3D Face Geometry Capture Using Monocular Video

Shubham Agrawal

May 20, 2019

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Thesis Committee:
Simon Lucey, *Chair*
Martial Hebert
Ming-Fang Chang, *Carnegie Mellon University*

Thesis proposal submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics

Abstract

Accurate reconstruction of facial geometry has been one of the oldest tasks in computer vision. Despite being a long-studied problem, many modern methods fail to reconstruct realistic looking faces or rely on highly constrained environments for capture. High fidelity face reconstructions have so far been limited to either studio settings or through expensive 3D scanners. On the other hand, unconstrained reconstruction methods are typically limited by low-capacity models. We aim to capture face geometry with high fidelity using just a single monocular video sequence of the face.

Our method reconstructs accurate face geometry of a subject using a video shot from a smartphone in an unconstrained environment. Our approach takes advantage of recent advances in visual SLAM, keypoint detection, and object detection to improve accuracy and robustness. By not being constrained to a model subspace, our reconstructed meshes capture important details while being robust to noise and being topologically consistent. Our evaluations show that our method outperforms current single and multi-view baselines by a significant margin, both in terms of geometric accuracy and in capturing person-specific details important for making realistic looking models.

To further the current work on single and multi-view 3D face reconstruction, we also propose a dataset of video sequences of individuals, specifically with the goal to improve deep-learning based reconstruction techniques using self-supervision as a training loss.

Acknowledgements

First and foremost, I would like to express my deepest thanks to my advisor, Dr. Simon Lucey. His patient guidance, continued encouragement, and immense knowledge were key motivating factors throughout my masters. His insights on what the right tool would be to solve a particular problem really helped me hone my own intuition as a researcher and helped guide me through my masters.

It also has been an absolute privilege to work with and learn from all the brilliant people in my masters program cohort. I have had the good fortune to meet the most helpful and humble people at CMU RI.

Last, but not the least, I would like to express my gratitude and indebtedness to my family for their love and support. Their support and hope brought me here in the first place.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Reconstructing faces has been a problem of great interest in computer vision and graphics with applications in a wide variety of domains, ranging from animation [29], entertainment [47], genetics, bio-metrics, medical procedures, and more recently, augmented and virtual reality. Despite the long body of work, 3D face reconstruction still remains an open and challenging problem, primarily because of the high level of detail required owing to our sensitivity to facial features. Even slight anomalies in the reconstructions can make the output look unrealistic and hence, the accuracy of reconstructed face models is of utmost importance.

While accurate scans of facial geometry can be obtained using structured light or laser scanners, these are often prohibitively expensive, typically costing tens of thousands of dollars. The seminal work of Beeler [7] showed that a studio setup of cameras could be used to capture face geometry accurately. Since then, a variety of work has focused on using Photometric stereo or Multi-view stereo techniques in studio settings for face reconstruction and performance capture [13, 21]. Although accurate in their reconstructions, these studio setups are not trivial to set up, typically requiring a calibrated camera setup along with controlled lighting and backgrounds. This makes them infeasible for capturing 'in-the-wild' subject faces in unconstrained settings, for instance, an end user of a virtual reality app.

To tackle the problem of unconstrained 3D face reconstruction, the community has mostly relied on three-dimensional morphable models (3DMMs) [9]. 3DMMs are low-dimensional linear sub-spaces of faces typically constructed using a small set of ground truth 3D scans that enable rapid approximation of face geometry, classically through a non-linear optimization over appearance and landmarks. Deep neural nets have more recently been used to fit morphable models using a single image. Generalization to in-the-wild images is often a concern for these methods. While the results are often visually appealing with texture, the reconstructions suffer from high geometric inaccuracies.

With the limited availability of 3D data for faces, using geometric cues from multiple views to improve the accuracy of reconstruction becomes necessary. Previous work has shown that a single template or 3DMM can be optimized using constraints from multiple views, using techniques like photometric stereo [45] or advances in automatic keypoint detection [28]. Recently, Hernandez [26] proposed an elegant multi-view constrained structure-from-motion scheme that explicitly optimized the coefficients of a 3DMM shape

to recover face geometry. However, the output still remains constrained to the underlying training data and low capacity of the 3DMM. This greatly limits its expressivity and is particularly undesirable for medical or bio-metric usage.

In this work, we attempt to answer the question "What's the most accurate reconstruction an end-user can obtain, without needing access to special equipment or studio setups?". To this end, we propose a pipeline for highly accurate yet robust face geometry capture, requiring nothing but a smartphone. We leverage recent advances in the fields of object and keypoint detection, direct methods for visual SLAM, and higher frame-rate capture functionality available on modern smartphones. This allows us to incorporate multi-view consistency, landmark, edge and silhouette constraints into a single optimization framework. We also explicitly train a model for ear detection to incorporate ear landmarks, an area that has almost entirely been ignored in face reconstruction works. This enables us to achieve state-of-the-art geometric accuracy among unconstrained face reconstruction techniques.

1.2 Challenges

Ground truth 3D data

Many recent approaches at face reconstruction have focused on the problem of single-view reconstruction using deep networks. Since single-view reconstruction is an ill-posed problem, these networks rely on learnt priors about face geometry for the reconstruction. However the major challenge here is the dearth of 3D ground truth training data available for faces, which is needed to train such models in a supervised manner. While datasets like Imagenet have enabled very impressive results on 2D computer vision tasks using deep learning based approaches, there is no such large-scale dataset publicly available for 3D faces. Thus most deep networks have to rely on synthetic on lower-fidelity 3D face data generated using proxy methods in order to train the networks. This in turn leads to poor generalization and often very smoothed out face geometry which does not really capture the distinctive person-specific detail that is crucial for downstream tasks on the reconstruction.

Shape priors

To handle the ill-constrained nature of 3D reconstruction form monocular sources, a lot of face reconstruction literature has focused on using priors of the face geometry to reduce the complexity of the reconstruction problem. The most popular of these priors has been based on the seminal work on 3D morphable models by Blanz and Vetter [9]. 3D Morphable models (3DMMs) provide a low dimensional representation of the face geometry, by modeling the linear subspace of a few training set meshes using a PCA decomposition. Thus, the task of recovering a face geometry just becomes solving for the PCA coefficients (α_s). i.e., a 3D face can be recovered as :

$$X = \bar{X} + P_s\alpha_s$$

Where $\bar{X}$ is the mean 3D shape.

However, as noted in several works, the PCA based 3DMM representation is severely constrained in its expressivity, and resulting reconstructions lack any detail. Further, it is important to have a very diverse set of meshes in the training data of the model, as the resulting mesh would be constrained to the linear subspace of the data. Thus, while 3DMMs

greatly reduce the complexity of the reconstruction problem, the tradeoff is upper limit on the fidelity of the reconstruction and generalization issues.

Pose estimation

For improving the accuracy and generalizability of reconstruction algorithms in the aforementioned data-constrained problem setup, using multiple views of the face becomes necessary. Under this setup, multiple images of a single individual's face are available, either taken during a single capture time, or spread across time (such as collections of celebrity images).
A major step in a multi-view formulation is estimating the poses of the cameras, either with respect to a reference camera, or the face. This is needed to geometrically constrain and fuse the information of the multiple views. Historically, face landmarks detectors have primarily been used for this purpose. These detectors, output 2d locations of a few semantic fiducial keypoints on the face, such as the eyes, lips and nose. If the landmarking is accurate, the task for pose estimation with respect to the face can be easily solved as a PnP problem. However, as can be seen in fig 1.1, due to the nature of the training data of such trackers, the landmarking is not really robust beyond a certain angle with respect to the frontal face. Many expressions are also not handled well. Since in the monocular 3D reconstruction problem, 3D structure and camera pose estimation is tightly coupled, relying on noisy landmarks for poses in turn reduced the accuracy of the inferred 3d structure.

1.3 Contributions

Our contributions are two-fold. First, we propose a 3D face reconstruction algorithm that takes a single video of a subject's face and reconstructs their face geometry, making high fidelity reconstructions accessible to users for downstream tasks like animation, printing, genetic analysis/modeling and bio-metrics. The reconstructed meshes have semantic correspondence and consistent topology, without being constrained to any model subspace. Second, we release a dataset 200 video sequences of 100 individuals shot at 120fps, where we collect two sequences per subject, under varying lighting conditions.

1.4 Thesis Outline

The thesis is organized as follows.
In Chapter 2, explore and describe the previous work done in the field of 3D reconstruction of faces. We divide the works into a few broad categories, and analyze the pros and cons of each strategy, and where the current state-of-the-art is.
Chapter 3 describes our Pose estimation strategy, to recover poses from the uncalibrated video clip.
In Chapter 4 we talk about the Multi-view stereo problem, and the algorithm we use to recover 3D structure of the face in the form of point cloud.
Chapter 5 details our mesh-fitting algorithm, which is key in capturing face geometry in a semantic mesh while being robust to noise.
In Chapter 6 we discuss an effective strategy based on the idea of mesoscopic augmentations that allows for recovering high-frequency details on the mesh using the face texture. In Chapter 7 we discuss quantitative and qualitative evaluation of our reconstruction pipeline.

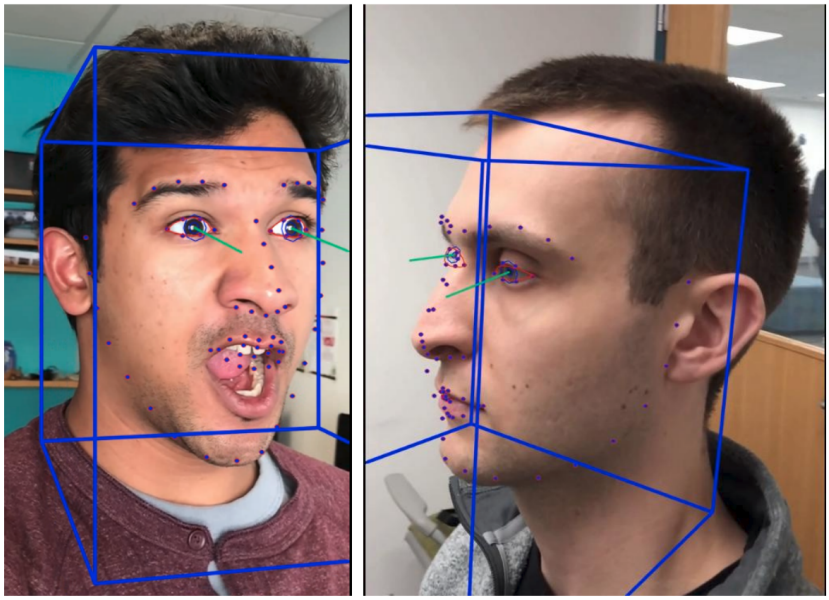

Figure 1.1: While machine learning based models for keypoint detection have really improved over the past few years, they are still fairly brittle to images with face angles beyond a certain threshold and face geometry beyond what they may have been trained with. This in turn means that relying on landmarks for pose estimation does not lead to accurate pose estimates

In order to improve existing methods, we discuss our proposed dataset, its collection and the motivation to construct such a dataset in Chapter 8.
Finally in the last chapter, conclusions, limitations and future work are discussed.

Chapter 2

Related Work

Prior work on 3D Face Reconstruction is substantially large. To make the analysis easier, we classify the works on the basis of the general 3D reconstruction approach they follow, namely : 3DMM recovery, SfM based Multi-view Reconstruction, Photometric Stereo, Single Image 3D Face Reconstruction

2.1 3D Morphable Models (3DMMs)

One ot the most seminal work in this field has been the 3D morphable model approach proposed by Blanz and Vetter [9]. The authors collected 200 scans human faces, performed a dense alignment on them using Procrustes analysis. The normalized vertex coordinates of the 200 meshes were stacked together as column vectors into a single matrix. They then performed a PCA decomposition on this matrix, thus obtaining a low dimensional representation of the face geometry, by modeling the linear subspace of this training data. Thus, the task of recovering a face geometry just becomes solving for the PCA coefficients (α_s). i.e., a 3D face can be recovered as :

$$X = \bar{X} + P_s\alpha_s$$

Where $\bar{X}$ is the mean 3D shape.

However, as noted in several works, the PCA based 3DMM representation is severely constrained in its expressivity, and resulting reconstructions lack any detail.

2.2 Single Image 3D Face Reconstruction

3D Morphable Models have successfully been used as prior for modeling faces from a single image [9, 12, 32, 44, 48, 54, 58]. Facial landmarks have commonly been used in conjunction with 3DMMs for the reconstruction [1, 15, 37, 58]. While landmarks are informative for 3D reconstruction, relying primarily on them results in generic looking meshes which lack recognizable detail. More recently, convolutional neural networks have been put to use for directly regressing the parameters of the 3D Morphable Model [33, 57]. To overcome the limited expressivity of 3DMMs, recent methods have tried to reconstruct unrestricted geometry, by predicting a volumetric representation [30], UV map [18], or depth map [51].

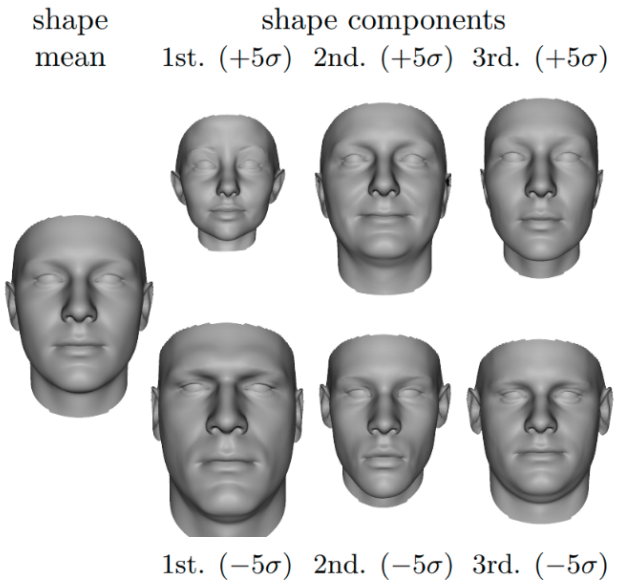

Figure 2.1: Visualization of 3DMM mesh, and variation along the first few principal components

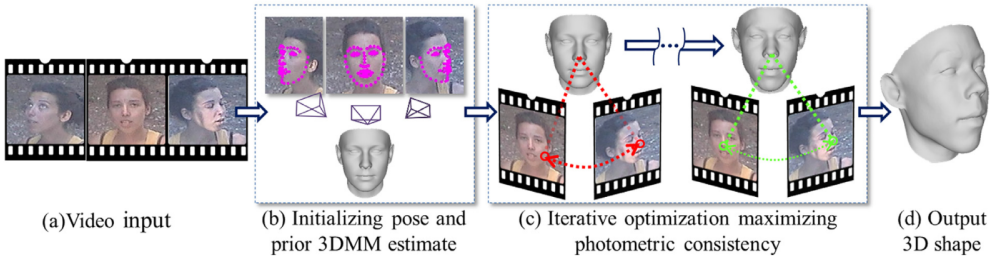

Figure 2.2: Overview of the pipeline of the state of the art multi-view algorithm of [26] et. al. based on prior constrained structure from motion. The method uses landmarks to initialize poses in a bundle-adjustment system that minimizes photometric consistency between frames, while optimizing 3d structure in the constrained 3DMM space.

However, the underlying training data of these methods has been limited to synthetic data generated using 3DMMs or course meshes fit using landmarks. Thus the ability of these methods to generalize to 'in-the-wild' images and face geometry is still quite limited. While single image reconstruction is of great research interest, we believe multi-view consistency is crucial for generating accurate 3D face representations, specially given the limited data available for faces. For a more comprehensive literature review of monocular 3D Face Reconstruction, we direct the readers to [59].

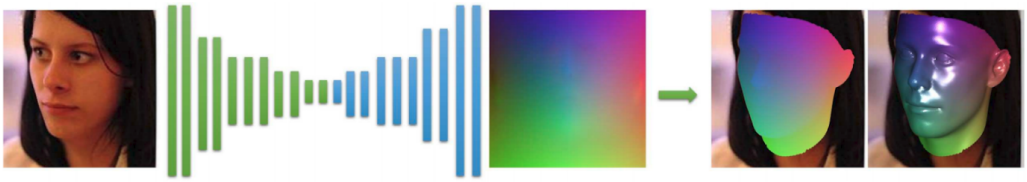

Figure 2.3: General approach taken by current SOTA image-to-image translation based deep networks for reconstruction. The training data of such networks is still limited by low quality synthetic data.

2.3 SfM based Multi-view Reconstruction

A lot of multi-view reconstruction methods employ a Structure-from-Motion pipeline [19, 23, 40] but with unconvincing results on unconstrained in-the-wild videos [26]. [11] and [52] use 3D Morphable Model [9] for fitting shapes on every frame after computing correspondences among them. This restricts the reconstruction to a low-dimensional linear subspace. The current state-of-the-art approach by Hernandez [26] uses 3DMM as a prior instead to search for correspondences among frames. This allowed them to achieve state-of-the-art results in unconstrained multi-view face reconstruction. However their method requires camera intrinsics to be known and the output is still constrained to a linear basis. We use this method as one of the baselines for comparison.

2.4 Photometric Stereo

Photometric stereo based methods have proven effective for large unconstrained collection of photos [36, 38, 45]. [38] generates a 2.5D face surface by using SVD to find the low rank spherical harmonics. Roth [45] expand on it to handle pose variations and the scale ambiguity prevalent in the former method. They further expand their work in [46] where they fit a 3DMM to 2D landmarks for every image and optimize for the lighting parameters rather than SVD based factorization. Suwajanakorn [53] use shape from shading coupled with 3D flow estimation to target uncalibrated video sequences. While these methods capture fine facial features, most of them rely on simplified lighting, illumination and reflectance models, resulting in specularities and unwanted facial features showing up on the mesh.

Chapter 3

Camera Pose Estimation

3.1 Introduction

Most multi-view face reconstruction methods have traditionally relied on pre-calibrated cameras (a studio setup) or used landmark trackers for estimating camera pose relative to a geometric prior, such as a template mesh or 3DMM. However, landmark trackers are less than reliable beyond a small angle from the front of the face, which reduces their utility for camera pose estimation. For our method, we aim to get sub-pixel accurate camera pose estimates using recent advances in direct methods for visual SLAM, based on the seminal work by Engel [16, 17]. Direct methods are particularly effective for faces, where a lot of corner points are not present for feature point detection and matching.

3.2 Our Approach

We take advantage of the fact that the input is a single continuous video sequence. We use the geometric bundle adjustment based initialization scheme proposed in [24] to get relative pose estimates for an initial baseline distance. Then, a LK tracker is used to track the camera frames in the video, and a keyframe is selected once the camera moves a certain baseline distance. The set of keyframe camera poses are optimized using photometric bundle adjustment to maximize photometric consistency between frames.

As in [16], PBA is a joint optimization of all model parameters, including camera poses, the intrinsics, and the radial distortion parameters. For a typical sequence, 50-80 keyframes with accurately known camera poses are obtained.

Independently of pose estimation, we use the publicly available Openface toolkit [4] for facial landmark detection. We fit the Basel 3DMM [9] to these landmarks and align it with the coordinate system of the keyframes. We use this coarse mesh in the next stage.

The advantages of decoupling camera pose estimation and face alignment are threefold: 1) Robustness to landmark tracker failures, which, despite many recent advances, is not robust at large angles 2) By not relying on the estimated coarse mesh for registering camera poses, errors in the shape estimation do not propagate to the camera poses. 3) Purely using photometric consistency allows us to achieve sub-pixel accuracy in estimating camera poses.

Chapter 4

Multi-view Stereo

4.1 Introduction

At the end of the PBA stage, we obtain a set of 50-80 keyframes whose camera poses are known with high accuracy, and a coarse face mesh fitted to the landmarks from Openface. Next, we use these keyframes to generate a dense point cloud of the face geometry using Multi-view stereo.

4.2 Our Approach

We use the parallelized multi-view PatchMatch implementation of Galliani [22] and use 12 source views for each reference view for depth inference. The core algorithm is based on randomized search. Starting with random depth estimated per pixel, the algorithm uses a cost metric based on photometric consistency to propagate good depth guesses. This is done for a fixed number of iterations. The multi-view PatchMatch estimates a depth map for each of the keyframes. We initialize the depths and search range using the coarse mesh.

To select which source views to use to infer the depth map of a reference view, we calculate a view selection score [55] for each pair of keyframes, $s(i,j) = \sum_{\mathbf{p}} \mathcal{G}(\theta_{ij}(\mathbf{p}))$, where $\mathbf{p}$ is a point common to both views and its baseline angle between the cameras $\mathbf{c}_i$ and $\mathbf{c}_j$ is $\theta_{ij}(\mathbf{p}) = (180/\pi)\arccos((\mathbf{c}_i - \mathbf{p}) \cdot (\mathbf{c}_j - \mathbf{p}))$. $\mathcal{G}$ is a piecewise Gaussian, as follows :

$$\mathcal{G}(\theta) = \begin{cases} \exp(-\frac{(\theta-\theta_0)^2}{2\sigma_1^2}), \theta \leq \theta_0 \\ \exp(-\frac{(\theta-\theta_0)^2}{2\sigma_2^2}), \theta > \theta_0 \end{cases}$$

For our method, we pick $\theta_0 = 10$, $\sigma_1 = 5$ and $\sigma_2 = 10$. We use the estimated coarse mesh to filter out noisy patches in the depth maps produced by the PatchMatch. We then project the depth maps to a single fused point cloud using the fusion strategy proposed in [22].

Example point clouds output at this step are visualized in Fig 4.1.

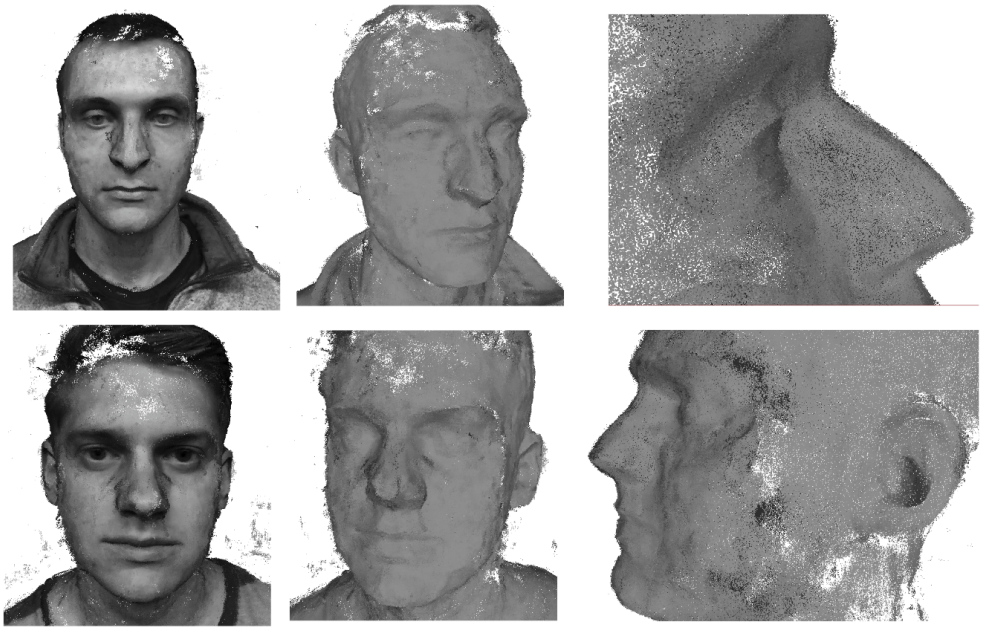

Figure 4.1: Example point clouds generated at the end of our Point cloud generation stage, with and without texture. The point clouds accurately capture the overall face geometry and details in areas like eyes and lips, that make the person recognizable. However, the point clouds have missing data as well as noise, which requires a robust mesh fitting approach

Chapter 5

Mesh Fitting

5.1 Introduction

Due to factors like non-ideal lighting, lack of texture and sensor noise of the smartphone, the obtained point cloud typically has noise and incompletions, with the points distributed around the 'true' surface. Techniques like Screened Poisson reconstruction or the depth map fusion strategy of [25] either return meshes with a lot of surface noise or extremely smoothed out details, depending on the regularization used (see Fig. 5.1). Further, for the reconstructed mesh to be of use in further downstream tasks such as animation, biometrics or as input to a learning algorithm, it is extremely desirable for the meshes to have a consistent topology.

Statistical ICP inspired techniques have proposed fitting a 3DMM to a point cloud [6, 8, 49] in the past. However, they often assume that the point cloud is from an RGB-D sensor and so has a single, 'clean' surface. Further, fitting a 3DMM defeats the purpose of not being constrained to an existing linear basis of shape. We thus adapt the non-rigid mesh fitting algorithm of [3], originally proposed for registering template meshes to 3D scanner data, to deform a template using a combination of constraints given by the point cloud, landmarks, mesh stiffness and edge constraints.

5.2 Point cloud constraints

The primary constraint for the mesh deformation comes from the 3D information captured in the point cloud. While well-studied techniques exist to register a template mesh to a 3D scanned mesh [3], registering a mesh to point clouds of the sort obtained from multi-view stereo techniques is more challenging. For example, simply fitting each vertex to its nearest-neighbor in the point cloud will cause the mesh to become extremely noisy, as there will be many outlier points.

To address this, we take advantage of the fact that for a template mesh, the vertex normals can be easily estimated. For each vertex, we select the points in its neighborhood, and for each point, we calculate its perpendicular distance to the normal of the vertex. Points within a small threshold distance are accepted while the rest are rejected (see Fig. 5.2.

For each vertex on the template mesh we obtain its desired location in 3D as the median of the accepted points.

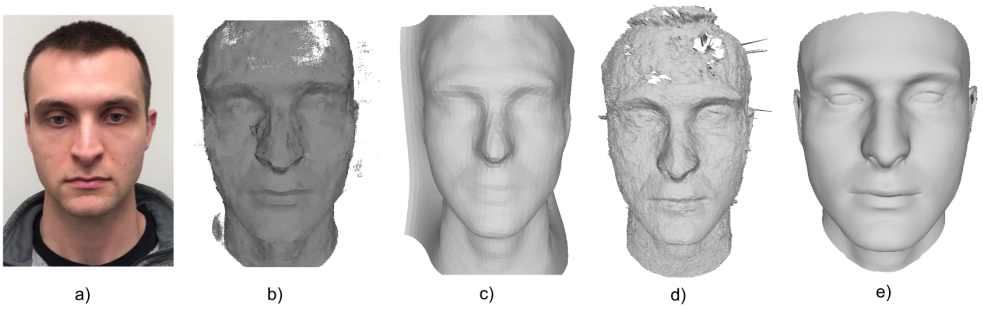

Figure 5.1: Comparison of mesh generation methods a) Sample image. b) Generated point cloud c) [35] can fill in gaps in the point cloud but at the cost of overly smooth meshes. d) Depth fusion method of [25] can preserve details, but is unable to handle missing data. e) Our approach reconstructs meshes with consistent topology and correspondence between vertices, while capturing details of the point cloud and being robust to noise and missing data.

5.3 Landmark constraints

The second source of information are the 68 2D landmarks obtained using the automatic landmarking solution of [4]. Landmarks are important for global alignment and scaling of the mesh, as well as ensuring all the reconstructed meshes are in semantic correspondence.

For the set of frames for which the landmarks have been annotated with high confidence by the tracker (typically close to frontal poses), we solve for the 3D locations of the landmarks by minimizing geometric reprojection error,

$$E_{X_j} = \sum_i \sum_j d(\pi(\theta_i, X_j), x_{ij})^2 \tag{5.1}$$

Where θ_i is the i-th camera's pose, X_j is the j-th landmark's coordinates in 3D, and x_{ij} is the 2D coordinate of the landmark returned by the landmark tracker for the i-th frame. For our purposes, we ignore the 18 landmarks corresponding to the face contour, and use the remaining 50 landmarks as constraints for the corresponding 3D vertices.

Historically, most landmark trackers have focused only on these 68 keypoints. As a consequence, many reconstruction techniques either focus only on reconstructing the frontal face region, or generate a full mesh but evaluate only on the frontal section. Ears and the side facial regions have mostly been ignored in previous works. Even learning-based face alignment techniques do not do well on the ears, as the underlying training data is based on the annotation/detection of these 68 landmarks.

To explicitly address this, we make use of a recent dataset of 'in-the-wild' ear images annotated with bounding boxes and landmarks [56]. We first train the deep object detection model of Redmon [43] for a single 'ear' class. We then train an ensemble of regression trees [34] for predicting landmarks using the bounding box detection as input. As seen in Fig 5.3, despite the limited training data size, we are able to achieve impressive robustness and accuracy in the landmark detection. We use a subset of the landmarks corresponding to the outer contour of the ear as additional landmark constraints in our mesh fitting. To

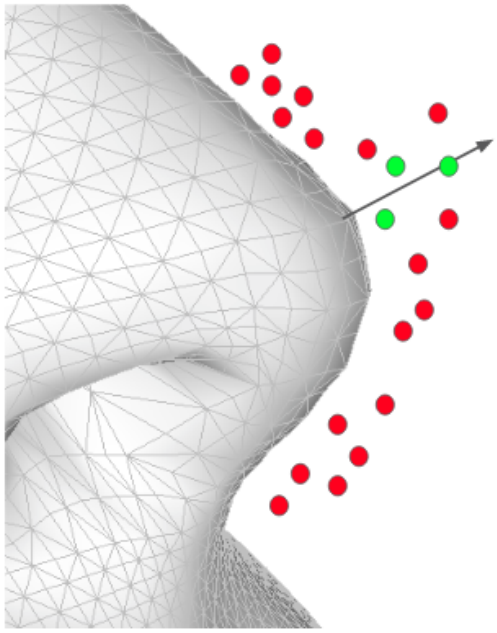

Figure 5.2: Exaggerated view of the point cloud constraints. For each vertex, the set of points within a small threshold of its normal (in green here) are found and their median used as the target 3D coordinate for the vertex.

the best of our knowledge, ours is the first face reconstruction method to explicitly address the ears, which in turn improves overall accuracy and metrics like the width of the face.

5.4 Edge constraints

Silhouette constraints have shown to be powerful cues in recent 3D reconstruction literature [2, 5]. For faces, views that are close to profile are particularly informative. However, since many single and multi-view approaches rely on landmarking for camera pose estimation, they fail to make use of silhouettes beyond a certain angle. By solving for the camera poses independently of landmarking, we can actually make use of extreme profile views. This proves to be helpful in capturing challenging areas for face reconstruction algorithms, such as the nose, lips and lower chin/neck region. We use a combination of Z-buffering [20] and backface-culling to estimate vertices that project an edge onto a given view. To find the corresponding edges in the RGB image, we use the Structured Forests edge detection approach proposed in [14]. For each vertex projecting an edge in the frame, its nearest neighbor is found in the edge map. This corresponding point is back-projected in 3D to obtain a 'target' location for the vertex in 3D.

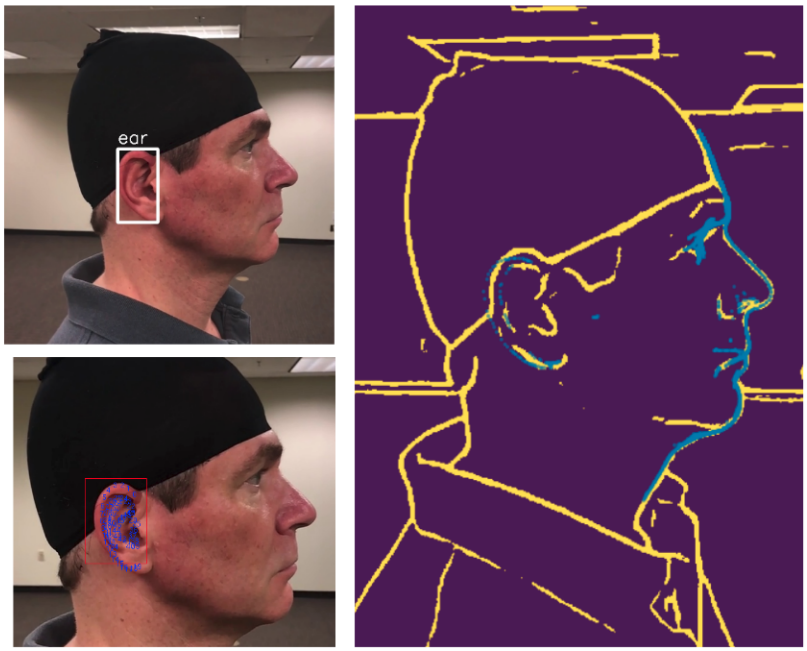

Figure 5.3: a) We train a bounding box regressor (above) and landmark detector (below) specifically for ears. This improves our reconstruction's overall accuracy while allowing us to capture ear size and contour. b) Visualization of edge constraints. Image edges in yellow, mesh vertices corresponding to edges projected in blue. Note that mesh vertices fit the ear well because of the ear landmark detection.

5.5 Non-Rigid Iterative Closest Points

With the combination of the cues from the point cloud, landmarks, and silhouettes, we obtain a set of constraints that we wish to use to deform the template mesh. For a template mesh M of fixed topology (V, E), this can be written as a weighted sum of energies we wish to minimize:

$$\arg\min_{V} E_{pcl} + \alpha E_{lms} + \beta E_{edges} + \gamma E_{reg}$$

where E_{reg} is a regularization energy arising from the mesh topology that restricts connected vertices to deform similarly. This system can naturally be expressed in the iterative linear system-based non-rigid mesh registration algorithm proposed by Amberg [3].

At each iteration, a linear system of the form $\mathbf{AX} = \mathbf{B}$ is solved, where $\mathbf{X}$ is a $4n \times 3$ matrix, containing the per-vertex 3x4 affine transform matrix. The matrix $\mathbf{A}$ captures information of the source template in terms of the mesh connectivity and vertex locations. The mesh connectivity acts as an adjustable 'stiffness' regularization, which controls how much neighboring vertices can move with respect to each other. The matrix $\mathbf{B}$ contains the corresponding 'target' locations in 3D, such as those obtained from the point cloud, landmarks

and edges.

The mesh regularization energy is modeled using a laplacian like formulation, weighted by a "stiffness" scalar, which determines how much local curvature can occur in the mesh. For the stiffness item, we define a node-arc incidence matrix M. If edge r connects the vertices (i, j) and $i < j$, the nonzero entries of M in row r are $M_{ri} = 1$ and $M_{rj} = 1$. Then the item can be rewritten as:

$$E_{reg}(X) := \|(M \otimes G)X\|_F^2 \tag{5.2}$$

Similarly, the other energy terms can be written as follows :

$$E_{lms}(X) := \|D_L X - U_L\|_F^2 \tag{5.3}$$

$$E_{edges}(X) := \|DX - U_m\|_F^2 \tag{5.4}$$

Now, the original cost function becomes a quadratic function:

$$\begin{aligned} E(X) &= \left\| \begin{bmatrix} \gamma M \otimes G \\ D \\ \alpha D_{lms} \\ \beta D_{edges} \end{bmatrix} X - \begin{bmatrix} 0 \\ U \\ \alpha U_{lms} \\ \beta U_{edges} \end{bmatrix} \right\|_F^2 \\ &= \|AX - B\|_F^2 \end{aligned} \tag{5.5}$$

which is a typical linear least square problem. And $E(X)$ takes on its minimum at $X = (A^T A)^{-1} A^T B$. Thus, For each iteration, given fixed correspondences and coefficents, we could determine the optimal deformation quickly. We use sksparse's cholesky decomposition functionality to do this inversion efficiently.

The mesh stiffness and the weights of the landmarks are gradually decreased, gradually moving from global stretching to local, data-driven deformations. After every few iterations, the point cloud and edge constraints are recalculated using the current locations of the vertices. For further details, we refer the reader to the original paper [3]. For our template, we use the Basel 3DMM mesh [9], simply because of its prevalence as an input or output of several face reconstruction algorithms.

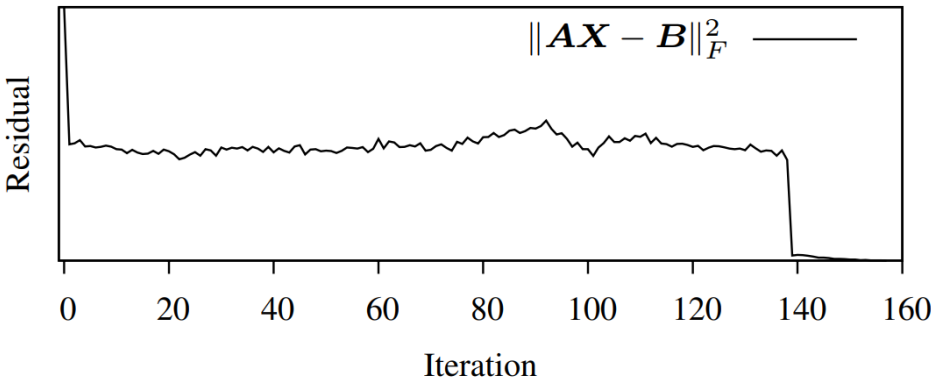

Figure 5.4: The figure shows the non-monotonic decrease of the residual. This non-convex nature prevents the use of a black-box optimiser. The figure shows the residual versus iteration during a registration. The residual increases between some steps, as the reliability weights increase when the template aligns itself with the target and more points find a correspondence. A general optimiser can not escape from the local minima, while the method we use is robust to this behaviour of the loss. In our method, convergence is determined when a threshold stiffness value is reached.

Chapter 6

Mesoscopic Augmentations

6.1 Introduction

A recent trend in the 3D face reconstruction research has been to emboss fine high-frequency details to the reconstruction with methods like shape-from-shading [42] or mesoscopic augmentations [7]. While not always reflecting the true geometry of the surface, these methods add realism to the mesh, which can be desirable for purposes like animation. Such methods can easily be applied to our reconstructions as well. We modify the mesoscopic augmentation method proposed in [51] so that the underlying geometry is preserved, and apply it to our meshes. Since these are based on a dark-is-deep assumption, we skip quantitative evaluation, and provide qualitative results in Fig.6.1. Details on the modified approach are provided in the supplementary.

Recently, Sela [51] showed impressive results by interpreting the idea of high frequency mesoscopic augmentations [7] through mesh heat flows. In our experiments, we found this method to not adapt well to "in-the-wild" images, distorting the mesh too much due to sensor noise/unconstrained lighting. We make some modifications to their method to add details without losing the underlying mesh structure.
Since this method is based on a "dark-is-deep" assumption and not necessarily founded in geometry, we skip quantitative evaluation for these results and simply provide qualitative comparisons between our reconstructions, with and without our augmentation scheme, compared to the augmentation scheme of [51].

6.2 Our Approach

Beeler [7] proposed using a high-pass filtered version of the texture to emboss a mesh with fine details, such as wrinkled and pores. The recent work of Sela [51] proposed using the mesh itself to obtain the high frequency component of the texture, using heat flow to model a low pass filterer version of the texture. In their results, this modification allows them to capture more medium-to-fine scale details, such as the nasolabial folds. However, we observed that their method also tends to distort the mesh, and also pick up on other high frequency noise such as sensor noise that is not desirable. We thus propose to augment our reconstructions as follows:
For a mesh with per vertex texture mapping τ_v, we calculate a low-pass filtered version of

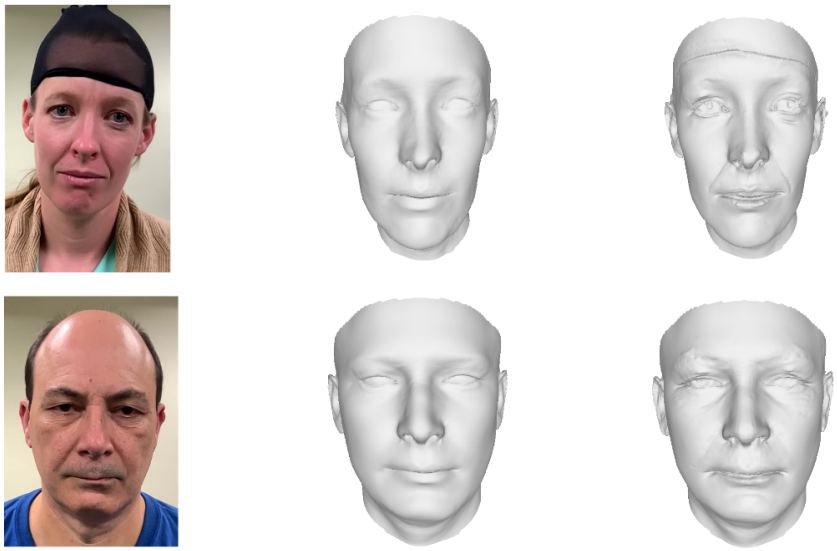

Figure 6.1: (Centre) Ours. (Right) Ours with modified mesoscopic augmentations.

the texture as :

$$\tau_{lp} = (M - dt.C)^{-1}.M\tau_v \tag{6.1}$$

Where M and C are the mesh mass matrix and cotangent Laplacian matrix. dt is set to a small constant value of 0.001. This has the effect of removing noisy effects like sensor noise from τ_{lp}

We then calculate a band-pass version of the texture, where we wish to capture the medium-high frequency details in the texture:

$$\mu_v = \tau_{lp} - (M - \Delta t.C)^{-1}.M\tau_v \tag{6.2}$$

Where $\Delta t = 0.01$.
Now, μ_v ,the band-pass version of the texture map is used to calculate the per vertex deformation, such that vertices which deviate more from the mean of the band pass texture are deformed more :

$$\vec{\delta_\mu}(v) = ||\mu_v(v) - \mu_m||.\vec{n}(v) \tag{6.3}$$

Where μ_m is the mean of μ_v, and $\vec{n}(v)$ is the normal vector of vertex v.

Although this per vertex deformation can be applied to the mesh directly, for "smoother" results, it can be plugged into the mesh fitting optimization as described in Sec 3.3.4 , so that the mesh fitting energy becomes:

$$\arg\min_V E_{pcl} + \alpha E_{lms} + \beta E_{edges} + \gamma E_{reg} + \lambda E_{meso}$$

Where E_{meso} is the distance between the current vertex location and the desired location calculated using $\delta_\mu(v)$

This simplified version of the original approach suggested by [7] allows us to capture details like eye-lids, lip corners and nasolabial-folds in the mesh.

Chapter 7

Experimental Results

For evaluating our approach, we collect a dataset of videos using an iPhone X, with a stationary subject and the camera moving from one profile to the other. The videos are shot at the 120fps setting and are typically 15-20 seconds long, depending on whether they are shot by the subject themselves or by an assistant. The background conditions are unconstrained, though we do ensure a mostly static scene to get accurate camera pose estimates from our Photometric Bundle Adjustment step.

7.1 Quantitative Evaluation

For 10 subjects among the videos we collected, we obtained high accuracy 3D face scans using an Arctic Eva structured light hand-held scanner. The scans were obtained immediately after the video was recorded with the subjects still in the same poses, to ensure no discrepancy in face geometry between the video and the scan. We use the videos of these subjects to reconstruct face meshes using the methods listed in Table 7.1. For methods that work on a single image, we use a close to frontal keyframe as input. For the edge-fitting based single view method of Bas [5], we select a frame at roughly 15 degrees to the front for the input, since that was reported to work best in their paper. For the multi-view methods, we either use the keyframes generated by our method or the whole video, depending on what the method uses as input. For all methods except PCSfM, the authors make their implementations public and we use those for evaluation. For PCSfM, we use our own implementation.

A challenge in fair quantitative evaluation arises from the fact that different methods reconstruct different amounts of the face area as defined by the Basel mesh, such as frontal only in pix2vertex [51] , front and side without ears in PRN [18], full Basel mesh for PCSfM [26] and arbitrary in SfM (using COLMAP [50]). To address this, we first register a standard Basel mesh to the ground truth 3D scans using Non-rigid ICP [3,10]. We then borrow a strategy from MVS benchmarks [31,39,55] to evaluate the reconstructions using **Accuracy** - the distance from the reconstruction's vertices to the ground truth, and **Completion** - the distance from the ground truth's vertices to the reconstruction. Thus, if a method reconstructs only the frontal face, it might do well in Accuracy and be penalized in Completion. We report the mean and median of these distances, averaged over the 10 subjects, in Table 7.1.

Method	Views	Accuracy (mm) ↓			Completion(mm) ↓		
		Mean	Std. Dev.	Median	Mean	Std. Dev.	Median
Mean Basel [9]	-	3.09	1.24	2.62	3.02	1.25	2.76
Landmark fitting [28]	Single	2.53	0.62	1.88	8.01	2.13	3.62
pix2vertex [51]	Single	3.49	0.76	2.76	25.33	4.62	16.34
PRN [18]	Single	2.63	0.84	2.30	6.27	2.17	3.24
Edge-fitting [5]	Single	3.06	1.28	2.63	3.02	1.25	2.75
M.view lm fit [27,28]	Multi	2.23	0.41	1.69	7.87	1.98	3.59
Roth [45]	Multi	3.31	1.03	2.65	7.65	1.86	3.67
SfM [50]	Multi	5.42	2.55	3.61	4.72	2.41	3.60
PCSfM* [26]	Multi	1.87	0.40	1.66	2.38	0.85	2.04
Ours w/o Edges	Multi	1.38	0.24	0.98	1.30	0.29	0.97
Ours w/o ear lms	Multi	1.33	0.27	0.96	1.41	0.36	1.07
Ours	Multi	**1.24**	0.26	**0.95**	**1.29**	0.29	**0.95**

Table 7.1: Quantitative results against ground truth scans. We evaluate the state of the art single and multi-view reconstruction methods. As is common in MVS benchmarks, we evaluate the reconstructions in terms of average distance from reconstruction to ground truth (accuracy) and distance from ground truth to reconstruction (completion). All numbers in mm; lower is better. * denotes that the method needs camera intrinsics to be known in advance.

We compare our methods against several recent single and multi-view reconstruction methods. As can be observed, single view methods typically have very poor performance in terms of accuracy and completion. As also noted in [26], certain methods that just reconstruct smooth meshes tend to have low numeric errors, even if the reconstruction lacks details important for making a person recognizable.

Our method clearly outperforms single and multi-view baselines, both in terms of accuracy and completion. We note that our median accuracy is around 0.95 mm, showing that for majority of the mesh we achieve sub-millimeter accuracy.

7.1.1 Ablation

We generate reconstructions without Edge constraints and without the ear landmarks respectively. Edges improve the accuracy of the reconstructions by improving the fit of areas like the jaw and nose, whereas the ear landmarks improve the information captured in the ears as well as overall mesh scale and width. Thus dropping either leads to a drop in accuracy. A significant drop in completion is also observed when removing the ear landmarking, because the reconstructed mesh moves away from the ears of the ground truth mesh.

7.2 Expressions

Our method captures the geometry of the face in a completely data-driven manner, and hence it also generalizes naturally to deformations caused by expressions (Fig 7.4). Since, is difficult to hold the same expression through a video sequence and then also obtain a corresponding ground truth 3D scan, we skip quantitative evaluation of this. We also note

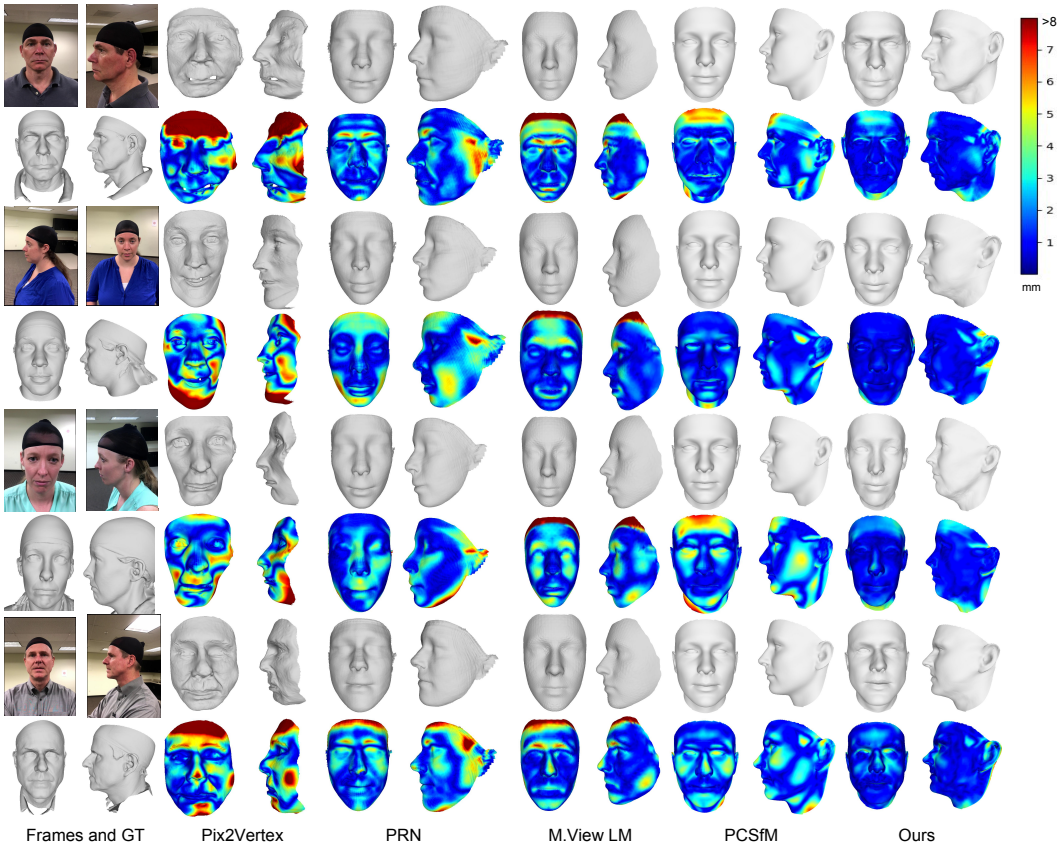

Figure 7.1: Qualitative comparison against reconstructions of various single and multi-view methods. Let to Right: sample frame and ground truth 3D, Pix2vertex [51], PRN [18], multi-view landmark fitting (4dface [28]), PCSfM [26], Ours. For each method the upper row shows the reconstructed mesh, front and profile, and the corresponding heatmap of error (Accuracy) is shown in the lower row

that since there is dense correspondence and consistent topology across meshes, various existing techniques like blendshapes [**?**] can be applied with our reconstructed meshes to generate animated, expressive face models.

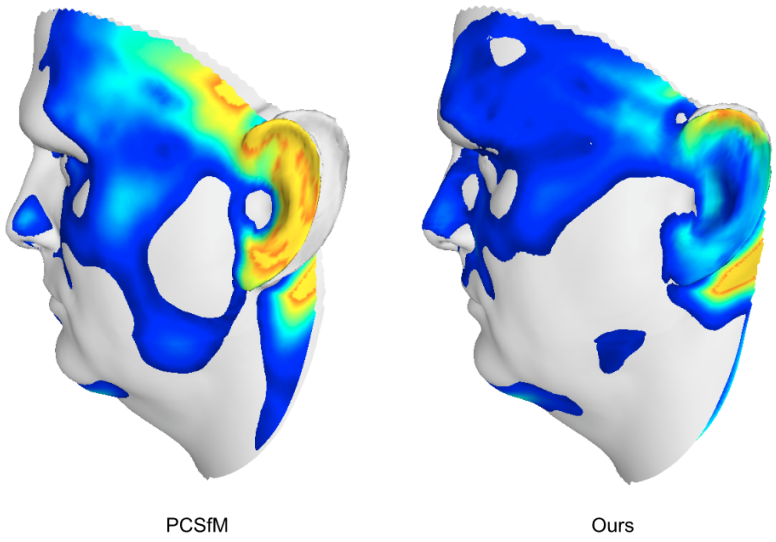

Figure 7.2: Effect of ear landmarking: Ground truth mesh (white) overlapped with error heatmaps of PCSfM(left) and ours(right). Landmarking the ears greatly improves our fitting and reduces the geometric error in our reconstructions

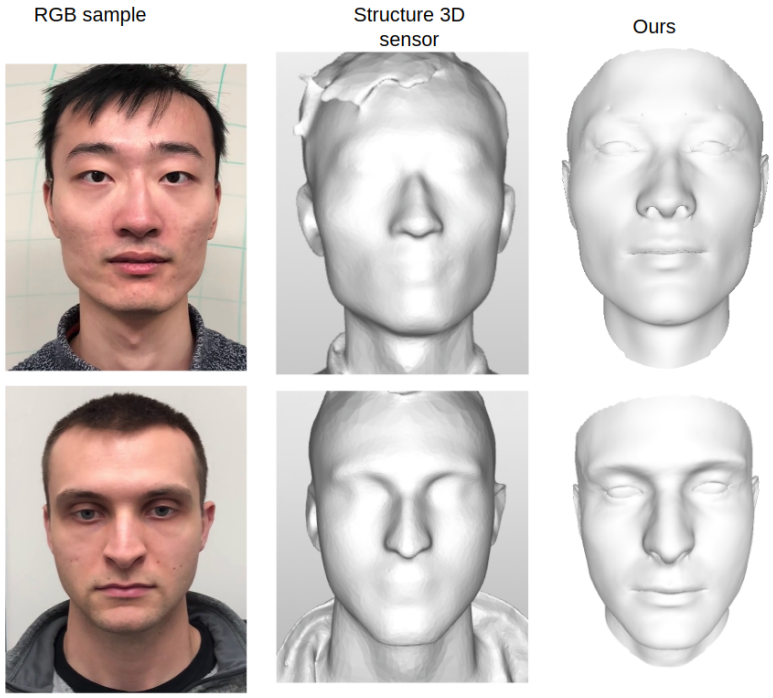

Figure 7.3: (Middle) Output from Structure RGB-D Sensor [41]. Details like the eyes, nose and lips are excessively smoothed out. (Right) Our reconstruction.

Figure 7.4: Our method naturally generalizes to any face geometry, including deformations caused by expressions.

Chapter 8

Dataset

Our results reaffirm that incorporating multi-view consistency in 3D reconstruction greatly improves the quality and reliability of the results. Incorporating geometric structural priors into deep learning based reconstruction has shown to be extremely effective [55], even with moderate amounts of training data. The dearth of multi-view data for faces (see Table 8.1) has prohibited progress in this space. We make our dataset of 100 subjects available, with 2 video sequences recorded per subject under different lighting and background conditions. For each video, we provide a set of 50-80 keyframes we used and our reconstructions (mesh, point clouds and surface normal maps) for reference. For a subset of the data we also provide high accuracy meshes obtained using a structured light scanner. For the rest of the scans, for each subject we validate the meshes to be self-consistent between the two sequences, within a small tolerance.

We found that there is a lack of datasets containing multi-view sequences of faces with consistent geometry in "in-the-wild" settings. To this end, we have constructed our own dataset of 200 sequences of 100 individuals. Each video sequence is shot from an iPhone X, at 1920x1080 resolution and 120fps. Each video sequences is 15-20 seconds long, containing a profile-to-profile sweep of the subject's face. We acquire 2 sequences of the same individual with different background and lighting conditions. For a subset of the dataset, we acquire high accuracy ground truth to serve as validation for testing various methods. For the remaining sequences, we provide our reconstructions as reference, where the meshes are validated to be self-consistent between two sequences of the same subject (Fig 8.1). While a lot of work has been done on learning-based single view face reconstruction, we wish to also encourage better multi-view methods for face reconstructions with this dataset. We hope that this dataset will help further the research and evaluation of unconstrained multi and single view reconstruction algorithms that should be both accurate and consistent. It will especially enable self-supervised methods that enforce consistency across views and between sequences.

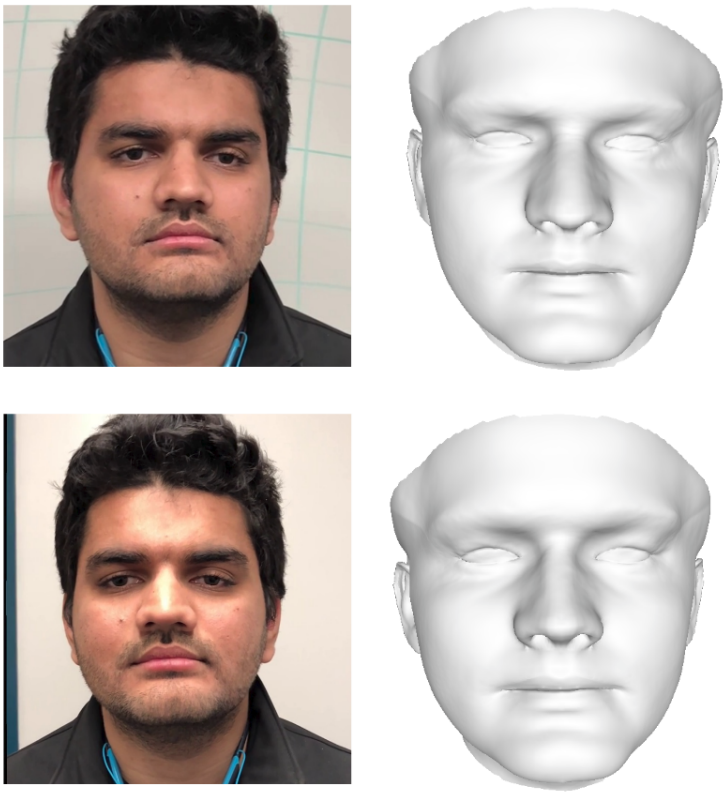

Figure 8.1: For each subject, we record two video sequences under different lighting and background. For the subject's where ground truth is not available, we self-validate the two reconstructed meshes to be consistent, within a small tolerance.

Dataset	# Subjects	# Poses
ND-2006	888	None
BU-3DFE	100	2
Texas 3DFRD	118	None
Bosphorus	105	13
CASIA-3D	123	11
MICC	53	3
UHDB11	23	12

Table 8.1: An overview of available 3D face datasets and the pose variation in RGB images available in them.

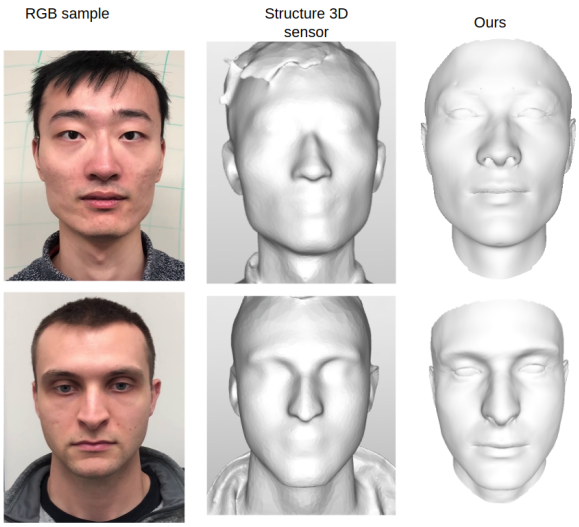

Figure 8.2: (Middle) Output from Structure RGB-D Sensor [41]. Details like the eyes, nose and lips are excessively smoothed out. (Right) Our reconstruction.

Chapter 9

Conclusion

In this work, we present a practical solution for an end user to capture accurate face geometry without using any specialized sensor. To do this we combine techniques of geometric computer vision as well as advances in visual SLAM and modern machine learning methods. We improve over the prior work in several aspects: Our optimization scheme allows integration of landmark, edge and point cloud constraints from multiple frames. Experiments demonstrate better face reconstructions, both quantitatively and qualitatively.

Since we optimize over an unrestricted geometry, our method is slower than many recent learning based methods. Further, our PBA based pose estimation is not robust to dynamic movements in the scene. Deep learning methods have proven to be effective in overcoming these shortcomings but this has not translated to face reconstruction research due to lack of data. We plan to address this in our future work and hope that our proposed pipeline and dataset will further research in this direction.

Bibliography

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

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018.

[3] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[4] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[5] A. Bas, W. A. Smith, T. Bolkart, and S. Wuhrer. Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In *Asian Conference on Computer Vision*, pages 377–391. Springer, 2016.

[6] M. Bazik and D. Crispell. Robust registration and geometry estimation from unstructured facial scans. *arXiv preprint arXiv:1708.05340*, 2017.

[7] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. In *ACM Transactions on Graphics (ToG)*, volume 29, page 40. ACM, 2010.

[8] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 293–300. IEEE, 2004.

[9] V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, 1999.

[10] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254, 2018.

[11] W. Brand. Morphable 3d models from video. In *Conference on Computer Vision and Pattern Recognition*, 2001.

[12] P. Breuer, K.-I. Kim, W. Kienzle, B. Scholkopf, and V. Blanz. Automatic 3d face reconstruction from single images or video. In *International Conference on Automatic Face & Gesture Recognition*, 2008.

[13] X. Cao, Z. Chen, A. Chen, X. Chen, S. Li, and J. Yu. Sparse photometric 3d face reconstruction guided by morphable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4635–4644, 2018.

[14] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1841–1848, 2013.

[15] P. Dou, Y. Wu, S. K. Shah, and I. A. Kakadiaris. Robust 3d face shape reconstruction from single images via two-fold coupled structure learning. In *British Machine Vision Conference*, 2014.

[16] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2018.

[17] J. Engel, T. Schöps, and D. Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[18] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[19] D. Fidaleo and G. Medioni. Model-assisted 3d face reconstruction from video. In *International Workshop on Analysis and Modeling of Faces and Gestures*, 2007.

[20] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes. Computer graphics - principles and practice, 2nd edition. 1990.

[21] G. Fyffe, K. Nagano, L. Huynh, S. Saito, J. Busch, A. Jones, H. Li, and P. Debevec. Multi-view stereo on consistent face topology. In *Computer Graphics Forum*, volume 36, pages 295–309. Wiley Online Library, 2017.

[22] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015.

[23] P. F. Gotardo, T. Simon, Y. Sheikh, and I. Matthews. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *International Conference on Computer Vision*, 2015.

[24] C. Ham, M.-F. Chang, S. Lucey, and S. Singh. Monocular depth from small motion video accelerated. In *2017 International Conference on 3D Vision (3DV)*, pages 575–583. IEEE, 2017.

[25] M. Hernandez, J. Choi, and G. Medioni. Near laser-scan quality 3-d face reconstruction from a low-quality depth stream. *Image and Vision Computing*, 36:61–69, 2015.

[26] M. Hernandez, T. Hassner, J. Choi, and G. Medioni. Accurate 3d face reconstruction via prior constrained structure from motion. *Computers & Graphics*, 66, 2017.

[27] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch. Fitting 3d morphable face models using local features. In *2015 IEEE international conference on image processing (ICIP)*, pages 1195–1199. IEEE, 2015.

[28] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[29] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3d avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)*, 34(4):45, 2015.

[30] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *International Conference on Computer Vision*, 2017.

[31] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014.

[32] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27(10), 2018.

[33] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[34] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[35] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):29, 2013.

[36] I. Kemelmacher-Shlizerman. Internet based morphable model. In *International Conference on Computer Vision*, 2013.

[37] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *Transactions on Pattern Analysis and Machine Intelligence*, 33(2), 2011.

[38] I. Kemelmacher-Shlizerman and S. M. Seitz. Face reconstruction in the wild. In *International Conference on Computer Vision*, 2011.

[39] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):78, 2017.

[40] Y. Lin, G. Medioni, and J. Choi. Accurate 3d face reconstruction from weakly calibrated wide baseline images with profile contours. In *Conference on Computer Vision and Pattern Recognition*, 2010.

[41] Occipital. https://structure.io/structure-sensorStructure Sensor.

[42] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. M. Bruckstein. Rgbd-fusion: Real-time high precision depth recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5407–5416, 2015.

[43] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[44] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *International Conference on 3D Vision (3DV)*, 2016.

[45] J. Roth, Y. Tong, and X. Liu. Unconstrained 3d face reconstruction. In *Conference on Computer Vision and Pattern Recognition*, 2015.

[46] J. Roth, Y. Tong, and X. Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[47] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from rgb input. In *European Conference on Computer Vision*, pages 244–261. Springer, 2016.

[48] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. Photorealistic facial texture inference using deep neural networks. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[49] D. C. Schneider and P. Eisert. Fitting a morphable model to pose and shape of a point cloud.

[50] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[51] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017.

[52] F. Shi, H.-T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33(6), 2014.

[53] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz. Total moving face reconstruction. In *European Conference on Computer Vision*, 2014.

[54] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Conference on Computer Vision and Pattern Recognition*, 2017.

[55] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[56] Y. Zhou and S. Zaferiou. Deformable models of ears in-the-wild for alignment and recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 626–633. IEEE, 2017.

[57] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[58] 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 *Conference on Computer Vision and Pattern Recognition*, 2015.

[59] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, 2018.