Unconstrained 3D Face Reconstruction

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

Abstract

This paper presents an algorithm for unconstrained 3D face reconstruction. The input to our algorithm is an "unconstrained" collection of face images captured under a diverse variation of poses, expressions, and illuminations, without meta data about cameras or timing. The output of our algorithm is a true 3D face surface model represented as a watertight triangulated surface with albedo data or texture information. 3D face reconstruction from a collection of unconstrained 2D images is a long-standing computer vision problem. Motivated by the success of the state-of-the-art method, we developed a novel photometric stereo-based method with two distinct novelties. First, working with a true 3D model allows us to enjoy the benefits of using images from all possible poses, including profiles. Second, by leveraging emerging face alignment techniques and our novel normal field-based Laplace editing, a combination of landmark constraints and photometric stereo-based normals drives our surface reconstruction. Given large photo collections and a ground truth 3D surface, we demonstrate the effectiveness and strength of our algorithm both qualitatively and quantitatively.

1. Introduction

Obtaining a user-specific 3D face surface model is useful for a variety of applications, such as 3D-assisted face recognition [7, 19, 28], 3D expression recognition [37], and facial animations [9]. Despite emerging 3D sensors to acquire 3D faces, accurately *reconstructing the 3D surface model from 2D images* is a long-standing computer vision problem.

There are numerous scenarios for face reconstruction. With access to the subject and a controlled environment, many techniques provide highly detailed models. For example, stereo imaging is applicable when two calibrated cameras capture images simultaneously, or photometric stereo with a controlled lighting array. The task becomes more difficult given unconstrained face images captured in an unstructured scenario. If the images are from a video sequence, structure from motion is useful when key points are tracked across the face. If the input is a collection of unconstrained face photos, photometric stereo works across areas of consistent albedo and estimates fine surface details.

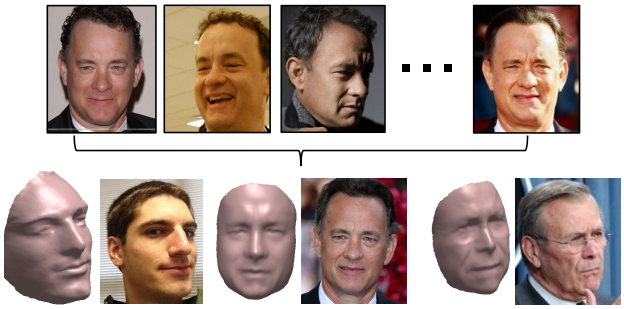


Figure 1. Given Tom Hanks' photo collection with pose, expression, and illumination variations, our system performs surface reconstruction, shown along with a real photo at the same viewpoint.

Among the aforementioned scenarios, this paper targets the case of the *unconstrained photo collection*. As shown in Fig. 1, given a collection of unconstrained face photos of one subject, we would like to reconstruct the 3D face surface model, despite the diverse variations in Pose, Illumination, and Expression (PIE). This is certainly a very challenging problem, as we do *not* have access to stereo imaging [36] or video [35, 12]. Kemelmacher-Shlizerman and Seitz developed an impressive photometric stereo-based method to produce high-quality face models from photo collections [23], where the recovering of a locally consistent shape was intelligently achieved by using a different subset of images. However, there are still two limitations in [23]. One is that mainly near-frontal images are selected to contribute to the reconstruction, while the consensus is that non-frontal, especially profile, images are highly useful for 3D reconstruction. The other is that due to surface reconstruction on a 2D grid, a 2.5D height field, rather than a full 3D model, is produced.

Motivated by the state-of-the-art results of photometric stereo-based methods, as well as amendable limitations, this paper proposes a novel approach to 3D face reconstruction. Our approach is also motivated by the recent explosion of face alignment techniques [27, 39, 21, 31], where the preci-

sion of 2D landmark estimation has been substantially improved. Specifically, given a collection of unconstrained face images, we first perform 2D landmark estimation [39] of each image. In order to prepare an enhanced 3D template for the photometric stereo, we deform a generic 3D face template such that the projections of its 3D landmarks are consistent with the estimated 2D landmarks on all images, and the surface normals are maintained. With the enhanced 3D template, 2D face images at all poses are back projected onto the 3D surface, where the collection of projections will form a data matrix spanning all vertices of the template. Since there are inevitably missing elements in the data matrix due to varying poses, matrix completion is employed and followed by the shape and lighting decomposition via SVD. With the estimated surface normals, we further deform the 3D shape such that the updated shape will have normals similar to the estimated ones, under the same landmark constraint and an additional boundary constraint. To illustrate the strength of our approach, we perform experiments on several large collections of celebrities, as well as one subject where the ground truth 3D model is collected. Both qualitative and quantitative experiments are conducted and compared with the state-of-the-art method.

In summary, this paper has made three contributions.

- A true 3D facial surface model is generated. During the iterative reconstruction, we perform the photometric stereo on the *entire 3D surface*, and the 3D coordinates of all vertices are updated toward the specific shape of an individual. As a benefit of a 3D surface model, our approach allows faces from *all* poses, including the profiles, to contribute to the reconstruction.

- Our surface reconstruction utilizes a combination of photometric stereo-based normals and landmark constraints, which leverages the power of emerging face alignment techniques. It also strikes a good balance of allowing facial details to deform according to the photometric stereo, while maintaining the consistency of the overall shape with 2D landmarks.

- In order to achieve the deformation of a template using normal estimates, we develop a novel Laplace mesh editing with surface normals as input, while prior mesh editing use mean curvature normal as input.

2. Related Work

Face reconstruction Face reconstruction, the process of creating a 3D model of a face from 2D image(s), is a difficult problem with many applications. In the biometrics community, 3D face models are invariant to pose, illumination, and background and allow warping of the image to new poses to improve face recognition. The graphics and animation community desire highly detailed models including skeletal structure for expressions. With a cooperative subject, there are range scanners, multi-camera stereo [4, 5], or photometric stereo with light arrays [16]. In-the-wild techniques that can operate on a single image include 3D morphable models [7], shape from shading [24], and warping to a reference shape [14]. Recently combinations of techniques are proposed to capture finer details, [40] uses a 3DMM combined with a deformable model to obtain wrinkle details and [32] combines a noisy stereo reconstruction with photometric stereo for generic objects and scenes.

Photometric stereo Photometric stereo is the process of recovering a 3D shape from a set of 2D images based on the differences in shading caused by lighting conditions. Some works require the knowledge of the lighting conditions [13, 16]. Others estimate the lighting conditions taking advantage of that, for a Lambertian model, the images lie on a low rank subspace [15, 42, 25, 2, 3, 38]. But these works operate in a constrained setting where point correspondences between images are known. Recently, a novel photometric stereo-based method reconstructs from a set of in-the-wild face images [23], as well as a video sequence [35]. With the exception of [13], these works reconstruct a depth map or 2.5D surface instead of a 3D surface since they tend to operate on an object with a *narrow range* of poses. By using an unconstrained photo collection with a wide range of poses, we can reconstruct a 3D face with more accurate depths due to wide-angle stereo, and also wraps accurately around the cheeks and chin.

Surface reconstruction Surface reconstruction is a challenging process that varies significantly depending on the nature of input (noise, outlier, etc.), output (mesh, skeleton, volume, etc.), and types of shape (man-made, organic, etc.). The majority of reconstruction algorithms take a point cloud as their input, including methods with surface-smoothness priors (*e.g.*, tangent planes [17], moving least squares [1], radial basis function [10], Poisson surface reconstruction [22]), visibility-based methods [11], data-driven methods [29], etc. More details on the topic are in the state of the art report [6]. One of the most widely used methods is the Poisson surface reconstruction [22] due largely to its efficiency and reliability. This method estimates a volumetric normal field based on the point cloud, and constructs a 3D Poisson equation akin to the 2D Poisson equation resulting from photometric stereo. However, the requirement for the point cloud input renders the 3D Poisson surface reconstruction not directly applicable to the normal field and landmark constraints in our case. Thus, we resort to a template deformation approach, and fit the template to this input while maintaining the global structure of the template 3D face. Our technique is based on the gradient domain methods called the Poisson/Laplace mesh editing [33, 41], where the mean curvature normal fields are given, and the surface is deformed under additional (land-

Table 1. Common notations.

Symbol	Dim.	Description
$\mathbf{I}$		image
q	scalar	number of landmarks
n	scalar	number of images
p	scalar	number of vertices
$\mathbf{Q}$	$3\times q$	3D landmarks
$\mathbf{W}_i$	$2\times q$	2D image landmarks
$\mathbf{P}_i$	2×3	image projection matrix
$\mathbf{F}$	$n\times p$	backprojected photo collection
$\mathbf{M}$	$n\times p$	matrix completed photo collection
$\mathbf{L}$	$n\times 4$	lighting matrix
$\mathbf{S}$	$4\times p$	shape matrix
$\mathbf{S}^t$	$4\times p$	template shape matrix
$\mathbf{X}_0$	$3p$	template shape vector
$\mathbf{X}$	$3p$	shape vector
$\mathbf{H}$	$3p$	mean curvature normal

mark) constraints. Our method differs from the existing variants of Laplace mesh editing in that we are only given the normal fields, and have to infer the mean curvature.

3. Proposed Algorithm

The proposed algorithm operates on a photo collection of an individual. No constraints are placed regarding poses or expressions for the images, but it is assumed that the collection contains a variety of, albeit unknown, lighting conditions. An initial generic face template mesh including labeled 3D landmark locations is also given. We assume weak perspective camera projection and Lambertian reflection.

Table 1 provides notations for the major variables used throughout the algorithm. Figure 2 illustrates the major components and pipeline of our proposed algorithm.

3.1. 2D Landmark Alignment

Proper 2D face alignment is vital in providing registration among images in the photo collection and registration with the 3D template, although the proposed approach is robust to a fair amount of error. We employ the state-of-the-art cascade of regressors approach [39] to automatically fit q (=68) landmarks onto each image. An example of the landmark fitting is given in Figure 2. Given an image $\mathbf{I}(x, y)$, the landmark alignment returns a $2\times q$ matrix $\mathbf{W}_i$.

3.2. Landmark Driven 3D Warping

The initial template face is not nearly isometric to the individual face, *e.g.*, the aspect ratio of the face may be different and, as such, it will not fit closely to the images even in the absence of expression. Therefore, it is highly desirable to warp the initial template toward the true 3D shape of the individual so that the subsequent photometric stereo can have a better initialization.

Since the estimated 2D landmarks provide the correspondences of q points between 3D and 2D as well as across images, they should be leveraged to guide the template warping. Based on this observation, we aim to warp the template in a way such that the projections of the warped 3D landmark locations can match well with the estimated 2D landmarks. The technique we use is based on Laplacian surface editing [33] and adapted for the landmark constraints. Specifically, in order to maintain the shape of the original template face while reducing the matching error from the 3D landmarks to the 2D landmarks, we minimize the following energy function,

$$\int_\Omega \|\Delta\mathbf{x} - \Delta\mathbf{x}_0\|^2 + \lambda_l \sum_i \|\mathbf{P}_i\mathbf{Q} - \mathbf{W}_i\|^2, \qquad (1)$$

where the first term measures the deviation of the Laplace-Beltrami operator Δ (trace of Hessian) of the deformed mesh $\mathbf{x}$ from that of the original mesh $\mathbf{x}_0$ integrated over the entire surface Ω; the second term measures the squared distance between the set of 3D landmarks weakly perspective projected through $\mathbf{P}_i$ and the 2D landmark locations $\mathbf{W}_i$ for image i; and λ_l is the weight for landmark correspondence. Note that the operator Δ measures the difference between a function's value at a vertex with the average value at the neighboring vertices, so the minimization of the first term helps maintain the geometric details.

To solve Eq 1, we discretize the surface patch Ω as a triangle mesh with p vertices, with the vertex locations concatenated as a $3p$-dimensional vector $\mathbf{X}$. Throughout the deformation process, we keep the connectivity of the vertices (i.e., which triplets form triangles) fixed and the same as the given template mesh. We deform the mesh only through modifications to the vertex locations. Eq 1 is thus turned into a quadratic function on $\mathbf{X}$,

$$E_{\text{warp}}(\mathbf{X},\mathbf{P}_i) = \|\mathcal{L}\mathbf{X} - \mathcal{L}\mathbf{X}_0\|^2 + \lambda_l \sum_i \|\mathbf{P}_i\mathbf{D}_i\mathbf{X} - \mathbf{W}_i\|^2,$$

where $\mathcal{L}$ is a discretization of Δ. Using linear finite elements, it is turned into a *symmetric* matrix with entries $\mathcal{L}_{ij} = \frac{1}{2}(\cot\alpha_{ij} + \cot\beta_{ij})$, where α_{ij} and β_{ij} are the opposite angles of edge ij in the two incident triangles (see Figure 3), known as the cotan formula [30], and $\mathbf{D}_i$ is the selection matrix picking out the landmarks that have a correspondence in image i, i.e., it is a diagonal matrix with 1's on the diagonal for the vertices corresponding to such a landmark and 0 everywhere else.

In order to find the minimizer, we first estimate an initial $\mathbf{P}_i$ via the corresponding pairs of 2D and 3D landmarks. With the projection matrices $\mathbf{P}_i$ fixed, the minimizer of the energy E_{warp} can be obtained by solving a linear system.

However the above procedure is not rotation invariant. As in [33, 41], we can resolve this issue by noting that

$$\Delta\mathbf{x} = -H\mathbf{n},$$

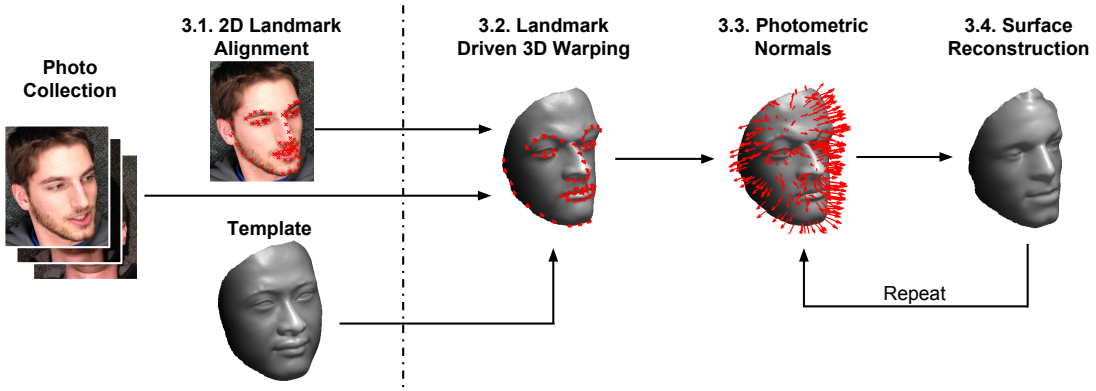


Figure 2. Overview of our 3D face reconstruction. Given a photo collection, a generic template face mesh, and 2D landmark alignment, we propose an iterative process to warp the mesh based on the estimated 3D landmarks and the photometric stereo-based normals.

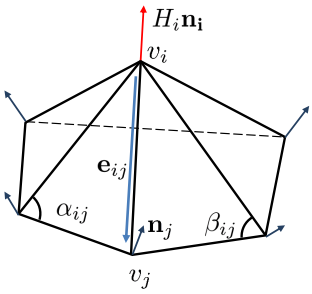


Figure 3. The mean curvature normal indicates how a vertex deviates from the average location of its immediate neighbors, which can be evaluated as the Laplacian of the position. The mean curvature H_i can be evaluated through $\mathbf{n}$.

which means that the Laplacian of the positions is the mean curvature H times the unit normal of the surface $\mathbf{n}$. The rotation-invariant geometric details are captured by the Laplacian operator and the mean curvature scalar H. Thus, to keep the original geometric detail while allowing it to rotate, we compute the original mean curvature H_0 (the discretization of which corresponds to the integral of the mean curvature in a neighborhood around each vertex), and update $\mathbf{n}^k$ according to the direction of $\mathcal{L}\mathbf{X}^k$ for the shape $\mathbf{X}^k$ at iteration k, and solve for

$$\mathbf{X}^{k+1} = \underset{\mathbf{X}}{\operatorname{argmin}}(\|\mathcal{L}\mathbf{X} + H_0\mathbf{n}^k\|^2 + \lambda_l \sum_i \|\mathbf{P}_i^k\mathbf{D}_i\mathbf{X} - \mathbf{W}_i\|^2),$$

which leads to a linear system,

$$(\mathcal{L}^2 + \lambda_l \sum_i \mathbf{D}_i(\mathbf{P}_i^k)^\intercal \mathbf{P}_i^k \mathbf{D}_i)\mathbf{X} = -\mathcal{L}H_0\mathbf{n}^k + \lambda_l \sum_i (\mathbf{P}_i^k)^\intercal \mathbf{W}_i. \quad (2)$$

In practice, the procedure of iteratively estimating $\mathbf{P}_i^k$ and $\mathbf{X}^{k+1}$ converges quickly in 10-20 iterations in our tests.

3.3. Photometric Normals

Fitting the landmarks allows for a global deformation of the template mesh toward the shape of the individual, but the fine details of the individual are not present. To recover these details, we use the photometric stereo with unknown lighting conditions, similar to the one described by Kemelmacher-Shlizerman and Seitz [23]. The approach in [23] estimates an initial lighting and shape based on the factorization of a 2D image set, and refines the estimate based on localized subsets of images that match closely to the estimate for a given pixel. One key difference is that in [23] the input to factorization is the frontal-projected 2D images of the 3D textures, rather than the collection of 3D texture maps themselves in our algorithm. That is, our photometric stereo is performed on the entire 3D surface. We present the photometric normal estimation as follows.

We assume a Lambertian reflectance along with an ambient term for any point x in an image,

$$\mathbf{I}_x = \rho_x(k_a + k_d\ell \cdot \mathbf{n}_x),$$

where k_a is the ambient weight, k_d is the diffuse weight, ℓ is the light source direction, ρ_x is the point's albedo, and $\mathbf{n}_x$ is the point's surface normal. We assemble the numbers into a row vector for the lighting $\mathbf{l} = [k_a, k_d\ell]$ and a column vector for the shape $\mathbf{s}_x = \rho_x[1, \mathbf{n}_x]^\intercal$, so that $\mathbf{I}_x = \mathbf{l}\mathbf{s}_x$.

3.3.1 Initial Normal Estimation

In this section, we assume point correspondence between the current mesh and each image in the collection. Thus, we store in $\mathbf{F}_{i,j}$ the reflectance intensity $\mathbf{I}_x$ corresponding to the projected location x of vertex j in image i. This correspondence is established by projecting the warped shape template onto the images via the $\mathbf{P}_i$ matrices from Section 3.2. For non-frontal images, there are vertices that are not visible due to the projection. When this occurs, we set $\mathbf{F}_{i,j} = 0$ and use matrix completion [26] to fill in the missing values to obtain $\mathbf{M}$. Our experiments show that given

an image set with diverse poses, missing data occurs at different areas of $\mathbf{F}$ and is handled well by matrix completion.

If we assemble $\mathbf{l}_i$ for image i into an $n\times4$ matrix $\mathbf{L}$, and the shape vector $\mathbf{s}_j$ for each vertex j into a $4\times p$ matrix $\mathbf{S}$, we have $\mathbf{M} = \mathbf{LS}$. To obtain the lighting and normal estimation $\mathbf{L}$ and $\mathbf{S}$ from $\mathbf{M}$, we use a typical photometric stereo technique knowing that a Lambertian surface will be rank-4 ignoring self-shadows. We factorize $\mathbf{M}$ via singular value decomposition (SVD) to obtain $\mathbf{M} = \mathbf{U\Lambda V^\intercal}$ and use the rank-4 approximation $\mathbf{M} = \tilde{\mathbf{L}}\tilde{\mathbf{S}}$ where $\tilde{\mathbf{L}} = \mathbf{U}\sqrt{\mathbf{\Lambda}}$ and $\tilde{\mathbf{S}} = \sqrt{\mathbf{\Lambda}}\mathbf{V^\intercal}$. $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{S}}$ are the same size as the desired lighting and shape matrices $\mathbf{L}$ and $\mathbf{S}$, but the factorization is not unique as any invertible 4×4 matrix $\mathbf{A}$ gives a valid factorization since $\mathbf{LS} = (\tilde{\mathbf{L}}\mathbf{A}^{-1})(\mathbf{A}\tilde{\mathbf{S}})$.

The ambiguity can be resolved up to a generalized bas-relief transform through integrability constraints, but [23] states that it may be unstable for images with expression variations. Thus, we follow the approach from [23], where we select images that are modeled well by the low rank approximation, *i.e.*, $\|\mathbf{M} - \tilde{\mathbf{L}}\tilde{\mathbf{S}}\|^2 < \epsilon$, and solve for $\text{argmin}_{\mathbf{A}} \|\mathbf{S}^t - \mathbf{A}\tilde{\mathbf{S}}\|^2$, where $\mathbf{S}^t$ is the shape matrix for the template shape. This allows us to then estimate the lighting and shape for the individual via $\mathbf{L} = \tilde{\mathbf{L}}\mathbf{A}^{-1}$ and $\mathbf{S} = \mathbf{A}\tilde{\mathbf{S}}$.

3.3.2 Albedo Estimation

The ambiguity recovery requires the template shape matrix $\mathbf{S}^t$ including a surface albedo component, which we estimate for the individual based on the photo collection. For a row $\mathbf{M}_i$ corresponding to image i, we know from our lighting assumption that each vertex is a linear combination of a shared light source direction and the surface normal scaled by albedo, *i.e.,* $\mathbf{M}_{i,j} = \rho_j \mathbf{L}_i \mathbf{n}_j$ for a vertex j, where ρ_j and $\mathbf{L}_i$ are unknown. We initialize all ρ_j to 1, and then solve iteratively until convergence for $\mathbf{L}_i$ by $\text{argmin}_{\mathbf{L}_i} \sum_j \|\rho_j \mathbf{L}_i \mathbf{n}_j - \mathbf{M}_{i,j}\|^2$ and then for ρ_j directly by $\rho_j = \mathbf{M}_{i,j}/(\mathbf{L}_i\mathbf{n}_j)$. We average all ρ estimates for the same set of images that are modeled well by the low rank approximation, thereby allowing us to compute $\mathbf{S}^t$ for use in the ambiguity recovery.

3.3.3 Local Normal Refinement

The initial normal estimation produces a smoothed result that is akin to the mean shape. We follow the procedure from [23] where different local regions of the face are refined by using different subsets of images. Thereby selecting a set of consistent images for each point with less expression variation to cause smoothing, *e.g.*, closed mouth.

The local image selection is similar to [23]. We select a subset of k images with the minimum distance $\|\mathbf{M}_j - \mathbf{LS}_j\|^2$ with $k\geq4$ images and enough to produce a low condition number of $\mathbf{L}_{k\times4}$. We recover the local shape $\mathbf{S}_j$ via

$$\min_{\mathbf{S}_j} \|\mathbf{M}_{k\times1} - \mathbf{L}_{k\times4}\mathbf{S}_j\|^2, \tag{3}$$

where we omitted the Tikhonov regularization term proposed by [23], as the Minkovski norm is not positive definite and should be properly treated through a Lagrange multiplier, but we found the above energy produces sufficiently close results.

3.4. Surface Reconstruction

Given the shape vectors $\mathbf{S}$ we can assemble the normals $\mathbf{n}$ by normalizing the last three components for each vertex. Then, we reconstruct the triangulated surface patch Ω with p vertices $\mathbf{X}$ that is consistent with fine details specified by $\mathbf{n}$. As in 3D landmark-driven warping, we keep the connectivity of the vertices intact.

Assuming that the template has a similar metric tensor (distance measure on the surface) to the output mesh, we can reconstruct the shape $\mathbf{X}$ from the normal field $\mathbf{n}$ through the mean curvature formula $\Delta\mathbf{x} = -H\mathbf{n}$, *i.e.*, we minimize

$$\|\mathcal{L}\mathbf{X} - \mathbf{H}\|^2,$$

where $\mathbf{H}$ is the mean curvature normal, obtained by collecting $-H_i\mathbf{n}_i$ into a $3p$-dimensional vector.

Since we are only given $\mathbf{n}$, we first estimate H_i, the integral of mean curvature around vertex i from $\mathbf{n}$ through the discretization of $H = \nabla A \cdot \mathbf{n}$, *i.e.*, the mean curvature is how fast the area changes when surface points move along the normal direction [34]. The discretization of the first variation of the area can be measured by the difference between $\mathbf{n}_i$ and $\mathbf{n}_j$ as follows,

$$H_i = \frac{1}{4A_i} \sum_{j\in N(i)} (\cot\alpha_{ij} + \cot\beta_{ij})\, \mathbf{e}_{ij} \cdot (\mathbf{n}_j - \mathbf{n}_i), \tag{4}$$

where $N(i)$ is the set of immediate neighboring vertices of i, A_i is the sum of the triangles' areas incident to i, $\mathbf{e}_{ij}$ is the edge from i to j (Figure 3). Note the cotan weights are the same as those in the Laplace operator. For more accurate results, we update the cotan weights in each global iteration.

One unique challenge in handling a 3D model instead of a height field is the boundary. On the boundary, the mean curvature formula degenerates into a 1D version

$$\mathcal{L}_b\mathbf{X} = \kappa\mathbf{b},$$

which is based on the 1D Laplace operator $\mathcal{L}_b$ with nonzero entries corresponding to boundary edges, with $\mathcal{L}_{b,ij} = 1/e_{ij}$, where j is one of the two boundary vertices adjacent to boundary vertex i, and κ is the geodesic curvature along the boundary and $\mathbf{b}$ is the cross product between the surface normal and the boundary tangent. Since the photometric

Algorithm 1: Unconstrained 3D face reconstruction

Data: photo collection, template $\mathbf{X}_0$
Result: 3D face mesh $\mathbf{X}$

compute 2D landmarks for all images (Sec. 3.1)
`// warp template through landmarks (Sec. 3.2)`
while *template deformation* $>$ *threshold* **do**
  estimate projection $\mathbf{P}_i$ for each image
  solve Eq 2
while *3D face vector* $\mathbf{X}$ *change* $>$ *threshold* **do**
  `// estimate` $\mathbf{n}$ `and` ρ `(Sec. 3.3)`
  re-estimate $\mathbf{P}_i$
  perform matrix completion on $\mathbf{F}$ to obtain $\mathbf{M}$
  estimate lighting, $\mathbf{L}$, and shape, $\mathbf{S}$, by SVD
  estimate albedo, ρ
  resolve ambiguity by estimating $\mathbf{A}$
  refine local normal estimate via Eq 3
  `// deform` $\mathbf{X}$ `with` $\mathbf{n}$ `(Sec. 3.4)`
  update $\mathbf{n}$ via Eq 6
  estimate mean curvature via Eq 4
  solve Eq 5

normal does not provide information about κ, we simply use $\kappa^k \mathbf{b}^k = \mathcal{L}_b \mathbf{X}^k$ where $\mathbf{X}^k$ is the estimated shape in k-th iteration.

We can finally put all the constraints and equations together into an overall energy,

$$\|\mathcal{L}\mathbf{X} - \mathbf{H}^k\|^2 + \lambda_b \|\mathcal{L}_b \mathbf{X} - \mathcal{L}_b \mathbf{X}^k\|^2 + \lambda_l \sum_i \|\mathbf{P}_i^k \mathbf{D}_i \mathbf{X} - \mathbf{W}_i\|^2,$$

leading to a linear system for $\mathbf{X}$ after we fix the projection matrices,

$$(\mathcal{L}^2 + \lambda_b \mathcal{L}_b^2 + \lambda_l \sum_i \mathbf{D}_i (\mathbf{P}_i^k)^\intercal \mathbf{P}_i^k \mathbf{D}_i) \mathbf{X} = \mathcal{L}\mathbf{H}^k + \lambda_b \mathcal{L}_b^2 \mathbf{X}^k + \lambda_l \sum_i (\mathbf{P}_i^k)^\intercal \mathbf{W}_i, \quad (5)$$

where λ_b is the boundary constraint weight. Figure 4 illustrates the effects of the above system in aligning the normals while optimizing the landmark locations.

Smoothing of shadowed regions. We additionally set a threshold θ to detect attached shadow regions through $\mathbf{L} \cdot \mathbf{n} < \theta$, where $\mathbf{L}$ is the average incoming light direction, and replace the entries in $\mathbf{n}$ corresponding to the vertices in those regions by $\mathbf{n}^k$. We then smooth the resulting normal field $\tilde{\mathbf{n}}$ by the following procedure. First, we construct a diagonal selection matrix $\mathbf{T}$, with $\mathbf{T}_{ii} = 1$ only if vertex i is not in attached shadow region. We then update $\mathbf{n}$ using the following linear system

$$(\mathbb{I} + w_d \mathbf{T}(\mathcal{L} + w_s \mathcal{L}^2)\mathbf{T})\mathbf{n} = (\mathbb{I} + w_d \mathbf{T}\mathcal{L}\mathbf{T})\tilde{\mathbf{n}}, \quad (6)$$

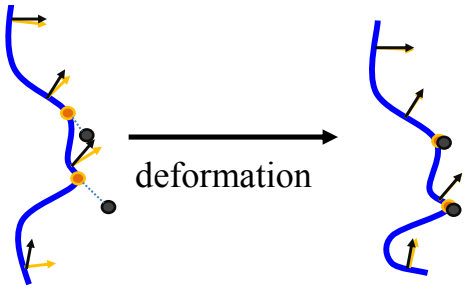


Figure 4. The effects of the deformation-based surface reconstruction. The black arrows indicate the photometric normal estimates and the orange arrows show the actual surface normal. The black dots are the target landmark locations and the orange dots are the corresponding vertices in the mesh.

where $\mathbb{I}$ is the identity matrix. The procedure fixes the non-shadowed region, blends the shadowed region through impainting with weight w_d, and smooths out the shadowed region with a weight w_s.

Finally, Algorithm 1 summarizes the overall procedure in our algorithm.

4. Experiments

In this section we present our experiments. We first describe the pipeline to prepare a photo collection for face reconstruction. We then demonstrate qualitative results compared with 2.5D reconstruction on the Labeled Faces in the Wild (LFW) database [20], and on celebrities downloaded from Bing image search. Finally, we compare quantitatively on a personal photo collection where we have the ground truth model captured via a range scanner.

4.1. Data Preparation

Photo collection pipeline For the celebrities, we use the Bing API to access up to the first $1,000$ image results by searching on their first and last names. We remove duplicate images from the retrieval results. The images are then imported into Picasa, which performs face detection and groups similar images. After manually naming a few groups, further images are suggested by Picasa and automatically added to the collection. In the end, about half of the images remain for each person since many search results are not photographs of the person of interest or are duplicates. A landmark detector estimates 68 landmarks in each image around the eyes, eyebrows, nose, mouth, and chin line. For the initial shape template, we use the space-time faces neutral face model [43], which we subdivide to create more vertices and thereby a higher resolution.

Ground truth models We use a Minolta Vivid 910 range scanner to construct ground truth depth maps for a personal photo collection. The scanner produces a 2.5D depth scan;

Figure 5. Visual comparison on Bing celebrities with images from [23] to the left of each of our viewpoints. Note how our method can incorporate the chin and more of the cheeks, as well as producing more realistic reconstructions especially in the detailed eye region.

so we capture three scans, one frontal, and two at ~ 45 degree yaw. We align the scans via Iterative Closest Point and merge them to produce a ground truth model.

4.2. Results

Qualitative evaluation We process the same celebrities as used in [23], George Clooney (476 photos), Tom Hanks (416), Kevin Spacey (316), and Bill Clinton (460), as well as the four individuals with the most images in LFW, George Bush (528), Colin Powell (236), Tony Blair (144), and Donald Rumsfeld (121). The resolution of LFW is 250×250 and we scale all Bing face regions to 500 pixels height. Figure 5 compares the results between our approach and the figures from [23]. Figure 6 shows our reconstructions on the LFW dataset. We see that our reconstruction provides more accurate fine details in areas with high mean curvatures, *e.g.*, the eyes and mouth, as well as allowing for reconstruction of the chin and cheeks when the surface normal points away from the frontal pose. Furthermore, the facial features in our results are less caricature-like than [23], but closer to the true geometry.

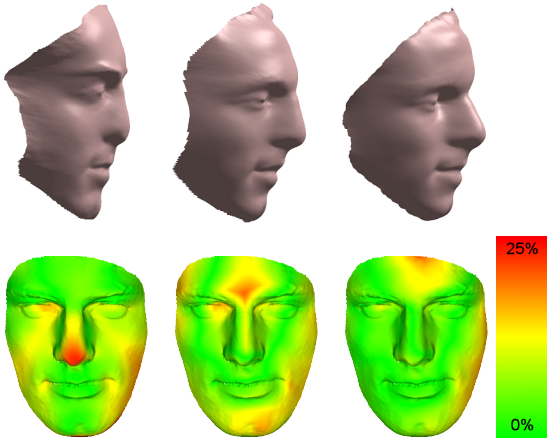
Table 2. Distances of the reconstruction to the ground truth.

Methods	2.5D	2.5 Improved	3D
Mean	7.86%	7.79%	**5.42%**
RMS	9.71%	9.04%	**6.89%**

Quantitative evaluation We also implement the 2.5D approach by warping our estimated photometric normals to a frontal view and integrating the depth. Since the 2.5D approach from [23] is not metrically correct as they mention, we also perform an improved 2.5D approach where we first use our landmark driven 3D warping as a preprocessing step to resolve the aspect ambiguities.

To compare the approaches numerically, we compute the shortest distance from each vertex in the ground truth to the closest point on the reconstructed surface face. Meshes are aligned by their internal landmarks according to the absolute orientation problem [18]. We report the mean euclidean distance and the root mean square (RMS) of the distance, after normalized by the eye-to-eye distance, in Table 2. Figure 7 shows a coloring of the template to visualize where on the face is close for reconstruction. The base 2.5D ap-

Figure 6. Results on subjects from the LFW dataset. The reconstructed 3D model, sample image from which we extract the texture, and a novel rendered viewpoint.

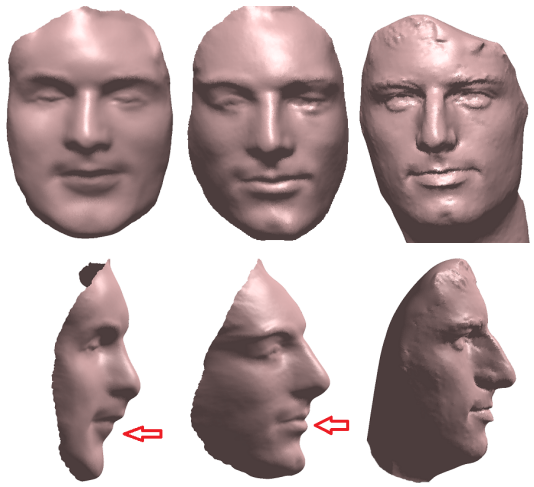


(a) (b) (c)

Figure 7. Distance from the ground truth to the face reconstructed via (a) 2.5D, (b) 2.5D improved, and (c) 3D reconstruction. Distance increases from green to red. Best viewed in color.

proach (a) has incorrect depth information at the nose since the shape ambiguity is recovered from the flatter initial template, the improved 2.5D approach (b) better approximates the depth, but the bridge of the nose protrudes too far, and our 3D reconstruction best matches with the ground truth across all the details of the face.

Usage of profile views One advantage of the landmark-based deformation approach combined with photometric stereo is the ability to use the profile images more effectively. With only photometric stereo, the extreme poses obscure parts of the face and also cause increased possibility for distortion when rendering in a frontal view. However, the profile views provide rich 3D landmark depth information. We run an experiment on the personal photo collection where we use 70 nearly frontal images with < 10 degree yaw, and then add 40 images with side view information of > 45 degree yaw. Figure 8 shows the improved depth of reconstruction and accurate mouth details from using additional side view images. Note that we manually labeled the ground truth for these images due to the deficiency of our 2D face alignment implementation when points are occluded, but there are detectors that work well even in these situations [8].



(a) (b) (c)

Figure 8. Comparison of (a) frontal only, (b) including side view for landmark warping, and (c) ground truth scan. The addition of side view improves the nose and mouth region (see arrows) while also allowing for reconstruction further back on the cheeks.

5. Conclusions

We presented a method for 3D face reconstruction from an unconstrained photo collection. The entire pipeline of iterative reconstruction is coherently conducted on the 3D triangulated surface, including texture mapping, surface normal estimation and surface reconstruction. This enables consuming faces with all possible poses in the reconstruction process. Also, by leveraging the recently developed image alignment technique, we use a combination of 2D landmark driven constraint and the photometric stereo-based normal field for surface reconstruction. Both qualitative and quantitative experiments show that our method is able to produce high-quality 3D face models. Finally, there are multiple directions to build on this novel development, including incorporating automatically detected 2D landmarks in the profile views, validating our approach on a diverse set of populations, and extending to non-face objects.

References

[1] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C. T. Silva. Computing and rendering point set surfaces. *Visualization and Computer Graphics, IEEE Transactions on*, 9(1):3–15, 2003. 2

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

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

[4] 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. 2

[5] 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. 2

[6] M. Berger, A. Tagliasacchi, L. Seversky, P. Alliez, J. Levine, A. Sharf, and C. Silva. State of the Art in Surface Reconstruction from Point Clouds. *EUROGRAPHICS star reports*, 1(1):161–185, Apr. 2014. 2

[7] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE T-PAMI*, 25(9):1063–1074, 2003. 1, 2

[8] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520, 2013. 8

[9] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.*, 33(4):43, 2014. 1

[10] J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans. Reconstruction and representation of 3D objects with radial basis functions. In *Proc. of the 28th annual conf. on Computer graphics and interactive techniques*, pages 67–76. ACM, 2001. 2

[11] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proc. of the 23rd annual conf. on Computer graphics and interactive techniques*, pages 303–312. ACM, 1996. 2

[12] 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

[13] 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

[14] T. Hassner. Viewing real-world faces in 3D. In *ICCV*, pages 3607–3614. IEEE, 2013. 2

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

[16] C. Hernández, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *IEEE T-PAMI*, 30(3):548–554, 2008. 2

[17] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. *SIGGRAPH Comput. Graph.*, 26(2):71–78, July 1992. 2

[18] B. K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *J. Optical Soc. of America A.*, 4(4):629–642, 1987. 7

[19] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3D reconstruction for face recognition. In *FG*, pages 843–848. IEEE, 2004. 1

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 6

[21] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014. 1

[22] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proc. of the 4th Eurographics symposium on Geometry processing*, 2006. 2

[23] I. Kemelmacher-Schlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, 2011. 1, 2, 4, 5, 7

[24] I. Kemelmacher-Shlizerman and R. Basri. 3D face reconstruction from a single image using a single reference face shape. *IEEE T-PAMI*, 33(2):394–405, 2010. 2

[25] 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

[26] Z. Lin, M. Chen, and Y. Ma. The augmented lagrage multiplier method for exact recovery of corrupted low-rank matrices. Technical Report UILU-ENG-09-2215, UIUC, Nov. 2009. 4

[27] X. Liu. Discriminative face alignment. *IEEE T-PAMI*, 31(11):1941–1954, 2009. 1

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

[29] M. Pauly, N. J. Mitra, J. Giesen, M. H. Gross, and L. J. Guibas. Example-based 3d scan completion. In *Symposium on Geometry Processing*, number EPFL-CONF-149337, pages 23–32, 2005. 2

[30] U. Pinkall and K. Polthier. Computing discrete minimal surfaces and their conjugates. *Experimental mathematics*, 2(1):15–36, 1993. 3

[31] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *CVPR*, 2014. 1

[32] B. Shi, K. Inose, Y. Matsushita, P. Tan, S.-K. Yeung, and K. Ikeuchi. Photometric stereo using internet images. *International Conference on 3D Vision (3DV)*, 2014. 2

[33] 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*, pages 175–184. ACM, 2004. 2, 3

[34] M. Spivak. A comprehensive introduction to differential geometry, vol. 5. *Publish or Perish*, 1979. 5

[35] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *ECCV*, pages 796–812. Springer, 2014. 1, 2

[36] 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, 2012. 1

[37] J. Wang, L. Yin, X. Wei, and Y. Sun. 3D facial expression recognition based on primitive surface feature distribution. In *CVPR*, volume 2, pages 1399–1406. IEEE, 2006. 1

[38] 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

[39] 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, 2, 3

[40] 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

[41] 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.*, 23(3):644–651, 2004. 2, 3

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

[43] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: high resolution capture for modeling and animation. *ACM Trans. Graph.*, 23:548–558, Aug. 2004. 6