal{B}_j$ for each vertex j and then re-minimize the photometric equation for that vertex's normal:

$$\underset{\mathbf{n}_j}{\operatorname{argmin}} \sum_{i \in \mathcal{B}_j} \|d_{ij}(\rho_j \mathbf{l}_i^\intercal \mathbf{n}_j - f_{ij})\|^2 + \lambda_n \|\mathbf{n}_j - \mathbf{n}_j^t\|^2. \tag{10}$$

All of the prior work uses the same scheme of local selection [14], [16] which we term square error localization. The subset is chosen such that the square error of the observed value for the image matches the estimated value for the specific vertex, $\mathcal{B}_j = \{i \mid \|\rho_j \mathbf{l}_i^\intercal \mathbf{n}_j - f_{ij}\|^2 < \epsilon\}$. This localization scheme makes its decision solely on the observed value at one particular vertex. It also uses the same loss function to select the local images as the global loss function used to initially estimate the albedo, lighting, and surface normals. This may be advantageous since it forces the localized result to remain close to the global result, while removing outliers. But since it only uses one pixel value in the image for selection, the result is sensitive to noise.

Fig. 5. Raw image, synthetic image under estimated lighting conditions, and SSIM used for local selection. Brighter indicates higher SSIM.

We seek to design a local selection scheme which is influenced by a larger area than a single pixel and uses a loss function consistent with human visual perception. Structural similarity (SSIM) is a measure of perceived quality between two images [59]. Initially used to measure the quality of digital television, it typically uses a raw uncompressed image as ground truth and compares against the encoded version as presented on a screen. SSIM is computed between two windows x and y of common size from different images using the following equation:

$$\text{SSIM}(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{11}$$

where μ_x and μ_y are the mean of x and y, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance, and c_1 and c_2 are constants used to stabilize the division. In Matlab implementation, the window is an isotropic Gaussian weighting with standard deviation γ of the surrounding pixels instead of a blocked window to avoid artifacts. SSIM was specifically designed to better match human visual perception than standard measurements such as mean square error (MSE) or peak signal to noise ratio (PSNR). We can also vary the window size (γ) of SSIM in order to enforce a larger area of local similarity than a single pixel. For this reason, we propose using SSIM for local selection instead of square error.

To select the subset of images for each vertex, we need to compute the SSIM at a vertex on the face model, $\mathbf{S}$, and not at a pixel in the image. To do this, we render a synthetic image using the estimated per image pose and lighting and global albedo and normal. We then compute the SSIM in the image space which gives us a different SSIM value for each pixel. Finally, we backproject the SSIM image onto the face model to create $\mathbf{S}$ in the same way we created $\mathbf{F}$ in Sec. 3.2.1. Figure 5 demonstrates this process for a single image. The local selection now becomes $\mathcal{B}_j = \{i \mid s_{ij} > \epsilon\}$.

3.4 Step 3: Surface Reconstruction

Given the localized surface normals $\mathbf{n}_j$ that specify the fine details of the face, we desire to reconstruct a new face surface $\mathbf{X}$ which matches the observed normals. This process is described in full detail in [16], and we briefly summarize the procedure here.

We use a Laplacian-based surface editing technique motivated by [60]. The Laplace-Beltrami operator is the divergence of a gradient field. Using linear finite elements, it can be discretized into $\mathcal{L}$, a symmetric matrix with entries $\mathcal{L}_{jk} = \frac{1}{2}(\cot\alpha_{jk} + \cot\beta_{jk})$, where α_{jk} and β_{jk} are the opposite angles of edge jk in the two incident triangles (see Figure 6), known as the cotan formula [61]. Geometrically,

Algorithm 1: Adaptive 3D face reconstruction

Data: Photo collection
Result: 3D face mesh $\mathbf{X}$
// Step 1
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
    // Step 2
    globally estimate $\mathbf{L}$, ρ, and $\mathbf{N}$ via Eq. 6
    local selection of images $\mathcal{B}$ via Sec. 3.3.4
    re-estimate surface normals $\mathbf{N}$ via Eq. 10
    // Step 3
    reconstruct surface $\mathbf{X}^{k+1}$ via Eq. 13
  until $\frac{1}{p}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 < \tau$
  subdivide surface

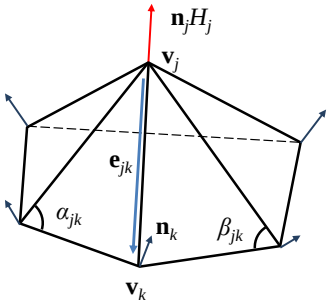


Fig. 6. The mean curvature normal indicates how a vertex deviates from the average location of its immediate neighbors, which can be evaluated as the Laplacian of the position. The mean curvature H_j can be evaluated through $\mathbf{n}$.

$\mathcal{L}$ measures the difference between a functions value at a vertex and the average value of the neighboring vertices. As in [60], [62], we note that $\mathbf{x}_j\mathcal{L} = -\mathbf{n}_j H_j$, where H_j is the integral of the mean curvature in the Voronoi region of vertex j. What this means for us, is that we can use the estimated surface normals to update the positions of the mesh assuming we can determine the mean curvature.

We follow the procedure from [16] to estimate H_j given a normal field. Using a discretization of $H = \nabla A \cdot \mathbf{n}$, *i.e.*, the mean curvature measures how fast the area changes when moving the surface along the normal direction. The first variation of the area can be measured through the difference between $\mathbf{n}_i$ and $\mathbf{n}_j$ as follows,

$$H_j = \frac{1}{4}\sum_{k\in N(j)} (\cot\alpha_{jk} + \cot\beta_{jk})\, \mathbf{e}_{jk}\cdot(\mathbf{n}_k - \mathbf{n}_j), \qquad (12)$$

where $N(j)$ is the one-ring neighborhood of j, and $\mathbf{e}_{jk}$ is the edge from j to k (Figure 6). Note the cotan weights are identical to those from the Laplace-Beltrami operator.

We put this together to perform surface reconstruction with an energy composed of three parts,

$$\underset{\mathbf{X}}{\operatorname{argmin}}\, E_n + \lambda_b E_b + \lambda_l E_l. \qquad (13)$$

Here $E_n = \|\mathbf{X}\mathcal{L} + \mathbf{N}\mathbf{H}^k\|^2$ is the normal energy derived from the Laplacian discussion where $\mathbf{H}^k$ is a diagonal matrix of the vertex mean curvature integrals H_j from the face model at k-th iteration. $E_b = \|\mathbf{X}\mathcal{L}_b - \mathbf{X}^k\mathcal{L}_b\|^2$ 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 $\mathcal{L}_{b,jk} = 1/|\mathbf{e}_{jk}|$ where $|\mathbf{e}_{jk}|$ is the edge length connecting adjacent boundary vertices j and k. To avoid numerical drift accumulated over iterations, we include landmark deviation energy $E_l = \sum_i \|s_i\mathbf{R}_i[\mathbf{X}]_{\text{land}_i} + \mathbf{t}_i - \mathbf{W}_i\|_F^2$, which uses the landmark projection error to provide a global constraint on the face. The weights λ_b and λ_l are necessary to match units and balance the influences of the three terms. Unlike [16] we do not need a shadow region smoothing, since we already introduce dependability and use the template normal as a regularizer during normal estimation.

In each iteration, the optimization is achieved with a sparse linear system,

$$(\mathcal{L}^2 + \lambda_b\mathcal{L}_b^2 + \lambda_l\sum_i s_i^2\mathbf{C}_i\mathbf{C}_i^\intercal)\mathbf{X} = \mathbf{N}\mathbf{H}^k\mathcal{L} + \lambda_b\mathbf{X}^k\mathcal{L}_b + \frac{\lambda_l}{n}\sum_i s\mathbf{R}_i^\intercal(\mathbf{t} - \mathbf{W}_i)\mathbf{C}_i^\intercal, \qquad (14)$$

where $\mathbf{C}_i \in \mathbb{R}^{p\times q}$ is a sparse selection matrix. Each column of $\mathbf{C}_i$ has a single 1 indicating the vertex index selected through landmark marching for the corresponding 2D landmarks for image i.

3.5 Adaptive Mesh Resolution

Additionally, we propose a coarse-to-fine scheme for reconstruction. Starting with a low resolution mesh allows the reconstruction process to find the low frequency features in an efficient manner. Then, as the resolution increases, we can decrease the surface normal regularization to find the higher frequency details, while increasing the landmark reconstruction constraint to ensure the low frequency details maintain their position.

Here we describe the engineering details of the approach and present how all the steps fit together in Algorithm 1. After personalizing the face model, we use ReMESH [63] to uniformly resample the personalized mesh $\mathbf{X}^0$ to a coarse mesh with $6,248$ $(= p)$ vertices. The resampling is done once offline on the mean shape and is transferred to a personalized mesh by using the barycentric coordinates of the corresponding triangle on the original mesh for each coarse mesh vertex. Within each resolution, steps 2 and 3 are repeated until the surface converges. After convergence, one step of Loop subdivision [64] increases the resolution of the mesh, multiplying the number of vertices by 4. Moving from the coarse to fine level, we increase the localization selectivity by altering ϵ and we lower the template normal regularization λ_n (Sec. 4.1.3) to rely more on the observed images. This helps the coarse reconstruction stay smooth and fit the generic structure while allowing the fine reconstruction to capture the details.

3.6 SSIM Quality Measure

Accurately measuring reconstruction quality in the absence of ground truth data is a difficult task. Even with a ground

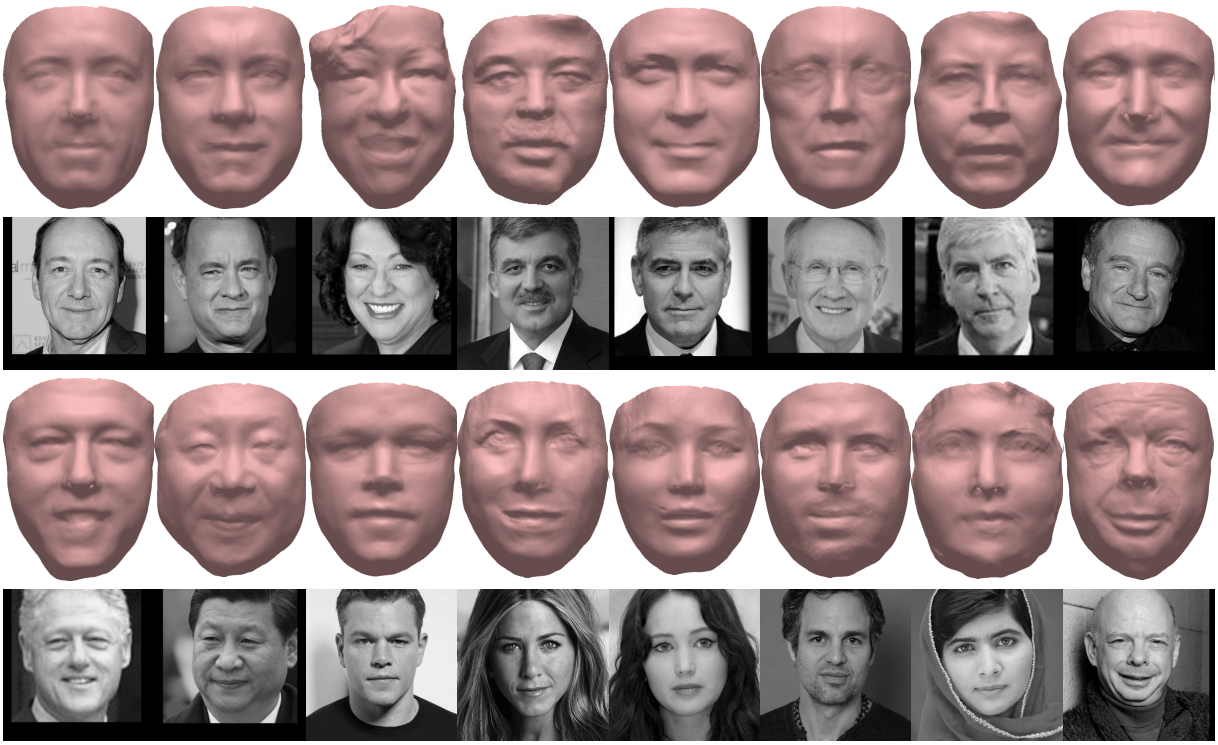


Fig. 7. Qualitative evaluation of 16 individuals from Internet photo collections. Note the diversity in ages, ethnicities and genders.

truth face scan, each popular surface-to-surface distance measurement has its flaws. Typically, surfaces are aligned through iterative closest point and for each vertex on the reconstructed surface, the error is reported as the minimum distance to the ground truth surface and not the corresponding semantic vertex. Such a measurement captures the overall similarity but places little emphasis on high frequency details like wrinkles. In light of this, the angle between the surface normals for the closest point is sometimes reported. However, when ground truth scans are not available, the surface reconstruction accuracy cannot be measured directly and must instead be measured indirectly. We desire the indirect measure to have two properties: 1) surface reconstruction errors should be evident in the score, 2) it should align with human perception of the reconstruction.

Considering the case where the only available information is the photo collection itself, we propose to measure the reconstruction accuracy indirectly by using the reconstructed model to render synthetic images under the same conditions as the real images and measuring the difference. If the image conditions (pose and illumination) and albedo are known, this will satisfy the first property since any change to the surface will result in a change to the rendered image. However, we are using the estimated conditions and albedo, and for a single image collection, it is trivial to change the albedo for any surface to produce an identical synthetic image to the real image. Fortunately, for multi-image collections (with different poses), property one is satisfied. To satisfy the second property, we use SSIM as the comparison measure because, as mentioned in Sec. 3.3.4, SSIM was developed to mimic human perception. We will verify this relationship in Sec. 4.2.2.

The SSIM quality measure for a reconstruction is given as follows. For each raw image from the photo collection, a synthetic image is rendered using the image-specific pose and lighting condition with the global albedo and surface estimate. The images are cropped tightly to the bounding box of the face in the synthetic image and the background of the synthetic image is filled in with the background from the raw image (Fig. 5). A single SSIM value from each image (mean of the pixel-wise SSIM scores) forms a set of scores for the collection. Two collections are compared directly by calculating if there is a significant difference between the two means of the collections using a p-value of 0.01. Globally, the mean SSIM for the set provides an overall quality of the reconstruction.

4 Experimental Results

We run a variety of experiments in order to qualitatively and quantitatively compare the proposed approach to prior face reconstruction work. For baselines, we only compare against other photo collection targeted approaches which use photometric stereo-based approaches [14], [16]. Stereo imaging and video-based approaches are not compared against since they can make use of the additional temporal information. Furthermore, since the proposed approach uses 3DMM fitting for Step 1 to personalize the template, we do not compare against other 3DMM fitting approaches, since any state-of-the-art 3DMM technique can be used in place for initialization. We also present results exploring the effectiveness of different parts of the reconstruction process.

4.1 Experimental Setup

4.1.1 Data Collection

We collect three distinct types of photo collections in this work. First, *Internet* photo collections. For these, we use the Bing image search API with a person's full name to fetch a

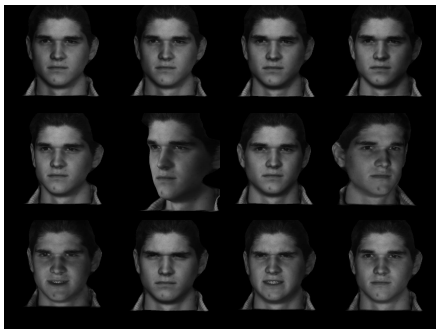


Fig. 8. Synthetic data with lighting (top), pose (middle), and expression (bottom) variation.

set of images. Occasionally images of the wrong person are included in these collections due to incorrect search results or more than one person being in an image. As long as this is infrequent, these images may be ignored through local selection. Second, *synthetic* images are rendered from subject M001 of the BU-4DFE database [65] using the provided texture and selecting random frames from the 6 expression sequences (Fig. 8). A Lambertian lighting model re-illuminates the face with light sources randomly sampled from a uniform distribution in front of the face. Third, *personal* photo collections. For these, we ask a person to provide a set of their own personal photos from social media or their phones photo gallery with it pre-cropped to remove other people from the images. In all cases, we use Bob [48] in Python to detect, crop, and scale faces as described in Sec. 3.1.1. Ground truth scans are captured for personal collections with a Minolta Vivid 910 range scanner at VGA resolution capturing 2.5D depth scans accurate to ± 0.03mm. Given frontal and both 45° yaw scans, we stitch them together using MeshLab to create a full 3D model.

4.1.2 Metrics

We use two different quantitative metrics. For Internet collections where we do not have a ground truth face shape, we use structural similarity (SSIM) as a proxy measurement of the reconstruction error, detailed in Sec. 3.6. For personal collections where we have a ground truth surface, we compute the average surface to surface distance. Both surfaces are roughly aligned by Procrustes superimposition of the 3D landmarks from the internal part of the face and ICP finalizes the alignment. The normalized vertex error is computed as the distance between a vertex in the reconstructed mesh and the closest vertex in the ground truth surface divided by the eye-to-eye distance. We report the average normalized vertex error. As a baseline, the mean face of the 3DMM has an average of 4.58% error to the ground truths.

4.1.3 Parameters

The parameters for the algorithm are set as follows: $\tau = 0.005$, $\lambda_l = 0.01$, $\lambda_b = 10$, $\lambda_n = [1, 0.1, 0.01]$, square error $\epsilon = [0.2, 0.08, 0.08]$, and SSIM error $\epsilon = [0.65, 0.65, 0.65]$ for coarse, medium, and fine resolution respectively.

4.2 Internet Results

4.2.1 Qualitative Evaluation

We begin by presenting qualitative results of the proposed method on a diverse set of subjects, spanning multiple ethnicity and both genders. While qualitative results are subjective and hard to compare with existing approaches, they do provide an overview of what types of details are captured in the reconstruction, whereas numerical surface-to-surface measurements sometimes lose perspective of the reconstruction quality. We strive not only for metrically correct reconstructions, but also for visually compelling reconstructions. In Figure 7, we present a large sample of reconstructions from Internet photo collections. The reconstructions are visually compelling and were generated using anywhere from 25 to 100 images per person. Note the ability to even reconstruct hairstyles, which are neither included in the 3DMM nor explicitly considered in our approach. However, we do see that facial hair often creates difficulty for the reconstruction since it is hard to establish correspondence with the same surface normal across images.

To visually place the proposed approach in comparison with prior work, we show reconstructions of the sample with four celebrities used in [14] and [16], George Clooney (99 photos), Kevin Spacey (143), Bill Clinton (179), and Tom Hanks (255). Figure 9 presents a side by side comparison between the various approaches. Due to the subjective nature of these reconstructions, we only comment on the fact that we reconstruct with the largest surface area, with similar visual appearance if not better. For example, by including areas outside of the internal face features, we capture the wrinkles to the sides of Clooney's eyes, and the smile lines on Spacey's cheek.

4.2.2 SSIM Quality Evaluation

We seek to validate the hypothesis that the proposed SSIM quality measure is consistent with human perception of reconstruction quality. To this end, we design the following experiment. For a total of 22 subjects, we collect Internet collections querying 100 images per person; after Python face detection filtering, this leaves us with 53 images per person on average. Face reconstruction is performed on the collection and the final shape and albedo are rendered under five viewpoints. The set of SSIM scores for each collection is obtained as described in Sec. 3.6. Since human perception of reconstruction quality is subjective, it is difficult to ask humans for a single number ranking the quality of each reconstruction. Therefore, we use an easier question where we present a pair of reconstructions and ask which "is a more visually compelling reconstruction". The human may answer "top", "bottom", or "equal" for a score of $1, -1, 0$ respectively. An example image pair is given in Fig. 10. Six random sets of 50 comparisons are given to different pairs of human evaluators. The average human-to-human correlation within each set is 0.63.

PageRank [66] can provide a global human ranking based on the comparisons. We create a graph with subjects as vertices and decisions as directed edges from the less compelling to the more compelling subject. PageRank produces a probability score for visiting a subject along random walks through the graph and has succeeded in sports teams

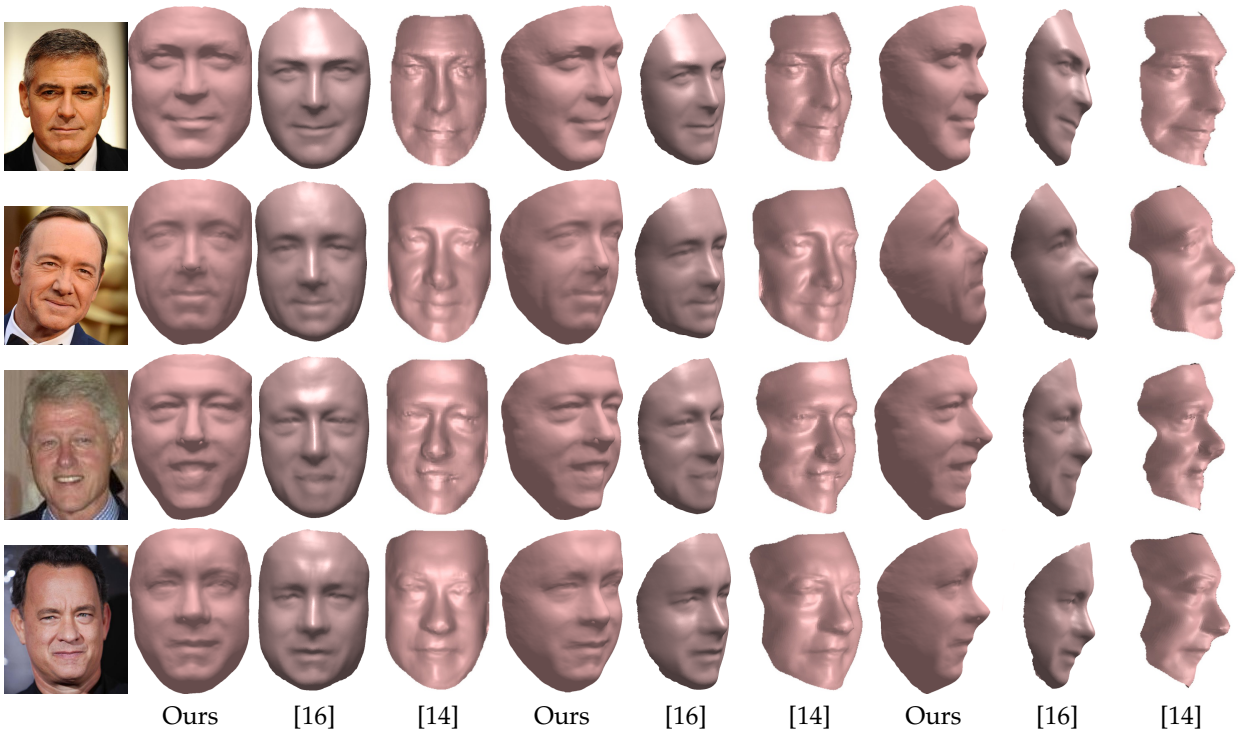


Fig. 9. Qualitative comparison on celebrities. The proposed approach incorporates more of the sides of the face and neck.

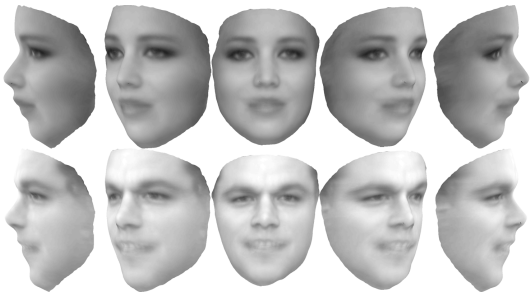


Fig. 10. Sample rendering used for human perception experiment.

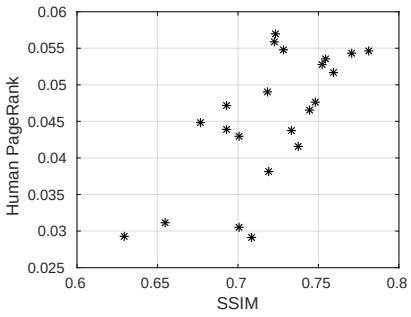


Fig. 11. Comparing human-based PageRank scores to SSIM.

Fig. 12. Best (top) and worst (bottom) reconstructions as determined by human (a) and SSIM (b).

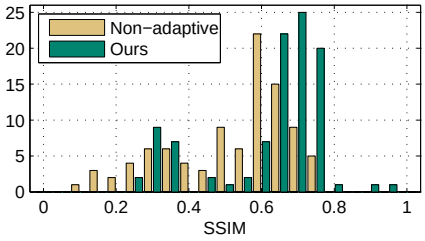


Fig. 13. Histograms of reconstruction performance.

ranking [67]. Figure 11 compares human and SSIM scores, with a correlation of 0.69 indicating SSIM is equivalent to a single human evaluation. Figure 12 shows the best and worst subjects as determined by human and SSIM.

SSIM allows large-scale comparison with prior work. Internet collections for 100 actors, singers, or politicians are captured by querying 50 images per person with an average of 28 images remaining after pre-processing. Comparing against [16], Fig. 13 plots histograms of SSIM quality scores. We see a clear improvement for the proposed method.

One interesting note is the bimodal distribution of the scores. One mode at 0.7 is similar to that observed in Fig. 11, and the other at 0.3 can be viewed as complete failures. We show an example collection in Fig. 14. While it is hard to identify a common trend, observed failure collections contain strong specular reflection, wide age range, cartoon images, and repeated images. Future work can explore identifying these conditions, and automatically filtering out the problematic images before reconstruction.

4.2.3 *Adaptability*

We look at adaptability with respect to two different factors. One, the number of images in the photo collection.

Fig. 14. A complete failure Internet image collection.

A major critique of prior SVD-based photometric stereo reconstructions is their dependence on a large number of images, typically over one hundred, which is too many for numerous applications. Two, the resolution of the images. By default, we have scaled all images to the same size $\sim$110 pixel eye-to-eye distance. We desire to know how well the proposed approach works for very low resolution images at $\sim$20 pixel eye-to-eye distance.

Figure 15(a) shows the adaptability of the reconstruction for George Clooney. As the number of images increases, the reconstruction becomes cleaner, but the overall details are still present with few images. We also test reconstruction for low resolution images and find it is able to capture wrinkles on the forehead since the sampling across multiple images acts as super-resolution.

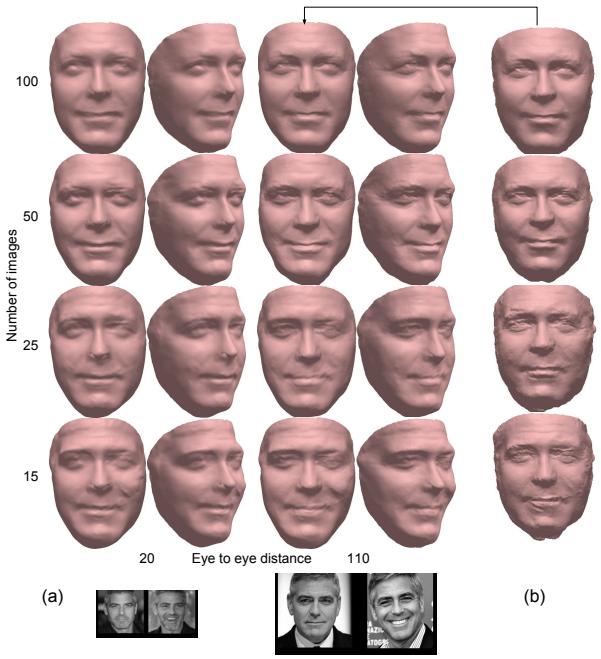


Fig. 15. (a) George Clooney with different numbers and quality images. (b) Reconstruction without coarse-to-fine process.

4.3 Synthetic Results

The synthetic dataset allows testing under known assumptions to see robustness to pose and expression *independently*. We generate three sets of 50 images: frontal with neutral expression, neutral expression with random yaw angles between $\pm 30^\circ$, and frontal with random expressions (Fig. 8). Error is reported as the surface-to-surface distance to the neutral expression model. Table 3 shows the proposed method outperforms prior work in all scenarios, and is more robust to pose than expression variation.

4.4 Personal Results

4.4.1 Local Selection

We explore the effects of local selection on the personal photo collections. There are 10 personal photo collections ranging from 6 to 50 images with a median of 24. Table 4 shows the different choices for local selection showing that local selection improves performance with SSIM performing better than square error. Exploring why SSIM performs better, Tab. 5 shows the performance based on the window size γ, or the size of the local area to consider. When γ is very small, it behaves similar to the square error method where only a single point on the face contributes to the selection. The error decreases as the selection area increases until it is too broad of an area.

4.4.2 Adaptability

We perform a *thorough experiment* comparing the adaptability of the proposed method to the SVD-based approach of [16]. We split each photo collection into 4 sizes, 25%, 50%, 75%, 100% of the images and use the high and low resolution. The results for all 10 photo collections are averaged together in Table 6. The proposed method performs better for all collection sizes, and adapts better to small collections. While the SVD-based approach degrades by 1.44%, the proposed method only degrades by 0.46%. Figure 17 shows substantial improvement to the chin and forehead. Figure 16 demonstrates reconstruction quality for different size photo collections. More images reduce noise and add details, but small collections are still recognizable.

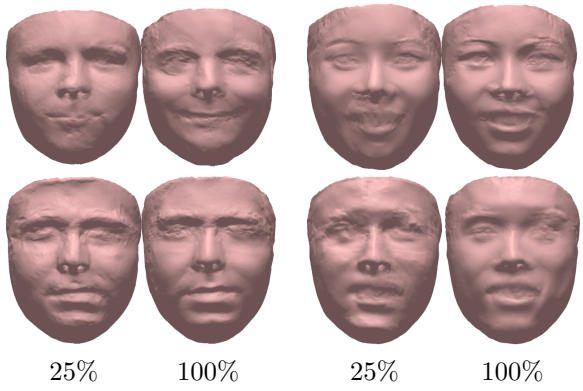


Fig. 16. Adaptability on personal photo collections. Sample reconstructions on quarter and full-size collections.

4.5 Discussions

Efficiency Written in a mixture of C++ and Matlab, the algorithm runs on a commodity PC with an Intel $i7$-$4770k$ 3.5 GHz CPU and 8 GB RAM. 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 2, 11, and 45 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 < 1.5 minutes.

TABLE 3
Synthetic Surface-to-Surface Error.

Method	Neutral	30° Yaw	Expression
Ours	**3.22%**	**3.82%**	**4.40%**
[16]	6.13%	7.48%	6.59%

TABLE 4
Local Selection Error.

Method	None	Square Error	SSIM
Error	4.57%	3.93%	**3.58%**

TABLE 5
SSIM Radius Error.

γ	0.5	1.5	2.5	3.5
Error	4.11%	4.81%	**3.58%**	4.86%

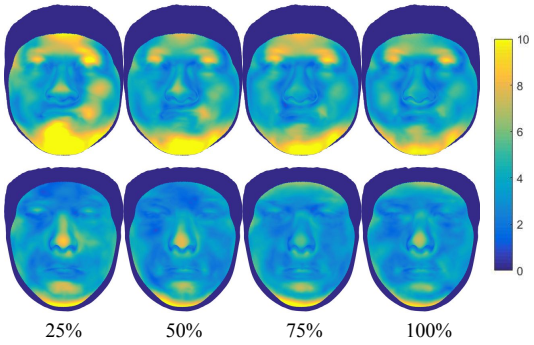


Fig. 17. Surface-to-Surface errors by location for personal reconstructions using [16] (top) and the proposed method (bottom).

TABLE 6
Personal Collection Adaptability.

% of Images	25	50	75	100
Ours - Low	4.10%	3.85%	3.78%	3.65%
Ours - High	4.04%	3.54%	3.53%	3.58%
[16] - Low	5.54%	4.78%	4.63%	4.34%
[16] - High	5.56%	4.77%	4.70%	4.10%

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, the Clooney set takes 4 iterations to converge for a total time of 3.7 minutes, more than double the time. Also, Fig. 15(b) shows that the resultant reconstructions are similar for large collections, but noisy for small collections since the coarse step allows for more template regularization.

5 Conclusions

We presented an approach for reconstructing a wrinkle-level 3D face model from an unconstrained 2D photo collection, adapting to the quantity and quality of images present. Incorporating prior face knowledge through a 3DMM and an adaptive regularization allows the method to work on smaller photo collections, and the novel structural similarity-based local selection improves performance in the presence of occlusions. Our coarse-to-fine scheme first reconstructs a smooth yet accurate model and then adds in the details present in the collection. The resulting reconstructions have applications for improving face recognition, landmark alignment, and consumer entertainment.

6 Acknowledgments

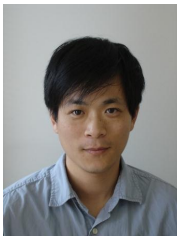
An earlier version of this work appeared in CVPR [68].

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*, vol. 54, no. 10, pp. 105–112, 2011.
[2] X. Liu and T. Chen, "Pose-robust face recognition using geometry assisted probabilistic modeling," in *CVPR*, vol. 1, 2005, pp. 502–509.
[3] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D facial expression recognition based on primitive surface feature distribution," in *CVPR*, vol. 2. IEEE, 2006, pp. 1399–1406.
[4] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. Graph.*, vol. 33, no. 4, p. 43, 2014.
[5] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/off: live facial puppetry," in *SCA*, 2009, pp. 7–16.
[6] 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*, 2015, pp. 787–796.
[7] D. Zeng, Q. Zhao, and J. Li, "Examplar coherent 3D face reconstruction from forensic mugshot database," *J. Image Vision Computing*, 2016.
[8] T. Beeler, B. Bickel, P. Beardsley, R. Sumner, and M. Gross, "High-quality single-shot capture of facial geometry," *ACM Trans. Graph.*, vol. 29, no. 3, 2010.
[9] 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*, vol. 114, no. 8, pp. 942–951, 2010.
[10] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Computer Graphics Proceeding, Annual Conference Series*. New York: ACM SIGGRAPH, 1999, pp. 187 – 194.
[11] A. Jourabloo and X. Liu, "Large-pose face alignment via CNN-based dense 3D model fitting," in *CVPR*, 2016.
[12] I. Kemelmacher-Shlizerman, "Internet-based morphable model," in *ICCV*, 2013.
[13] R. Newcombe, D. Fox, and S. Seitz, "Dynamicfusion: Reconstruction and tracking on non-rigid scenes in real-time," in *CVPR*, 2015, pp. 343–352.
[14] I. Kemelmacher-Shlizerman and S. M. Seitz, "Face reconstruction in the wild," in *ICCV*, 2011.
[15] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, Aug. 1981, pp. 674–679.
[16] J. Roth, Y. Tong, and X. Liu, "Unconstrained 3D face reconstruction," in *CVPR*, 2015.
[17] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *ICCVW*, 2013, pp. 392–396.
[18] "Konica minolta vivid 9i non-contact 3D digitizer," http://www.dirdim.com/pdfs/DDI_Konica_Minolta_Vivid_9i.pdf.
[19] "Iiid scanner: Primesense carmine 1.09," http://www.robotshop.com/media/files/pdf/datasheet-sca001_1.pdf.
[20] 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.*, vol. 30, no. 4, pp. 75:1–75:10, 2011.
[21] C. Hernández, G. Vogiatzis, and R. Cipolla, "Multiview photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 548–554, 2008.
[22] V. Blanz and T. Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, 2003.
[23] L. Gu and T. Kanade, "A generative shape regularization model for robust face alignment," in *ECCV*, 2008.
[24] L. Zhang and D. Samaras, "Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, 2006.
[25] A. Jourabloo, X. Yin, and X. Liu, "Attribute preserved face de-identification," in *ICB*, 2015.
[26] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *CVPR*, 2016.
[27] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz, "Total moving face reconstruction," in *ECCV*. Springer, 2014, pp. 796–812.
[28] P. Garrido, L. Valgaerts, C. Wu, and C. Theobalt, "Reconstructing detailed dynamic face geometry from monocular video," *ACM Trans. Graph.*, vol. 33, no. 6, p. 158, 2013.

[29] 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*, vol. 20, no. 3, pp. 413–425, 2014.
[30] F. Shi, H.-T. Wu, X. Tong, and J. Chai, "Automatic acquisition of high-fidelity facial performances using monocular videos," *ACM Trans. Graph.*, vol. 33, no. 6, 2014.
[31] C. Cao, D. Bradley, K. Zhou, and T. Beeler, "Real-time high-fidelity facial performance capture," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 46:1–46:9, 2015.
[32] A. E. Ichim, S. Bouaziz, and M. Pauly, "Dynamic 3D avatar creation from hand-held video input," *ACM Trans. Graph.*, vol. 34, no. 4, p. 45, 2015.
[33] L. A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3D face alignment from 2D videos in real-time," in *FG*, 2015. [Online]. Available: articles/Jeni15FG_ZFace.pdf
[34] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR*, Jun. 2016.
[35] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, "3D scanning deformable objects with a single RGBD sensor," in *CVPR*, 2015, pp. 493–501.
[36] W. Peng, C. Xu, and Z. Feng, "3D face modeling based on structure optimization and surface reconstruction with b-spline," *Neurocomputing*, vol. 179, pp. 228–237, Feb. 2016.
[37] C. Yang, J. Chen, N. Su, and G. Su, "Improving 3D face details based on normal map of hetero-source images," in *CVPRW*. IEEE, 2014, pp. 9–14.
[38] P. Snape, Y. Panagakis, and S. Zefeiriou, "Automatic construction of robust spherical harmonic subspaces," in *CVPR*, 2015.
[39] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 139–144, 1980.
[40] H. Hayakawa, "Photometric stereo under a light source with arbitrary motion," *J. Optical Soc. America A.*, vol. 11, no. 11, pp. 3079–3089, 1994.
[41] 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*, vol. 35, pp. 203–222, 1999.
[42] K. Lee, J. Ho, and D. Kriegman, "Nine points of light: Acquiring subspaces for face recognition under variable lighting," in *CVPR*, 2001, pp. 129–139.
[43] R. Basri and D. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, 2003.
[44] R. Basri, D. Jacobs, and I. Kemelmacher, "Photometric stereo with general, unknown lighting," *Int. J. Comput. Vision*, vol. 72, no. 3, pp. 239–257, 2007.
[45] 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*, 2010, pp. 703–717.
[46] B. Shi, K. Inose, Y. Matsushita, and P. Tan, "Photometric stereo using internet images," in *3DV*, 2014, pp. 361–368.
[47] M. Piotraschke and V. Blanz, "Automated 3d face reconstruction from multiple images using quality measures," in *CVPR*, 2016.
[48] 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*. ACM Press, 2012, pp. 1449–1452.
[49] K. E. Spaulding, A. C. Gallagher, E. B. Gindele, and W. Ptucha, Raymond, "Constructing extended color gamut images from limited color gamut digital images," U.S. Patent 7 308 135, 2007.
[50] J. Deng, Q. Liu, J. Yang, and D. Tao, "M3 CSR: multi-view, multi-scale and multi-component cascade shape regression," *J. Image Vision Computing*, vol. 47, pp. 19–26, Mar. 2016.
[51] H. Fan and E. Zhou, "Approaching human level facial landmark localization by deep learning," *J. Image Vision Computing*, vol. 47, pp. 27–35, Mar. 2016.
[52] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *CVPR*. IEEE, 2011, pp. 545–552.
[53] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *J. Image Vision Computing*, vol. 47, pp. 3–16, Mar. 2016.
[54] C. Qu, E. Monari, T. Schuchert, and J. Beyerer, "Adaptive contour fitting for pose-invariant 3D face shape reconstruction," in *BMVC*, 2015.
[55] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *AVSS*, 2009.
[56] B. Chu, S. Romdhani, and L. Chen, "3D-aided face recognition robust to expression and pose variations," in *CVPR*, 2014, pp. 1899–1906.
[57] A. M. Bruckstein, R. J. Holt, T. S. Huang, and A. N. Netravali, "Optimum fiducials under weak perspective projection," *Int. J. Comput. Vision*, vol. 35, no. 4, pp. 223–244, 1999.
[58] D. Frolova, D. Simakov, and R. Basri, "Accuracy of spherical harmonic approximations for images of lambertian objects under far and near lighting," in *ECCV*, 2004, pp. 574–587.
[59] Z. Wang, A. C. Bovik, and H. R. Sheikh, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
[60] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel, "Laplacian surface editing," in *Proc. of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*. ACM, 2004, pp. 175–184.
[61] U. Pinkall and K. Polthier, "Computing discrete minimal surfaces and their conjugates," *Experimental mathematics*, vol. 2, no. 1, pp. 15–36, 1993.
[62] Y. Yu, K. Zhou, D. Xu, X. Shi, H. Bao, B. Guo, and H.-Y. Shum, "Mesh editing with poisson-based gradient field manipulation," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 644–651, 2004.
[63] M. Attene and B. Falcidieno, "ReMESH: An interactive environment to edit and repair triangle meshes," in *SMI*, 2006, pp. 271–276.
[64] A. Jacobson *et al.*, "gptoolbox: Geometry processing toolbox," 2015, http://github.com/alecjacobson/gptoolbox.
[65] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *FG*, 2008.
[66] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: bringing order to the web," 1999.
[67] A. Y. Govan and C. D. Meyer, "Ranking national football league teams using google's pagerank," in *MAM*, 2006.
[68] J. Roth, Y. Tong, and X. Liu, "Adaptive 3d face reconstruction from unconstrained photo collections," in *Proc. IEEE Computer Vision and Pattern Recognition*, Las Vegas, NV, June 2016.

Joseph Roth received the Ph.D. degree with the Computer Vision Lab from the Department of Computer Science and Engineering at Michigan State University in 2016. He received the B.S. in Computer Science from Grand Valley State University with greater honors in 2010. His research interests are computer vision and face modeling. He is now a machine perception researcher with Google.



Yiying Tong received the Ph.D. degree from the University of Southern California in 2004. He is an Associate Professor at Michigan State University. Prior to joining MSU, he was a postdoctoral scholar at Caltech. His research interests include discrete geometric modeling, physically based simulation/animation, and discrete differential geometry. He received the US National Science Foundation (NSF) Career Award in 2010. He is a member of the IEEE.

Xiaoming Liu is an Assistant Professor in the Department of Computer Science and Engineering at Michigan State University (MSU). He received the Ph.D. degree in Electrical and Computer Engineering from Carnegie Mellon University in 2004. Before joining MSU in Fall 2012, he was a research scientist at General Electric Global Research Center. His research areas are face recognition, biometrics, image alignment, video surveillance, computer vision and pattern recognition. He has authored more than 100 publications, and has filed 22 U.S. patents. He is a member of the IEEE.