UNCONSTRAINED 3D FACE RECONSTRUCTION FROM PHOTO COLLECTIONS

By

Joseph Roth

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science — Doctor of Philosophy

2016

ABSTRACT

UNCONSTRAINED 3D FACE RECONSTRUCTION FROM PHOTO COLLECTIONS

By

Joseph Roth

This thesis presents a novel approach for 3D face reconstruction from unconstrained photo collections. An unconstrained photo collection is a set of face images captured under an unknown and diverse variation of poses, expressions, and illuminations. The output of the proposed algorithm is a true 3D face surface model represented as a watertight triangulated surface with albedo data colloquially referred to as texture information. Reconstructing a 3D understanding of a face based on 2D input is a long-standing computer vision problem. Traditional photometric stereo-based reconstruction techniques work on aligned 2D images and produce a 2.5D depth map reconstruction. We extend face reconstruction to work with a true 3D model, allowing us to enjoy the benefits of using images from all poses, up to and including profiles. To use a 3D model, we propose a novel normal field-based Laplace editing technique which allows us to deform a triangulated mesh to match the observed surface normals. Unlike prior work that require large photo collections, we formulate an approach to adapt to photo collections with few images of potentially poor quality. We achieve this through incorporating prior knowledge about face shape by fitting a 3D Morphable Model to form a personalized template before using a novel analysis-by-synthesis photometric stereo formulation to complete the fine face details. A structural similarity-based quality measure allows evaluation in the absence of ground truth 3D scans. Superior large-scale experimental results are reported on Internet, synthetic, and personal photo collections.

This thesis is dedicated to my beautiful wife Lynnette, whose encouragement along the way was paramount to making it thus far.

ACKNOWLEDGMENTS

"The fear of the Lord is the beginning of knowledge." - Solomon, Proverbs

Throughout my Ph.D. studies, I have come to understand deeper God's blessing of man with the gift of knowledge and the ability to explore and understand the workings of His creation.

I have been privileged to have Dr. Xiaoming Liu as my advisor. His desire to see me succeed has pushed me to obtain far more than I could imagine alone. The late nights spent writing papers together, attention to the smallest details, and desire the push the bounds of knowledge have inspired my dedication to excellence. I am deeply indebted for his refinement of my writing and presentation skills. I am also grateful to Dr. Yiying Tong for his collaboration and filling in the gaps of my computer graphics knowledge. I immensely appreciate his contribution and enjoy researching alongside of him. I would also like to thank the remainder of my committee members, Dr. Arun Ross, Dr. Anil K. Jain, and Dr. Hayder Radha for their valuable insights and contributions along the way.

I am grateful to my Computer Vision Lab members, Xi Yin, Amin Jourabloo, Morteza Safdarnejad, Yousef Atoum, Jamal Afridi, and Luan Tran for the excellent working atmosphere. The willingness to answer any questions and late nights working together causes all of our work to flourish. I will also never forget the times of celebration and entertainment together that keeps our sanity.

A word of thanks for the larger biometrics community at Michigan State University in the PRIP Lab and i-PRoBe Lab with their insights and questions at our seminars. Specifically, a big thanks to Charles Otto and Lacey Best-Rowden for their assistance with my experiments.

A word of appreciation to Katherine Trinklein, Courtney Kosloski, Linda Moore, Cathy Davi-

son, and Debbie Kruch for their administrative assistance. Special thanks to Katy Luchini Colbry for assisting through organizing professional education and her mother-like care.

Finally, I would like to thank my family for their prayers and encouragement along the way. The largest thanks for my beautiful wife Lynnette.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Contributions

3D reconstruction from photographs is a long standing problem with much interest in computer vision. The goal of 3D reconstruction is to infer depth information from 2D inputs such as photos or videos. Beginning with reconstruction of rigid desktop objects [73, 62, 28] in highly constrained environments [36, 17, 86], 3D reconstruction has advanced to outdoor environments [80, 41, 76] and even large-scale objects [55, 68, 1] from in-the-wild images. Even so, most reconstruction techniques operate on *rigid* objects since the 3D structure will not change over time and common points identified in different images can directly determine the surface under perspective geometry.

One object in particular, the face, is highly studied, since obtaining a user-specific 3D face surface model is useful for applications in 3D-assisted face recognition [14, 42, 59], 3D expression recognition [88], facial animations [21], avatar puppeteering [90], and more. For instance, accurate face models have been shown to significantly improve face recognition by allowing the rendering of a frontal-view face image with netural expression, thereby suppressing intra-person variability [101]. However, face reconstruction is especially challenging since the face is highly *non-rigid* and the structure changes with expressions.

In this thesis, I present an approach to solving the specific problem of unconstrained face reconstruction from photo collections. *Unconstrained* means there is no prior knowledge about the image capturing conditions. These conditions are often referred to as PIE for variations in Pose, Illumination, and Expression. Pose variation refers to the orientation and position of the person with respect to the camera. Portrait photographs are mainly taken from a frontal orientation, but

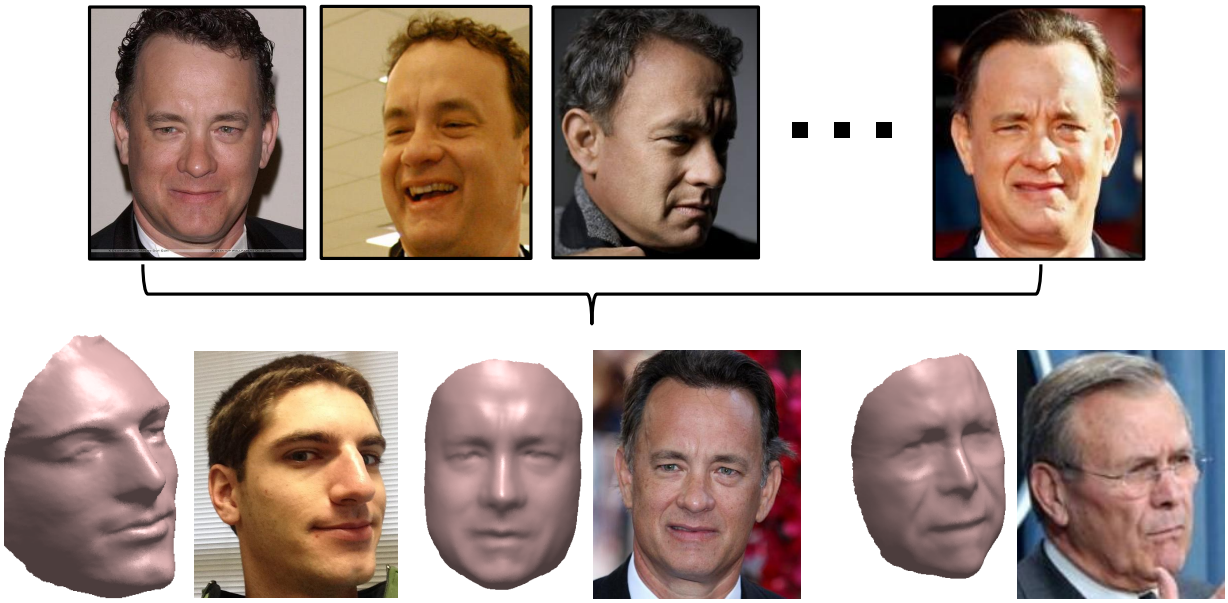


Figure 1.1: Given an unconstrained photo collection of Tom Hanks containing images of unknown pose, expression, and lighting, face reconstruction seeks to create an accurate 3D model of his face. Two additional example reconstructions along with a single representative image are shown as well.

unconstrained photos may be taken from any orientation, including the side, revealing or obstructing different parts of the face. Illumination refers to the lighting conditions in terms of numbers, positions, orientations, or colors. Poorly illuminated images may cast shadows or wash out parts of the image by being too bright. Expression variations such as smiling, laughing, or frowning can dramatically change the structure and appearance of a face. *Face reconstruction* is defined as the process of creating a detailed 3D model of a person's face from 2D input.

1.1 Photo Collections

A *photo collection* is a set of images of the same individual captured from potentially different cameras at unknown times. This is in contrast to a video that is a set of images captured from the same camera in quick succession. Photo collections present particularly difficult challenges since

no temporal information may be used and different camera lenses may distort images differently.

There are a few recent works exploring the specific problem of reconstructing faces from photo collections. The seminal work by Kemelmacher-Shlizerman and Seitz [53] uses a photometric stereo approach to create a depth map of the frontal view from the photo collection. To do this, they first warp all images to a frontal view so they are in correspondence at the pixel level. Then they use Singular Value Decomposition (SVD) to jointly extract the lighting conditions and the surface normals. Photo collections with a variety of expressions exhibiting different surface normals will produce an averaged or smooth reconstruction. To compensate, they refine the surface normals by selecting a subset of images that are locally consistent for each pixel. The surface is reconstructed by integrating the pixel-grid of surface normals.

The normals are further refined by selecting a subset of images for each point which are locally consistent. The surface normals are then integrated to produce a 2.5D depth map of the subject. It is extended in a few different directions, one in [94] where they use the surface normals from frontal faces to improve the fitting of a 3DMM. And two, in [51, 77] where the technique is used to generate a 3DMM

1.2 Thesis Contributions

In this thesis, the challenging problem of unconstrained face reconstruction from photo collections is addressed. There are two major problems with prior approaches. One, a restrictive 2.5D depth map is reconstructed instead of a true 3D model. Two, a large collection of images is required due to the SVD-based approach to photometric stereo. I make the following contributions in order to solve these outstanding problems.

- A novel Laplace mesh editing technique using surface normals as the input is proposed to

allow reconstruction of the 3D face model. Prior approaches require the mean curvature normal as input, however we introduce an estimation of the mean curvature based on the normal field. This enables reconstruction when only the surface normals are known as is the case in photometric stereo methods. While we only apply this method to face reconstruction, it is applicable to general mesh editing.

- A true 3D facial surface model is reconstructed from photo collections. State-of-the-art reconstruction from photo collections aligns all images to a frontal pose before reconstructing a depth map. By formulating the problem on a 3D surface instead of a 2D pixel grid, we can utilize faces from all poses in the reconstruction process. 3D models can capture more details and have broader applications than 2.5D reconstructions.

- Photometric stereo is solved in a joint Lambertian image rendering formulation, with an adaptive template regularization that allows for graceful degradation to a small number of images. Current face reconstruction uses an SVD-based approach to solving photometric stereo, requiring large photo collections.

- A pose based dependability measure and structural similarity based local selection are proposed to use the best images in the collection for reconstructing different parts of the face. Not all images are useful for face reconstruction. When all images are used for every part of the face, a smooth reconstruction is obtained. Previous work assuming large collections aggressively eliminated $\sim 90\%$ of the images in the collection for each face detail. By designing a novel local selection procedure, we are able to use half of the images and only discard parts that would contribute negatively to the reconstruction. This less aggressive approach is necessary for small photo collections.

- A structural similarity based quality measure is introduced to evaluate reconstruction perfor-

mance in the absence of ground truth scans. Performance evaluation is crucial to comparing reconstruction techniques. However, many photo collections lack ground truth face scans and people must rely on qualitative evaluation. We propose a quantitative evaluation criteria for use on Internet photo collections, where ground truth data cannot exist.

Chapter 2

Background and Related Work

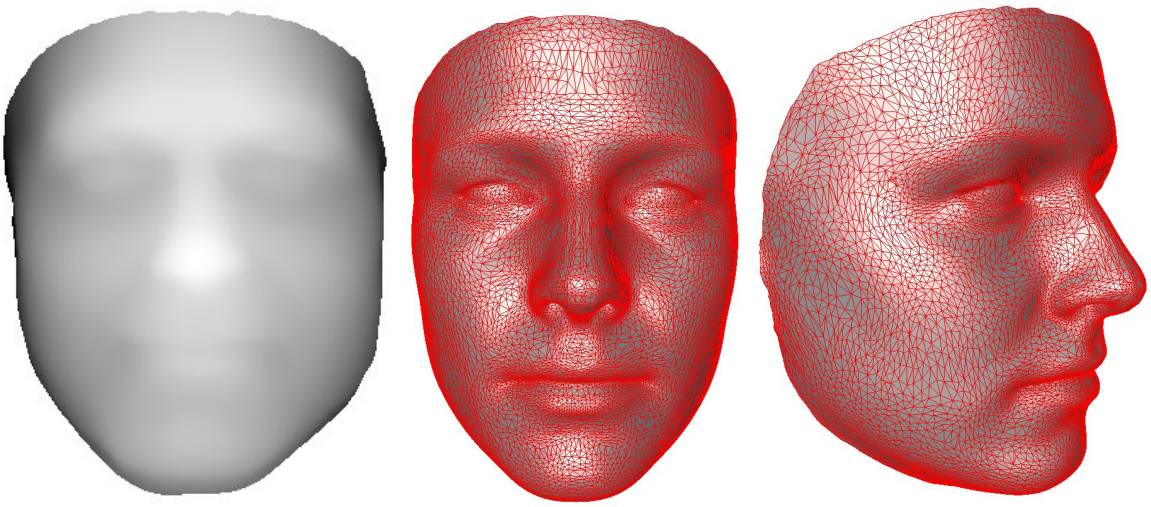
Now that a basic understanding of the problem is known, I will present some background information and related work necessary for fully understanding face reconstruction.

2.1 Scene Parameters

The inverse process of reconstruction is called rendering. Starting with a 3D model composed of a surface and its texture and using a set of scene parameters describing the object positions, lighting, and camera projection, a 2D image may be rendered. Many assumptions are made due to either limited understanding about reflectance properties of different surfaces or more often for computational efficiency. I now present some of these assumptions for the different parts necessary to render an image.

2.1.1 Object model

An object model consists of both a representation of the surface shape and the texture. Object surfaces are continuous but are usually approximated with discrete structures. A limited model of an object is to express the surface as a depth map. Depth maps are represented as an image consisting of a single channel containing the distance of the surface from a specific view point. This simplifies some tasks by removing the need for camera projection and only storing the visible information. However, such a model is limited because the object only contains information from



(a) (b)

Figure 2.1: Comparison of object models, depth map (a) and triangulated mesh (b). The depth map has a fixed orientation since it is simply an image, but the triangulated mesh may be rendered at any orientation.

one view or side of the face, and has limited to no knowledge about the sides or chin of the face that are typically occluded. These depth maps are sometimes called 2.5D models since they only contain partial information in the third dimension. Depth maps are frequently used in conjunction with a regular image where they are called, RGB-D for the standard red, green, and blue channels with the addition of the depth channel.

A true 3D model is better represented as a triangulated mesh and is able to capture details around the entire surface. A triangle mesh is composed of vertices, edges, and triangle faces. The vertices are a position in 3D space that may contain other information such as texture or a normal vector. An edge is a connection between two vertices. And a triangle face is a closed set of three edges. Typically faces share each edge with a single other face in order to make a continuous surface, but there can be boundaries which form holes in the surface. Figure 2.1 shows the differences between a 2.5D and 3D model. Notice how the 3D model is able to wrap around the sides of the face and fold upon itself at the nostrils. This particular mesh has a single hole

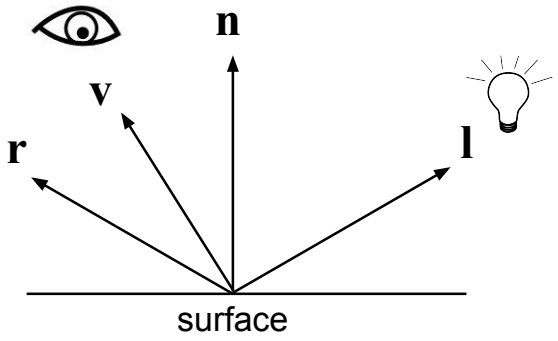


Figure 2.2: Unit vectors used in lighting models. **l** is the light source direction, **n** is the surface normal, **v** is the viewpoint direction, and **r** is the reflection of the light source.

and boundary around the back of the head, but is continuous across the face region. Under a triangulated mesh representation, the object contains information about all parts of the face and may be rotated and rendered under novel views.

2.1.2 Lighting

Without lighting or illumination, all scenes would appear solid black. Photons of light emit from light sources and interact with surfaces in the scene until finally reaching either the human eye or the camera sensor which forms the colors in an image depending on the number of photons which reach the sensor. The interactions with surfaces are complicated and not fully understood. Some surfaces the light simply bounces off, others like glass light bends as it passes through, and others like fog can partly occlude light while allowing some to pass through. Human skin is a very challenging surface since it is partly translucent where some light reflects off the surface, while some of it goes through the surface allowing us to see blood veins underneath the skin. Much work exists to create realistic skin rendering [16].

Since lighting is complicated, we often make simplifying assumptions. One common model

for lighting is the Lambertian model which models diffuse light, and is expressed as,

$$I_d = k_d \mathbf{l}^\intercal \mathbf{n}. \tag{2.1}$$

Where I_d is the diffuse intensity, k_d is the intensity of the light source, $\mathbf{l}$ is the light source direction from the surface to the source, and $\mathbf{n}$ is the surface normal. This lighting assumption is not dependent on the point of view and handles matte surfaces well. However, polished surfaces demonstrate changing illumination dependent on the observers point of view. Think of the bright spot on a bald person's head that moves when you change positions. These are modeled with the Phong illumination model which includes a specular reflection component,

$$I_s = k_s(\mathbf{r}^\intercal \mathbf{v})^\alpha. \tag{2.2}$$

Where I_s is the specular illumination, k_a is the specular constant, α is the power of the cosine angle which determines how glossy the surface is, $\mathbf{v}$ is the view vector from the surface to the point of view, and $\mathbf{r}$ is the reflection of the light source, $\mathbf{r} = 2(\mathbf{l}^\intercal \mathbf{n})\mathbf{n} - \mathbf{l}$. Figure 2.2 shows a visualization of the different vectors used in the lighting models.

2.1.3 Camera Model

To understand an image, not only do we need to model how light interacts with the surface, but we also need to model how a camera captures the light in the scene. The camera model describes how to project real world 3D points onto a 2D pixel location in the image. In camera terms, how does the photon of light traveling from the surface reach a particular element on the CCD sensor. The most commonly used model is the pinhole camera model, which assumes the sensing plane is

located behind a barrier with a single point allowing light to pass. With this model, the projection is explained by a 3x4 projection matrix [33] P, up to a scale, that can be decomposed into an intrinsic and extrinsic matrix.

$$P = \underbrace{\begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \times \left[\begin{array}{ccc|c} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ \underbrace{r_{31} \quad r_{32} \quad r_{33}}_{R} & & & \underbrace{t_z}_{T} \end{array}\right] \tag{2.3}$$

The intrinsic matrix K describes the pinhole and image plane properties: focal length (f_x, f_y), principal point (c_x, c_y), and skew s. The extrinsic matrix describe the position of the camera in world coordinates: rotation R and translation T. Note, that typically further assumptions are made, $f_x = f_y$, $s = 0$, and (c_x, c_y) is the image center to leave only 8 parameters including the scale. As described, this model projection equation is also known as perspective projection.

The perspective projection model is a simplification of typical cameras that have a lens distorting the light from a wider range than a pinhole. In these cases, radial distortions need to be considered particularly for high resolution photographs or wide-angle lenses. However, typically the distortion will be removed prior to using the images in any pipeline by resampling the image with known parameters.

A simpler camera model is using weak perspective projection. In a weak perspective projection model, all rays from the sensor plane are parallel, immediately removing skew, focal length, and the principal point from consideration. The weak perspective model is therefore only the first 2 rows of the extrinsic matrix and has 6 parameters including the scale. A comparison of the two describe models is given in Figure 2.3. The weak perspective projection works when the object is relatively far away from the camera compared to the focal length and the size of the object. For

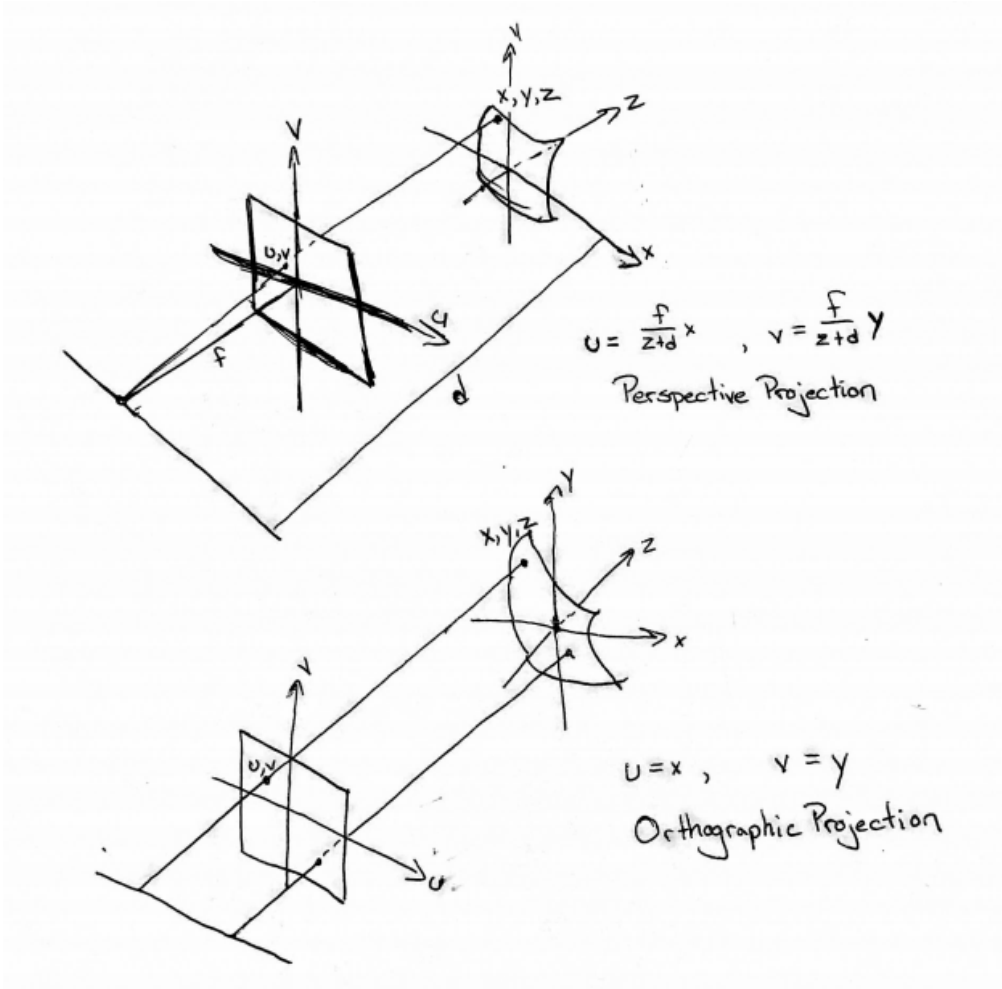


Figure 2.3: Comparison of perspective and weak perspective projection. With perspective projection, the intrinsic camera parameters and distance to the object affect the location of the point on the image plane. With weak perspective projection, points along orthogonal lines to the image manifest in the same location regardless of depth.

faces, images taken greater than arm length demonstrate negligible perspective distortion.

2.2 Surface Reconstruction

Surface reconstruction is a challenging process that varies significantly depending on the type of input (images, point cloud, normal field, etc.), the quality of the data (noise, outlier, etc.), the desired output (mesh, skeleton, volume, etc.), and the type of shape (man-made, organic, etc.). To the graphics community, surface reconstruction almost always refers to fitting a shape to a point cloud, potentially with additional side information like surface normals. The majority of reconstruction algorithms take a point cloud as their input, including methods with surface-smoothness priors (*e.g.*, tangent planes [39], moving least squares [3], radial basis function [23], Poisson surface reconstruction [50]), visibility-based methods [25], data-driven methods [64], etc. More details on this particular topic are in the state of the art report [12].

One of the most widely used methods is the Poisson surface reconstruction [50] 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. A surface is then fit globally by solving the Poisson equation. This approach is robust to noise and allows a global solution unlike many heuristic based approaches. However, our proposed method will begin with a normal field and not a point cloud, which renders the 3D Poisson surface reconstruction not directly applicable. Thus, we resort to a template deformation approach, where we deform a template face mesh to match the observed surface normals 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 [78, 97], where the mean curvature normal fields are given, and the surface is deformed under additional landmark 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.

In computer vision, however, surface reconstruction usually refers to producing a 3D understanding of an object from 2D images. Some techniques will consider RGB-D input which can be viewed as a point cloud with texture information. In the specific case where the RGB-D data is a video, the task of reconstruction is called *dynamic fusion* [61]. Starting with only 2D photographs is more challenging than a point cloud. The goal in this scenario can better be described as *estimating the most probable 3D shape that explains an image under a set of assumed materials, viewpoints, and lighting conditions*. Generally, the problem is ill-posed without assumptions since there are multiple combinations of 3D objects, materials, viewpoints, and lighting that can produce identical images. However, under different sets of assumptions, highly detailed and accurate reconstructions are possible. Many different cues provide insight into the object shape such as, image focus, texture, shading, and stereo correspondence. These last two cues have demonstrated robustness for reconstruction and are generally the most common approaches.

2.2.0.1 Multi-view Stereo

Multi-view stereo takes a photo collection obtained under different viewpoints and tries to reconstruct a plausible 3D geometry that explains the images under a set of reasonable assumptions, foremost being object rigidity. An overview of the procedure is as follows. First, estimate the camera parameters; both the extrinsic (position and orientation) and the intrinsic (focal length and sensor scale). This is generally done using Structure-from-Motion [33] to compute the camera models from a photo collection offline. Typically, a pinhole camera model is used to fully describe how a 3D point on the object is mapped onto a 2D pixel location in the image. Second, identify common points between different images. With known camera models, any pair of points found in

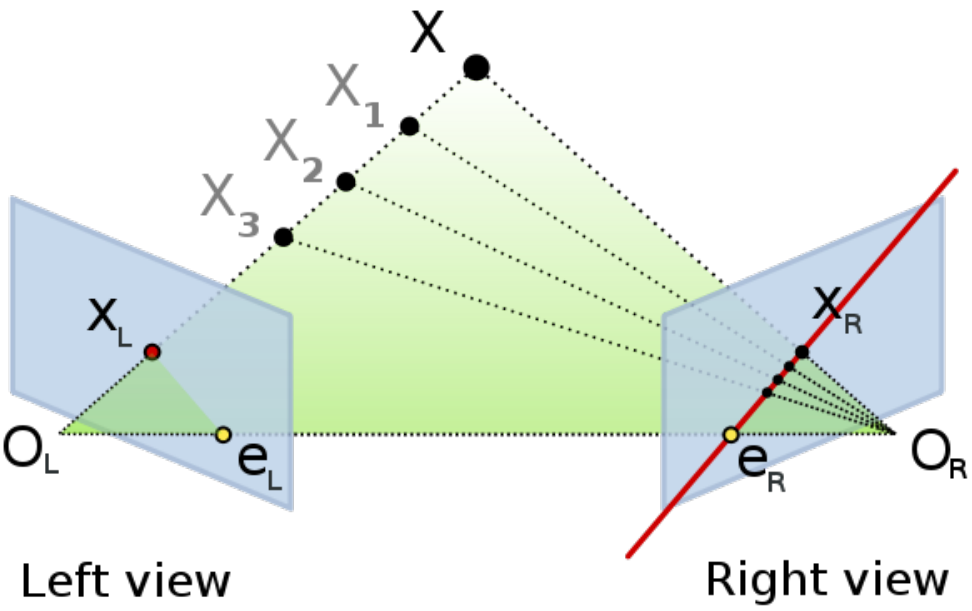


Figure 2.4: Epipolar geometry, the basis for multi-view stereo reconstruction. If the left camera, $\mathbf{O}_L$, observes a point $\mathbf{X}_L$ and there exists another camera, $\mathbf{O}_R$, with known relative orientation, the point must fall along epipolar line in the right view. If the matching point is determined in the right view, $\mathbf{X}_R$, then the position of that point is uniquely determined in 3D space, $\mathbf{X}$.

both cameras uniquely determines an object point in 3D space (Figure 2.4). Finally, the 3D locations of all common points form a point cloud which may then be solved using traditional graphics surface reconstruction methods.

2.2.0.2 Photometric Stereo

While multi-view stereo uses simultaneous images taken under different viewpoints, *photometric stereo* uses sequential images taken from the same viewpoint under different lighting conditions. When making reasonable assumptions about the material and light properties, the surface normals of the object may be determined. Intuitively, a part of an object oriented towards the light will appear brighter than a part oriented away from the light, therefore we may infer the angle between the surface normal and the light source direction. Multiple different lighting directions are required

Figure 2.5: Sample images used for photometric stereo. Multiple images from a fixed position are captured under a variety of lighting conditions. The surface normals can be determined due to the brighter appearance when facing the light source.

to resolve the ambiguity about the true surface orientation and to ensure every part of the object is oriented towards the light in one image. A normal field is constructed using the pixel grid of the image, and the shape is determined by integrating the normal field. Shape-from-Shading in this manner produces a depth map representation of the object since all images were captured from the same view.

The easiest assumptions are a weak perspective camera projection model with a Lambertian lighting model. Under these conditions, the seminal work on photometric stereo [91] demonstrates the ability to reconstruct a surface from 3 images with known lighting conditions. Even current methods still use known lighting conditions for cooperative subjects [32].

Later it was discovered that even without knowledge of the light source photometric stereo can take advantage of the low rank nature of the lighting assumption [35, 98, 56, 7, 8, 92].. Spherical harmonics, a complete set of orthogonal functions on the surface of a sphere (Figure 2.6, were proposed to help solve photometric stereo. Instead of using the bidirectional reflectance distribution function to model light, spherical harmonics could be used. For a simple Lambertian surface (Eq. 2.1), the 1st order spherical harmonics is a direct match with the values 1, n_x, n_y, and n_z. The Phong illumination model does not directly translate, but 2nd order spherical harmonics can model

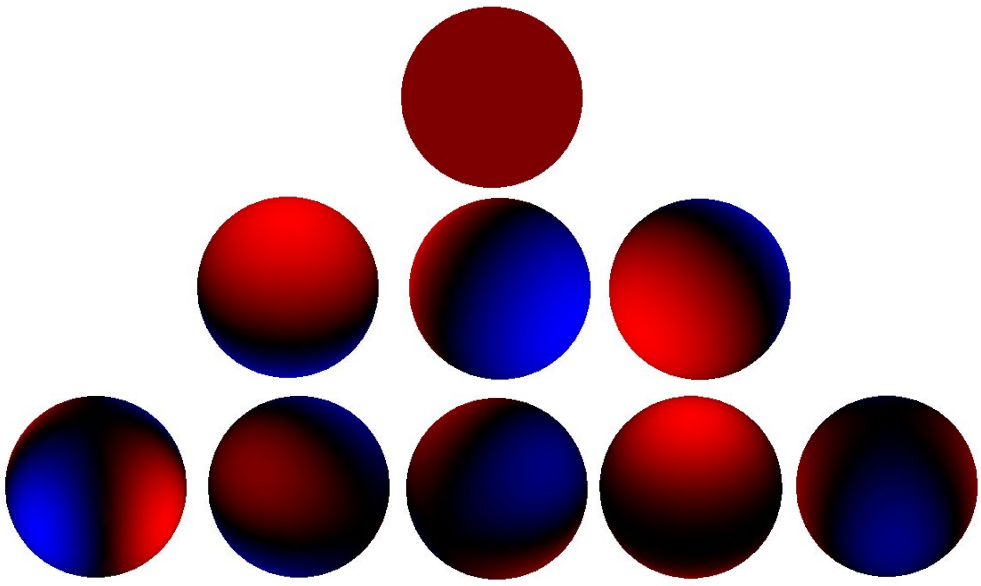


Figure 2.6: Visual representation of 0th, 1st, and 2nd order spherical harmonics. Red is positive and blue is negative with the intensity indicating the magnitude of the function.

99.2 of the lighting energy [29].

Viewing the problem with spherical harmonics allows for a creative solution through singular value decomposition (SVD). Take a matrix where each row is an image and each column is a corresponding pixel across all images, then the rank-4 SVD is a robust way of simultaneously estimating the surface normals and lighting condition with a Lambertian assumption. A rank-9 decomposition can handle more complicated lighting situations. After SVD, an integrability constraint or prior knowledge of the object is required to resolve ambiguity of the decomposition. Such approaches require a sufficient number of images to obtain an accurate reconstruction, especially for non-rigid objects like the face where expression variation can disturb the low rank assumption.

It is also important to note, that these photometric stereo techniques reconstruct from a common viewpoint and therefore produce a 2.5D surface instead of a true 3D surface like multi-view stereo. There are a few works combining multi-view stereo and photometric stereo to produce more accurate 3D surfaces [38, 75] even estimating arbitrary non-linear camera response maps.

Table 2.1: Overview of Face Reconstruction Approaches.

	Input	Approach	Detail
Constr.	Point cloud	Range scanner	±0.03mm max
	Synchronized images	Multi-view stereo [9]	0.088mm mean, pore
	Time-multiplexed	Photometric stereo [32]	wrinkle
Unconstr.	Single image	3DMM [13]	smooth
		CNN [48]	smooth
	Video	Optical flow tracking [51]	wrinkle
	RGB-D Video	Dynamic fusion [61]	wrinkle
	Photo collection	Photometric stereo [53]	wrinkle

Figure 2.7: Example reconstructions at different levels of detail. From left to right: pore, wrinkle, and smooth.

Note that these reconstruction techniques are generic and work for any arbitrary object that satisfies the material and lighting assumptions. I now move to describing face reconstruction methods that may take prior knowledge about faces in order to simplify the problem.

2.3 Face Reconstruction Methodologies

As mentioned before, the face has seen a recent growth of research for creating detailed 3D models. There are numerous methods for face reconstruction that depend on the level of control over the capturing environment and can produce different qualities of reconstructions suitable for different

tasks. Table 2.1 provides an overview of the most common scenarios and approaches. Figure 2.7 provides examples of the levels of detail different reconstruction approaches can produce. In the remainder of this section, we will provide a brief overview of the different scenarios, their most common approaches, and some of their applications.

2.3.1 Constrained

The most accurate scenario for face reconstruction is under a *constrained* environment. Constrained means there is a cooperative environment for capturing the imagery. In other words, the subject of interest may be placed in a known pose and expression, the lighting conditions may be fully known, or the camera placement and calibration may be known. In a constrained scenario, assumptions may be made about the reconstruction process which allow highly detailed and accurate models to be reconstructed.

2.3.1.1 Range Scanner

A range scanner is a special piece of equipment designed to capture a point cloud directly. There are a variety of approaches including projecting a structured infrared light pattern onto a surface and viewing its deformation, or using a laser to either measure the time-of-flight or triangulation. Regardless of the approach, a range scanner does the processing within the hardware and returns either a point cloud or depth map of the scene.

If you have money to spend, the $75,000 Konica Minolta Vivid 9i can capture details with $\pm$0.03mm accuracy across a broad range of depths. The system is suitable for manufacturing processes and reverse engineering parts. If the part is too large, it even includes software to automatically stitch together multiple scans to create a single object. To capture the entirety of an object, it includes a rotating platform. For face applications, these scans are considered as ground

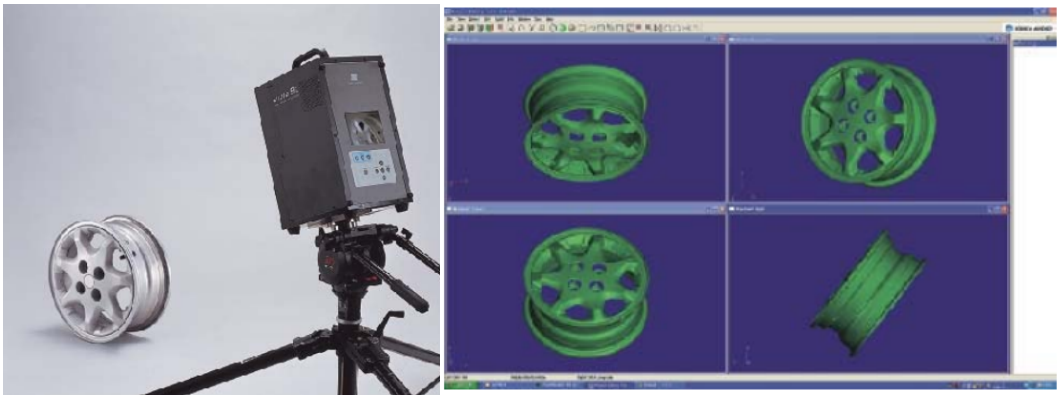


Figure 2.8: The Konica Minolta Vivid 9i is a non-contact range scanner capable of capturing multiple views of an object and creating a 3D reconstruction accurate to $\pm$0.03mm.

truth and are often used for evaluation of reconstruction approaches or creation of probabilistic face models. One downside to this level of accuracy is that the scan time takes 2 seconds for a single capture.

For a more reasonably priced $1,441, you may purchase the IIIDScan PrimeSense device. PrimeSense powers the Microsoft Kinect entertainment system. This particular PrimeSense device captures with $\pm$0.5mm accuracy, which may not work for manufacturing, but will still capture wrinkles on human faces. What it lacks in accuracy, it more than makes up in scan time, since this device can capture in 30 frames per second, making it the applicable for dynamic fusion.

2.3.1.2 Multi-View Stereo

Humans have depth perception based in part by having two eyes that can triangulate objects in space. Multi-view stereo imaging uses this same principle by placing two or more cameras with known positions and intrinsic camera matrices. When a common point is identified in both views, the 3D position is uniquely determined based on the intersection of the rays originating at each camera and passing through the common point. Beeler *et al*. present state-of-the-art methods for reconstructing faces based on a consumer setup of cameras [9] or stereo video [10]. Their work

first calibrates the cameras based on a calibration sphere with known fiducials in order to determine the camera matrices and relative positions in space. Once calibrated, frames are captured within 0.1 seconds of each other from all cameras and a multi-view stereo approach reconstructs faces with pore and wrinkle level detail. Working on an image pyramid from low to high resolution, it searches the epipolar lines between pairs of cameras based using the normalized cross correlation to identify common points between cameras. The common points are further refined based on photometric and surface consistency. From the common points, a 3D point cloud may be obtained using the known camera parameters. Poisson surface reconstruction fits a surface to the point cloud which is then further refined to add in the pore level details.

These reconstructions are known to be extremely accurate and are often used as ground truths in place of laser scans since the equipment is relatively cheap. Using a 3D printed mask instead of a face [9] demonstrates an average error of 0.088mm using a setup of 7 DSLR cameras. A real human subject will have higher error since the mask had consistent albedo and no specularity or sub-surface reflection present in human skin. However, they come at a cost of computation time as it takes 20 minutes to reconstruct a single image, and the subject must be stationary in a specialized room.

2.3.1.3 Photometric Stereo

The principles behind general photometric stereo were described in detail in Sec. 2.2.0.2. One of the challenges of photometric stereo is the time delay between images necessary to change lighting. For rigid objects, this is not an issue, but for non-rigid faces, it is desirable to simultaneously capture all images. In order to capture different lighting conditions simultaneously, [37, 85] use three different colored lights (red, green, and blue) to illuminate the scene from different viewpoints. Since this is the minimum number of illumination variations to determine the surface normal,

extra care must be taken for regions of shadows unlike approaches using four light sources [6]. Another approach designed for humans is [32] where near infrared light is used for illumination since it is not as obtrusive as visible light and will not disturb the subject during illumination.

For the best results, combining multi-view stereo and photometric stereo can achieve extremely detailed results.

2.3.2 Unconstrained

When the subject is not cooperative and no prior knowledge about the lighting or cameras is known, face reconstruction becomes *unconstrained*. Under these scenarios, prior knowledge about face shapes becomes highly important to create a compelling reconstruction.

2.3.2.1 3D Morphable Model

This seminal work on unconstrained face reconstruction works on a single image and was proposed by Blanz and Vetter [14]. They designed a 3D Morphable Model (3DMM) which extended classic active appearance models from 2D face alignment to fit a 3D model onto an image based on the texture information. The 3DMM is created by taking the ground truth scans of 200 individuals and putting them into correspondence. The assumption is that an arbitrary face can be expressed as a linear combination of the scanned faces.

$$\mathbf{S}_{\text{mod}} = \sum_{i=1}^{200} \alpha_i \mathbf{S}_i, \quad \mathbf{T}_{\text{mod}} = \sum_{i=1}^{200} \beta_i \mathbf{T}_i, \quad \sum_{i=1}^{200} \alpha_i = \sum_{i=1}^{200} \beta_i = 1. \tag{2.4}$$

The face space can then be compressed and expressed as a statistical distribution by using Principal Component Analysis (PCA) [47],

$$\mathbf{S}_{\text{mod}} = \bar{\mathbf{S}} + \sum_{i=1}^{199} \alpha_i \mathbf{S}_i, \quad \mathbf{T}_{\text{mod}} = \bar{\mathbf{T}} + \sum_{i=1}^{199} \beta_i \mathbf{T}_i, \tag{2.5}$$

where the probability for coefficients $\vec{\alpha}$ becomes,

$$p(\vec{\alpha}) \sim exp[-\frac{1}{2}\sum_{i=1}^{199}\left(\frac{\alpha_i}{\sigma_i}\right)^2], \tag{2.6}$$

where σ_i^2 is the eigenvalue of the shape covariance.

Given the model as prior knowledge about expected face shape and texture, the goal once again is to infer pose, lighting, shape, and texture parameters of the scene based on a single image. The initial formulation is an analysis by synthesis approach where the error is the distance between the real image and the models rendered image along with the likelihood of observing the coefficients. This formulation is non-convex and tends to fit to local minimums, so they have manual annotation of the lighting and initial pose parameters to help the model fit to the face.

To attempt to fix the non-convex energy function, [71] extended the fitting procedure to include edges, specular highlights, and texture constraints. Morphable models have further been decomposed into identity and expression basis instead of a single combined shape basis [84, 22]. As time progressed, 2D landmark alignment, the process of finding keypoints such as the eyes, nose, and mouth in images improved. Recent works, fit morphable models based on the projection error of the 2D landmark alignment which is very efficient and produces reasonable results [95, 2].

The 3DMM is ubiquitous in face reconstruction approaches and is a major component of a large percentage of different approaches. One major complaint with the 3DMM is the small training size

of subject. A recent paper [15] builds a new model from $10,000$ people, that when released should help approaches generalize better to a wide range of face shapes.

2.3.2.2 Single Image

Besides the 3DMM, there are a number of other techniques for performing reconstruction from a single image. [52] takes a single reference model, manually aligns it to the image, and morphs it based on a shape from shading approach. The solution is a 2.5D recovery of the frontal part of the face.

A pair of recent works perform reconstruction using a CNN to fit a 3DMM instead of the model-based fitting [100, 48]. These works focus on the problem of dense face alignment, where every pixel on the face is assigned a location on a canonical face, but in the process they fit a cascade of regressors to solve the projection parameters and 3DMM shape coefficients. In doing so, they actually produce a reconstruction for the image as well.

2.3.2.3 Video-based Reconstruction

There is a lot of interest in video-based reconstruction. We briefly highlight a few of the key works. In [30], sparse 2D landmarks in the video are tracked and a 3DMM is fit to each frame, the model is then refined using optical flow and temporal consistency. This approach works reasonably well on in the wild videos, but struggles with specular reflections. In [81], instead of fitting a 3DMM, a surface reconstructed similar to our proposed method from a photo collection is used as input, the pose for each frame is estimated and similar to [30] optical flow refines the alignment of the model and deforms to the expressions present, and a temporal consistency is maintained both forwards and backwards in the video. In [22], a 3DMM is separated into identity and expression basis and then fit to a still image, finally, a Kinect sensor is used to find the expression coefficients

and transfer the expression of a user to puppeteer the fit 3DMM. In [46], a cascaded regressor is learned on a large set of face scans to perform 3D face alignment for a single image or a video. The resulting alignment has only 1024 vertices and does not capture wrinkle level details of the face. In [20], a low resolution mesh is tracked in a video similar to prior works, and the wrinkle details are estimated based on regressors from local patches. This allows visually appealing wrinkle level reconstructions in real time, but the details are not metrically correct for the observed face. In [44], an individual uses a cell phone to record their face in a neutral expression from all poses, structure from motion fits a static face model from the video. Then the individual records themselves making predefined expressions in order to build a blend shape model. This interactive approach allows them to create wrinkle level avatars which may be controlled by other videos. In [82] a smooth model is fit to two videos in real time and the expressions from one are transferred to the other for real time puppeteering.

One major application for video reconstruction is puppeteering. Avatar puppeteering is the process where a person controls the actions of virtual person called an avatar through their own actions, usually as recorded by a video camera, but also through other means such as strain sensors pressed on a face. The popular movie *Being John Malkovich* is an example of a puppeteer. This application requires a detailed model of the avatar, which can either be completely computer generated by an artist, or reconstructed based on a real person. The model needs to be rigged in order to deform to match the desired expressions. To control the model from a video camera, face tracking fits the avatar model to most closely match the pose and expression of the puppeteer in the video at each frame while also incorporating temporal consistency. This face tracking shares some similarities with face reconstruction, especially when the avatar is a model of puppeteer. These systems are most interested in creating a visually compelling avatar and the efficiency of the algorithm to run in real time, but the metric accuracy of the reconstruction is secondary to the visual

cohesiveness.

2.3.2.4 Photo Collections

Photo collection methods were discussed in detail in Section 1.1.

Photo collections are more challenging than videos since no temporal constraints are available. In theory, they possess more information than a single image, but a naive use of a photo collection can actually produce a smoothed reconstruction since the face is non-rigid and may change between images. Photo collections are highly relevant since most people have a personal collection of photos. For biometrics purposes, a recent face database [54] introduces collection-based matching procedures due to the increasingly common occurrence of capturing multiple images of a suspect in forensic applications. An accurate 3D model of the face can improve face recognition.

2.4 Applications

Just as there are many different methods for reconstruction, there are numerous applications. While the scope of this work is mainly academic in nature in terms of how accurate of a reconstruction is possible from a photo collection, I do want to mention some of the potential applications of face reconstruction.

2.4.0.0.1 Medical Person specific 3D face models are used in the medical field. They are used to visualize the patients exact structure and design form fitting pieces. For example, the University of Michigan aligns 3D scans over time in order to track topographical changes that may be caused by different diseases. Models are used in surgical correction, orthopedic correction, and correcting craniofacial anomalies like cleft palettes. For medical applications, precise 3D

structure is important, and range scanners or constrained stereo setups are required to produce sub-millimeter level accuracy.

2.4.0.0.2 Face recognition Face recognition is known to be directly applicable from 3D to 3D. But recently, works demonstrate the effectiveness of using a 3D model to improve traditional 2D to 2D face recognition. Models are used to normalize pose, expression, and lighting in images in order to render an improved image that will have higher face recognition accuracy. It is unclear what level of detail is required in the model for optimal face recognition performance. Even using a hyper-ellipsoid as a face model improves face recognition [59]. Researchers have shown that using a generic 3D face can improve face recognition [34]. One initial objective of the 3DMM was to improve face recognition, and it has been shown to be effective for normalization as well [101]. However, there is need for a comparative evaluation determining the relative effectiveness of better person specific models.

2.4.0.0.3 Commercial video editing Person specific face models may be used to modify videos. They may be used to post-process a video by adding virtual makeup and changing lighting, or they can be used to create a completely CGI scene for a movie like avatar. For video editing, a blend shape model is desired, where the face may be deformed to match a different set of expressions. The blend shape may be a generic set of expressions obtained from a 3DMM, or a person-specific set of expressions captures of the true actor. For high definition commercial films, pore level accuracy and person-specific expressions are desired. Since the output will be high resolution and displayed on theater screens, consumers expect a high quality and accurate rendering.

2.4.0.0.4 Virtual communication This can be viewed as a low-detail version of video editing. By reconstruction or tracking a face in real time, an artificial model may be rendered on a different

screen to provide communication. Think video chat, but instead of sending a video, you send the parameters of the face reconstruction. This provides an immediate benefit of compression, but also allows for entertaining changes to the face. Snapchat provides filters to change the appearance of your own face, Face2Face [82] shows how to control a different persons face, or you could control virtual avatar. For these applications, realism is more important than accuracy, as oftentimes the face will be made a caricature anyways.

2.5 Organization

The remainder of the thesis is outlines as follows. In Chapter 3, I present the extension of the SVD-based reconstruction into a true 3D surface. Using a novel normal field-based Laplace editing technique, the proposed approach deforms a triangulated mesh to match the observed surface normals. In Chapter 4 the work is extend to adapt to small photo collections. Using a 3DMM as initialization and an adaptive energy minimization formulation for photometric stereo allows reconstructions from even a few images. I also introduce the SSIM-based quality evaluation metric. In Chapter 5 I present a discussion on potential applications and suggestions for future development.

Table 2.2, presents a list of the most common notations used through the thesis.

Table 2.2: 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
$\mathbf{X}$	3xp	3D shape model
p	scalar	number of mesh vertices
$\mathbf{N}$	4xp	surface normal matrix
$\mathbf{L}$	4xn	lighting matrix
$\mathbf{D}$	nxp	dependability matrix
$\mathscr{L}$	pxp	sparse Laplacian
s	scalar	scale
$\mathbf{R}$	2x3	projected rotation matrix
$\mathbf{t}$	2x1	translation vector
$\vec{\rho}$	1xp	albedo
$\mathbf{F}$	nxp	image correspondence
H_j	scalar	mean curvature

Chapter 3

Unconstrained 3D Face Reconstruction

3.1 Introduction

As shown in Fig. 1.1, 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 [83] or video [81, 30]. Kemelmacher-Shlizerman and Seitz developed an impressive photometric stereo-based method to produce high-quality face models from photo collections [53], 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 [53]. 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, in this chapter, I propose a novel approach to 3D face reconstruction where a number of crucial innovative components are designed and developed. Our approach is also motivated by the recent explosion of face alignment techniques [58, 93, 49, 70], where the precision of 2D landmark estimation has been substantially improved. Specifically, given a collection of unconstrained face images, we first perform 2D landmark estimation [93] of each image. In order to prepare an enhanced 3D template for the photometric stereo, we deform a generic 3D face template

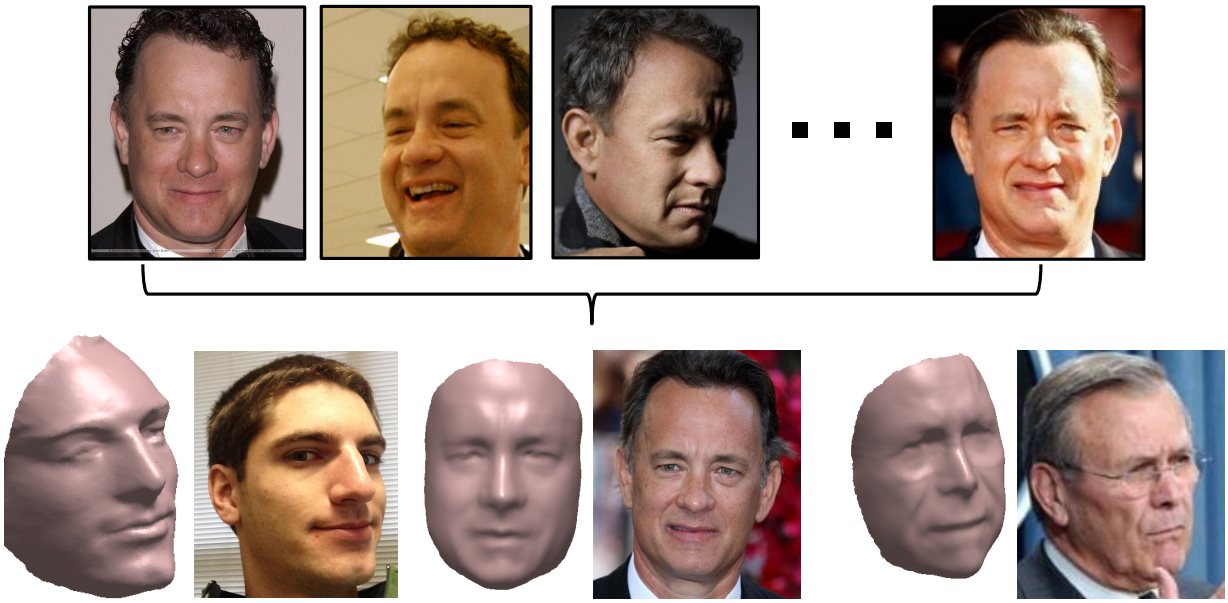


Figure 3.1: Given an unconstrained photo collection of Tom Hanks containing images of unknown pose, expression, and lighting, face reconstruction seeks to create an accurate 3D model of his face. Two additional example reconstructions along with a single representative image are shown as well.

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 chapter 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.

3.2 Proposed Algorithm

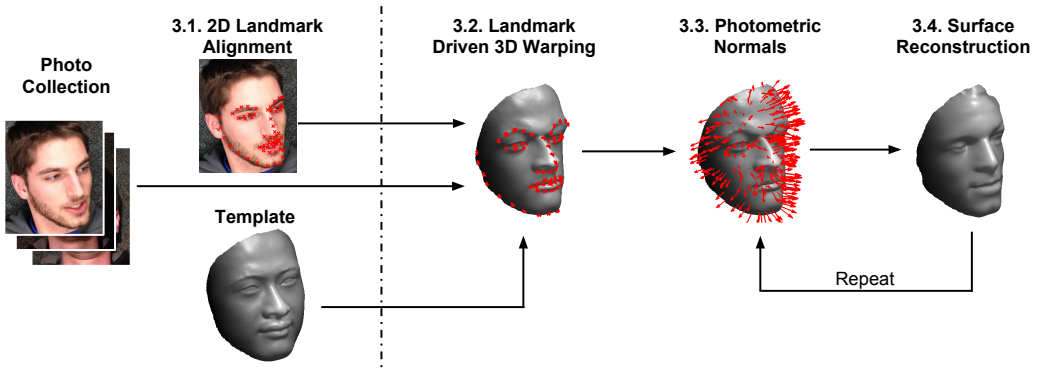


Figure 3.2: Overview of our 3D face reconstruction. Given a photo collection, a generic template face mesh, and 2D landmark alignment, we propose an iterative process to warp the mesh based on the estimated 3D landmarks and the photometric stereo-based normals.

The proposed algorithm operates on a photo collection of n images of an individual. No con-

straints 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.

Figure 3.2 illustrates the major components and pipeline of our proposed algorithm. Throughout the algorithm description, we present multiple reconstructions where we omit the part being discussed to demonstrate the effects and needs of each different part of the reconstruction process.

3.2.1 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 [93] to automatically fit q (=68) landmarks onto each image. An example of the landmark fitting is given in Figure 3.2. Given an image $\mathbf{I}(u,v)$, the landmark alignment returns a $2\times q$ matrix $\mathbf{W}_i$.

3.2.2 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.

Figure 3.3 demonstrates the motivation for why landmark driven warping is important. Not only does the warping process give the correct aspect ratio, but it also enables a clearer reconstruction since better correspondence may be established throughout the collection.

(a) (b) (c)

Figure 3.3: Effects of landmark driven warping on final face reconstruction. (a) Full reconstruction process from warped template for George Clooney, (b) reconstruction from initial template without warping, and (c) image of Clooney used in reconstruction.

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 [78] 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,

$$\int_{\Omega} \|\Delta \mathbf{x} - \Delta \mathbf{x}_0\|^2 + \frac{\lambda_l}{n} \sum_i \|\mathbf{R}_i \mathbf{X}\mathbf{C} - \mathbf{W}_i\|_F^2, \tag{3.1}$$

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 annotated on the mesh $\mathbf{XC}$ weakly perspective projected through $\mathbf{R}_i$ and the 2D landmark locations $\mathbf{W}_i$ for image i;

and λ_l is the weight for landmark correspondence. $\mathbf{C}$ is a $p \times q$ selection matrix where each column selects a single manually annotated landmark vertex from the mesh, *i.e.*, it is a sparse matrix with a 1 in each column at the vertex index for the corresponding landmark. 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 3.1, we discretize the surface patch Ω as a triangle mesh with p vertices, with the vertex locations concatenated as a $3 \times p$-dimensional matrix $\mathbf{X}$. 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 3.1 is thus turned into a quadratic function on $\mathbf{X}$,

$$E_{\text{warp}}(\mathbf{X}, \mathbf{R}_i) = \|\mathbf{X}\mathscr{L} - \mathbf{X}_0\mathscr{L}\|_F^2 + \frac{\lambda_l}{n}\sum_i \|\mathbf{R}_i\mathbf{X}\mathbf{C} - \mathbf{W}_i\|_F^2, \tag{3.2}$$

where $\mathscr{L}$ is a discretization of Δ. Using linear finite elements, it is turned into a *symmetric* matrix with entries $\mathscr{L}_{ij} = \frac{1}{2}(\cot\alpha_{ij} + \cot\beta_{ij})$, where α_{ij} and β_{ij} are the opposite angles of edge ij in the two incident triangles (see Figure 3.4), known as the cotan formula [66].

In order to find the minimizer, we first estimate an initial $\mathbf{R}_i$ via the corresponding pairs of 2D and 3D landmarks. With the projection matrices $\mathbf{R}_i$ 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 [78, 97], we can resolve this issue by noting that

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

which means that the Laplacian of the position 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

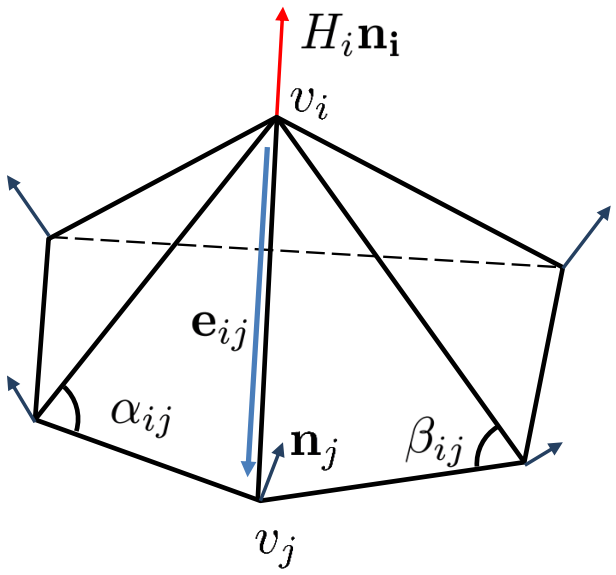


Figure 3.4: The mean curvature normal indicates how a vertex deviates from the average location of its immediate neighbors, which can be evaluated as the Laplacian of the position. The mean curvature H_i can be evaluated through $\mathbf{n}$.

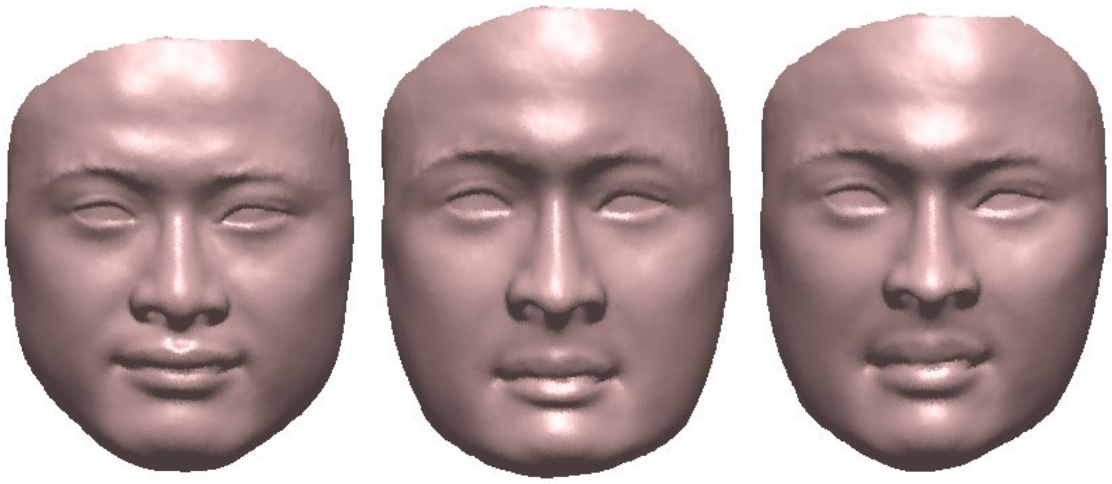
and the mean curvature scalar H. Thus, to keep the original geometric detail while allowing it to rotate, we compute the original mean curvature H_0 (the discretization of which corresponds to the integral of the mean curvature in a neighborhood around each vertex), and update $\mathbf{n}^k$ according to the direction of $\mathbf{X}^k\mathscr{L}$ for the shape $\mathbf{X}^k$ at iteration k, and solve for

$$\mathbf{X}^{k+1} = \underset{\mathbf{X}}{\operatorname{argmin}}(\|\mathbf{X}\mathscr{L} + \mathbf{N}^k\mathbf{H}_0\|_F^2 + \frac{\lambda_l}{n}\sum_i \|\mathbf{R}_i^k\mathbf{X}\mathbf{C} - \mathbf{W}_i\|_F^2), \tag{3.3}$$

where $\mathbf{N}$ is a $3\times p$ matrix of all vertex surface normals and $\mathbf{H}_0$ is a diagonal matrix with H_0 for each vertex. This leads to a linear system,

$$(\mathscr{L}^2 + \lambda_l\mathbf{C}\mathbf{C}^\intercal)\mathbf{X} = -\mathbf{N}^k\mathbf{H}_0\mathscr{L} + \frac{\lambda_l}{n}\sum_i (\mathbf{R}_i^k)^\intercal\mathbf{W}_i\mathbf{C}^\intercal. \tag{3.4}$$

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



(a) (b) (c)

Figure 3.5: Example of template deformation. (a) Initial generic template, (b) template warped for Justin Trudeau, and (c) template warped for Kiera Knightley.

Figure 3.5 shows the initial template as well as examples after warping has completed. Notice that we maintain the appearance of the initial template since we try to keep the same curvature. The only change is that the keypoints on the face have been moved to better align with the observed landmarks in the collection, which changes the aspect ratio, height of nose, etc.

3.2.3 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 [53]. The approach in [53] 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 [53] 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 $\mathbf{I}_x = \mathbf{l}\mathbf{s}_x$.

3.2.3.1 Initial Normal Estimation

In this section, we assume point correspondence between the current mesh and each image in the collection. We create an $n \times p$ correspondence matrix $\mathbf{F}$ by storing in f_{ij} 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 $\mathbf{R}_i$ matrices from Section 3.2.2. For non-frontal images, there are vertices that are not visible due to the projection. When this occurs, we set $f_{ij} = 0$ and use matrix completion [57] 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 $n \times 4$ matrix $\mathbf{L}$, and the shape vector $\mathbf{s}_j$ for each vertex j into a $4 \times p$ matrix $\mathbf{S}$, we have $\mathbf{M} = \mathbf{L}\mathbf{S}$. 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 $\mathbf{M} = \mathbf{U}\Lambda\mathbf{V}^\intercal$ and use the rank-4 approximation $\mathbf{M} = \tilde{\mathbf{L}}\tilde{\mathbf{S}}$ where $\tilde{\mathbf{L}} = \mathbf{U}\sqrt{\Lambda}$ and $\tilde{\mathbf{S}} = \sqrt{\Lambda}\mathbf{V}^\intercal$. $\tilde{\mathbf{L}}$ 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 [53] states that it may be unstable for images with expression variations. Thus, we follow the approach from [53], where we select images that are modeled well by the low rank approximation, *i.e.*, $\|\mathbf{M} - \tilde{\mathbf{L}}\tilde{\mathbf{S}}\|^2 < \varepsilon$, 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 $\mathbf{S} = \mathbf{A}\tilde{\mathbf{S}}$.

3.2.3.2 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.*, $m_{ij} = \rho_j \mathbf{l}_i \mathbf{n}_j$ 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 $\text{argmin}_{\mathbf{l}_i} \sum_j \|\rho_j \mathbf{l}_i \mathbf{n}_j - m_{ij}\|^2$ and then for ρ_j directly by $\rho_j = m_{ij}/(\mathbf{l}_i \mathbf{n}_j)$. 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.

3.2.3.3 Local Normal Refinement

The initial normal estimation produces a smoothed result that is akin to the mean shape. We follow the procedure from [53] 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 [53]. We select a subset of k images with the minimum

distance $\|\mathbf{m}_j - \mathbf{L}\mathbf{s}_j\|^2$ with $k \geq 4$ images and enough to produce a low condition number of $\mathbf{L}_{k\times 4}$. We recover the local shape $\mathbf{s}_j$ via

$$\min_{\mathbf{s}_j} \|\mathbf{m}_{k\times 1} - \mathbf{L}_{k\times 4}\mathbf{s}_j\|_F^2, \tag{3.5}$$

where we omitted the Tikhonov regularization term proposed by [53], 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.

3.2.4 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

$$\|\mathbf{X}\mathscr{L} + \mathbf{N}\mathbf{H}\|_F^2. \tag{3.6}$$

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 [79]. 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,

$$H_i = \frac{1}{4A_i} \sum_{j \in N(i)} (\cot \alpha_{ij} + \cot \beta_{ij})\, \mathbf{e}_{ij} \cdot (\mathbf{n}_j - \mathbf{n}_i), \tag{3.7}$$

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 3.4). 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.

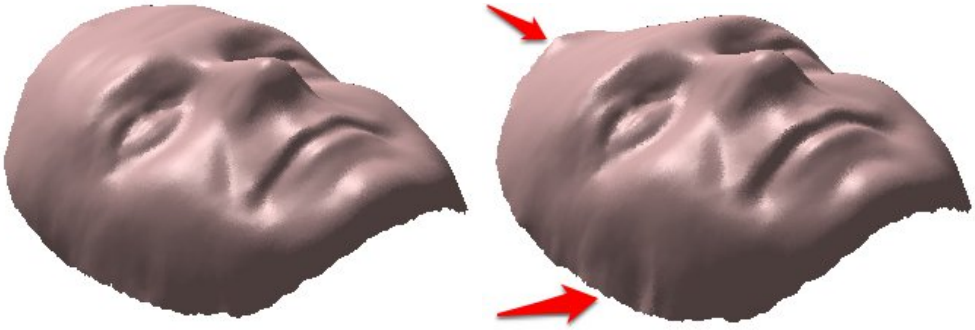
One unique challenge in handling a 3D model instead of a height field is the boundary. On the boundary, the mean curvature formula degenerates into a 1D version

$$\mathbf{X}\mathscr{L}_b = \mathbf{b}\kappa, \tag{3.8}$$

which is based on the 1D Laplace operator $\mathscr{L}_b$ with non-zero entries corresponding to boundary edges, with $\mathscr{L}_{b,ij} = 1/e_{ij}$, 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 normal does not provide information about κ, we simply use $\mathbf{b}^k \kappa^k = \mathbf{X}^k \mathscr{L}_b$ where $\mathbf{X}^k$ is the estimated shape in k-th iteration.

Figure 3.6 shows the need to include a boundary constraint. Since the Laplace operator is degenerate along the boundary, it attempts to pull the reconstructed surface in on itself. This is manifest especially along the forehead where there are no nearby landmarks to help constrain the reconstruction, but is also observable along the rest of the boundary where it causes folds in the surface.

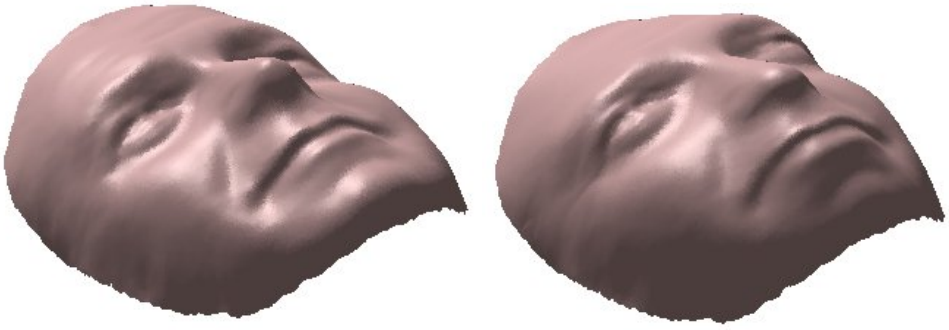
We also maintain the landmark constraint, $\lambda_l \sum_i \|\mathbf{R}_i^k \mathbf{X}\mathbf{C} - \mathbf{W}_i\|_F^2$, in the reconstruction, in order



(a) (b)

Figure 3.6: Effects of boundary constraints on face reconstruction. (a) Full reconstruction of Clooney and (b) boundary constraint removed.

to keep the correspondence between the reconstructed face and the images in the collection. It also helps prevent small errors in the surface normal estimation from accumulating across the face and creating drift. Figure 3.7 demonstrates the result of omitting the landmark constraint from the reconstruction.



(a) (b)

Figure 3.7: Effects of landmark constraint on face reconstruction. (a) Full reconstruction of Clooney and (b) landmark constraint removed.

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

$$|\mathbf{X}\mathscr{L} + \mathbf{N}\mathbf{H}^k\|_F^2 + \lambda_b\|\mathbf{X}\mathscr{L}_b - \mathbf{X}^k\mathscr{L}_b\|_F^2 + \frac{\lambda_l}{n}\sum_i \|\mathbf{R}_i^k\mathbf{X}\mathbf{C} - \mathbf{W}_i\|_F^2, \tag{3.9}$$

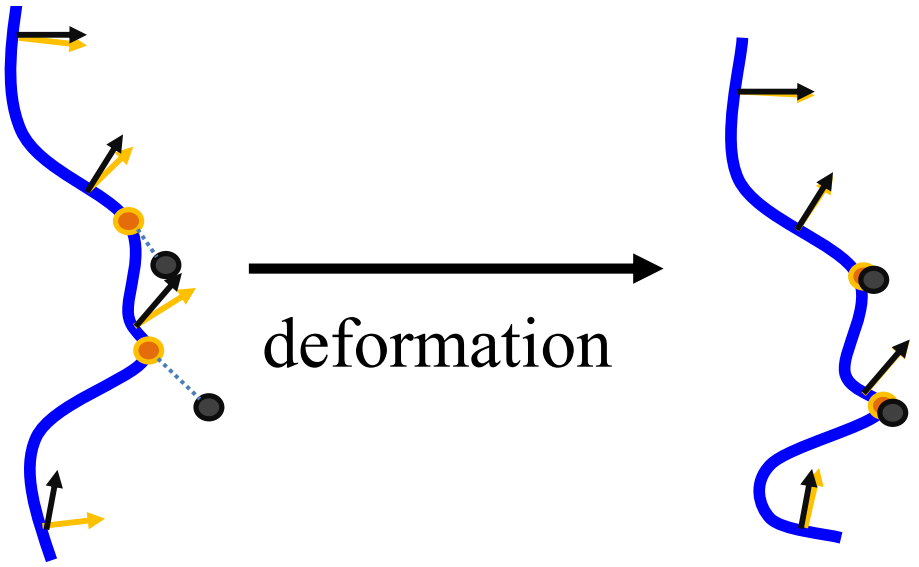


Figure 3.8: The effects of the deformation-based surface reconstruction. The black arrows indicate the photometric normal estimates and the orange arrows show the actual surface normal. The black dots are the target landmark locations and the orange dots are the corresponding vertices in the mesh.

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

$$(\mathscr{L}^2 + \lambda_b \mathscr{L}_b^2 + \lambda_l \mathbf{C}\mathbf{C}^\intercal)\mathbf{X} = -\mathbf{N}\mathbf{H}^k \mathscr{L} + \lambda_b \mathbf{X}^k \mathscr{L}_b + \frac{\lambda_l}{n} \sum_i (\mathbf{R}_i^k)^\intercal \mathbf{W}_i \mathbf{C}^\intercal, \tag{3.10}$$

where λ_b is the boundary constraint weight. Figure 3.8 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

$$(\mathbb{I} + w_d \mathbf{T}(\mathscr{L} + w_s \mathscr{L}^2)\mathbf{T})\mathbf{n} = (\mathbb{I} + w_d \mathbf{T}\mathscr{L}\mathbf{T})\tilde{\mathbf{n}}, \tag{3.11}$$

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. 3.2.1)
// warp template through landmarks (Sec. 3.2.2)
while *template deformation* $>$ *threshold* **do**
    estimate projection $\mathbf{R}_i$ for each image
    solve Eq 3.4
while 3*D face vector* $\mathbf{X}$ *change* $>$ *threshold* **do**
    // estimate $\mathbf{n}$ and ρ (Sec. 3.2.3)
    re-estimate $\mathbf{R}_i^k$
    establish correspondence $\mathbf{F}$
    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 3.5
    // deform $\mathbf{X}$ with $\mathbf{n}$ (Sec. 3.2.4)
    update $\mathbf{n}$ via Eq 3.11
    estimate mean curvature via Eq 4.18
    solve Eq 3.10

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.

3.3 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 [43], 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.

Figure 3.9: Visual comparison on Bing celebrities with images from [53] 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.

3.3.1 Data Preparation

3.3.1.0.1 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

Figure 3.10: Results on subjects from the LFW dataset. The reconstructed 3D model, sample image from which we extract the texture, and a novel rendered viewpoint.

model [99], which we subdivide to create more vertices and thereby a higher resolution.

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

Table 3.1: Distances of the reconstruction to the ground truth.

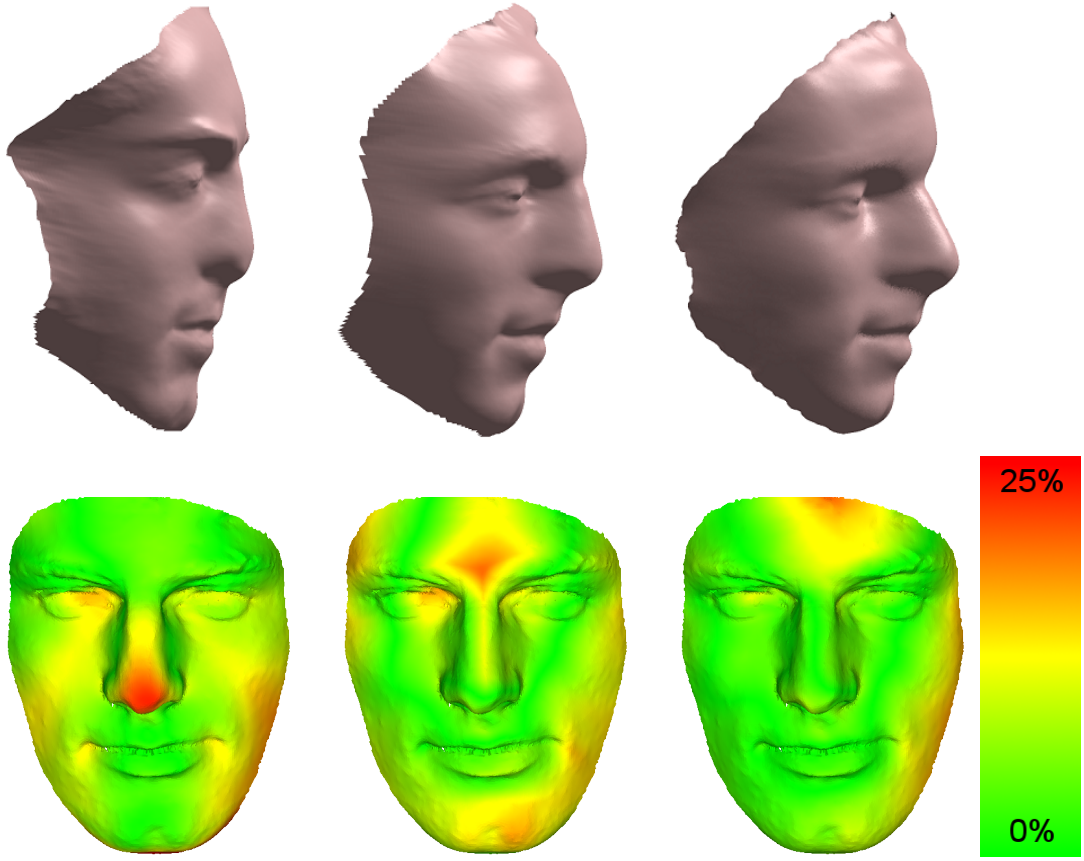
Methods	2.5D	2.5 Improved	3D
Mean	7.86%	7.79%	**5.42**%
RMS	9.71%	9.04%	**6.89**%

3.3.2 Results

3.3.2.0.1 Qualitative evaluation We process the same celebrities as used in [53], 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 3.9 compares the results between our approach and the figures from [53]. Figure 3.10 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 [53], but closer to the true geometry.

3.3.2.0.2 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 [53] 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 [40]. We report the mean euclidean distance and the root mean square (RMS) of the distance, after normalized by the eye-

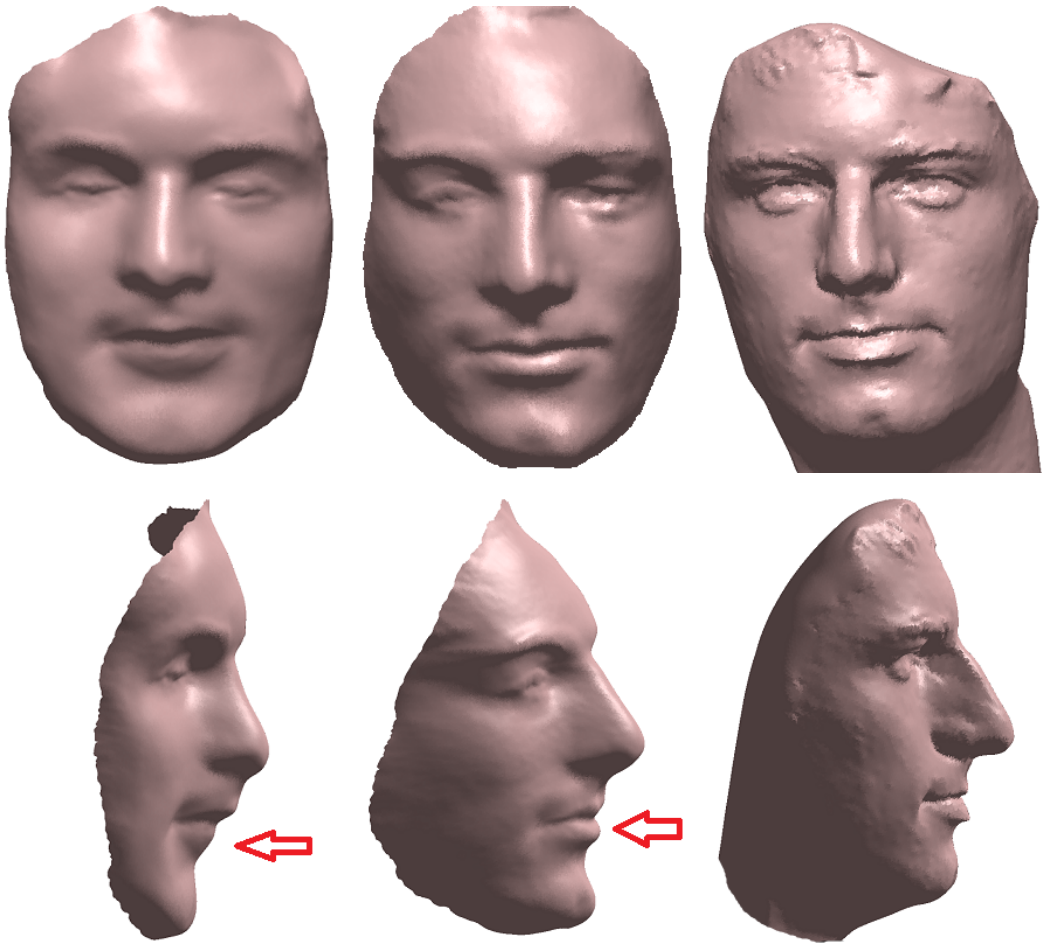


(a) (b) (c)

Figure 3.11: 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.

to-eye distance, in Table 3.1. Figure 3.11 shows a coloring of the template to visualize where on the face is close for reconstruction. The base 2.5D approach (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.

3.3.2.0.3 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



(a) (b) (c)

Figure 3.12: 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.

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 3.12 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 [19].

3.3.2.0.4 Additional Reconstructions To further demonstrate the performance and generalization of the proposed approach, we present a larger set of individuals processed through the Bing and Picassa pipeline. Note that for these individuals, the photo collections contain images mainly within ± 30 degrees yaw due to the lack of profile images returned through the Bing API.

I mainly leave the figures to speak for themselves, but comment on a few small points.

- Most reconstructions are convincing, and the identity is easily recognizable.
- Prominent wrinkles are captured for some people including Harry Reid and Robin Williams.
- Vin Diesel's common use of sunglasses appears in the reconstruction.
- Jim Carrey's extreme expression variations cause the reconstruction to fail.
- Denzel Washington has high amounts of specularities that are not modeled by our lighting assumption possibly contributing to his poor reconstruction.

3.4 Summary

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

Figure 3.13: Visualization of many successful examples.

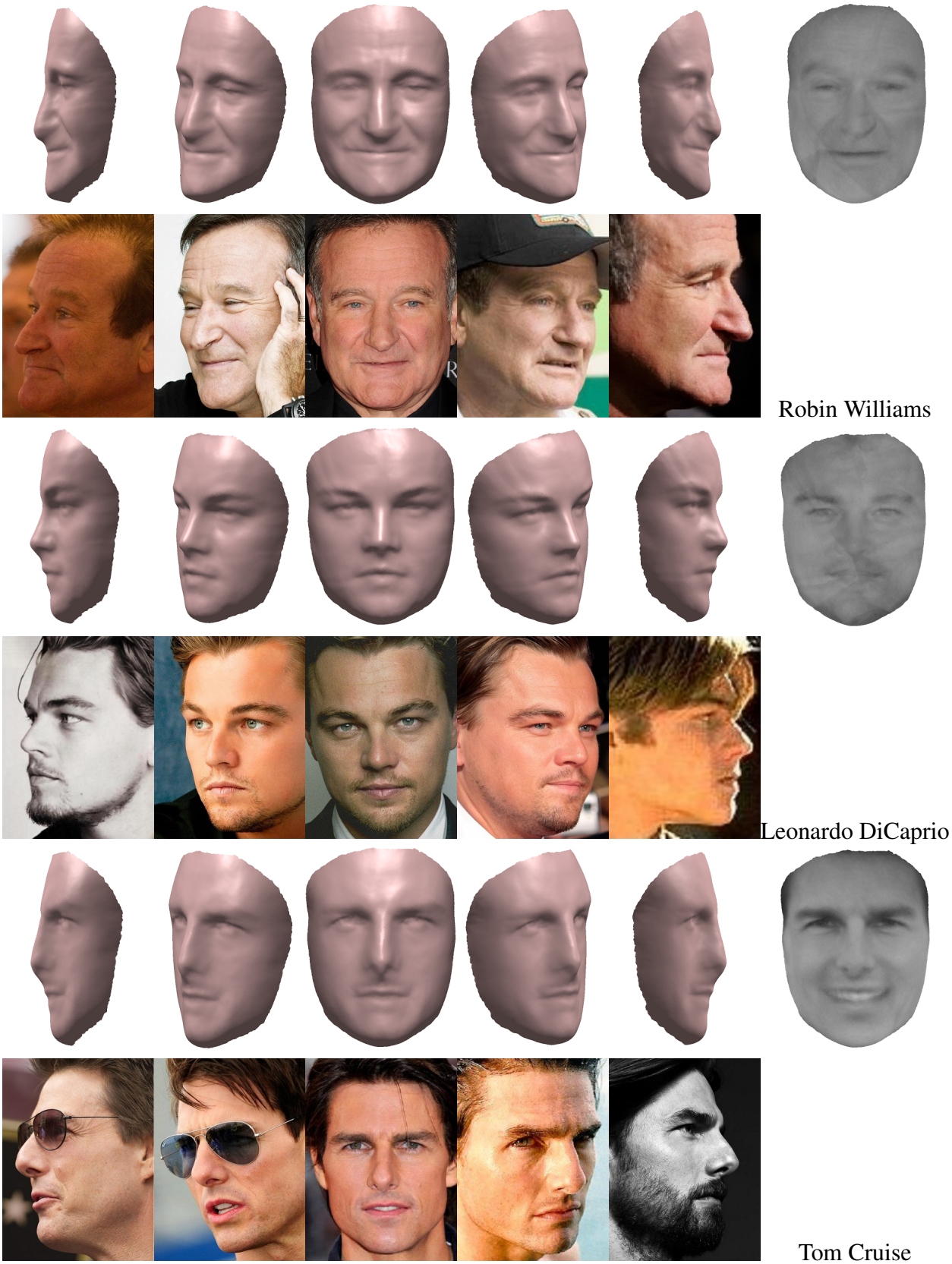


Figure 3.14: Visualization of many successful examples.

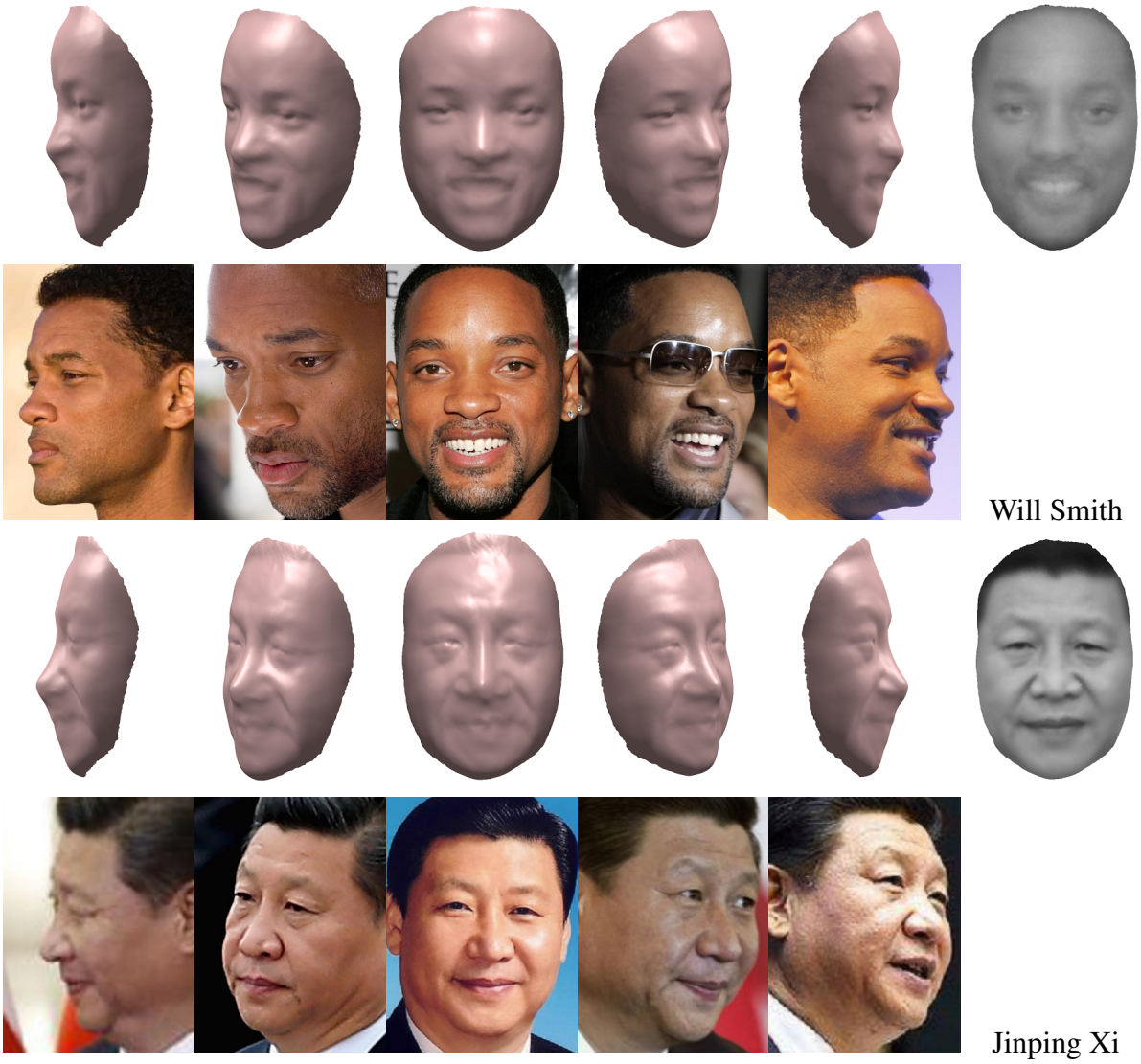
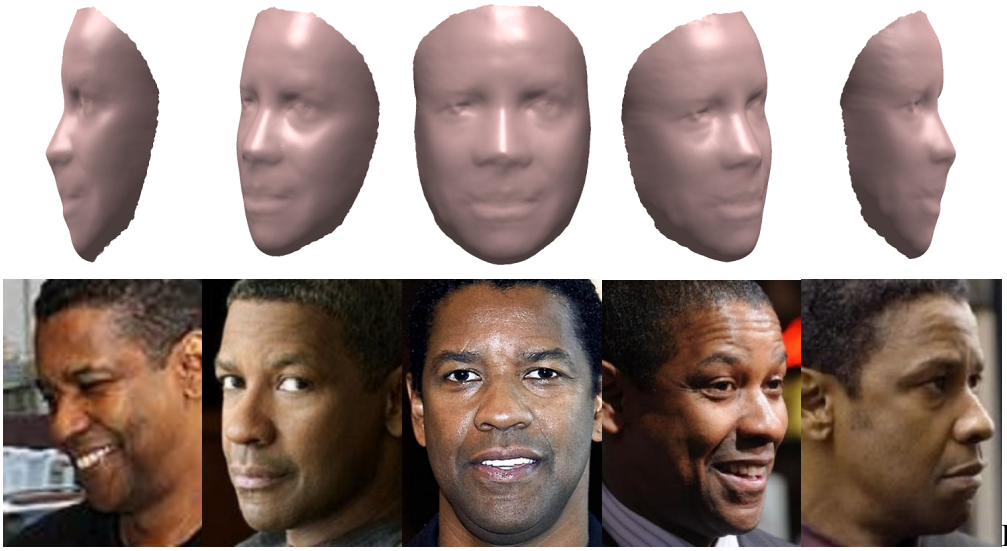
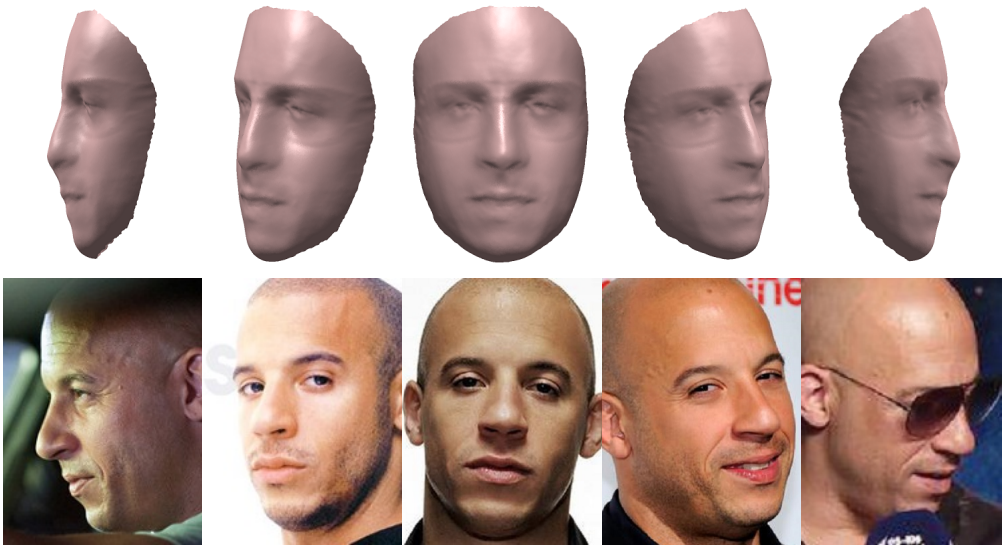


Figure 3.15: Visualization of many successful examples.


Denzel Washington


Vin Diesel


Jim Carrey

Figure 3.16: Visualization of difficult examples.

automatically detected 2D landmarks in the profile views, validating our approach on a diverse set of populations, and extending to non-face objects.

Chapter 4

Adaptive 3D Face Reconstruction from Unconstrained Photo Collections

4.1 Introduction

In this chapter we continue to improve unconstrained face reconstruction by making it more adaptive to the number of images in the collection as well as the quality of the images. The reconstruction approach presented in the Chapter 3 still has limitations. Frontal images were required for [53], and even though the previous chapter can use non-frontal images, we will demonstrate that its performance drops significantly with large pose variation. We propose a pose-based dependability measure to explicitly handle non-frontal images. Another limitation is 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 over one hundred image collections 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 part of the face. We propose an energy minimization solution with an adaptive template regularization to reconstruct small collections, even down to a single image. And we propose to use a larger region for local selection to allow use of $\sim$50% of images.

To perform face reconstruction, given a collection of unconstrained face images, we first align

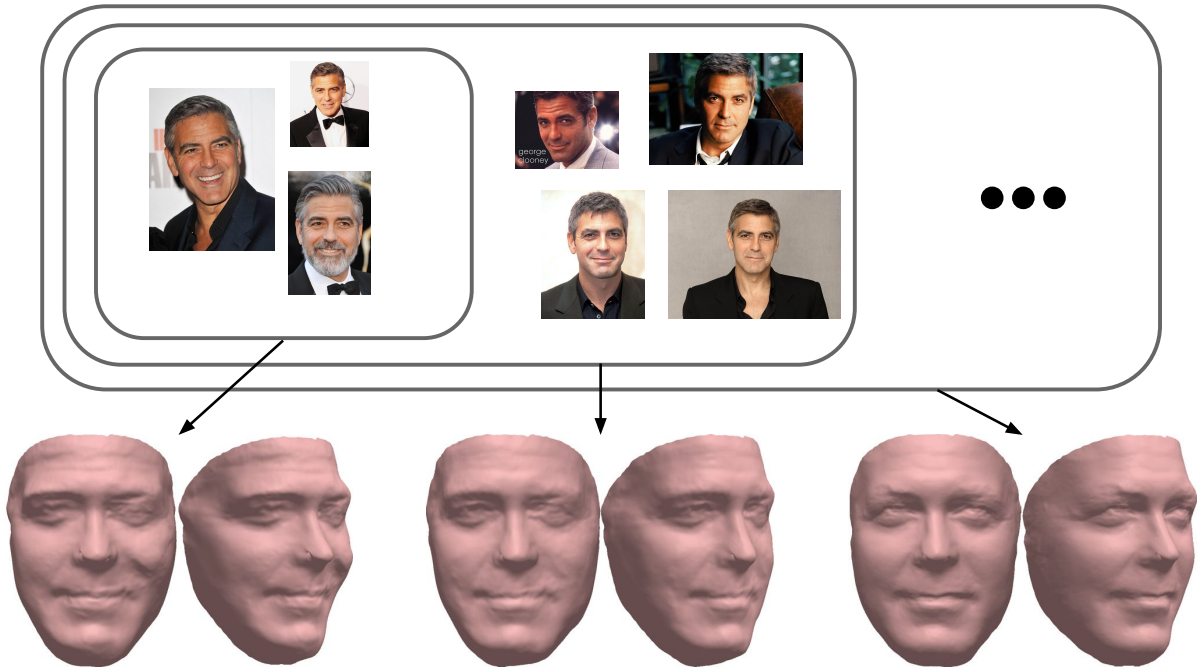


Figure 4.1: The proposed system reconstructs a detailed 3D face model of the individual, adapting to the number and quality of photos provided.

2D landmarks [93] 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. A global estimation of the albedo, lighting, and surface normals is performed using a dependability weighting based on the pose of each image. The surface normal estimate is improved locally using a novel structural similarity feedback to identify a collection of images which are similar for each part of the face. 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 summary, this chapter makes the following main contributions.

◇ A 3D Morphable Model is fit jointly to 2D landmarks 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 more confident face parts.

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

◇ The use of structural similarity as a quality measurement for performance evaluation in the absence of ground truth.

4.2 Quality Measures

One crucial task for reconstruction is measuring the quality of the reconstruction. The quality is useful both for comparing between different reconstruction methodologies as well as for providing feedback to a specific reconstruction technique. For example, if we know that one image contributes poorly to the reconstruction quality, it may be better to remove the image or weight it less during the reconstruction process. There are many different measures used in the literature for face reconstruction.

4.2.1 Image Distance Square Error

The image distance is the square error between the original image (real) and a rendering of the reconstruction (syn).

$$q_{\text{image}} = \sum_{u,v} \|I_{\text{real}}(u,v) - I_{\text{syn}}(u,v)\|^2. \tag{4.1}$$

The image distance is widely used in many reconstruction approaches due to its inclusion of all of the model parameters including object shape, orientation, camera model, lighting, and albedo. It is also straightforward to compute and easy to understand.

However, there are some major drawbacks to the image distance quality measure. Since the rendering is projected into the original image, it is impossible to penalize parts of the face not occluded by the synthetic image. For example, if the shape of the face is estimated too narrow, there will not be a synthetic rendering across the entire cheek compared with the real image and there will be a 0 distance for these regions even though there is clearly reconstruction error.

On the other extreme, q_{image} may also be high even for very high quality reconstructions. For example, if the shape and projection of the faces are perfect, but the albedo is wrong, the error will be high. Or if the projection is shifted, there will also be high error since the pixel correspondence between the real and synthetic image no longer have shared semantic meaning. These problems are due to using a single pixel correspondence to compute the error as well as many small errors contributing larger than a single large error.

Despite the problems with q_{image}, it is very relevant as a quality measure, especially during the fitting procedure. It provides a complete measure of all parameters in the fitting and is easy to compute the derivative for minimization.

4.2.2 Mahalanobis Distance

The Mahalanobis distance measures the relative distance of a face from a statistical distribution of faces. By taking a training set of faces and using principle component analysis (PCA), the multivariate Gaussian probability density function of face shapes is estimated. This distance is given as,

$$q_{\text{mahal}} = \sum_i \left(\frac{\alpha_i^{\text{id}}}{\sigma_i^{\text{id}}} \right)^2, \tag{4.2}$$

for the shape alone, where α^{id} is the coefficients of the shape projected into the PCA space and σ^{id} is the standard deviation from the PCA.

The Mahalanobis distance is often used as a regularizer during fitting since it relates to the faceness of the shape or how probable the shape is a face. When used as a regularizer often the distance for the other parameters such as albedo, projection, and lighting are added for fitting a 3D morphable model.

However, q_{mahal} is not good for comparing between reconstruction techniques, since it does not take into consideration the true ground truth shape of the individual and only measures the distance to the distribution of a training set.

4.2.3 Mean Euclidean Distance

The mean Euclidean distance is a good measure when a ground truth shape is present.

It measures the average distance of all points on one shape to another shape. This is usually discretized to only consider the vertices of the shapes. Given a template shape $\mathbf{Y}$ and reconstructed shape $\mathbf{X}$, when the shapes are given in vertex correspondence it is,

$$q_{\text{eucl}} = \frac{1}{p} \sum_{i=1}^{p} \|\mathbf{x}_i - \mathbf{y}_i\|_2. \tag{4.3}$$

However, correspondence is typically not known between the two shapes, in which case, the distance at each vertex is typically measured as the minimum distance to the other shape.

Note that q_{eucl} has a few major drawbacks. One, it is highly sensitive to rigid transformations of the shapes. Any slight transformations of one shape, (*e.g.*, a global scale) causes a significant effect on q_{eucl}. Therefore, it is important to perform rigid alignment of the faces first through iterative closest point alignment. Two, the distribution of point differences is likely to have a right tail and a single value for q_{eucl} may not be informative of detail recovery.

4.2.4 Hausdorff Distance

The Hausdorff distance is similar to the L-infinity distance and is defined as,

$$q_{\text{haus}} = \max\{ \sup_{x \in X} \inf_{y \in Y} d(x,y), \sup_{y \in Y} \inf_{x \in X} d(x,y) \}, \tag{4.4}$$

where sup is the supremum and inf is the infimum. In other words, it measures the maximum distance between any point on surface $\mathbf{X}$ and the closest point on $\mathbf{Y}$ and vice versa. When dealing with surfaces extremely close to each other, $q_{texthaus}$ is a good measure to differentiate between different reconstructions. For example, in computer graphics, it is used to measure the level of detail when downsampling objects far away in the scene.

However, it is a poor quality metric for current face reconstruction methods since they are far away from the ground truth, and only one point out of place will increase the measure.

4.2.5 Surface Normal Distance

One problem with the Euclidean distance measure is that points on the surface may be close to each other, but still be dissimilar. Particularly when there is no direct correspondence between the

surfaces and the closest point must be used. For example, if there is a small wrinkle in one surface, it will have very small Euclidean distance, to the other surface, but is clearly a larger error. If we look at the surface normal difference between the wrinkle and the smooth surface, we will see a much larger difference. In this regards, a surface normal difference can measure the finer details of a reconstruction, at the expense of ignoring the true surface to surface distance.

In [67] the normal metric is defined as,

$$q_{\text{normal}} = \frac{1}{p}\sum_{i=1}^{p} \arccos(\mathbf{n}_i \cdot \mathbf{n}'_i), \tag{4.5}$$

where $\mathbf{n}_i$ is the normal of the reconstructed face and $\mathbf{n}'_i$ is the normal of the average face. Their definition is used as a regularizer for the face shape. It is observed that when the Mahalanobis distance is used as a regularizer, it requires either (1) it allows non-plausible faces with a low weight or (2) the weight is too large and produces smooth faces when a more optimal result could be obtained for some images. Therefore, Mahalanobis has a tradeoff between quality and robustness. In [67], they demonstrate the effectiveness of using the surface normal measure to regularize the reconstruction.

One issue with q_{normal} is the variance of different parts of the face. For example, the nose and lips may have large differences in the surface normal while still maintaining a reasonable face compared to the cheeks which should remain smooth. To compensate, they introduce a weight for each vertex defined as,

$$\hat{\omega}_i = 1 - \frac{\bar{\phi}_i - \bar{\phi}_{\min}}{\bar{\phi}_{\max} - \bar{\phi}_{\min}}, \tag{4.6}$$

where $\bar{\phi}_i$ is the average normal deviation in the training set for their morphable model.

The normal distance could also be used to evaluate reconstruction without vertex correspondence by selecting the closest vertex on the ground truth surface to compare the normal difference.

4.2.5.0.1 Summary There are clear advantages and disadvantages to the different quality measures as they can distinguish between different type of failures in the reconstruction. q_{eucl} and q_{haus} can only be used with a ground truth face scan. q_{normal} and q_{mahal} are only appropriate for regularization because they are person independent. That leaves only q_{image} as a relevant measure for photo collections without a ground truth scan. However, q_{image} is very sensitive to numerous factors and may not be informative of shape.

Based on the limitations of current quality measures for evaluation of photo collections, we propose a new quality measure based on structural similarity (SSIM) that encompasses all parameters of the reconstruction like q_{image}, but considers a larger area than a single pixel for comparison so it is more robust to small changes in alignment.

4.3 Algorithm

We now present the details of the proposed approach, describing the motivational differences from prior works. Figure 4.2 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 chapter are provide in Table 2.2. 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.

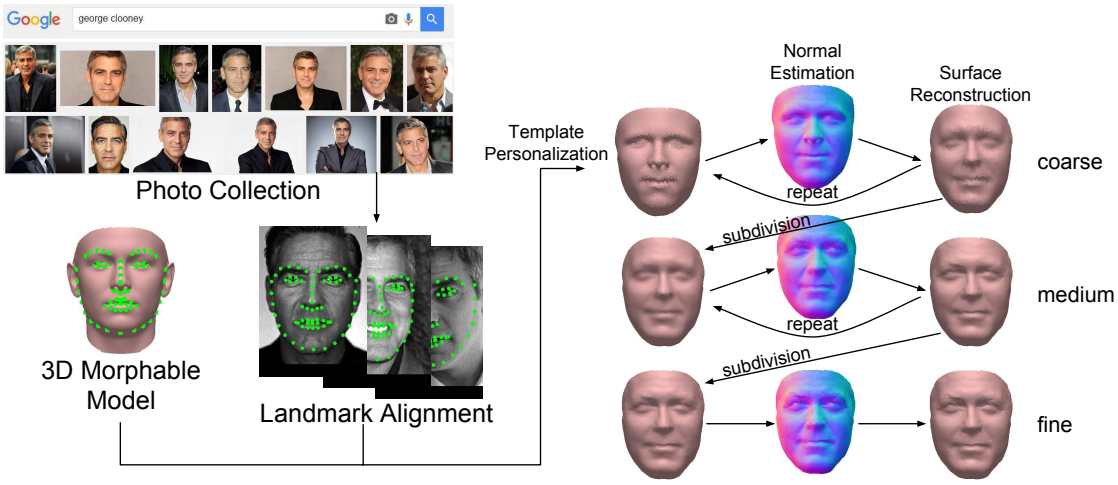


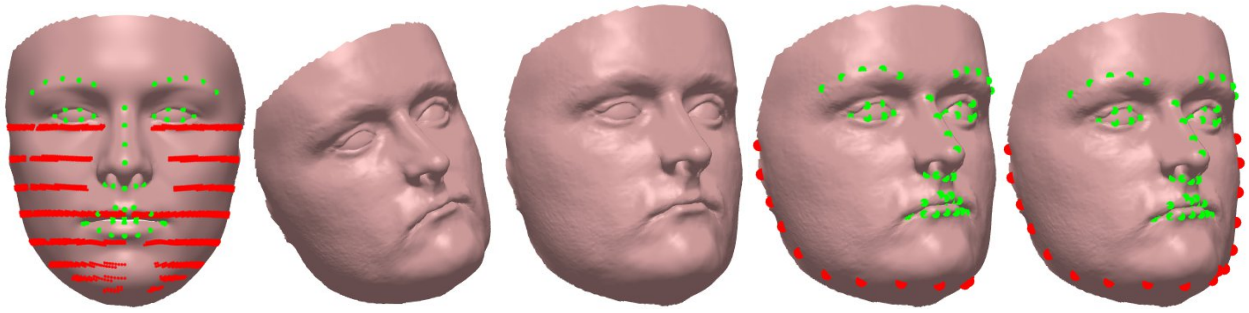
Figure 4.2: Overview of face reconstruction. Given a photo collection, we apply landmark alignment and use a 3DMM to create a personalized template. Then a coarse-to-fine process alternates between normal estimation and surface reconstruction.

4.3.1 Inputs and Preprocessing

4.3.1.1 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, *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 [4] 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 450×450 to ensure 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, humans can distinguish differences in low intensity better than high intensity, so most



(a) (b) (c) (d) (e)

Figure 4.3: 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.

cameras use a non-linear gamma encoding of images in order to provide a subjectively equal step in brightness for humans. For this work, we apply a single industry standard gamma correction to convert each image into the linear intensity scale.

4.3.1.2 Landmarks

Landmarks are the locations of common key features such as the eyes, nose, or mouth on a face. In recent years, the automatic detection of landmarks [26, 27, 48] has seen rapid improvement due to large labeled datasets such as LFPW [11] and 300-W [74]. To estimate 2D landmarks, we employ the state-of-the-art cascade of regressors approach [93] to automatically fit $q = 68$ landmarks denoted as $\mathbf{W} \in \mathbb{R}^{2\times q}$ onto each image. Figure 4.3 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.

4.3.1.2.1 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 [21, 69, 101]. In this work, we follow the proposed *landmark marching* method from [101]. 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. A visualization of the process is presented in Fig 4.3.

4.3.2 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 [53] or a Structure from Motion-based (SfM) deformation of a single face mesh [72]. These models have two main limitations. One, the model is a single ethnicity / 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 more prior information to help form a personalized template for a wide range of subjects with few images.

4.3.2.1 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 a 3DMM in the form,

$$\mathbf{X} = \bar{\mathbf{X}} + \sum_{k=1}^{199} \mathbf{X}_k^{\mathrm{id}} \alpha_k^{\mathrm{id}} + \sum_{k=1}^{29} \mathbf{X}_k^{\mathrm{exp}} \alpha_k^{\mathrm{exp}}, \tag{4.7}$$

where $\mathbf{X} \in \mathbb{R}^{3\times p}$ is the 3D face composed of the mean shape $\bar{\mathbf{X}}$, a set of identity bases $\mathbf{X}^{\mathrm{id}}$, and a set of expression bases $\mathbf{X}^{\mathrm{exp}}$, with coefficients $\vec{\alpha}^{\mathrm{id}}$ and $\vec{\alpha}^{\mathrm{exp}}$. We use the 3DMM from [101] where the identity comes from the Basel Face Model [65] and the expression comes from Face Warehouse [22]. The separation of the bases into expression and identity is based on the method from [24].

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 [14] 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 [101]. 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 $\mathbf{W}$, the model parameters are estimated by minimizing the projection error of the 3DMM to the landmarks,

$$\arg\min_{s,\mathbf{R},\mathbf{t},\vec{\alpha}^{\text{id}},\vec{\alpha}^{\text{exp}}} \|\mathbf{W} - (s\mathbf{R}[\mathbf{X}]_{\text{land}} + \mathbf{t})\|_F^2 + E_{\text{reg}}, \tag{4.8}$$

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. 4.9) for the 3DMM coefficients. However, as discussed in Sec. 4.3.1.2, the pose must be known to march the external 3D landmarks along their paths to establish correspondence with $\mathbf{W}$. But in the current formulation, the pose is solved jointly with the 3DMM coefficients.

We follow [101], and solve Eq. 4.8 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 $\mathbf{t}$) [18], 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 4 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 α^{id} but a unique set of expression α_i^{exp} and pose parameters per image. The modified error function fully expressed is,

$$\begin{aligned}\arg\min_{s_i,\mathbf{R}_i,\mathbf{t}_i,\vec{\alpha}^{\text{id}},\vec{\alpha}_i^{\text{exp}}} \sum_{i=1}^{n} \frac{1}{n} \|\mathbf{W}_i - (s_i\mathbf{R}_i[\bar{\mathbf{X}} + \sum_{k=1}^{199} \mathbf{X}_k^{\text{id}}\alpha_k^{\text{id}} + \sum_{k=1}^{29} \mathbf{X}_k^{\text{exp}}\alpha_{ki}^{\text{exp}}]_{\text{land}_i} + \mathbf{t}_i)\|_F^2 \\ + \sum_{k=1}^{199} \left(\frac{\alpha_k^{\text{id}}}{\sigma_k^{\text{id}}}\right)^2 + \sum_{k=1}^{29} \left(\frac{\frac{1}{n}\sum_{i=1}^{n} \alpha_{ki}^{\text{exp}}}{\sigma_k^{\text{exp}}}\right)^2,\end{aligned} \tag{4.9}$$

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.

4.3.2.1.1 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 image location of 3D vertex j.

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 further creating 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 used to add in the fine details of the face.

4.3.3 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.

4.3.3.1 Lighting Model

Computer graphics takes a modeled scene and renders a realistic synthetic image. Whereas, computer vision solves 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 models 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 intensity at a projected point is defined by a linear combination of lighting parameters and the surface normal,

$$\mathbf{I}(u,v) = \rho_j \left(k_\mathrm{a} + k_\mathrm{d} \left(l^x n_j^x + l^y n_j^y + l^z n_j^z \right) \right), \tag{4.10}$$

where ρ_j is the surface albedo at vertex j, n_j^x, n_j^y, n_j^z is the unit surface normal at vertex j, k_a is the ambient coefficient, k_d is the diffuse coefficient, and l^x, l^y, l^z is the unit light source direction of the image. For simplicity, we combine the lighting coefficients and direction into a vector $\mathbf{l} = \langle k_\mathrm{a}, k_\mathrm{d} l^x, k_\mathrm{d} l^y, k_\mathrm{d} l^z \rangle^\intercal$, 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 of the surface.

Ref [29] 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, [7] demonstrates that shape reconstruction accuracy using 1st order is 95-98% while 2nd order is 97-99%. So, while a

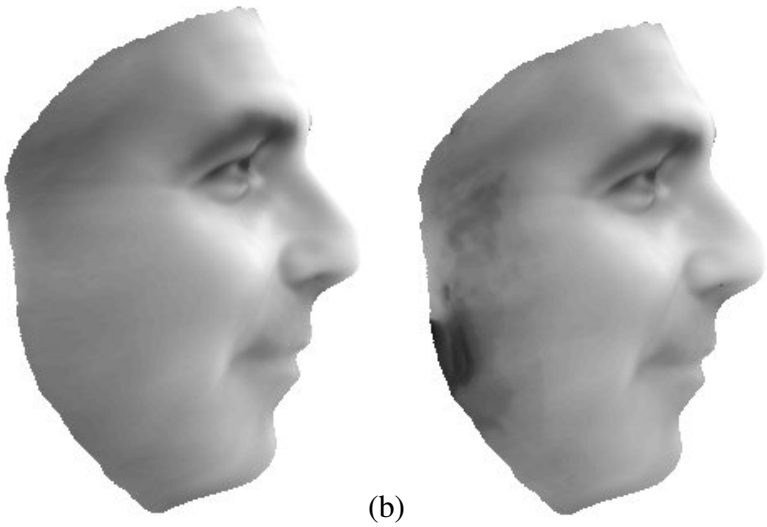
more complex lighting assumption may potentially increase the accuracy by a single percentage, it introduces non-linearity into the solution process. Therefore, we use the 1st order assumption in this work, but in the future, if we allow other nonlinearities in the model a 2nd order assumption could be made.

Prior work jointly solved for the Lambertian formulation using singular value decomposition (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 [53, 72]. 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 error. However, we will show that small collections are susceptible to any correspondence errors from misalignment or expressions. 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,

$$\underset{\rho_j,\mathbf{L},\mathbf{N}}{\operatorname{argmin}} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} \|f_{ij} - \rho_j \mathbf{l}_i^\intercal \mathbf{n}_j\|^2 + \lambda_n \|\mathbf{n}_j - \mathbf{n}_j^t\|^2 \right), \tag{4.11}$$

where $\mathbf{n}_j^t$ is the current surface normal of the face mesh at vertex j. The template regularization helps keep the face close to the initialization. But, since the summation is not averaged, as more photos are added to the collection, the regularization has less overall weight since λ_n is independent of collection size and 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



(a) (b)

Figure 4.4: 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.

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, it is important to find a good method for determining what images to use for each part of the face.

4.3.3.2 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 a unit camera vector perpendicular to the

image plane as the measure of dependability to handle self-occlusion and sampling artifacts. Other problems such as expression and and external occlusion we leave for local selection (Sec. 4.3.3.4).

What does this dependability measure accomplish? First, self-occluded parts of the face are given a weight of 0. Second, parts of the image more susceptible to pose estimation errors are given lower weights. As a vertex's normal approaches perpendicular to the camera, slight perturbations of the pose will cause a larger change in what u, v are sampled in the image. Whereas, a vertex pointing towards the camera is more stable and should be more dependable. Fig. 4.4 shows the albedo estimation with and without dependability. We update Eqn. 4.11 to,

$$\underset{\rho_j,\mathbf{L},\mathbf{N}}{\operatorname{argmin}} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} \|d_{ij}(f_{ij} - \rho_j \mathbf{l}_i^{\mathsf{T}} \mathbf{n}_j)\|^2 + \lambda_n \|\mathbf{n}_j - \mathbf{n}_j^t\|^2 \right). \tag{4.12}$$

What is not modeled by this dependability choice? First, any external occlusion, such as sunglasses. Second, landmark alignment errors. Third, expression differences. We will address these issues with the localization step introduced in Sec. 4.3.3.4.

What is not modeled by this dependability choice? First, any external occlusion, such as sunglasses. Second, landmark alignment errors. Third, expression differences. We will address these issues with the localization step introduced in Sec. 4.3.3.4.

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. 4.12. Note that Eq. 4.12 is not jointly convex, but it 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 is an

over-constrained least squares with the solution,

$$\mathbf{l}_i^{\intercal} = (\mathbf{f}_i \circ \mathbf{d}_i)/(\tilde{\rho} \circ \mathbf{N} \circ \mathbf{d}_i), \tag{4.13}$$

where $\circ$ is the Hadamard or entrywise product and $\tilde{\rho}$ is $\vec{\rho}$ repeated 4 times to become the same size as $\mathbf{N}$. Similarly, albedo has a closed form solution,

$$\rho_j = (\mathbf{d}_j^{\intercal}\mathbf{L}^{\intercal}\mathbf{n}_j)/(\mathbf{d}_j^{\intercal}\mathbf{f}_j). \tag{4.14}$$

Finally, the normals are solved via,

$$\mathbf{n}_j = (\mathbf{B}^{\intercal}\mathbf{B} + \lambda_n \mathbf{I})^{-1}(\mathbf{B}^{\intercal}(\mathbf{f}_j \circ \mathbf{d}_j) + \lambda_n \mathbf{n}_j^t), \tag{4.15}$$

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

4.3.3.4 Local Selection

As mentioned in Sec. 4.3.3.2, the dependability measure only handles small landmark alignment error, but does not consider expression changes, occlusions, or other potential correspondence errors. To handle these other forms of error, we use a local selection process as proposed in [53] 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 $\mathscr{B}_j$ for each vertex j and then re-minimize the

photometric equation for that vertex's normal:

$$\underset{\mathbf{n}_j}{\operatorname{argmin}} \sum_{i \in \mathscr{B}_j} \|d_{ij}(\rho_j \mathbf{l}_i^{\intercal} \mathbf{n}_j - f_{ij})\|^2 + \lambda_n \|\mathbf{n}_j - \mathbf{n}_j^t\|^2. \tag{4.16}$$

All of the prior work uses the same scheme of local selection [53, 72] 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, $\mathscr{B}_j = \{i \mid \|\rho_j \mathbf{l}_i^{\intercal} \mathbf{n}_j - f_{ij}\|^2 < \varepsilon\}$. 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 because 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, it can be distracted by noise.

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 [89]. 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:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{4.17}$$

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 the Matlab implementation, the window is an isotropic Gaussian weighting with standard deviation γ of the surrounding pixels instead of a

Figure 4.5: Raw image, synthetic image under estimated lighting conditions, and SSIM used for local selection. Brighter indicates higher SSIM.

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. 4.3.2.1.1. Figure 4.5 demonstrates this process for a single image. The local selection now becomes $\mathscr{B}_j = \{i \mid s_{ij} > \varepsilon\}$.

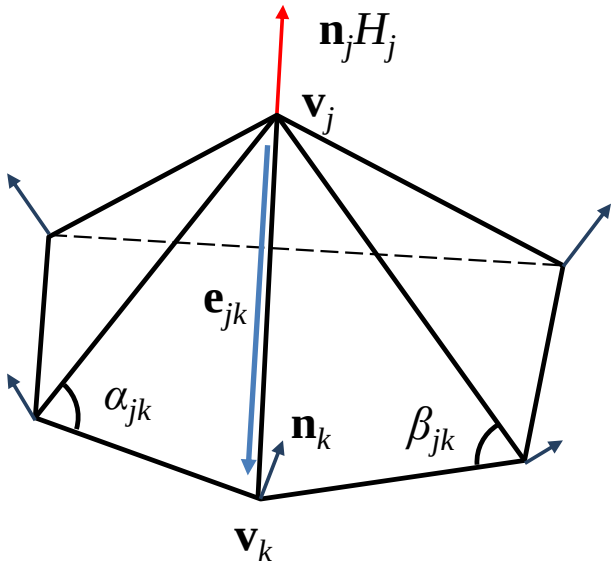


Figure 4.6: The mean curvature normal indicates how a vertex deviates from the average location of its immediate neighbors, which can be evaluated as the Laplacian of the position. The mean curvature H_j can be evaluated through $\mathbf{n}$.

4.3.4 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. The process is similar to Chp. 3, but we summarize it again here.

We use a Laplacian-based surface editing technique motivated by [78]. The Laplace-Beltrami operator is the divergence of a gradient field. Using linear finite elements, it can be discretized into $\mathscr{L}$, a symmetric matrix with entries $\mathscr{L}_{jk} = \frac{1}{2}(\cot\alpha_{jk} + \cot\beta_{jk})$, where α_{jk} and β_{jk} are the opposite angles of edge jk in the two incident triangles (see Figure 4.6), known as the cotan formula [66]. Geometrically, $\mathscr{L}$ measures the difference between a functions value at a vertex and the average value of the neighboring vertices. As in [78, 97], we note that $\mathbf{x}_j\mathscr{L} = -\mathbf{n}_j H_j$, where H_j is the integral of the mean curvature at 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 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,

$$H_j = \frac{1}{4A_j} \sum_{k \in N(j)} (\cot \alpha_{jk} + \cot \beta_{jk})\, \mathbf{e}_{jk} \cdot (\mathbf{n}_k - \mathbf{n}_j), \tag{4.18}$$

where $N(j)$ is the one-ring neighborhood of j, A_j is the sum of triangle areas incident to j, $\mathbf{e}_{jk}$ is the edge from j to k (Figure 4.6). 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,

$$\underset{\mathbf{X}}{\operatorname{argmin}}\, E_n + \lambda_b E_b + \lambda_l E_l. \tag{4.19}$$

Then $E_n = \|\mathbf{X}\mathscr{L} + \mathbf{N}\mathbf{H}^k\|^2$ 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 current face model. $E_b = \|\mathbf{X}\mathscr{L}_b - \mathbf{X}^k\mathscr{L}_b\|^2$ 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 $\mathscr{L}_{b,jk} = 1/|\mathbf{e}_{jk}|$ where $|\mathbf{e}_{jk}|$ is the edge length connecting adjacent boundary vertices j and k. And $E_l = \sum_i \|s_i\mathbf{R}_i[\mathbf{X}]_{\text{land}_i} + \mathbf{t}_i - \mathbf{W}_i\|_F^2$, which uses the landmark projection error to provide a global constraint on the face, without which, the integration of the normals can have numeric drift across the surface of the face. Unlike Chp. 3 we do not include a shadow region smoothing since we use the template normal as a regularizer during normal estimation.

Algorithm 2: 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. 4.9 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 $\mathbf{L}$, $\vec{\rho}$, and $\mathbf{N}$ via Eq. 4.12
    local selection of images $\mathscr{B}$ via Sec. 4.3.3.4
    re-estimate surface normals $\mathbf{N}$ via Eq. 4.16
    // Step 3
    reconstruct surface $\mathbf{X}^{k+1}$ via Eq. 4.19
  until $\frac{1}{p}\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F^2 < \tau$
  subdivide surface

The solution ends up being linear,

$$(\mathscr{L}^2 + \lambda_b \mathscr{L}_b^2 + \lambda_l \sum_i s_i^2 \mathbf{C}_i \mathbf{C}_i^\intercal)\mathbf{X} = \mathbf{N}\mathbf{H}^k \mathscr{L} + \lambda_b \mathbf{X}^k \mathscr{L}_b + \frac{\lambda_l}{n} \sum_i s \mathbf{R}_i^\intercal (\mathbf{t} - \mathbf{W}_i)\mathbf{C}_i^\intercal, \tag{4.20}$$

where $\mathbf{C}_i \in \mathscr{R}^{p \times q}$ 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 correspond to the 2D landmarks for image i.

4.3.5 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 high 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 2. After personalizing the face model, we use ReMESH [5] to uniformly resample the personalized mesh $\mathbf{X}^0$ to a coarse $6,248$ $(= p)$ vertices. The resampling is done once offline on the mean shape and is transferred to a personalized mesh by using the barycentric coordinates of the corresponding triangle. Within each resolution, steps 2 and 3 are repeated until the surface converges. After convergence, one step of Loop subdivision [45] 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. 4.4.1.0.3) 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.

4.3.6 SSIM Quality Measure

Accurately measuring reconstruction quality in the absence of ground truth data is a difficult task. Even with a ground truth scan of the face, the popular surface-to-surface distance measurement has its flaws since there is no direct correspondence between surfaces. Typically, surfaces are aligned (*e.g.*, 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 the 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. One, surface reconstruction errors should be evident in the score. Two, 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 so 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. 4.3.3.4, SSIM was developed to mimic human perception. We will verify this relationship in Sec. 4.4.2.2.

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. 4.5). 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.

4.4 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 designed approaches which use photometric stereo-based approaches [72, 53]. 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.

4.4.1 Experimental Setup

4.4.1.0.1 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 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 [96] using the provided texture and selecting random frames from the 6 expression sequences (Fig. 4.7). 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 [4] in Python to detect, crop, and scale faces as described in Sec. 4.3.1.1. 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 $\pm$0.03mm. Given frontal and both 45° yaw scans, we stitch them together using MeshLab to create a full 3D model.

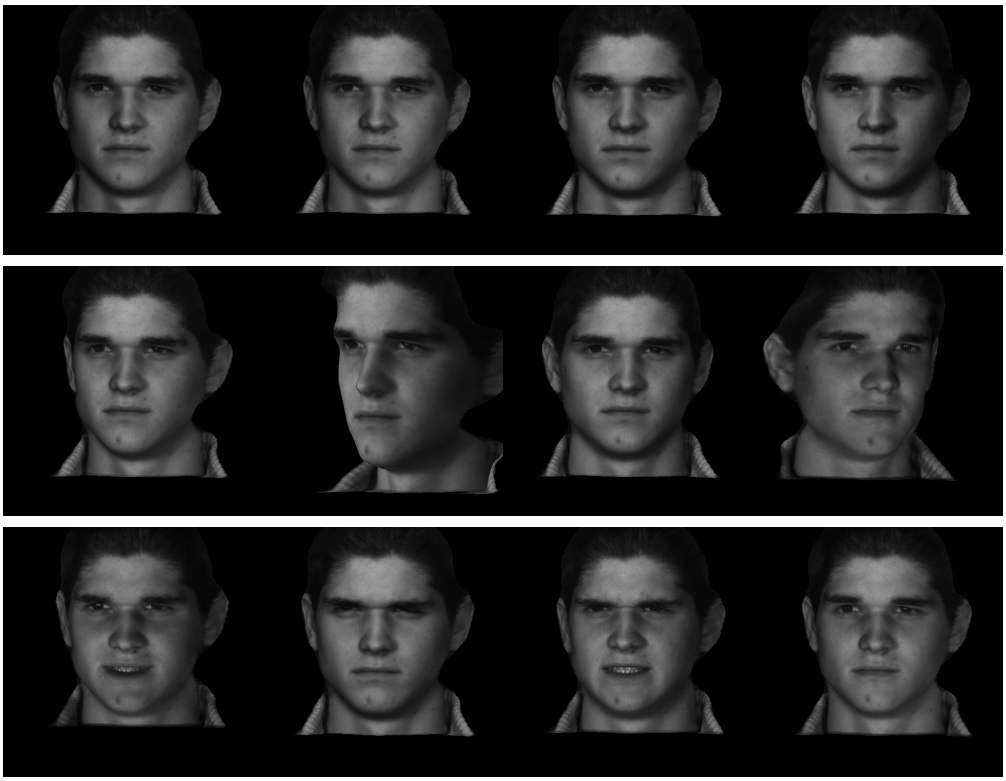


Figure 4.7: Synthetic data with lighting (top), pose (middle), and expression (bottom) variation.

4.4.1.0.2 Metrics For quantitative metrics, we use two different measurements. 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, described in detail in Sec. **??**. For personal collections where we have a ground truth surface, we compute the average surface to surface distance. Both surfaces are aligned by Procrustes superimposition of the 3D landmarks from the internal part of the face. The normalized vertex error is computed as the distance between a vertex in the 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.

4.4.1.0.3 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 $\varepsilon = [0.2, 0.08, 0.08]$, and SSIM error $\varepsilon = [0.65, 0.65, 0.65]$ for coarse, medium, and fine resolution respectively.

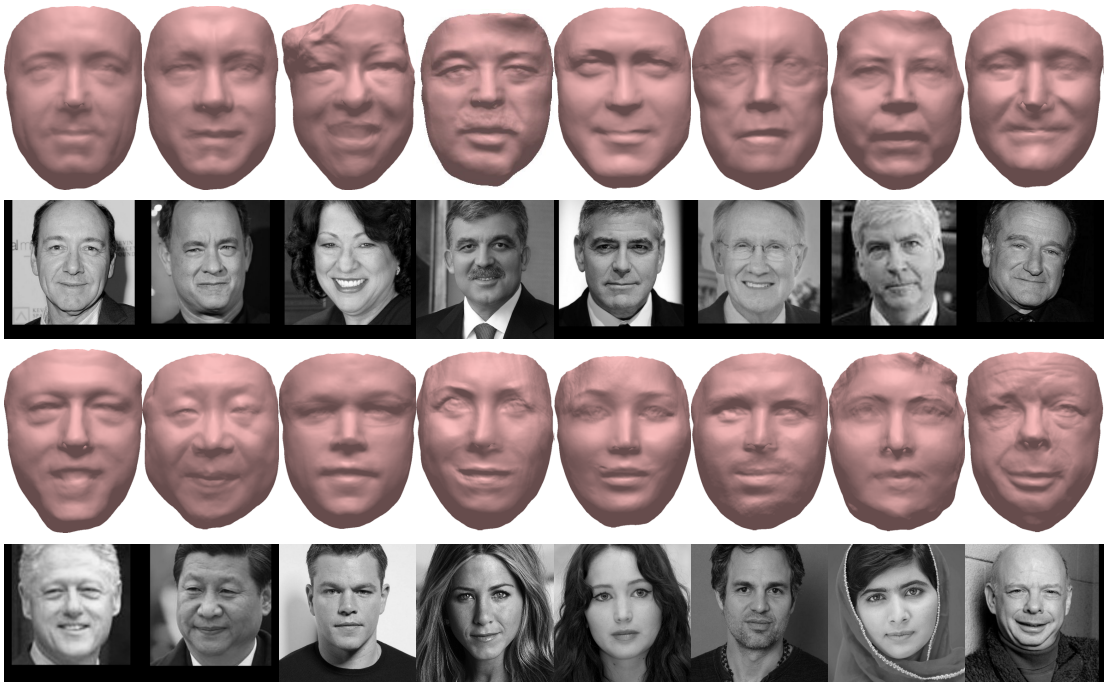


Figure 4.8: Qualitative evaluation of a diverse set of individuals from Internet photo collections.

4.4.2 Internet Results

4.4.2.1 Qualitative Evaluation

We begin by presenting qualitative results of the proposed reconstruction method on a diverse set of subjects, spanning ethnicity and gender. While qualitative results are subjective and hard to compare 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 4.8, 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 not included in the 3DMM nor were directly considered in our approach. However, we do see that facial hair often creates difficulty for the reconstruction since it is hard to establish

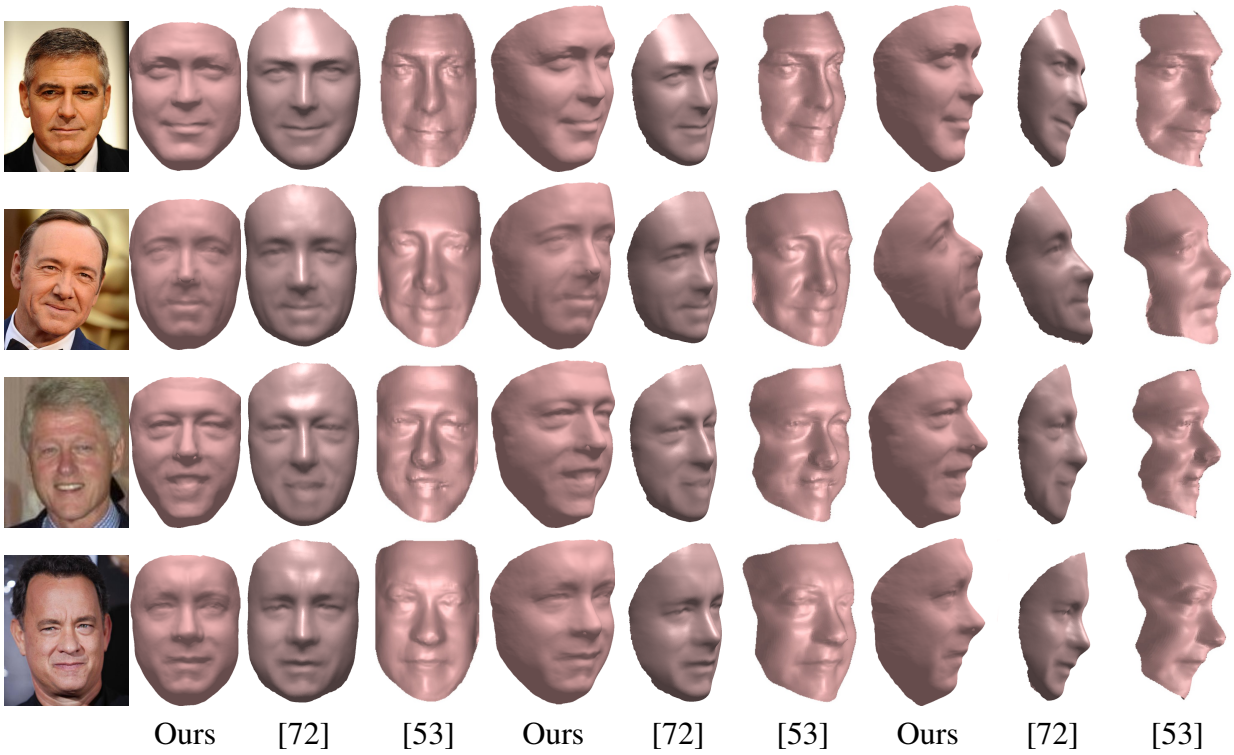


Figure 4.9: Qualitative comparison on celebrities. The proposed approach incorporates more of the sides of the face and neck.

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 four celebrities used in [53] and [72], George Clooney (99 photos), Kevin Spacey (143), Bill Clinton (179), and Tom Hanks (255). Figure 4.9 presents a side by side comparison between the various approaches. Due to the subjective nature of these reconstructions, we simply say they are similar, but we include the largest surface area. 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.

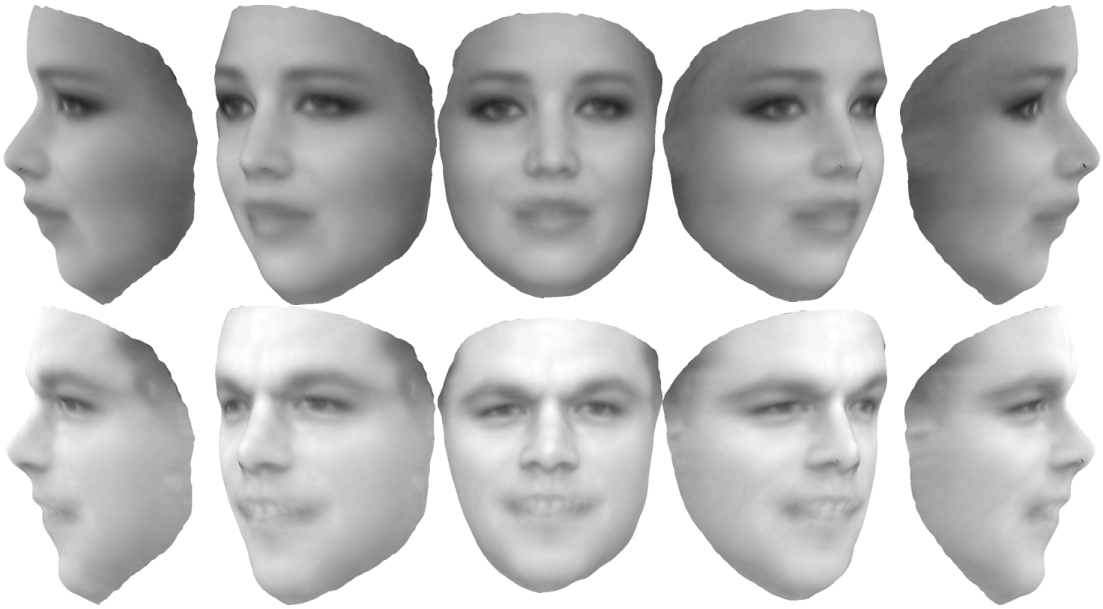


Figure 4.10: Sample rendering used for human perception experiment.

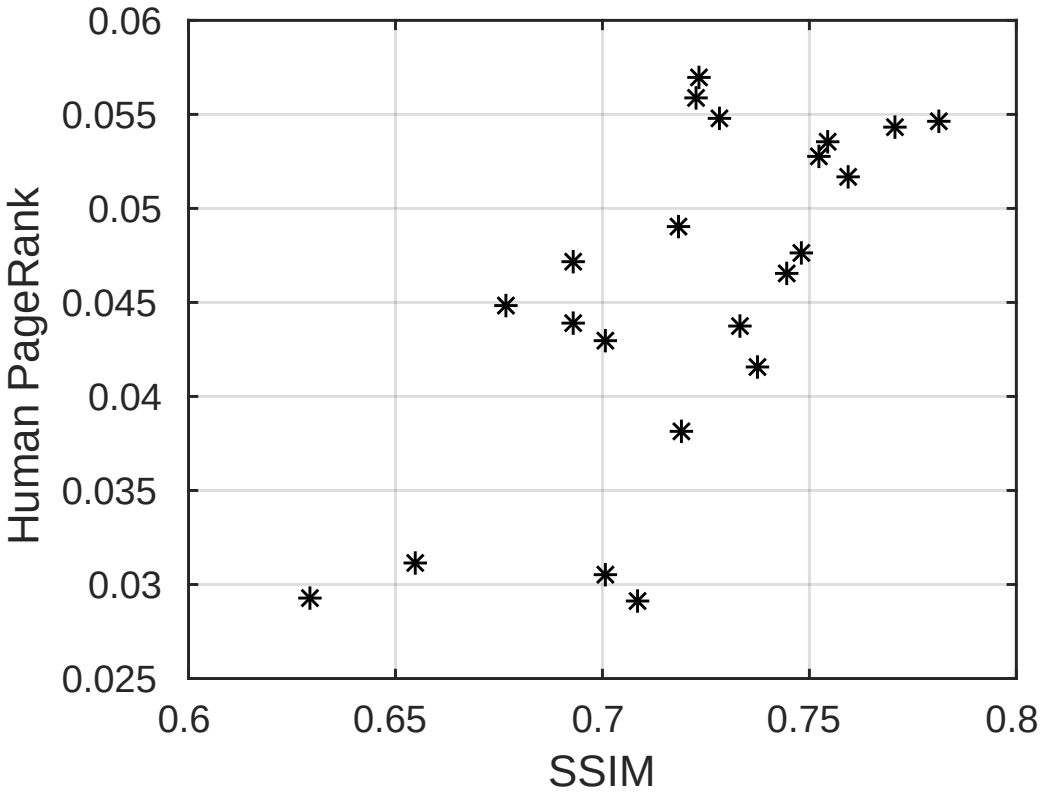


Figure 4.11: Human-based PageRank scores for SSIM.

4.4.2.2 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 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 is rendered under 5 viewpoints. The set of SSIM scores for each collection is obtained as described in Sec. 4.3.6. Human perception of reconstruction quality is subjective and it is a hard task to request humans to provide a single number ranking the quality of each reconstruction separately. Therefore, we design 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 Figure 4.10. 6 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.

(a) (b)

Figure 4.12: Best (top) and worst (bottom) reconstructions as determined by human (a) and SSIM (b).

PageRank [63] can provide a global human ranking. 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 been shown to accurately rank sports teams [31]. We compare human and SSIM scores in Figure 4.11, with a correlation of 0.69 indicating SSIM is equivalent to a single human evaluation. Figure 4.12 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 [72], Fig. 4.13 plots a histogram of SSIM quality score. 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. 4.11, and the other at 0.3 can be viewed as complete failures. We show an example collection in Fig. 4.14. 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.

4.4.2.3 Adaptability

We look at adaptability with respect to two different factors. One, the number of images in the photo 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 $\sim$20 pixel eye-to-eye distance.

Figure 4.15(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

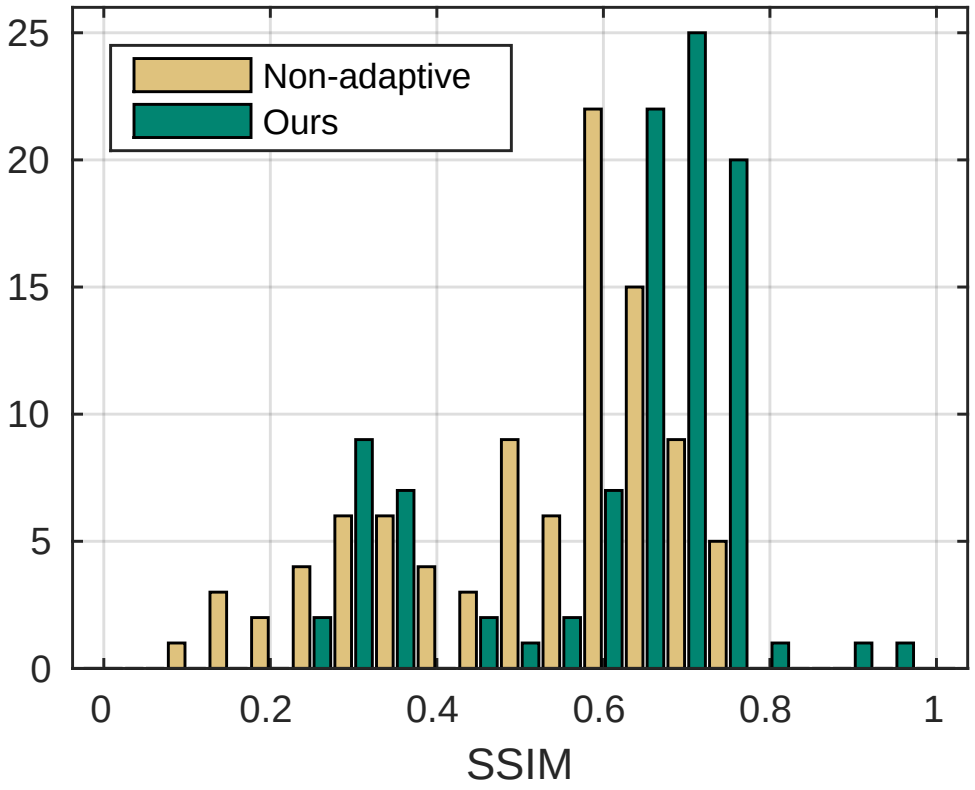


Figure 4.13: Histogram of reconstruction performance.

Figure 4.14: An Internet image collection that results in a complete failure reconstruction.

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.

Table 4.1: Synthetic Surface-to-Surface Error.

Method	Neutral	30° Yaw	Expression
Ours	**3.22%**	**3.82%**	**4.40%**
[72]	6.13%	7.48%	6.59%

4.4.3 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 ±30°, and frontal with random expressions (Fig. 4.7). Error is reported as the surface-to-surface distance to the neutral expression model. Table 4.1 shows the proposed approach outperforms prior work in all scenarios. We see the proposed algorithm is more robust to pose than expression variation.

4.4.4 Personal Results

4.4.4.1 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 4.2 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. 4.3 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.

Table 4.2: Local Selection Error.

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

Table 4.3: SSIM Radius Error.

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

4.4.4.2 Adaptability

We perform a *thorough experiment* comparing the adaptability of the proposed method to the SVD-based approach of [72]. 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 4.4. We make a few notes. First, the proposed method performs better for all collection sizes. Two, the proposed method better adapts to small collections. While the SVD-based approach degrades by 1.44%, the proposed method only degrades by 0.46%

Table 4.4: 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%
[72] - Low	5.54%	4.78%	4.63%	4.34%
[72] - High	5.56%	4.77%	4.70%	4.10%

4.4.5 Discussions

4.4.5.0.1 Efficiency Written in a mixture of C++ and Matlab, the algorithm runs on a commodity PC with an Intel *i*7-4770*k* 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.

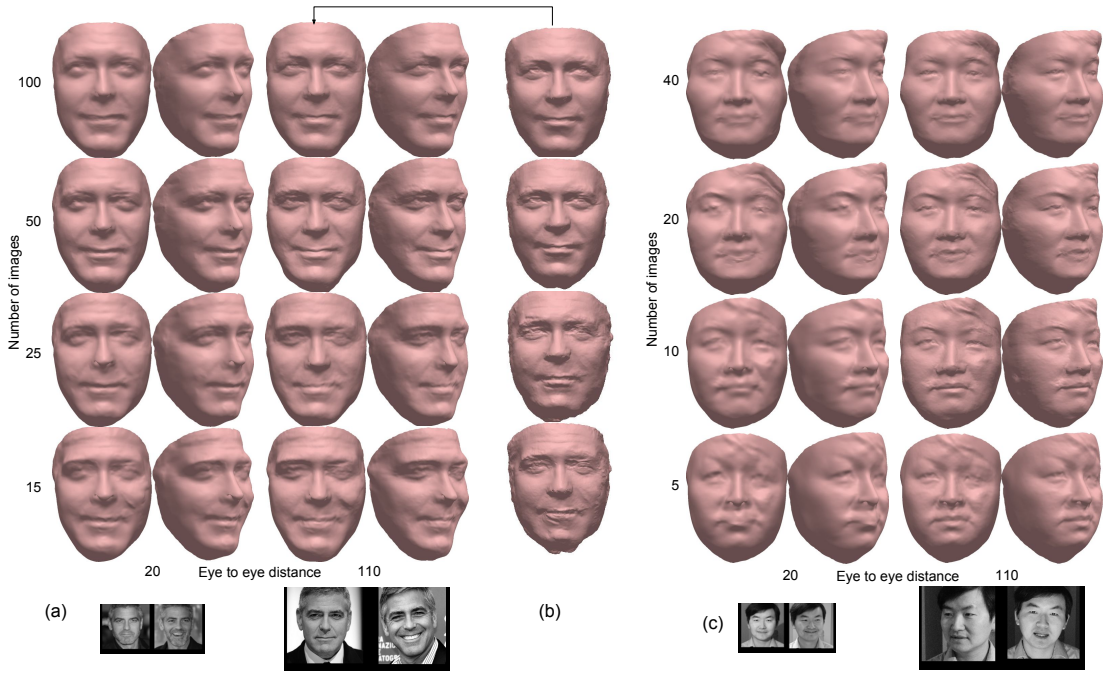


Figure 4.15: (a) George Clooney with different quality images. (b) Reconstruction without coarse-to-fine process. (c) Personal collection with different quality images.

4.4.5.0.2 Coarse to Fine The coarse-to-fine scheme benefits both efficiency and quality. If the coarse-to-fine scheme is not used and instead the reconstruction starts at the fine resolution, it takes 4 iterations to converge for a total time of 3.7 minutes more than double the time. Also, Fig. 4.15(b) shows the resultant reconstructions which are similar for large amounts of images, but noisy for small collections since the coarse step allows for more template regularization.

4.5 Summary

We presented a method for reconstructing a 3D face model from an unconstrained 2D photo collection which adapts to lower quality and fewer images. By using a 3DMM to create a personalized template which adaptively influences reconstruction in a coarse to fine scheme, we can efficiently create a more accurate model than prior work as demonstrated by experiments on synthetic and real-world data. There are numerous paths for future work, *e.g.*, fusing 3DMM and photometric

stereo-based reconstructions so it can gracefully degrade down to a single image, and automatically identifying the detail level of reconstruction possible from the photo collection.

Chapter 5

Conclusions and Future Work

Reconstructing faces based on unconstrained photo collections is extremely challenging due to the highly non-rigid nature of human faces and the lack of known information about the pose and lighting for the images. Throughout this work, I have presented an approach to reconstruct a true 3D surface including wrinkle level details that can adapt to small photo collections.

5.1 Limitations

While it is desirable for the proposed reconstruction process to work under any situation, there are still some limitations of the current approach.

5.1.0.0.1 Landmark reliance To establish correspondence across the images, landmarks alone are used when estimating the pose of the face model to each image. It is known that 2D landmark alignment approaches have error. When failed landmark alignment occurs, under the best case scenario, it never selects any pixels from that image for local selection and the surface normal estimation is unaffected. However, a more likely scenario, is that the misaligned parts of the image contribute an incorrect surface normal and introduce error in the reconstruction process.

5.1.0.0.2 Expression variation The only handling of expression exists within the initial 3DMM template fitting. After fitting, the average expression from the collection is used throughout the remainder of the process. Differences in expression cause non-rigid deformations of the face that

affects alignment and the surface normals of the face. We make no specific assumptions about expressions and leave the local selection process to filter out differences in expression.

5.1.0.0.3 Specular reflection The Lambertian lighting model is known to be a poor assumption for shiny surfaces. When humans sweat, the skin clearly exhibits specular reflections, particularly on the nose and forehead. A more complex lighting assumption is necessary to accurately handle these scenarios.

5.1.0.0.4 Continuous surface The mesh we use is closed except for the back of the head. In practice, the majority of the face is a continuous surface, but there are areas that the model breaks down. The eyelids and mouth both have discontinuities that are unable to be modeled with the current mesh. Particularly when the dominant expression is smiling, where the lips should be separate from the teeth, the closed mesh forms a smooth continuity between the different physical surfaces.

5.1.0.0.5 Hair Both facial hair and longer hair on the top of the head present difficulties. Correspondence between images is crucial for the photometric stereo approach. However, any perturbation of hair follicles between images causes a different surface normal to be observed. As such, facial hair usually presents as noise in the reconstruction, and long straight hair will show up as vertical lines on the surface of the forehead.

5.2 Future Work

There are a number of potential questions that arise from this work. I will address some of the most common or pertinent.

5.2.1 Texture Basis

The 3DMM contains a texture basis, how can this be incorporated into the reconstruction process? It is true that the 3DMM contains texture basis along with the shape basis. One question is how much information is contained in the appearance, and how likely is an arbitrary face to be well represented in the basis. It is known from active appearance models, that the shape distribution for faces is significantly more compact than the appearance. In [60], they use more than 3 times as many appearance vectors as shape to capture the same amount of energy in their training set. It is unclear how well using the same 200 faces for texture alone can generalize for unseen textures. Furthermore, a low-rank modeling is unlikely to be sufficient for the non-linear manifold of facial textures. For instance, different ethnicities or genders will likely form a multi-modal distribution.

I tried restricting albedo to a linear combination of the provided texture. That is, changing the albedo to,

$$\rho = \bar{\rho} + \sum_{k=1}^{199} \beta_k \rho_k, \tag{5.1}$$

and solving for $\vec{\beta}$. For larger collections, this was too restrictive and actually degraded the reconstruction quality. The hypothesis is that the small training set of textures is not sufficient to represent an arbitrary face.

Another possibility is to use the albedo in a regularization instead of a restriction. Similar to [52] where they reconstruct a single image by keeping the albedo close to the average albedo, *i.e.*,

$$E_{alb} = \|\Delta G * \rho - \Delta G * \bar{\rho}\|^2, \tag{5.2}$$

where $\Delta G*$ denotes convolution with the Laplacian of a Gaussian to smooth out the albedo constraint. This idea is reasonable, and could be implemented in our system in order to improve the

reconstruction for single image collections.

Currently, for a single image collection, the albedo estimation absorbs all difference due to normal differences from the 3DMM fit template. This is due to the regularization on the surface normal, but no constraint on the albedo. The result, is that the reconstruction is unchanged after 3DMM fitting for a single image. Including an albedo regularization will allow changes for a single image, perhaps allowing some wrinkles to appear in the reconstruction.

5.2.2 Multiple Reconstructed Shapes

Can more than one shape be recovered from the collection? Or taken to the extreme, can a different reconstruction be recovered for each image in the collection to match the individual expressions? This is a difficult task. One main assumption in our algorithm is consistency of the surface normals across the collection. When the face deforms for different expressions, it actually violates this assumption. We use local selection to choose a consistent set of images for each part of the face.

To reconstruct different models for each photo, we could deform the model using the expression coefficients for each image, and treat the wrinkles as a common bump map. However, it is known that different expressions produce different wrinkles, so this would not be accurate and would likely appear fake.

If enough images are present in the collection, it could be possible to cluster the images by expression, form a different independent reconstruction for each expression, and use those reconstructions to create a person specific blend shape. This would allow different wrinkles for different expressions, but would require very large photo collections.

5.2.3 Face Recognition Application

How does the reconstruction improve face recognition? This is a good question, that needs more research. As a cursory answer, in [101], they demonstrate the ability of 3D face reconstruction from a single image to create a normalized image. Their normalized image, is the image rendered in a frontal viewpoint with the expression removed. Such normalization demonstrates improvement on the Labeled Faces in the Wild dataset.

Our approach can be viewed as a generalization of [101] since it will find the same reconstruction for a single image. The question then, is can it improve for multiple image collections over a simple max or mean fusion of individual results. That is, does the improved reconstruction from multiple images further improve face recognition.

To answer this question, we use the IJB-A dataset [54]. IJB-A contains extremely challenging images with high amounts of occlusion, lighting, and pose variation. It also contains sets of images to be used instead of single images. IJB-A consists of image collections and sample videos for 500 people. There are a minimum of 5 images per person. We design a few experiments on this database to evaluate the effectiveness of the reconstruction.

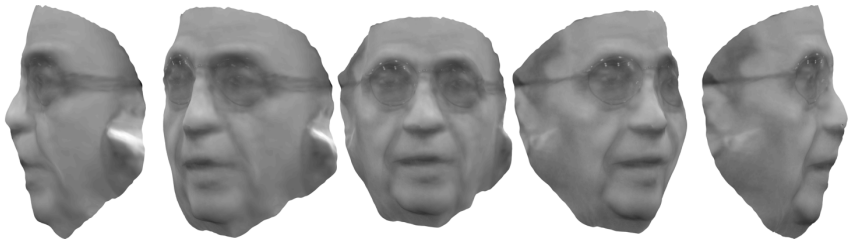


Figure 5.1: Sample synthetic rendering of subject from photo collection using estimated albedo.

First, we withhold 2 images per person as probes and use the remaining images to form gallery photo collections. We do not consider the videos since they contain consistent lighting that does not satisfy our photo collection assumptions. We perform reconstruction on each gallery collection and render synthetic images under a variety of poses using the estimated albedo from the collection

(Fig. 5.1). For face recognition we use the deep neural network from [87] that is state-of-the-art for face identification at scale. An identification experiment is run, where each probe is compared against the gallery with and without including the synthetic images. Results are provided in Fig. 5.2. There is almost no change when including the synthetic images. The main reason is due to the quality of the estimated albedo from the challenging photo collections. Establishing correspondence across the collection is challenging, and Fig. 5.1 shows one of the better estimations still containing artifacts. Most of the synthetic images have low match scores to any real probe image.

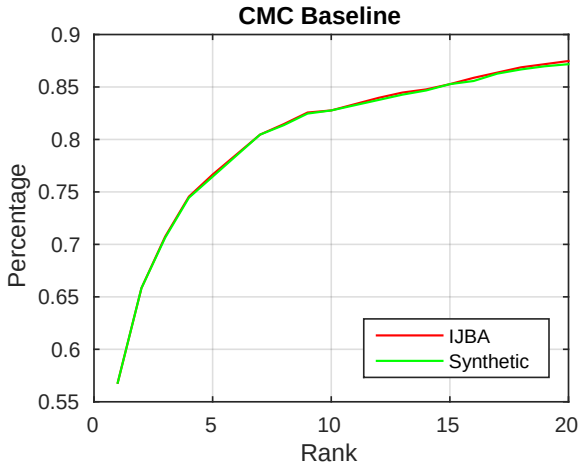


Figure 5.2: CMC curve comparing identification of IJB-A and adding synthetic images rendered from the proposed reconstruction. The synthetic images make an insignificant decrease in identification accuracy.

We also tried a filtering approach for the gallery. IJB-A contains some extremely poor images. Using the structural similarity (SSIM) quality measure for an image, we are able to evaluate, how well the image matches the reconstruction of the person. Images with low SSIM scores probably will not match well to this subject and may be discarded. Figure 5.3 shows the identification performance on the full collection and using a subset where images lower than the global mean SSIM 0.68 or the worst 15% of images from a single collection are filtered out of the gallery. Under this condition, the size of the gallery may be reduced substantially while still maintaining

the same identification performance.

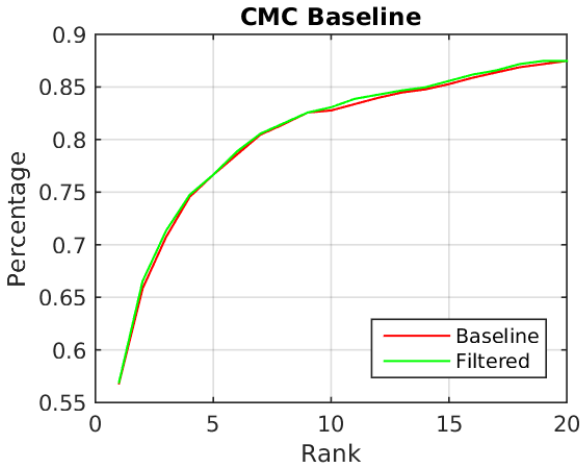


Figure 5.3: CMC curve comparing baseline IJB-A to a filtered subset of IJB-A where the images removed had low structural similarity score based on the reconstructions. The filtered images have an insignificant increase in identification accuracy.

Overall, these recognition results are encouraging, but leave room for future work to explore how best to use the reconstruction results from a photo collection.

APPENDIX

APPENDIX

OTHER PUBLICATIONS

Typing Behavior based Continuous User Authentication

We hypothesize that an individual computer user has a unique and consistent habitual pattern of Typing Behavior, independent of the text, while typing on a keyboard. This habit manifests itself visually through the shapes and motions of the hands, as well as audibly through the sounds produced when pressing and releasing keys. Given a webcam pointing towards a keyboard, we develop real-time computer vision and acoustic algorithms to automatically extract the habitual patterns from the video and audio streams and continuously verify the identity of the active computer user.

Unlike conventional authentication schemes, continuous authentication has a number of advantages, such as longer time for sensing, ability to rectify authentication decisions, and persistent verification of a user's identity, which are critical in applications demanding enhanced security. Traditional biometric modalities such as face and fingerprint have various drawbacks when used in continuous authentication scenarios such as privacy concerns and interference with normal computer operation. We propose typing behavior as a non-intrusive privacy-aware biometric modality that utilizes standard interactions with the keyboard peripheral.

Visual Typing Behavior

To capture the unique and consistent hand movements from typing, we use a simple webcam pointed at the keyboard. Given the video, the proposed system segments the hands from the back-

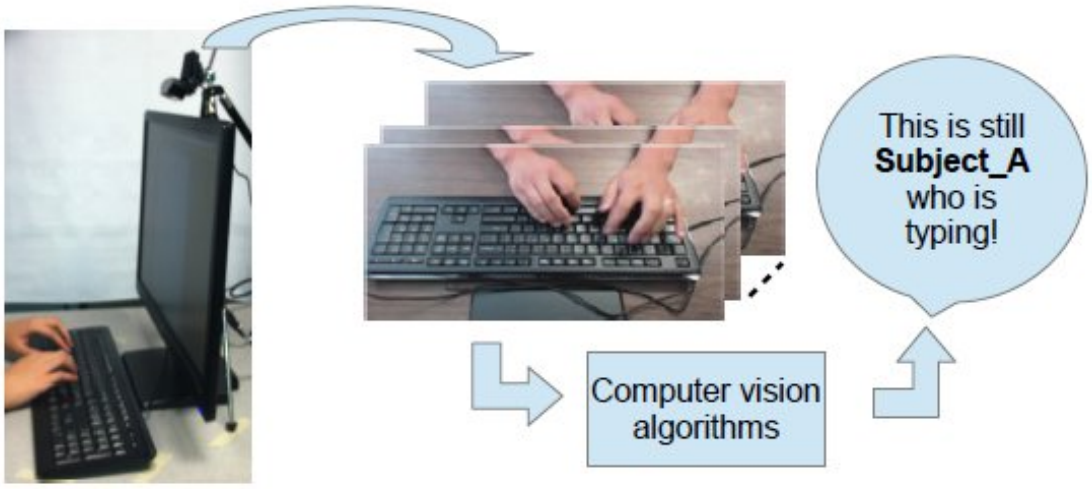


Figure 4: Overview of visual typing behavior. A webcam captures a video of the hands while typing on a keyboard and uses computer vision algorithms to detect and segment the hands and uses the shape information over time to verify the current computer user.

ground and separates the left and right hand. A shape context based on the silhouette is extracted from each hand, which is combined with the hand position relative to the keyboard. We also propose a novel extension extension of Bag of Phrases, *Bag of Multi-dimensional* Phrases, where a probe video finds corresponding hand shapes across the temporal domain, independently for each hand.

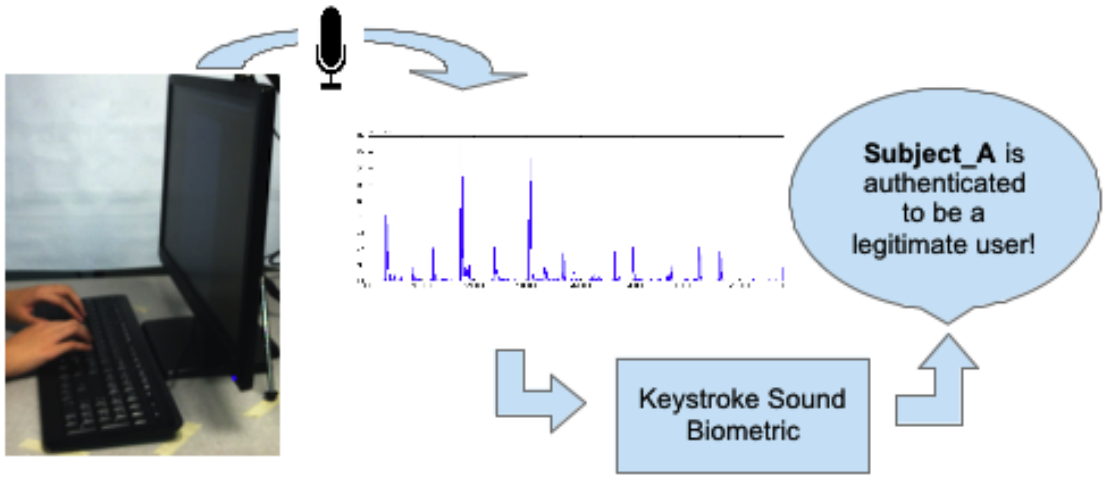


Figure 5: Overview of acoustic typing behavior. A microphone captures the sound produced from keypresses and extracts different informative features to determine the computer user.

Acoustic Typing Behavior

Given the keyboard sound recorded by the webcam, our system extracts discriminative features and performs matching between a gallery and a probe sound stream. Motivated by the concept of digraphs used in modeling keystroke dynamics, we learn a virtual alphabet from keystroke sound segments, from which the digraph latency within pairs of virtual letters as well as other statistical features are used to generate match scores. The resultant multiple scores are indicative of the similarities between two sound streams, and are fused to make a final authentication decision.

Typing Behavior Dataset

We collect a first-of-its kind keystroke database in two phases. Phase 1 includes 56 subjects typing multiple same day, fixed and free text, sessions. It includes the acoustics and video information. Phase 2 includes 30 subjects typing mutliple free text sessions on different days across months. It includes the video information as well as keystroke timing information for use with conventional keystroke dynamics.

This dataset is released in two different forms. Acoustics: 45 subjects from phase 1. Visual: Full dataset. This dataset is provided for non-commercial use.

Publications

1. Joseph Roth, Xiaoming Liu, Arun Ross, and Dimitris Metaxas, "Investigating the Discriminative Power of Keystroke Sound," IEEE Transactions on Information Forensics and Security, Vol. 10, No. 2, pp. 333-345, Feb. 2015.

2. Joseph Roth, Xiaoming Liu, and Dimitris Metaxas, "On Continuous User Authentication via Typing Behavior," IEEE Transactions on Image Processing, Vol. 23, No. 10, pp. 4611-4624,

Oct. 2014.

3. Joseph Roth, Xiaoming Liu, Arun Ross, and Dimitris Metaxas, "Biometric Authentication via Keystroke Sound," in Proceedings of the 6th IAPR International Conference on Biometrics (ICB) 2013, Madrid, Spain, June 4-7, 2013.

Person Re-Identification

Person re-identification seeks to locate the same individual across multiple non-overlapping cameras within a short time frame. It is an enabling technique for video surveillance and has many applications such as tracker linking, person retrieval, searching missing children in public spaces, *etc*. One way of performing person re-identification is to use a set of human describable attributes of the person such as, male, red shirt, has backpack, to return a set of images based on a verbal description.

In this work, I present an algorithm for *jointly* learning a set of mid-level attributes from an image ensemble by locating clusters of dependent attributes. Human describable attributes are an active research topic due to their ability to transfer between domains, human understanding, and improvement to identification performance. Joint learning may allow for enhanced attribute classification when there is inherent dependency among the attributes. We propose an agglomerative clustering scheme to determine *which* sets of attributes should be learned jointly in order to maximize the margin of performance improvement. We find the proposed algorithm improves the attribute classifier accuracy as well as the person re-identification task.

1. Joseph Roth and Xiaoming Liu, "On the Exploration of Joint Attribute Learning for Person Re-identification," in Proceedings of the 12th Asian Conference on Computer Vision (ACCV), Singapore, Nov. 1-5, 2014.

Hair Recognition in the Wild

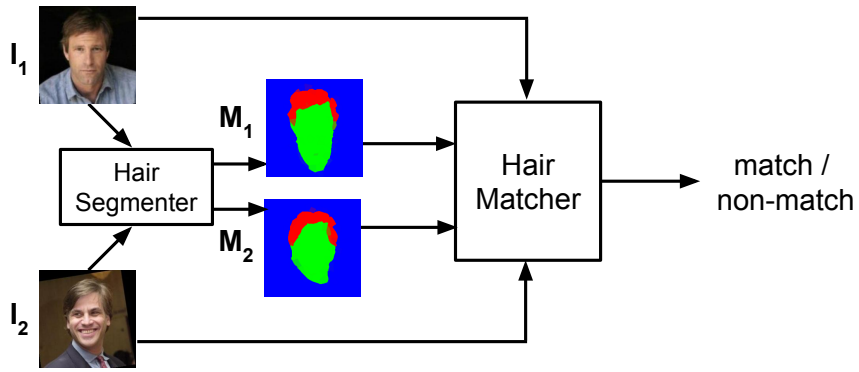


Figure 6: A hair matcher takes two images and their corresonding segmented hair mask and determines if they belong to the same subject or different subjects.

This work presents an algorithm for identity verification using only information from the hair. It is well known that humans utilize hair for identification, especially under challenging situations when the face is occluded, but little work exists to replicate this artificially. We propose a learned hair matcher using shape, color, and texture features derived from localized patches through an AdaBoost technique with abstaining weak classifiers when features are not present in the given location. The proposed hair matcher achieves 71.53% accuracy on the LFW dataset. Hair also reduces the error of a commercial off-the-shelf face matcher through simple score-level fusion by 5.7%.

1. Joseph Roth and Xiaoming Liu, "On Hair Recognition in the Wild by Machine," in Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2014), Quebec City, Canada, July 27-31, 2014.

BIBLIOGRAPHY

BIBLIOGRAPHY

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

[2] O. Aldrian. Inverse rendering of faces with a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(5):1080–1093, May 2013.

[3] 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.

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

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

[6] S. Barsky and M. Petrou. The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(10):1239–1252, 2003.

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

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

[9] 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.

[10] 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.

[11] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pages 545–552. IEEE, 2011.

[12] 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.

[13] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Computer Graphics Proceeding, Annual Conference Series*, pages 187 – 194, New York, 1999. ACM SIGGRAPH.

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

[15] J. Booth, A. Roussos, S. Zafeirious, A. Ponniah, and D. Dunaway. A 3D morphable model learnt from 10,000 faces. In *CVPR*, pages 5543–5552, 2016.

[16] G. Borshukov and L. J. P. Realistic human face rendering for "the matrix reloaded". In *ACM Siggraph 2005 Courses*, page 13, 2005.

[17] J.-Y. Bouguet. Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/`.

[18] A. M. Bruckstein, R. J. Holt, T. S. Huang, and A. N. Netravali. Optimum fiducials under weak perspective projection. *Int. J. Comput. Vision*, 35(4):223–244, 1999.

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

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

[21] 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.

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

[23] 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.

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

[25] 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.

[26] 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*, 47:19–26, Mar. 2016.

[27] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning. *J. Image Vision Computing*, 47:27–35, Mar. 2016.

[28] O. Faugeras and R. Keriven. Variational principles, surface evolution, pde's, level-set methods, and the stereo problem. *IEEE Trans. Image Process.*, 7(3):336–344, 1998.

[29] D. Frolova, D. Simakov, and R. Basri. Accuracy of spherical harmonic approximations for images of lambertian objects under far and near lighting. In *ECCV*, pages 574–587, 2004.

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

[31] A. Y. Govan and C. D. Meyer. Ranking national football league teams using google's pagerank. In *MAM*, 2006.

[32] 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.

[33] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[34] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015.

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

[36] C. Hernández and F. Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004.

[37] C. Hernández, G. Vogiatzis, G. Brostow, B. Stenger, and R. Cipolla. Non-rigid photometric stereo with colored lights. In *ICCV*, 2007.

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

[39] 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.

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

[41] A. Hornung and L. Kobbelt. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. In *CVPR*, pages 503–510, 2006.

[42] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3D reconstruction for face

recognition. In *FG*, pages 843–848. IEEE, 2004.

[43] 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.

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

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

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

[47] I. Jollife. *Principal Component Analysis*. Springer-Verlag, 1986.

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

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

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

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

[52] I. Kemelmacher-Shlizerman and R. Basri. 3D face reconstruction from a single image using a single reference face shape. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(2):394–405, 2010.

[53] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *ICCV*, 2011.

[54] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a. In *CVPR*, 2015.

[55] P. Labatut, J.-P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *ICCV*, 2007.

[56] 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.

[57] 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.

[58] X. Liu. Discriminative face alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(11):1941–1954, 2009.

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

[60] P. Martins. Active appearance models for facial expression recognition and monocular head pose estimation. Master's thesis, University of Coimbra, 2008.

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

[62] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(4):353–363, 1993.

[63] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. 1999.

[64] 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.

[65] 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.

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

[67] M. Piotraschke and V. Blanz. Automated 3d face reconstruction from multiple images using quality measures. In *CVPR*, 2016.

[68] M. Pollefeys, D. Nister, J. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. Kim, and P. Merrell. Detailed real-time urban 3d reconstruction from video. *Int. J. Comput. Vision*, 78(2):143–167, 2008.

[69] C. Qu, E. Monari, T. Schuchert, and J. Beyerer. Adaptive contour fitting for pose-invariant 3D face shape reconstruction. In *BMVC*, 2015.

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

[71] S. Romdhani and T. Vetter. Estimating 3D shape and texture using pixel intensities, edges, specular highlights, texture constraints and a prior. In *CVPR*, 2005.

[72] J. Roth, Y. Tong, and X. Liu. Unconstrained 3D face reconstruction. In *CVPR*, 2015.

[73] R. Rsai. Multiframe image point matching and 3-d surface reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 5(2):159–174, 1983.

[74] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *J. Image Vision Computing*, 47:3–16, Mar. 2016.

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

[76] S. Sinha, P. Mordohai, and M. Pollefeys. Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In *ICCV*, pages 1–8, 2007.

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

[78] 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.

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

[80] C. Streecha, R. Fransens, and L. Van Gool. Wide-baseline stereo from multiple views: a probabilistic account. In *CVPR*, 2006.

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

[82] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, June 2016.

[83] 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.

[84] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, July 2005.

[85] G. Vogiatzis and C. Hernández. Practical 3d reconstruction based on photometric stereo. In *Computer Vision*, pages 313–345. Springer, 2010.

[86] G. Vogiazis and C. Hernandez. Automatic camera pose estimation from dot pattern. `http://george-vogiatzis.org/calib/`.

[87] D. Wang, C. Otto, and A. K. Jain. Face search at scale. *IEEE Trans. Pattern Anal. Mach. Intell.*, June 2016.

[88] 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.

[89] Z. Wang, A. C. Bovik, and H. R. Sheikh. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[90] T. Weise, H. Li, L. Van Gool, and M. Pauly. Face/off: live facial puppetry. In *SCA*, pages 7–16, 2009.

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

[92] 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.

[93] 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.

[94] 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.

[95] F. Yang, J. Wang, E. Shechtman, L. Bourdev, and D. Metaxas. Expression flow for 3D-aware face component transfer. *ACM Trans. Graph.*, 30(4), July 2001.

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

[97] 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.

[98] A. L. Yuille, D. Snow, R. Epstein, and P. N. Belhumeur. Determining generative models of objects under varying illumination: Shape and albedo from multiple images using SVD and integrability. *Int. J. Comput. Vision*, 35:203–222, 1999.

[99] 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.

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

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