3D Face Reconstruction from Monocular Video and its Applications In the Wild

Rohith Krishnan Pillai

CMU-RI-TR-20-58

Decemeber 2020

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
László A. Jeni, *Co-Chair*
Jeffrey F. Cohn, *Co-Chair*
Louis-Philippe Morency
Chaoyang Wang
Carnegie Mellon University

Submitted in partial fulfillment of the requirements for the degree of Masters of Science in Robotics.

Keywords: 3D Face reconstruction, 3D ground truth dataset, Multi-frame fusion, Facial normalizations, Human-centric computer vision

Abstract

3D face reconstruction is a very popular field of computer vision due to its applications in social media, entertainment and health. However, ever since the introduction of 3D morphable models as facial priors, 3D face reconstruction has been dominated by reconstruction from single images due to its ease and proximity to 3D face alignment. Even so, single image reconstruction methods suffer from inconsistent reconstructions across time and view points. Hence a natural extension is to reconstruct 3D face shape from videos. Because of recent methods in single image reconstruction setting the standards for state-of-the-art reconstruction, we introduce a method to fuse single image reconstructions across multiple frames to create a more accurate reconstruction. Furthermore, the lack of structured video datasets that fully captures the face and provide 3D ground truth scans, led us to develop and release the 3DFAW-Video dataset and challenge. We also introduce a symmetric distance metric for benchmarking reconstructions on the 3DFAW-Video dataset that is less affected by reconstruction density. Finally, we discuss the usage of 3D face reconstructions in two different applications to be deployed 'in-the-wild'. In particular, we illustrate applications in mask-sizing from a metric face 3D reconstruction, and present 3D face normalization as a technique to improve vision based non-contact heart rate estimation methods.

Acknowledgments

I want to begin by thanking my advisors Dr. László Jeni and Professor Jeffrey Cohn. Dr.Jeni's guidance and mentor-ship over the years have greatly helped me grow as both an individual and an academic. His patience and gracious encouragement have been a defining characteristic of my time in graduate school. I am extremely grateful to him for the many projects that he trusted in me, and which have impressed on me the necessity and advantages of interdisciplinary work. Although my interactions with Professor Cohn have been fewer, his encouragement and passion for his work have forever inspired me. I also thank him for including me in the Affect Analysis Group and providing me with it's resources without which my research would not have been possible.

I have immense gratitude to all the members (past and current) of my thesis committee who have been extremely flexible and supportive through all the testing times in recent months. I would also like to thank all the great teams in have been a part of during my masters such as the BMGF project team, CFDRC team and other members of the Affect Analysis Group. I would also like to mention all the inspiring professors and students that I've had the pleasure to meet and work with here at CMU, and the RI community that have helped me through even the worst of times.

To my family, I would like to express my gratitude for the constant support, love and guidance over the years. I would especially like to thank my sister who graduated as a doctor during a global pandemic and worked tirelessly to ensure the good health of her patients while also for looking out for me, and helping me out when I needed her.

Last, but not least, I would like to thank the many institutions whose grants have made possible the plethora of projects that I have had the pleasure to work in, such as the US Department of Defense (grant W911SR-17-C0060), National Science Foundation (grants CNS-1629716 and CNS-1629898), and Bill & Melinda Gates Foundation (grant INV-002154).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

3D reconstruction is a very powerful technique for intuitively modeling the physical world for digital manipulation and simulation. With the spread of mobile technology containing multiple cameras and low cost sensors, human sensing has become a popular field. Today face tracking and reconstruction methods are used everywhere from entertaining social media filters to face modeling before plastic surgeries. Hence, 3D face reconstruction is an extremely versatile task that has wide range of applications.

We focus on the task of 3D face reconstruction using 2D monocular camera videos and highlight a couple of real-world applications. Current methods in the field are saturated with single-image based reconstruction techniques. They are a rapidly evolving group of methods due to it's proximity to other related tasks such as 3D face tracking and alignment. Albeit fast, such techniques inherently separate the temporal and multi-view consistency, due to reliance on a image-level reconstruction. Furthermore, the lack of 3D ground truth datasets is a big limitation in evaluation and pushing the state of the art. Hence, we realize the need for a structured monocular multi-frame dataset that does not currently exist. And lastly, almost all existing methods for 3D face reconstruction confine themselves to the task and don't explore other auxiliary tasks or challenges to application in the real-world.

This hence brings to light these question worth investigating:

1. How do we integrate the multi-view consistency of the subject from multiple single image reconstructions? What other details of the face shape can be learned additionally to the single image reconstructions by introducing a temporal component (i.e. using 2D monocular videos)?
2. How do we benchmark and evaluate the quality of our reconstructions?
3. What other applications and auxiliary tasks does a multi-view consistent 3D reconstruction allow, and how do these methods perform on real-world applications?

1.2 Challenges

There quite a few factors that makes 3D face reconstruction a challenging task. The majors ones are as discussed below.

1.2.1 Appearance Diversity

One of the major challenges in 3D face reconstruction is the enormous diversity in appearance of the population. Most new methods based on machine learning use datasets that are not entirely representative of the target population due to the bias in the existing datasets. Furthermore, handling occlusions, such as a facial hair, a hand, or other objects in between the subject and the camera can also make the task of 3D face reconstruction extremely challenging. It is also to be noted that various methods which rely on shading information for reconstruction can also be hampered by dynamic lighting or complex lighting environments. However, when trying to bridge the gap between single and multi image 3D face reconstruction, a major issue is the non-rigidity of the subject's face. When dealing with multiple frames of the subject in the wild, it is unlikely that the subject holds the same facial expression over time. While, single-image based approaches entirely bypass this issue, the multi-frame methods have to either actively control it, or ensure that the data collected is devoid of such variations in expressions as a hard constraint.

1.2.2 Use of Face Models

While the classical computer vision pipeline of 3D reconstruction uses multiple views for reconstruction using techniques such as Structure from Motion(SfM), or Photometric Stereo (PS), which depends instead on lighting cues on the image for reconstruction, a whole subgroup of methods reply on 3D facial priors in the form of a 3D morphable model such as the Basel Face model[25] to be fit to the face. While this allows the method to do a single image reconstruction, it also introduces limits on the total class of faces that can be represented from the model. Hence, such methods are hit their limits of accuracy when dealing with large non-symmetries on the face which are not expression based.

1.2.3 Lack of Ground Truth Datasets

Most recent methods rely on deep learning based approaches to generate the 3D reconstruction from images. While synthetic or model based data does not alter the nature of classical 3D reconstruction methods, with deep learning approaches these can become severe limitations. Many of the widely used datasets for 3D face reconstruction such as 300W-3D [54], only provide the 3D morphable model parameters for reconstruction, and not the ground truth 3D meshes. Such datasets with ground truth 3D meshes are far and few between, and tend to use images as opposed to video data. This is also an issue during evaluation, where all the methods are not compared to the ground truth in the wild data, but rather on model fitting with minor improvements.

1.3 Contributions

In this work we focus on the task of 3D face reconstruction from monocular video by leveraging the rapid improvements in single-image reconstruction based methods. More specifically, we explore how to extend single-image reconstruction to a multi-frame based approach. However, the lack of ground truth 3D datasets for such a task also needed to be filled, if we were to accurately evaluate the quality of the 3D reconstructions.

Hence the contributions of this work are the following:

1. Introduction of a novel method for extending single-image reconstructions by multi-frame reconstruction fusion.
2. Creation and release of the 3DFAW-Video dataset containing ground truth 3D meshes, 2D monocular video and associated evaluation metric.
3. Exploration of in-the-wild applications of 3D face reconstruction in 3D mask-sizing and heart rate estimation using 2D monocular mobile videos.

1.4 Outline

The thesis document is split into chapters detailing the background, methods, and results etc. The related works chapter provides a good starting place for those not familiar with the 3D face reconstruction field. The first half of the chapter explores various classical computer vision techniques in 3D reconstruction in general and about facial prior models such as 3D morphable models used by many methods. We also take a look at the ZFace[27] 3D face tracking software, as well as inspect a few deep learning based methods.

In the Datasets chapter, we detail the various datasets used by our methods in training and evaluation. In particular, we describe the the newly released 3D ground truth 3DFAW-video dataset. We also touch on the related challenge held in conjunction with ICCV 2019.

In the Methods chapter, we discuss the position map data representation and the cosine fusion method and illustrate how it works.

The Results chapter, contains the evaluation of the method that we introduce. It also contains the evaluation protocol and metric, and the results for 3DFAW-Video challenge that was held with comparisons between the top few methods and ours. We also discuss the improvement of the reconstruction of our method over the base PRNet model, and highlight the experiment on the effect of number of images used for reconstruction.

The Applications chapter, illustrates the power of 3D face reconstruction methods in the context of a 3D face mask sizing task and for heart rate estimation from 2D monocular facial videos.

Chapter 2

Related Work

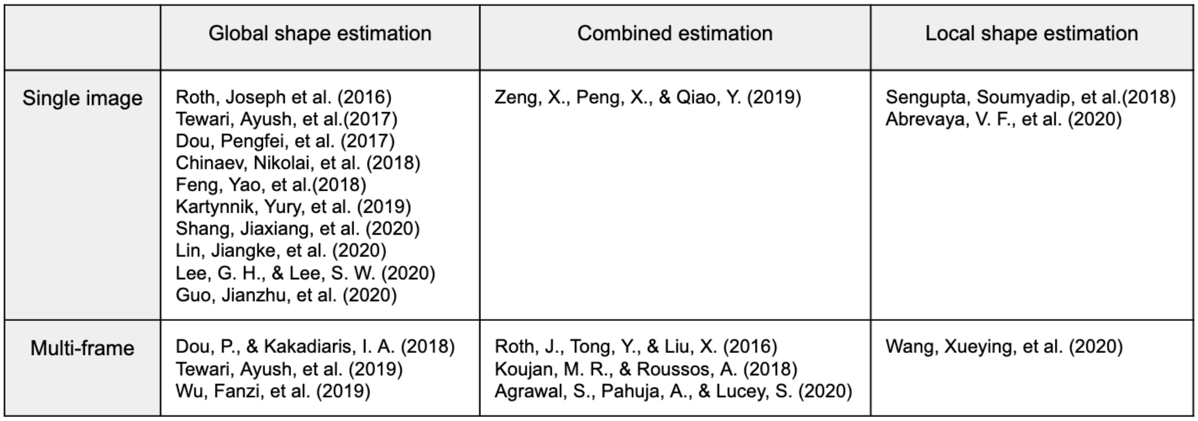
3D face reconstruction is a rapidly evolving field, with increasing level of interest from the entertainment industry for facial motion capture to health care industry for planing facial cosmetic surgical procedures. Due to the enormity of the number of methods in the field, there exists numerous ways of classifying the space. The major distinctions between key methods are discussed in this section along with a few examples of each.

One of the broadest way to distinguish methods are by class of techniques employed by them. In recent years , with the adoption of deep learning as a mainstay for computer vision tasks, there is a clear distinction to be made between the classical computer vision techniques and learning based methods of new. While classical methods often revolved around geometric and physics based vision, learning based techniques use large amounts of data to find the intrinsic patterns that lead to the derivation of the structure of the object. Hence, the datasets used to train a learning based method is as important as it's architecture, and the constraints it uses to learn network parameters.

Another classification of existing methods can be made on the precision(level of details) of results they produce. There are broadly three different groups of methods for 3D face reconstruction, depending on the amount of detail each of the methods are able to produce. They can be:

1. Global(Coarse) shape estimation methods
2. Local(Fine) detail estimation methods
3. Joint(Coarse & Fine) shape estimation methods

Global shape estimation methods focus on primarily reconstructing the face to a very low level of detail, while establishing the primary overall shape of the face. Many of these methods rely on facial priors to ensure that the reconstruction has a high facial coverage, and reduces the errors in regions around the major facial landmarks such as the eyes, nose, mouth regions. Such methods can densify their sparse reconstructions using Catmull-Clark subdivisions or similar techniques. Local detail estimation methods in contrast work on extracting higher levels details



	Global shape estimation	Combined estimation	Local shape estimation
Single image	Roth, Joseph et al. (2016) Tewari, Ayush, et al.(2017) Dou, Pengfei, et al. (2017) Chinaev, Nikolai, et al. (2018) Feng, Yao, et al.(2018) Kartynnik, Yury, et al. (2019) Shang, Jiaxiang, et al. (2020) Lin, Jiangke, et al. (2020) Lee, G. H., & Lee, S. W. (2020) Guo, Jianzhu, et al. (2020)	Zeng, X., Peng, X., & Qiao, Y. (2019)	Sengupta, Soumyadip, et al.(2018) Abrevaya, V. F., et al. (2020)
Multi-frame	Dou, P., & Kakadiaris, I. A. (2018) Tewari, Ayush, et al. (2019) Wu, Fanzi, et al. (2019)	Roth, J., Tong, Y., & Liu, X. (2016) Koujan, M. R., & Roussos, A. (2018) Agrawal, S., Pahuja, A., & Lucey, S. (2020)	Wang, Xueying, et al. (2020)

Table 2.1: Categorical representation of recent methods in 3D face reconstruction described by the precision of the reconstructions and the modality of the reconstruction (single-image vs multi-frame)

from the face, by photometric stereo, normal map estimation or related methods. These methods however, tend to suffer from self-occlusions due to facial head-pose changes as they usually rely on single image detail restoration. But there are a few methods that try to bridge the gap, by jointly estimating both global (coarse) and local (fine) shape from images. Such methods deal with optimizing their methods along the level of detail versus reconstruction noise trade-off.

Furthermore, from the perspective of using monocular videos, the existing methods for 3D face reconstruction can be segregated into two modalities of study. The first is single-image based reconstruction, which outputs a one is to one map between the frame and the reconstruction. Such methods are the most common in the field since the introduction of the 3D morphable models [5].

The second modality is multi-frame based reconstruction. This is essentially the task of reconstructing a single 3D face mesh from multiple frames spread over the time dimension. It is important to note the difference between multi-frame based reconstruction and multi-view based reconstruction. The major difference we will emphasize is that multi-frame based reconstruction does not make any assumptions that the frames were collected at the same instance in time, while multi-view based reconstruction does. In other words, multi-view would require 2 or more cameras with synchronized data capture, while multi-frame refers to single monocular camera captured data. This means that multi-frame reconstruction is a more generalized reconstruction modality and closer to reconstruction from image sets.

Table 2.1, shows the spread of the various methods in the field along on 2 dimensions: the precision(level of details) of the various methods and the modality of the reconstruction. This representation shows a good example of the number of methods preferring to rely on single image based reconstruction as compared to multi-frame approaches.

2.1 3D Morphable Models (3DMMs)

One of the corner stones of face modeling since its introduction in [5], 3D morphable models (3DMMs) have dominated the field as the principle 3D facial prior model in 3D face reconstruction and related tasks. The 3D morphable models are created by conducting a Principle Component Analysis (PCA) on a large dataset of 3D face scans, to get the principle component basis for the face shape in the dataset. In subsequent models such as [25], the same idea was expanded to skin texture as well. Hence, we can express a face shape(S) and face texture (T), as given in equation 2.1, with the mean shape ($\bar{S}$) and texture ($\bar{T}$) adjusted along the shape/texture principle components X_s and X_t, by the shape/texture parameters α_s and β_t. It is important to note that while the equation 2.1 in the backwards direction is the generative task of creating a face shape S, given it's α_s, β_t parameters, the equation 2.1 in the forward direction allows for fitting the 3DMM to find the α_s, β_t parameters that can as closely resemble the face shape S as possible. Some examples of the variations in shape that can be modeled by the popular 3DMM, Basel Face Model(BFM'09) [25], are shown in Figure 2.1. More recent 3DMMs such as BFM'19[19], FaceWarehouse[6], and FLAME[32] also account for facial expression changes. For those inclined, [14] provides a detailed background and history of 3DMMs and their use in face modeling.

$$\begin{aligned} S &= \bar{S} + X_s\alpha_s \\ T &= \bar{T} + X_t\beta_t \end{aligned} \tag{2.1}$$

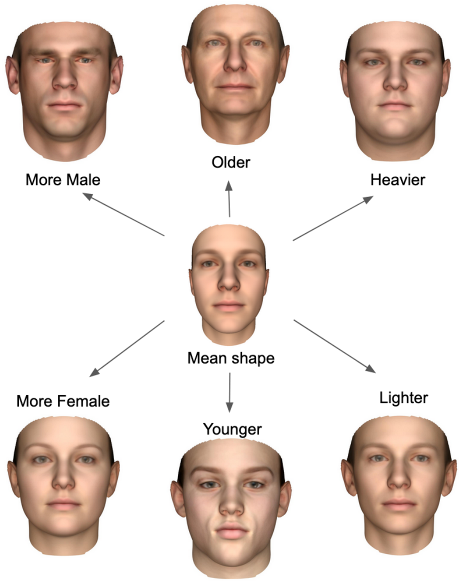


Figure 2.1: The Basel Face model (BFM'09) with the mean face in the center, along with several variations in the shape, generated by varying the shape parameter along specific principle components.

2.2 Classical Methods

In this section we discuss a few classical methods used for general 3D face reconstruction, and related methods that use some variation of such techniques.

2.2.1 Structure from Motion (SfM)

One of the staple methods for 3D reconstruction from image sets have always been Structure from Motion (SfM). Structure from Motion constructs both the 3D points clouds of the objects visible in the image set as well as simultaneously calculating their camera poses. This allows for reconstruction from un-calibrated cameras and image sets. SfM has been used as a major technique in methods such as [18] [20] [34]. However, SfM reconstructions need to be further processed to remove spurious points by smoothing or model fitting, leading to a coarse reconstruction. For instance, [2] uses a variant of SfM such as Bundle Adjustment (BA) along with multi-view stereo and 3D morphable models to reconstruct the global facial structure.

2.2.2 Shape from Shading (SfS)

Another related classical method for 3D shape extraction from images is the broad task of Shape from Shading (SfS). Shape from shading aims to conduct an inverse rendering of an image, i.e. separate an image into its reflectance(albedo), and corresponding shading components containing both the object geometry and the illumination. Although an ill-posed problem, with the right illumination assumptions, SfS can reconstruct high level of surface details. Photometric Stereo (PS), is one of the many techniques used for SfS. For example, [7] uses a Color PS pipeline with controlled near point lights setup along with a 3D morphable model for self-calibration, in order to retrieve the geometry of the face. In contrast, [33] uses a photo set of the subject, binned by head pose, and aligned using collective flow, before using PS to reconstruct and merge the reconstructions together without 3D morphable models.

2.3 Learning based methods

2.3.1 Single-Image Reconstruction

A related and important task for most 3D reconstruction pipeline is that of face detection and tracking. In fact, some methods like the ZFace 3D face tracker [27] provides a dense 3D face alignment on 2D monocular videos, which can be then further densified using a technique such as Catmull-Clark subdivision to produce a 3D face reconstruction. The method relies on extracting local Histogram of Oriented Gradient (HOG) features for a set of initial markers and refining the 2D marker alignment using a trained cascade regressor. These 2D markers are then used to iteratively fit a custom denser 3D model until convergence, and corrected for excess jitter using the previous tracked frame. The current 2D/3D alignment is then passed on as initial estimates for the next frame. However, the ZFace tracker limits the 3D alignment to a select internal facial region, which prevents it's use as a full 3D face reconstruction method. Hence, ZFace can be seen as a tracker that does single-image, global shape estimation.

Single image based reconstruction methods have become extremely popular in recent years, with the advent of larger datasets, ample computing power and advances in deep learning. Some methods like [13] use a CNN based architecture with fusion of intermediate layers to regress the 3DMM parameters of the face from a single image. Other architectures such as the more compact MobileFace [8] also used similar 3DMM fitting for real-time 3D reconstruction. MoFA [45], uses a convolutional autoencoder to create a semantic code vector that parameterizes facial expression, shape, skin reflectance, camera pose and scene illumination which can then be decoded to retrieve the 3D facial geometry.

3DDFA [54] used a cascaded CNN model to regress the 3DMM parameters as well as the camera pose iteratively till convergence. However, the model does not deal with extreme poses well, which was improved in the 3DDFAv2 [21] method by meta joint optimization across the regressed 3DMM parameters and the landmark estimation. The 3DDFA method also introduced the Normalized Coordinate Code (NCC) representation for the 3D face geometry, which was modified to include the camera pose in PRNet [16] to form the position map representation. Much like 3DDFA, PRNet estimates the 3D reconstruction as well the dense landmarks together, but unlike other methods, it does not explicitly regress a 3DMM. PRNet uses image to image translation, i.e. translates the input image to the position map image representation of geometry of the face. However, the data that was used in training the PRNet model is based on 3DMMs and hence the reconstructions have a bias towards the 3DMM space.

Some recent methods such as [42], also use image to image translation, but extend it to learn an implicit multi-view consistency while using 3DMMs for the facial geometry. This is done by considering 3 different views to either side of the face to constraint the the center image based 3D reconstruction using multiple multi-view consistency losses. Due to the heavy use of face priors such as 3DMMs, these single image methods tend to produce only the global shape estimates.

Other methods focus on the task of extracting fine details of the face from the images using normal map representations for geometry. While normal maps allow for great detail and provides a known correspondence to the image, it does not change the 3D shape of the face. The normal maps instead is only able to affect the rendered output of an image and hence is a passive representation of the geometry. Nevertheless, methods such as [1], is able to recover extremely fine details of facial images using this representation. The method uses a double autoencoder architecture with a shared latent representation between them, along with a deactivable skip connection that allows for the model to be trained on various paired and unpaired, image/normal map datasets. SfSnet [41], however uses various residual block sub-networks to separate the input images into their albedo, normal maps, and illumination components. This provides additional inverse rendering properties that allow for secondary tasks such as image relighting.

Finally, methods such as DF^{2}Net[51] implements a joint estimation of shape at both the global and local scale. The method splits the shape estimation based on the levels of detail, with the D-net providing a dense coarse depth estimate of the face using 3DMM fitting, which is refined by the hypercolumn F-net, before the final Fr-net adds details that cannot be modeled by 3DMM priors. This allows the network to also be stopped at any detail levels as needed depending on the use case.

2.3.2 Multi-Frame Reconstruction

While single-image reconstruction methods lack temporal and multi-view consistency, the multi-frame reconstruction techniques take these in to consideration to produce a single 3D face reconstruction. For instance, [37] is an early method that first localizes the facial landmarks and ignores the frames with self-occluded landmarks or large deviation from the neutral face, along with the head pose to select key frames from a video. Then using the [15] technique, the landmarks are used as a sparse feature set to conduct n-view 3DMM fitting.

Since one of the major issues with multi-frame reconstruction approaches are facial expression changes between frames, [12] proposed separately regressing the 3DMM facial shape, and expression for each frame. The DRFAR method approaches this challenge using a RNN to estimate the face shape common to all the frames in the video, while a CNN regresses just the facial expression per frame. However, the method suffers from application in the wild due to being trained entirely on a synthetic dataset. The FML model [46], similarly separates the shape and expression by ensuring orthogonality between the regressed identity basis and blendshape expression basis. The parameters for both these basis are learned using a shared identity network between frames, and the geometric losses calculated using a differential render. The method also constraints the reconstructions across frames using multi-frame consistency losses based on non-rigid SfM-like photometric consistency, 2D sparse landmark consistency and a skin appearance sparsity metric.

Other multi-frame reconstruction methods use differential renders for an analysis by synthesis approach to constraint the reconstructions across views, but also assume little to no expression changes across frames. For example, MVF-Net [49], selects 3 different frames of the subject, one in the frontal head pose, and the other 2 on either sides. The VGG features of these 3 images are concatenated to regress the 3DMM parameters for the face, as well as using another fully connected layer to estimate the pose of the cameras for each frame separately. These poses are then used in conjunction with a differentiable renderer to provide the photo and the alignment losses.

Few existing methods tackle local shape estimation in the multi-frame modality, such as FacePSNet [48]. While the method uses multiple images, it relies on images captured under near-field lights rather than at different views for the 3D reconstruction. Using a single image, a CNN is used to regress the coarse 3DMM shape and camera pose parameters, and an initial face normal map is generated. This is then concatenated with features from N input images to a convolutional decoder that outputs the normal map with the fine details reconstructed, which the initial 3DMM network cannot model. The method is also trained only on synthetic images. However, The biggest drawback of such a method is the need for near point light conditions lending it to be unusable with in-the-wild data.

2.4 Key Trends and Takeaways

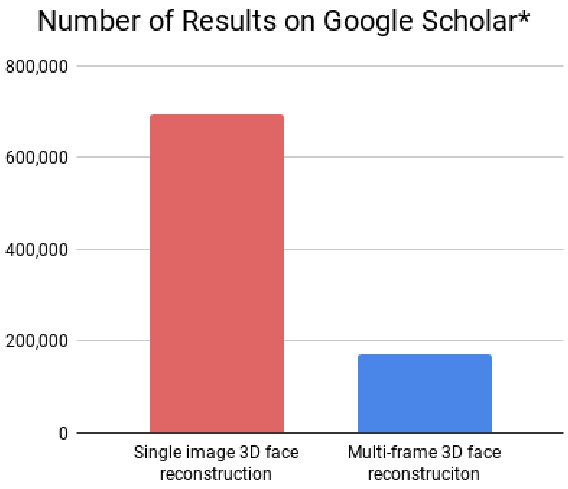


Figure 2.2: (Top)The number of methods in each modality of 3D face reconstruction from Google scholar search results.

The field of 3D face reconstruction is rapidly growing with the increase in facial modeling for applications. 3DMM based models have become a mainstay in the field, as it provides great flexibility in the number of images needed for facial alignment, and shape estimation. However, while 3DMMs provide a good rough global shape, their biggest drawback is the lack of detail that they provide, and hence most recent methods have joint estimation or subsequent methods to retrieve local shape details from the input images.

Furthermore, the field is also dominated by single-image reconstruction methods. The top Figure 2.3[1] illustrates the saturation of single image based methods in 3D face reconstruction, from the number of search results on each modality on Google Scholar. The number of hits for single image approaches are approximately 3 times the number of multi-frame 3D face reconstruction approaches. This shows that the large majority of research in the field has moved towards favoring the single-image modality over the multi-frame modality. This is counter-intuitive to those in other 3D reconstruction sub-fields where more frame/images of an object are usually used for reconstruction, and single image approaches are the minority. This disparity in modalities can be attributed to the effect of having strong 3D face priors such as 3DMMs, which is unavailable in general 3D reconstruction.

[1] * Results from search on Google Scholar with time frame set to 'any time' and including citations and patents

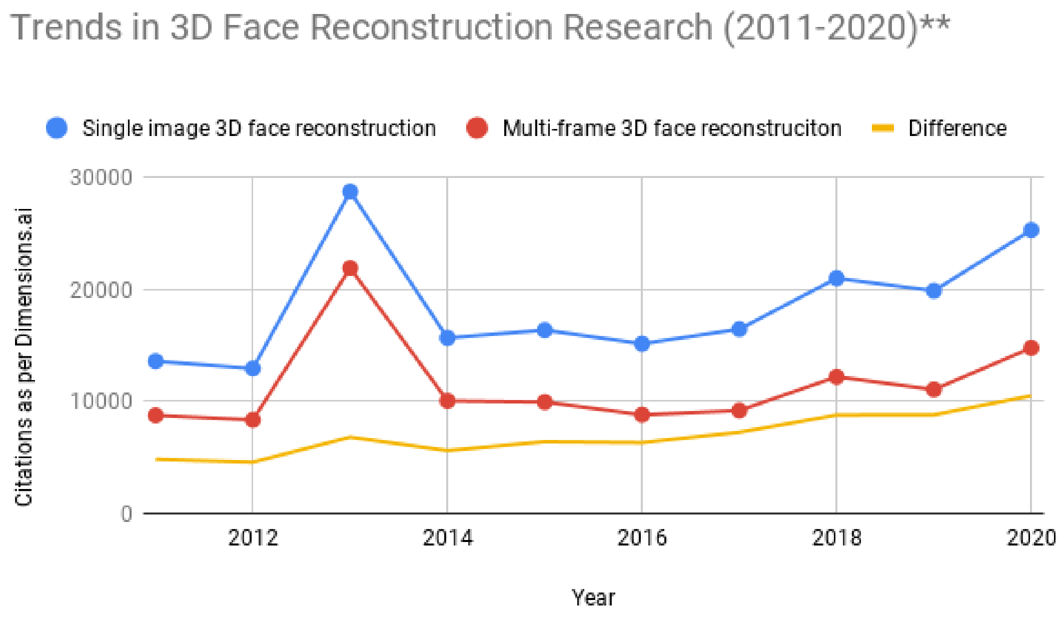


Figure 2.3: The trend in the number of citations from *dimensions.ai* in the last decade between the 2 modalities, and their widening gap in their popularity's.

Similar results can be seen in the number of citations on 3D face reconstructions over the last decade, in Figure 2.3[2], where the number of citations of single-image reconstruction methods consistently leads multi-frame approaches. This could possibly be due to such methods' abilities to also solve the closely related task of 3D face alignment. With the overwhelming number of methods favoring such a modality, and from Figure 2.3 chart the difference in popularity between the 2 modalities being lead significantly by single image approaches, it can be assumed that the field will continue to move in such a direction. Therefore there exists need for a meta-method that is able to fuse multiple single image reconstructions across multi-frame to remove the temporal or multi-view inconsistencies present inherently in these models. Hence, we set to explore how such a task of fusing multiple single image reconstructions to improve the overall quality of reconstructions can be executed.

[2] ** Results from search on *Dimensions.ai* based on 'full data', keywords search in documents

Chapter 3

3DFAW-Video Dataset and Challenge

3.1 Existing Datasets

While 3D face reconstruction methods have been around for a long time, the number of 3D face reconstruction datasets that provide 3D ground truth is only a handful. Most existing datasets that are used for training 3D face reconstruction models are based on single images and provide algorithmic ground truth data. Since the task of creating a fixed 3D face reconstruction from single images provide the 3D alignments for free, many 3D alignment datasets have been modified for the 3D reconstruction task. These datasets provide the ground truth reconstruction as a 3DMM fit to the image as opposed to ground truth 3D scans of subjects. For example, 300W-3D[55] combines several single-image datasets to create a large scale dataset with approximately 60,000 images. Similarly, AFLW2000-3D [54], AFLW-LFPA [28], and 3D Menpo [50] datasets are also large single image datasets frequently used for training 3D reconstruction models but contain only 3DMM fitted ground truth. However, since these reconstructions are generated using 3DMM fitting, the methods training on such dataset will be severely limited to accuracy of the 'ground-truth' mesh fitting methods used.

Although the existing datasets are dominated by large single-image datasets with 3DMM fitting, 3D ground truth datasets do have their niche in the 3D face reconstruction field. Most 3DMM fitted datasets are usually adapted from datasets annotated for 3D face alignment. Some of these also collect the images from the internet, and hence are able to build large datasets although the images themselves are uncontrolled camera poses and have wide variations in lighting. In comparison, 3D ground truth data requires a 3D imaging system and volunteers, which make data collection difficult. Hence, most 3D ground truth datasets also are much smaller in comparison in terms of the number of subjects. This also is a boon as it allows for flexibility in the procedure to capture specifically structured data, such as videos with the full range of head orientation changes. Due to these constraints and the use of real ground truth 3D data, such datasets are usually used for bench-marking, model evaluation or in challenges.

Some of the commonly used 3D ground truth datasets can be seen in Table 3.1. Here, we use the terms controlled camera motion to mean a fixed camera pose across images/video with

Dataset	Modality	Camera-Motion/Pose	Range	Dataset Size (subjects)	Ground Truth
NoW dataset[40]	Single-images	Uncontrolled	Partial range	100	3D-scans (3dMD)
Stirling ESRC [17]	Single-images	Controlled	Full range	100+	3D-scans (Di3D)
FaceWarehouse [6]	Single-images	Controlled	Partial range	150	RGB-D Kinect + 3DMM fitting
ICT-3DRFE [44]	Single-image	Controlled	Frontal Only	23	3D-scans (Light-Stage 5)
BP4D+ [53]	Videos	Controlled	Partial range	140	3D-scans (Di3D)
MICC Florence [3]	Videos	Controlled	Full range	53	3D-scans (3dMD)
3DFAW-Video (Ours) [31]	**Single-images, Videos**	**Controlled, Uncontrolled**	**Full range**	**66**	**3D-scans (Di3D)**

Table 3.1: The existing 3D face ground truth datasets compared on various dimensions.

respect to the background, while unconstrained camera motion refers to the camera pose changing over images or frames of a video. In addition, the range refers to how much of the face was captured by the images/videos.

The major split between datasets can be made by the modalities they cater to. For example Stirling ERSC [17] and NoW [40] are single image datasets with uncontrolled camera poses, partial facial range capture, and approximately 100 subjects each. FaceWarehouse [6] is the odd dataset that uses 3DMM fitting, but has a larger dataset size at 150 subjects. The ICT-3DRFE [44] provides single images with controlled camera pose and 3D scans captured using a light-stage which allows for relighting. In comparison, datasets such as BP-4D-Spontaneous[52], BP4D+ [53] and MICC Florence [3] provide 2D monocular videos, introducing the temporal dimension. BP-4D, like ICT-3DRFE and FaceWarehouse also provides annotation for facial action units (AUs), landmarks, as well as head pose. However, MICC Florence additionally provides videos in 3 different resolutions and zoom levels, with controlled camera motion. Additionally, MICC Florence dataset's 'cooperative videos' subset provide full range face capture due to the subject changing their head-pose to expose all possible regions on the face.

However, none of the existing 3D face datasets have inherent structure that most 3D reconstruction techniques could easily exploit. Videos provide a temporal dimension, and is a natural and seamless extension for 3D reconstruction from single images. However, where most other 3D ground truth datasets stop at having videos to support the multi-frame reconstruction modality, we introduce a known structure to the data by ensuring that all the videos captured of the subjects are from profile to profile, giving full range face capture. This structure is a guarantee that all regions of the face will be directly observable from the video, hence allowing true comparisons for multi-frame 3D reconstruction methods. Furthermore, 3DFAW-Video dataset, similar to MICC Florence, provides multiple camera resolutions and differing environments, while still increasing the number of subjects. The 3DFAW-Video dataset also unlike other existing 3D ground truth datasets, provides both controlled as well as uncontrolled camera motion videos. This allows our dataset to test the robustness of various methods to camera motion.

3.2 3DFAW-Video Dataset

The 3DFAW-Video Dataset that we created and released has 3 major components:

1. High resolution 2D monocular video from a DI3D Imaging system (Controlled camera motion, Full range capture)
2. Unconstrained 2D monocular video from an iPhone (Uncontrolled camera motion, Full range capture)
3. High resolution 3D ground truth mesh, also capture by a DI3D Imaging system

The driving similarity by design between the two different monocular videos(Hi-Res & iPhone 6 videos) collected is the full range face capture of each subject. These videos are hence able to provide the maximum amount of information on all regions of the face that are required for a full face 3D reconstruction. The 3D scanned ground truth we provide were created by combining multiple frame-level 3D scans, manually. The Figure 3.1 shows the 3 different components that make up the 3DFAW-Video dataset. The procedure for data collection, and the dataset construction are elaborated in the following subsections.

3.2.1 Data Acquisition

In order to collect the High resolution videos, our collaborators at Binghampton University, used a custom built setup. The DI3D 3D imaging system was used for capturing a high resolution 2D monocular video using a high resolution RGB camera. The camera rig was setup to have 2 monocular cameras, as a calibrated stereo pair, vertically above and below the RGB camera. This setup allowed for dense passive stereo photogrammetry to be used in order to retrieve a 3D scan of the subject's face per frame. The RGB camera allows the 3D scans to be textured. In addition the illumination was carefully controlled with 2 symmetrically arranged lights lighting the subject from either oblique angles, and a solid dark background and matching colored clothing worn by the subject. This allowed for easy separation of the subject's face and neck regions from the background and clothing by filtering by color.

The high resolution videos and the 3D scans have controlled camera motion but have full range capture. In other words, the camera is stationary but the subject is asked to rotate their head from one profile to the other. The 3D scans that are captured by the DI3D system at each frame is made up of 30K-50K vertices with a precision close to 0.2mm RMS. The frequency of both the RGB monocular video and the 3D meshes were a constant 25Hz. The high resolution RGB videos have a resolution of 1040x1392 pixels, and most typically only about $5 \sim 10$s long.

The second class of videos in the 3DFAW-Video dataset are captured entirely on an iPhone 6. Unlike the high resolution video, we also allow for the video background to be in a less controlled indoor environment, with ambient lighting. Although not fully 'in-the-wild' due to the uniformity of the all the videos with respect to their environments, we try emulate the camera motion factor to make it more in line with typical smartphone video capture. The videos collected have uncontrolled camera motion as the camera moves around the stationary subject from profile to profile. It is important to note that there were not external gimbals used to stabilize the camera motion and hence the videos exhibit jitter, out of frame motion, and speed changes in the motion of the camera, all of which are typical of handheld video capture.

The demographics of any dataset are crucial to the generalizability of the various models that use it. Hence we disclose our sampled population demographic in the hopes that it will condition any further research conducted using the 3DFAW-Video dataset and shed light on the biases it would introduce. The dataset has a total of N=66 subjects, with the mean age of 19.74 and standard deviation of 2.3, and an absolute range from 18-28 years of age. The ethnicity of the dataset population is shown in Table 3.2. The data was collected to ensure that the diversity in gender was close to equal, having 36 females and 30 males. It has to be noted that our dataset is heavily biased towards younger adults as well as skewed ethnically with a majority white sample population. Each participant in the dataset provided their informed consent for the distribution and use of their video/images for non-commercial research.

Ethnicity	No. of Subjects
African American	1
Asian	20
Latino/Hispanic	7
White	35
Others	3

Table 3.2: The racial distribution of the dataset population.

3.2.2 Dataset Generation

Since the data from the DI3D imaging system provides individual frame-level 3D scans of the face, these had to be manually merged together to create the final ground truth mesh. Each sequence of videos from the high resolution camera contains approximately 20K vertices, and 35-40K faces. However, these scans are only able to recover the geometry of the regions of the face that are directly visible to the camera system. This leaves us with multiple partial coverage

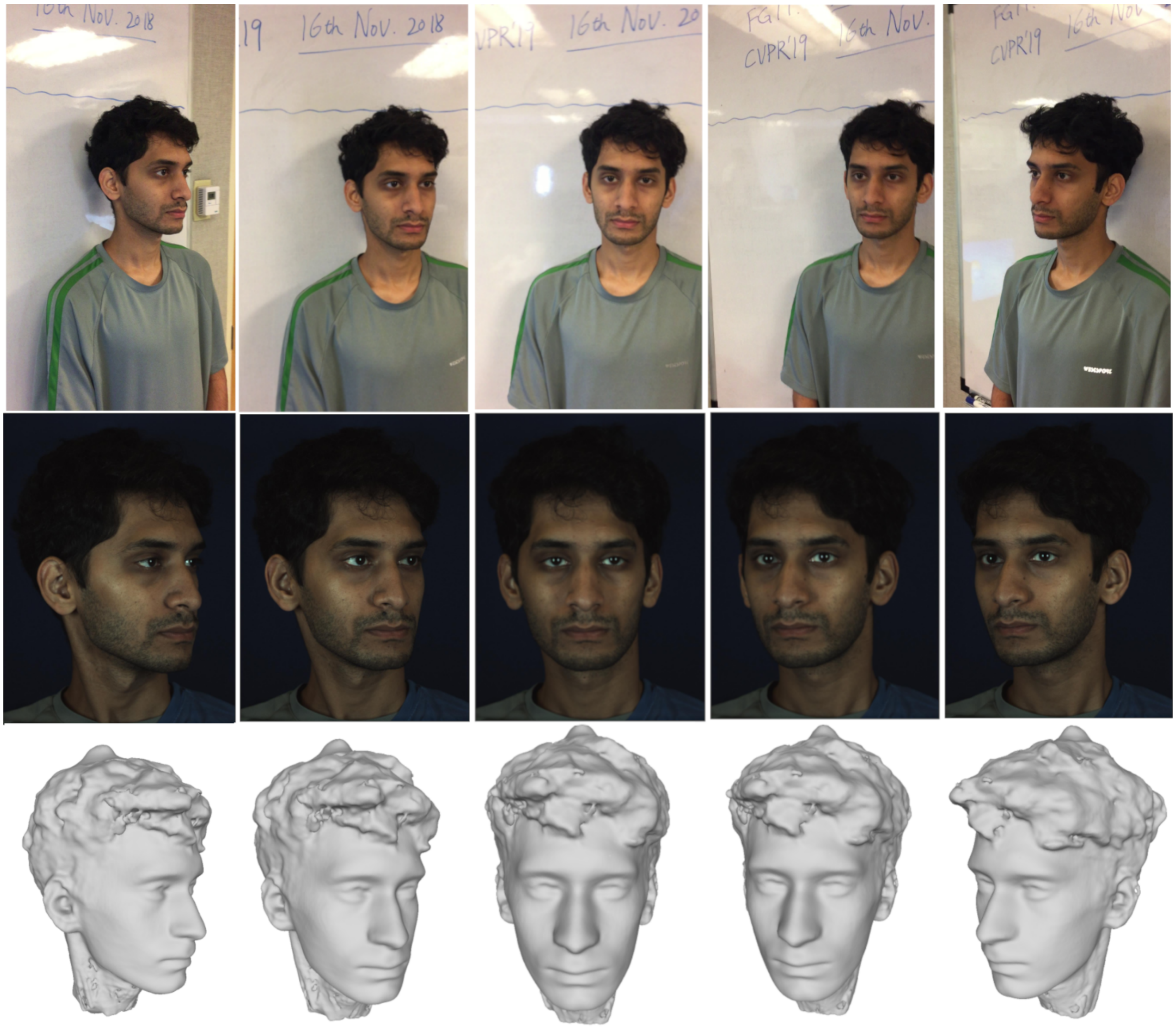


Figure 3.1: (Top) Selected frames from the uncontrolled camera motion, full range video collect on the iPhone 6. (Middle) Selected frames from the controlled camera motion, full range video collect from the DI3D imaging system's RGB camera. (Bottom) The final ground truth 3D model of the full head of the subject created by merging multiple frame level 3D scans from the DI3D imaging system's stereo camera.

3D scans that need to merged to produce a complete 3D mesh of the face. Each video from the high resolution system is approximately $5 \sim 10$s long and have approximately 130-150 individuals scans.

To begin the merging of multiple partial 3D scans of the subjects' face, we first manually cleaned all the scans using MeshLab [9], to remove the regions around the boundary of the mesh which typically produced high errors due to wrong projections. Then using the high resolution video approximately 10 frames are chosen which have little to no facial deformations(including blinks) from the neutral face. These 10 frames are also as equally spaced apart from each other as possible to ensure that they retain the full range of the video.

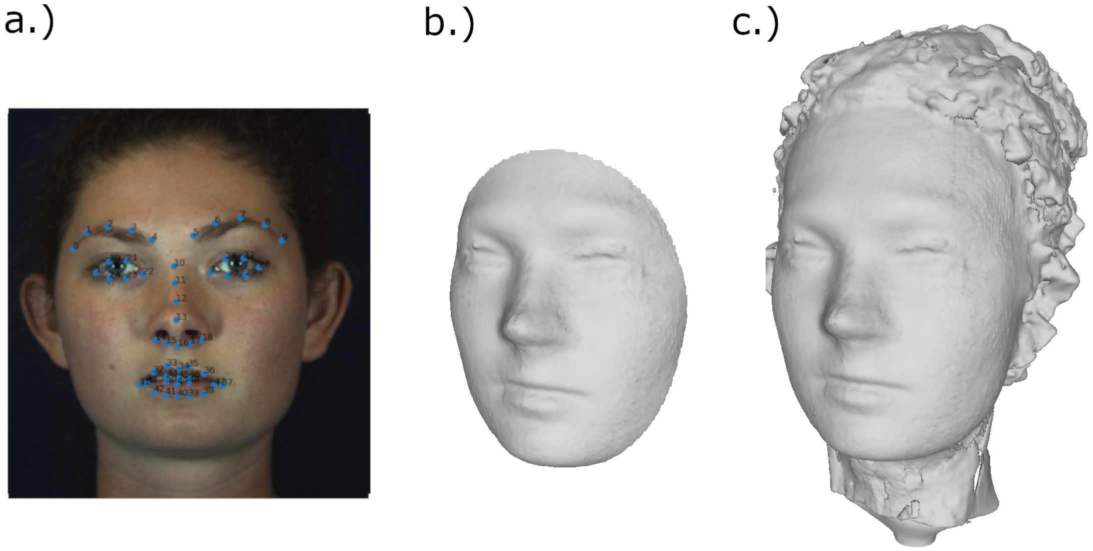


Figure 3.2: (a) The 51 facial landmarks provided by the dataset (b) The cropped ground truth mesh used in the challenge (c) The full watertight ground truth mesh.

Once all the 10 meshes have been selected as described, they are manually roughly aligned using the CloudCompare[1] software. Then each of the meshes are iteratively finely registered to the frontal mesh using Iterative Closest Point(ICP) algorithm without scaling. This fine registration is repeated until convergence defined by a RMSE difference below $1\text{x}10^{-4}$mm. Then the meshes are all merged together, and a Screened Poisson reconstruction[29] conducted to create a smooth water-tight mesh. This is the final 3D ground truth recovered from the DI3D imaging system after multi-partial scan fusion, shown in Figure 3.2c. While texture is available, we do not merge them, as the dataset focuses primarily on face geometry.

3.2.3 Data Folds

Since most newer techniques in 3D face reconstruction are learning based, these require an explicit split of the dataset into training, validation and testing folds. The 3DFAW-Video dataset is split into 3 subject-independent folds as shown in Table 3.3. The training and validation set together contain about 60% of the dataset, and the test set contains the remaining 40%. The training set contained the high resolution videos (controlled camera motion full range), the iPhone 6 videos (uncontrolled camera motion full range), and the restricted 3D ground truth mesh. The restricted 3D ground truth mesh is shown in Figure 3.2b, and is a crop of the water-tight ground truth mesh to a radius of 95mm around the tip of the nose. This qualified mesh was created

[1] http://www.cloudcompare.org/

Data Fold	Total Meshes	Stratification	No. Subjects
Train	26	Both	26
Validation	14	HiRes	7
		iPhone	7
Test	26	HiRes	13
		iPhone	13

Table 3.3: The subject independent data folds of the 3DFAW-Video dataset. The validation and test sets restrict the class of video that is provided (Either high resolution or iPhone 6)

to be released for the associated 3DFAW-Video challenge, as it followed other similar 3D face reconstruction evaluation protocols[11]. Furthermore, a set of 51 facial landmarks which make up the inner regions of the face from the Dlib library [30] was also provided for the frontal frame (Figure 3.2a). This annotation was provided to facilitate exploiting the profile-to-profile structure of the video by providing the frontal frame to begin the reconstruction.

Both the validation and test sets were constructed to be more challenging. Unlike with the training set, the test and validation sets only contained either the high resolution video or the iPhone 6 videos, along with the 51 landmarks. The 3D ground truth was withheld to be used in the evaluation of the various methods. 26 subjects make up the test set with an equal 13 subjects providing only one of the 2 videos types. Similarly the validation set contained 14 subjects with 7 in each video class. This was done to ensure that methods don't train on only the high resolution videos and disregard the iPhone 6 videos. During the data stratification, care was taken to balance the number of females/males across the various data splits. Our dataset and the training and validation splits can be downloaded from our challenge website[2].

3.3 3DFAW-Video Challenge

To test the use of 3DFAW-Video dataset as a benchmark for the multi-frame 3D reconstruction task, we released the dataset with the accompanying 2^{nd} 3D Face Alignment in the Wild Challenge (3DFAW-Video)[31], at ICCV 2019 held in Seoul, Korea. The challenge concluded with 4 different methods qualifying by submitting a detailed description of their method. The best method used a Structure from Motion (SfM) based reconstruction which is then used to fit a 3DMM. While the method reduced the error quantitatively, the qualitative results were very noisy with spurious local details. These results are discussed in more details in the results chapter. The second and third place methods both employed deep learning techniques. For example, [43] used an mesh retrieval process to pick the best reconstruction based on photometric loss between PRNet[16], MVFnet[54] and 3DDFA[54] predictions, while [38] used a siamese network to concatenate multi-frame features to regress a 3DMM. Their results along with the evaluation metric and procedure is described in the results chapter.

[2] https://3dfaw.github.io/

Chapter 4

Fusion based Multi-frame Reconstruction

With the creation of the 3DFAW-Video dataset, the task of 3D face reconstruction from 2D monocular structured videos can be benchmarked against 3D ground truth. Single image reconstruction methods dominate the field, and perform quite well for global shape reconstruction. However, provided video data, such methods are not able to exploit the temporal nor the multiple camera angles to improve the quality of their reconstructions. Hence, we propose a method that uses multiple frames of the video along with the structure of the continuously varying head pose to create a single reconstruction fusing the multiple single image reconstructions.

4.1 Assumptions

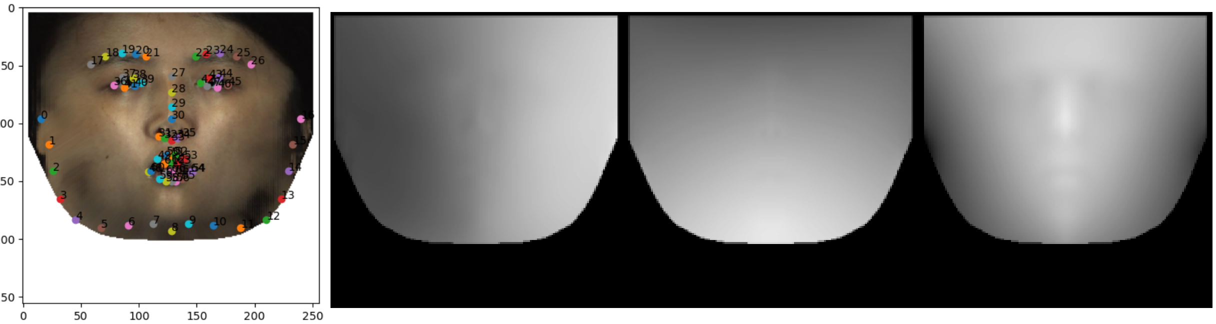
This method makes a few assumptions in order to fuse the multiple reconstructions. One of the core assumptions is that the reconstructed points on the face are most accurate when they are directly facing towards the camera. This is quite intuitive to see as surfaces that are orthogonal to the camera would have a degenerate projection in the image, and those that are occluded would face in the same direction as the camera itself. This allows for the camera pose and surface normal pairs to be used as a proxy for the confidence in the point's position in the reconstruction. Hence, picking the most accurate reconstructed points from each input frame becomes crucial for extending the method for image sets or videos. Moreover, we also assume that most single image reconstructions show a wide variation in their shape estimates when applied to an image set with large changes in headpose and illuminations. This assumption highlights the futility of a fusion if all the 3D face reconstructions per frame where so robust as to produce the same geometry, thereby adding no additional details in any novel view, and hence invalidating the need for information fusion across frames. However, the strongest assumption our method requires is the existence of a known dense correspondence function between the 3D reconstructions in different frames. This dense correspondence allows the method to ensure we track the same points across views to be able to assign them weights for recovering the best refinement of the face shape as possible.

4.2 Image Position Representation

Our method builds on the PRNet [16] model as it provides a few advantages over most other single image reconstruction methods. The PRNet model is not explicitly based on 3DMM fitting and can be trained directly on ground truth data. Moreover, the PRNet model covers a large portion of the face, from jawline to most of the ears. This is useful for fusing single image reconstructions as it allows for the final model to also be comparably expansive.

However, one of the biggest advantages with using PRNet is the associated position map representation, which is able to embed the whole face's 3D point cloud in a 2D image. The Figure 4.1b shows an example of the position map representation, where the 3 channels of the position map containing the spatial x,y,z coordinate values of the reconstructed face point cloud are show separately. Similarly, the color values from the original input image are also extracted to create the texture map as seen in Figure 4.1a. Hence, all the necessary information for a colored point set can be represented using just two 2D images (Position map and texture map) with perfect correspondences between them. Due to this compact, fixed 2D representation for a 3D face point cloud, extracting the landmarks for each of the images also becomes trivial and more importantly points correspondence between multiple reconstruction position maps is inherently available.

Finally, with the 3D face point cloud embedded in a 2D manifold representation as an image, we can leverage the power of frameworks such as image to image translation using CNNs for regressing the 3D structure from images. This is a great advantage over traditional representation of 3D structures using voxels, custom vector representations, or graph based representations which increase the complexity of models.



(a.)

(b.)

Figure 4.1: (a.) The 68 facial landmarks that are provided by the PRNet model, displayed on the texture map. (b.) An example of a position map representation with all 3 of the channels separately representing the x, y, z coordinates each.

4.3 Face Normal Cosine Method for Multi-Frame Fusion

To fuse multiple single image reconstructions, we use the position map representation as it is convenient to work in the 2D space and then extract the 3D points once the final position map is created. Our method revolves around using the cosine distance between the camera and surface normals of the reconstructions to fuse them together.

First, N different key-frames are selected such that there is always a frontal view frame that is included, and the other $(N-1)$ frames are from views from either side of the face. These N key-frames are then used to generate the position maps for their respective reconstructions. Figure 4.6a shows an example of 3 selected key-frames, the frontal, left and right sides respectively. The 3D reconstructions are full face meshes on their own, but each of the position maps encode not only the shape of the face but also the camera pose, as the point sets are oriented in the headpose in the input image. Each of the position maps have a resolution of 256x256 pixels but only a fixed 43,867 pixels of the total 65,536 pixels in the position map representation encodes meaningful facial information. Therefore, these are the only pixels from all the N position maps that we will use in our fusion.

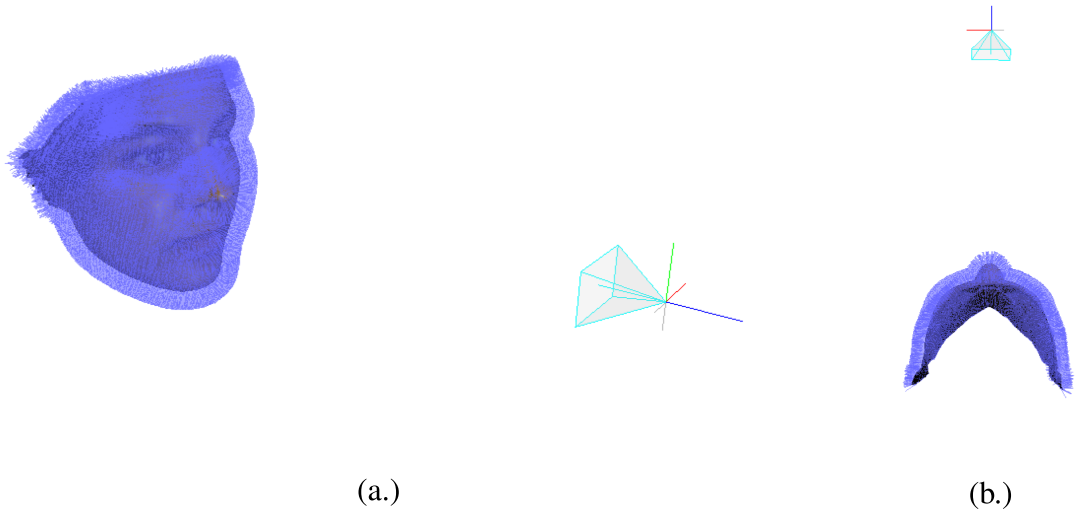


Figure 4.2: (a.) The camera from the frontal view and the blue lines indicating the vertex normals of the single image reconstructed mesh. (b.) The same mesh with the normals in blue and the camera pose, shown from the top view

The following terminology will be used to explain the cosine fusion method:

- $f \in [0, ..., N-1]$: The index of a key-frame/position map, of which there are N.
- $v \in [0, ..., m-1]$: The index of a vertex in a position map, there are $m = 43,867$ vertices in our implementation.
- P^f : The position map from the single-image reconstruction on f^{th} key-frame.

- P_p^f : The x,y,z positions of the vertices for the f^{th} position map.
- P_n^f : The vertex normals for the f^{th} position map.
- P_t^f : The RGB texture of the vertices for the f^{th} position map
- $\Phi(P_p^f)$: The x,y,z positions of the vertices for the f^{th} position map, after being frontalized to a canonical frontal pose.
- c^f : The camera orientation extracted from the f^{th} position map.
- $d_{cos}^f(v)$: The cosine distance for the v^{th} vertex in the f^{th} position map.
- $w^f(v)$: Weightage for the v^{th} vertex in the f^{th} position map.

Our method is detailed as follows. According to our core assumption of the accuracy of the reconstruction being the best for the points that are facing towards the camera, we first extract the camera orientation c^f, for each of the N reconstructions, and calculate the vertex normals of the position map, P_n^f, using the mesh triangular faces. Figure 4.2 shows the relation between the camera pose and the mesh along with the vertex normals of the reconstructed mesh. Figure 4.6b shows an example of the normal maps of the 3D reconstructions for each of the $N = 3$ key-frames.

For each of the position maps which correspond to f^{th} key-frame, we compute the cosine distance $d_{cos}^f(v)$ between the camera normal c^f and the vertex normal for the vertex v in the mesh P_n^f, as shown in Equation 4.1. This cosine distance metric, is geometrically significant as the cosine distance calculates the angle between 2 vectors, and hence this is able to inherently capture the orientation towards the camera in a smooth distance space. The Figure 4.6c visualizes the cosine distance on the 3 key-frames, where yellower color denotes a higher distance. The cosine distance for each of m vertices in each the meshes are then inverted using their max intra-cosine distance values as shown in Equation 4.2 to produce the per-vertex weights for each of the N key-frames. This inversion of the distance, will assign a higher weight to smaller distances. Figure 4.6d illustrates the final per vertex weights for each of the 3 key-frames, with yellower color indicating larger weights. As shown in the figure, the cosine similarity metric results in a smooth weighting across views, regardless of the number of key-frames.

$$d_{cos}^f(v) = \frac{P_n^f(v) \cdot c^f}{\|P_n^f(v)\| \|c^f\|} \tag{4.1}$$

$$w^f(v) = \max_{j=0}^{m-1}\{d_{cos}^f(j)\} - d_{cos}^f(v) \tag{4.2}$$

For each of the N key-frames, their position maps are frontalized, by solving a least square problem between each position map and a canonical position map. These frontalized position maps $\Phi(P_p^f)$, allow for each of the vertex positions to be in the same frame of reference before fusion. The final fusion of the multiple $\Phi(P_p^f)$s is done by a weighted averaging of normalized weights for each of the m vertices, as shown in Equation 4.3. This produces our final resulting fused point set, Q_p, in the frontalized coordinates, with the fused spatial information from the N

selected key-frames. Similarly, the same procedure can be applied, this time directly to the vertex textures, P_t^f, to fuse the texture across the N key-frames using the Equation 4.4 to produce a fused texture map Q_t.

$$Q_p(v) = \frac{\sum_{f=0}^{N} w^f(v)\Phi(P_p^f)(v)}{\sum_{f=0}^{N} w^f(v)} \tag{4.3}$$

$$Q_t(v) = \frac{\sum_{f=0}^{N} w^f(v)P_t^f(v)}{\sum_{f=0}^{N} w^f(v)} \tag{4.4}$$

This procedure provides the 3D mesh incorporating both the shape and the texture across N frames, as seen in Figure 4.4. In practice the N key-frames can be selected based on the confidence in the frames being roughly the same facial expression and lighting, since these factors are not addressed by the cosine fusion method.

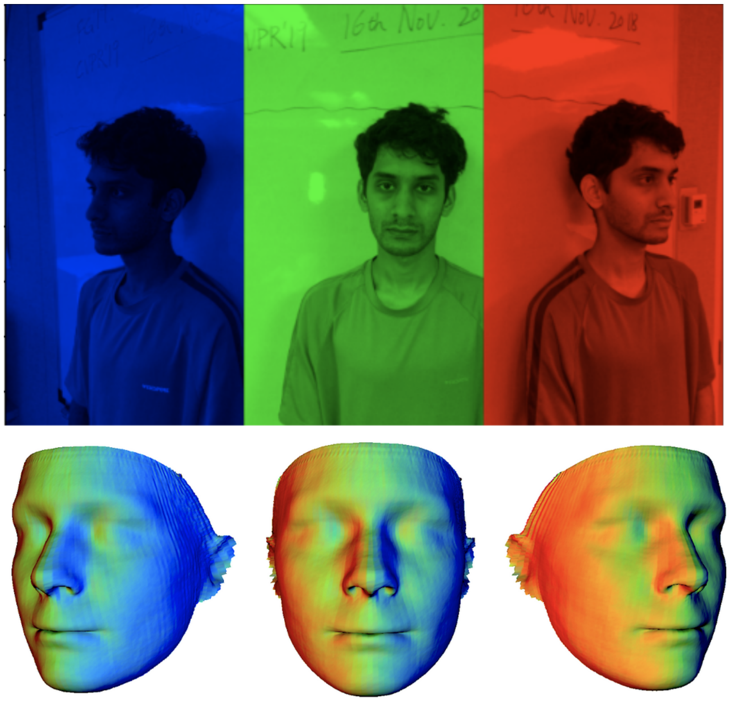


Figure 4.3: The fused mesh from the cosine fusion method color coded to show the combination of the information used from each key-frame.

Figure 4.3 shows the resultant mesh with each vertex colored by the combination of the weights and the solid red, green, blue colors assigned to each of the 3 key-frames. It can be seen that the cosine fusion method is drawing information from the frontal frame for the nose tip and forehead regions, while the 2 side key-frames contribute heavily to the shape on their respective sides of the face. However, it is interesting to note that in various complex shaped regions of the face such as the mouth, nostrils and the eyes the information across all the 3 key-frames contribute towards their final shape. These regions gain the most from the multi-frame fusion, and in the results section we will show that these are the regions where our cosine fusion method is able to outperform and improve over the base PRNet method.

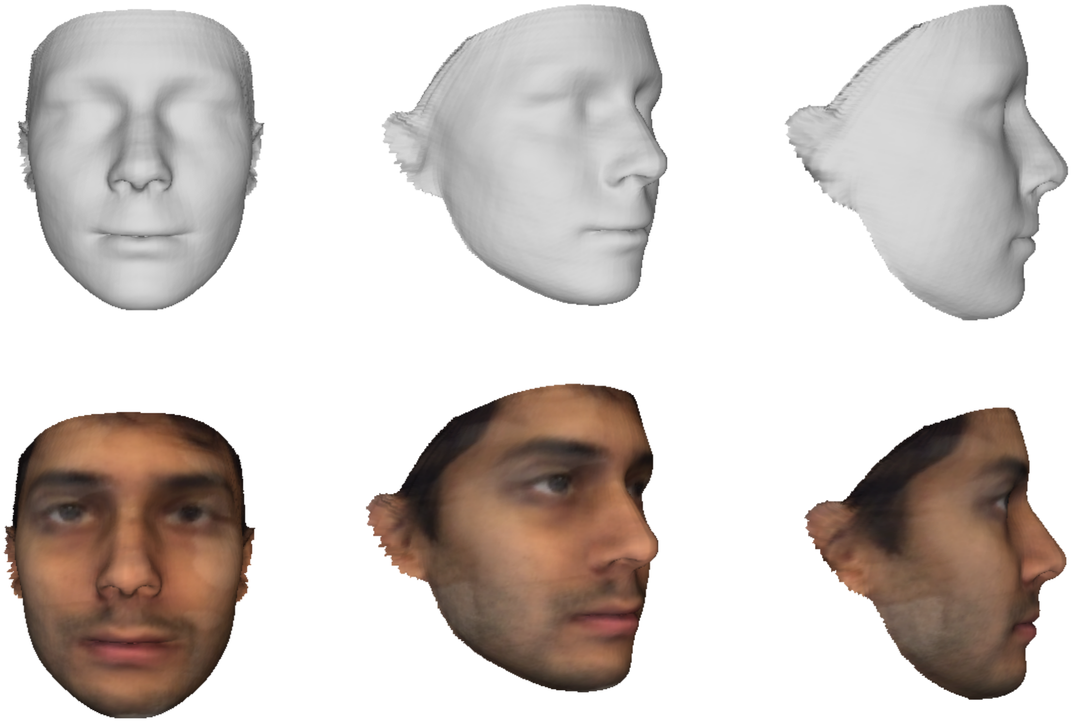


Figure 4.4: The final cosine fusion method result using the 3 frames as seen in Figure 4.6. The top row shows the mesh shape and the bottom row shows the mesh with texture also calculated using the method, across few views.

4.4 Fine Detail Estimation

Since the cosine fusion method is only calculating a shape based on multiple global shape estimates, the resultant mesh is also only globally accurate. In order to add finer details to the reconstruction there are 2 different approaches. For example, it is possible to train an image to image translation model to regress the displacement map to recover the fine details, by training on 3D ground truth datasets such as 3DFAW-Video. This would physically alter the point coordinates in 3D space of the reconstruction, improving results. However, the more popular approach is to use normal maps either through methods like SfS, or models such as [1] which produce stunning levels of details when rendered, but don't alter the geometry of the 3D mesh. (Figure 4.5(Left))

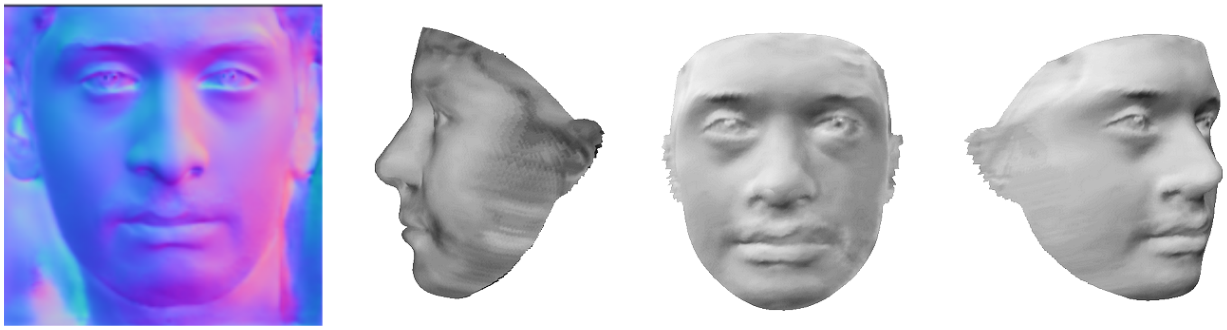
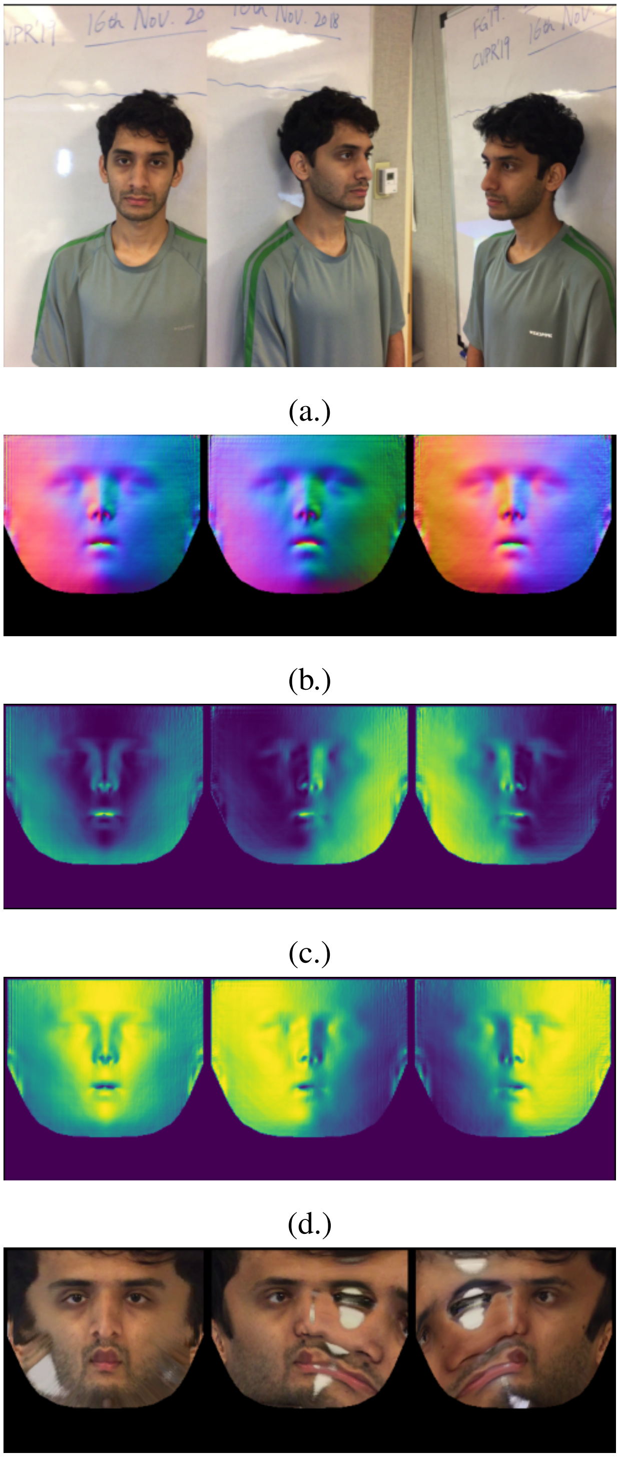


Figure 4.5: (Left) The normal map estimated by [1]. (Right) Rendered detail augmentation on the cosine fusion method mesh using the same normal map.

The results of applying the frontal frame normal map estimates, created using [1], on our cosine fusion method's resultant mesh can be seen in Figure 4.5(Right). Even though the details seem to add more skin surface realism, these meshes are geometrically exactly the same as the cosine fusion method output. However, it needs to be seen if the same details can be estimated instead with the use a displacement map which would alter the underlying geometry of the meshes produced. We leave this to future work, to explore the feasibility of such a method.



(a.)

(b.)

(c.)

(d.)

(e.)

Figure 4.6: From top to bottom: (a.) The selected N=3 frames from the video, (b.) The vertex normal map of the position maps/meshes reconstructions, (c.) The cosine distance between camera orientation and vertex normals (yellow is higher distance), (d.) The weights based on the cosine distance on each of the vertices (yellow is higher weights), (e.) The texture maps for the 3 reconstructions which will also be fused using the weights

Chapter 5

Results

5.1 Evaluation Protocol

The 3DFAW-Video dataset was introduced to benchmark various methods against 3D ground truth data in the 3DFAW-Video challenge held in conjunction with ICCV 2019. To be consistent with similar 3D face reconstruction challenges, we use an evaluation procedure borrowed from [11]. To evaluate the error between 2 meshes, the ground truth and the predicted mesh, the predicted mesh is first cropped to just the internal region of the face, within a 95mm radius of the nose tip. This allows us to compare the reconstruction with the ground truth that is also a similarly cropped mesh. Furthermore, since different method produce meshes that have varying levels of coverage of the face, the cropping ensures that it does not have any effect on the evaluated error of the prediction. The cropped meshes are then roughly aligned using their facial landmarks, and registered using rigid ICP. These registered meshes are then used to calculate the reconstruction error using the novel, ARMSE metric as explained in the following subsection. For efficiency and reducing computational load on the servers, a sampled down set of vertices are used to calculate the reconstruction ARMSE scores.

5.2 Evaluation Metric

3D-RMSE has been used as the error metric of choice for most other 3D ground truth datasets. However, 3D-RMSE only calculates the root mean squared distance between the vertices of the predicted mesh to the ground truth (also called accuracy metric). This can be an issue when comparing multiple predicted mesh with varying mesh densities to the same ground truth mesh, as a highly dense mesh with low errors and a sparse mesh with higher errors can produce the similar 3D-RMSE values. When comparing arbitrary reconstructions to a ground truth mesh, the effect of mesh density must be handled, and thus we introduce, average root mean square error(ARMSE), our novel error metric which is follows directly from our motivation.

As discussed the average root mean square error(ARMSE), tries to reduce the influence of mesh density, by extending the 3D-RMSE metric to be a symmetric error metric. While the 3D-

RMSE score only calculated distance from the predicted mesh to the ground truth, the ARMSE score is bidirectional and takes a normalized mean between the predicted mesh to ground truth distance and the vice versa, ground truth to predicted mesh distances. This reduces the effect of the mesh densities as the number of vertices in the ground truth is constant for all methods. The ARMSE metric differs from other bi-directional distance metrics such as Chamfer distance, as we consider distances between meshes rather than point clouds.

The average root mean square error(ARMSE), uses the equation 5.1 to calculate the distance from a source mesh A and the target mesh B, i.e. the vertex-to-mesh distance ($E(A_i, B)$). This equation calculates the distance between a vertex on the source mesh A_i, to the surface of the target mesh B, as the minimum of the L_2 distances from the vertex A_i to any vertex B_v, edge B_e, or face B_f on the target mesh.

$$E(A_i, B) = min(\|A_i - B_v\|_2, \|A_i - B_e\|_2, \|A_i - B_f\|_2) \tag{5.1}$$

Similar to the 3D-RMSE, these vertex-to-mesh distances ($E(A_i, B)$) from the source to target meshes are squared and aggregated for every vertex of the source mesh and the divided by the total number of source vertices N_a. Taking the square root provides the directed RMSE score ($D(A, B)$), from the source to the target mesh. This is captured in the equation 5.2.

$$D(A, B) = \sqrt{\frac{1}{N_a} \sum_{i}^{N_a} E(A_i, B)^2} \tag{5.2}$$

$$ARMSE(X, Y) = \frac{100(D(X, Y) + D(Y, X))}{2I} \tag{5.3}$$

The ARMSE score between the predicted mesh X and the ground truth mesh Y, is calculated by the equation 5.3. Here, the directed RMSE scores from the predicted mesh to the ground truth ($D(X, Y)$) and from the ground truth back to the predicted mesh ($D(Y, X)$) are averaged. The resulting score is then normalized by the outer inter-ocular distance I which is retrieved from the ground truth mesh and scaled by a factor of 100. This normalization allows the ARMSE scores between subjects to be compared with each other, making the metric easier to analyse. Furthermore, the scaling helps increase human readability of the ARMSE score as it expresses the reconstruction error as percentage of the outer inter-ocular distance.

5.3 Comparisons with 3DFAW-Video Challenge

The Table 5.1 provides the final leader-board results for the competition, ranked by their mean ARMSE scores on the testing set of the 3DFAW-Video dataset. The methods used by the participants in the competition are discussed briefly in the Challenge section of the 3DFAW-Dataset

Rank	Team	mean ARMSE
1.	Zheng	1.6962
2.	Shao et al.[43]	1.8642
3.	Maldonado et al. [38]	2.1429
4.	Chen	2.1865
5.	PRNet w/ Cosine fusion (ours)	2.4058
6.	PRNet	2.5606

Table 5.1: The mean ARMSE scores of the different methods on the 3DFAW-Video test set.

& Challenge chapter. In the major comparison between PRNet with Cosine fusion and the other methods used in the challenge, our method does not seem to preform well. This is most likely due to the fact the we currently don't add any mesoscopic detail to our predicted meshes, and simply rely on the core method. However, with additional details this performance could drastically improve. Furthermore, the more competitive methods involved various different models being combined which is also a technique that could bring in some robustness to the front.

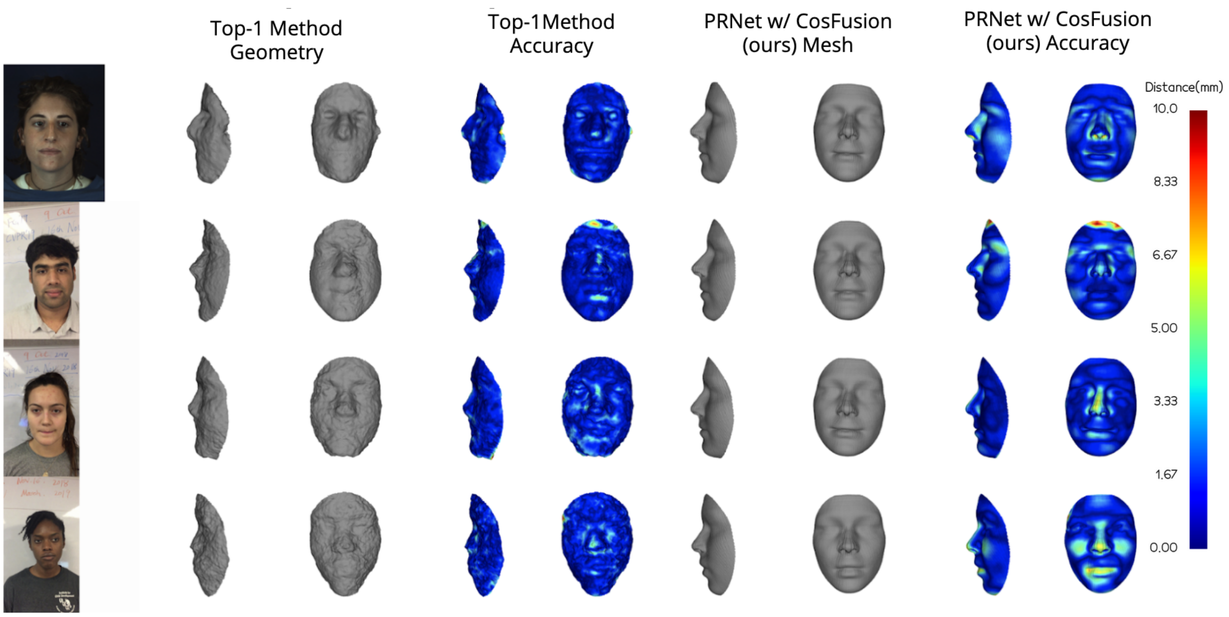


Figure 5.1: An example comparison between the Top-1 ranked method in the challenge and PRNet w/ cosine fusion method reconstructions. The heatmaps are shown to represent the distance (mm) between the ground truth meshes and the reconstructions (Accuracy metric).

Although our reconstruction with the cosine fusion method does not outperform the challenge methods, it has to be noted that the reconstructions that our method provides is of much higher quality qualitatively. Figure 5.1 shows the geometry and the heatmaps of the top-1 ranked method in the challenge, against our PRNet with cosine fusion method results. The top-1 ranked method used a SfM based approach to fuse the multi-view data to get a rough structure and further refined it with the use of a 3DMM. However, the overall mesh created from such an approach although able to minimize the quantitative ARMSE scores, are qualitatively extremely noisy with visually discernible local defects.

This indicates that although the ARMSE metric might be more robust to quantitative density differences between predicted meshes, it suffers from its inability to penalize qualitative errors in the predicted meshes. Hence, the ARMSE score is not a good score to compute the qualitative accuracy of a predicted mesh. Therefore, combining the ARMSE metric along with a quantification of the qualitative fit using another metric such as the Fréchet Inception Distance (FID) [23], should make comparing 3D face reconstructions more robust to unnatural local noise and defects in the mesh.

5.4 Improvements over Single Image Reconstruction

However, it can definitely be seen that comparing the base PRNet reconstructions with our PRNet w/ cosine fusion, that our method shows good improvements. The table 5.1 shows that we are able to reduce the mean ARMSE over the test set. This is shows that there is merit in studying the fusion of various single image 3D reconstructions as a method to improve multi-frame consistency. Furthermore, our method can in fact be extended over to various other single image reconstruction models that pass our assumptions since many such models produce a fixed representation based reconstruction and thus provide semantic consistency over video data.

Figure 5.2, shows the comparison of the base PRNet reconstructions against the PRNet with our cosine fusion method on 5 key-frames at 30 degrees separation for 9 subjects from the 3DFAW-Video test set. Since the test set contains both HiRes videos as well as iPhone videos, the figure shows the example results for videos of both classes. While the improvements in the geometry of the reconstructions are harder to see, the heatmaps comparing the error with the ground truth meshes show clearly that the overall error reduced in major regions of the face. The heatmaps show significant improvements mostly along the central vertical axis of the face. The cosine fusion method meshes reduces the errors on the forehead, the nose, and chin regions. We can also see improvements in the cheek regions for many of the meshes.

The cosine fusion method is also able to reduce the ARMSE error standard deviations from 0.9919 for the base PRNet method to 0.9037 with the cosine fusion method. Overall, combining PRNet with the cosine fusion method produced a 6.05% improvement in mean ARMSE, over the base PRNet reconstructions. Therefore, the information fusion between multiple frames using cosine fusion method does seem to help in reducing the errors in many of the key anthropometric landmarks of face.

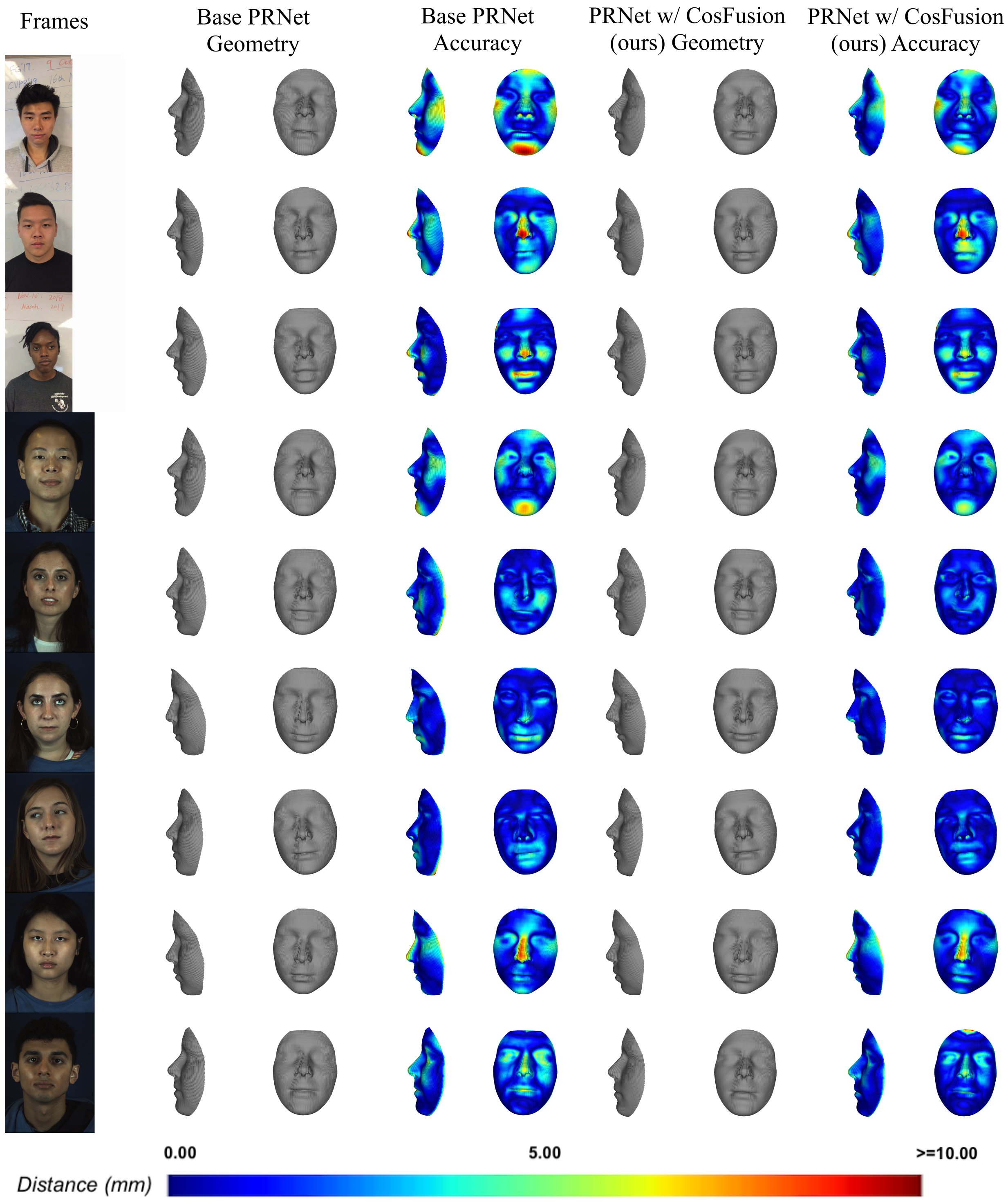


Figure 5.2: An example comparison between base PRNet and PRNet w/ cosine fusion method reconstructions. The heatmaps are shown to represent the distance (mm) between the ground truth meshes and the reconstructions (Accuracy metric). The reconstructions are based on the frames from the different video resolutions of the 3DFAW-Video test set.

Method (#Frames, Angle)	Mean ARMSE	Std ARMSE	Mean Accuracy	Std Accuracy	Mean Completion	Std Completion
PRNet	2.5606	0.9919	2.2575	0.6414	2.8638	1.5168
PRNet w/ CosFusion(3, 45)	2.4150	0.9231	2.1265	0.5475	2.7035	1.4630
PRNet w/ CosFusion(5, 30)	**2.4058**	**0.9037**	**2.1202**	**0.5190**	2.6914	1.4474
PRNet w/ CosFusion(7, 10)	2.5080	0.9412	2.2279	0.5649	2.7881	1.4926
PRNet w/ CosFusion(15, uniform)	2.4226	0.8980	2.1398	0.5197	2.7054	1.4415
PRNet w/ CosFusion(20, uniform)	2.4081	0.8974	2.1257	0.5144	**2.6905**	**1.4443**
PRNet w/ CosFusion(30, uniform)	2.4193	0.9053	2.1353	0.5241	2.7033	1.4517

Table 5.2: The results of the experiments with changing the number of frames (given in parentheses) used for the Cosine fusion method as compared to the base PRNet. The accuracy metric is the distance from the mesh to the ground truth while the completion is the distance from the ground truth to the mesh. Note, for all metrics lower numbers are better.

5.5 Experiments with Number of Frames

Since our method is a larger meta-reconstruction technique that is based on the single image reconstructions, it naturally raises the question on the effect of the number of frames used for the procedure. In traditional computer vision techniques such as SfM, larger number of inputs is always appreciated. The more views/frames of the object provide further constraints to help recover the 3D shape of the object. However, adding more images can increase the computation time and slow down the whole procedure.

For this experiment, the benchmarking was done on the 3DFAW-Video dataset's test set, as it provides us with a head-pose structured video, from which we could systematically sample N different frames for use in the reconstruction technique. We use both the HiRes and the iPhone video resolutions for this task, depending on the subject. The videos are all run through the ZFace tracker to get an initial head pose estimate, which is used to select the frontal frame and the other frames to each side. The set of the total number of selected frames tested are 3 frames(45 degrees apart), 5 frames(30 degrees apart), 7 frames(10 degrees apart), and 15, 20, 30 frames all uniformly spaced out between the maximal profile to profile head orientation changes. The Table 5.1, shows all the results including auxiliary metrics of accuracy, distance from the predicted mesh to the ground truth and completion metric which is the reverse distance from the ground truth to the predicted mesh. A cleaner, graphical summary can be seen in Figure 5.3, with just the mean metric scores compared for all the variations experimented.

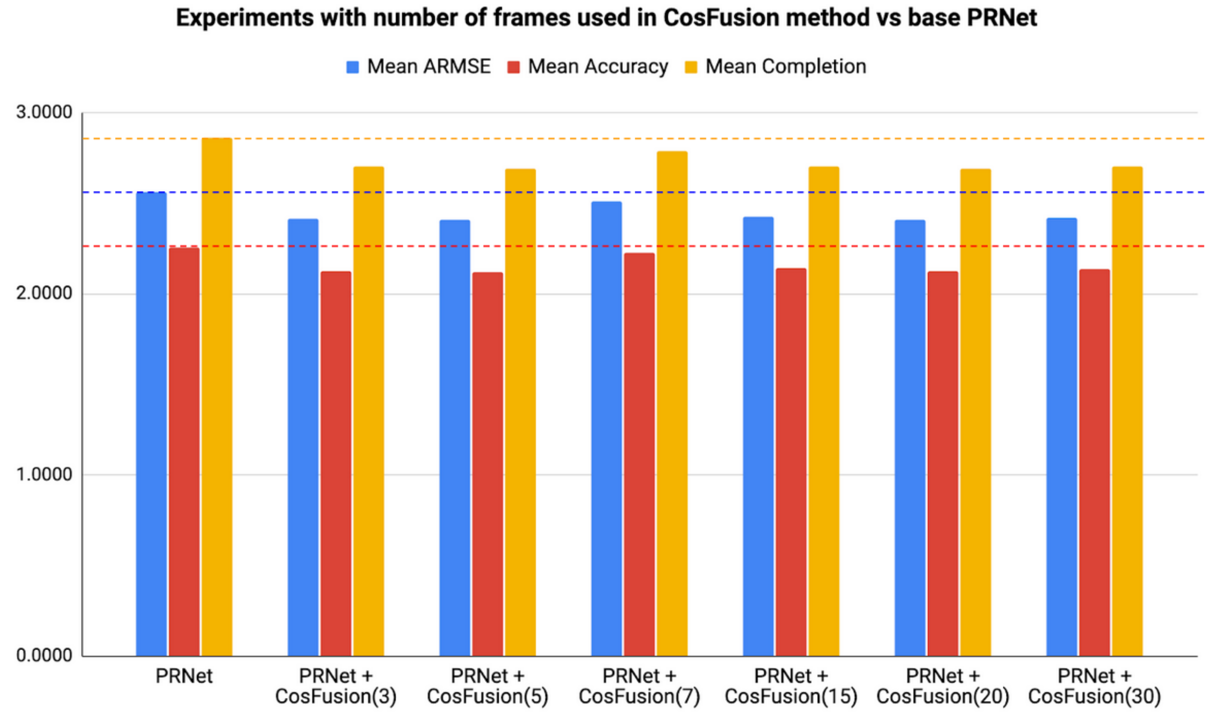


Figure 5.3: The results of the experiments with changing the number of frames used for the Cosine fusion method as compared to the base PRNet. Note, lower bars are better.

From the results it is not very obvious if there exists a strong correlation between more data and a better estimate. However, in general the larger number does reduce the errors although it does seem to plateau past 20 frames. The experiment with 15 images does seem to be anomalous in the general pattern observed. This could have been due to the set of frames that were selected introducing more noise due to blurry motions etc. But it is worth mentioning that selecting the 'right' frames, that can minimize the jitter of the single image reconstructions is a topic that need to be studied more closely.

Furthermore, the experiment also shows us that the cosine method suffers from similar issues as the classical Structure from Motion (SfM) approach. Having a larger angle of separation, wider baseline, between images provides greater amount of information for reconstruction fusion, however, it comes at the cost of lower overlapping regions between them. Conversely, with a smaller angle of separation, shorter baseline, the information gained is lower while the overlapping regions increase. The trade off is one that needs to be investigated further to find the optimum baseline width for the cosine fusion method.

Chapter 6

Applications In-the-Wild

Although 3D face reconstruction is not a new field in computer vision, there are increasingly newer applications for the technique being introduced year after year. While 3D face reconstruction research has seen it's uses in popular social media applications such as Snapchat and Instagram filters, most other uses of 3D face modeling are not as commonly seen in the general public. The couple of applications we discuss are based on projects focusing on in-the-wild data, and have applications beyond the just their immediate scope.

6.1 3D Mask Sizing using Mobile Videos

One of the most important piece of equipment for first-responders, soldiers and even clinicians are accurately and tightly fitting face masks. Since the environment in which they work can be hostile, leaks between the face and the mask could be fatal. However, the procedure to get such a mask fitted accurately is time consuming and requires specialized equipment, trained technicians, and can require multiple fitting sessions. Furthermore, the manual measurements are error prone due to human error and requires physically trying out multiple masks. Yet, even after a sizing is conducted, it must repeated for other masks as the sizing measurements are mask specific.

Hence, we introduce the task of automating the mask sizing using solely two specialized 2D monocular videos collected from a mobile device along with the phone's accelerometer readings to conduct 3D face reconstruction to metric scale. The metric 3D face reconstruction can then be used to get the distance between any points on the face as needed, to allow the whole procedure to be fast, hassle-free, and a non-contact mask sizing approach. An overview of our method can be found in Figure 6.1.

6.1.1 Face Reconstruction

To begin the metric face reconstruction, the user, with a custom mobile application built for the data collection, takes 2 different videos in a relatively well lit environment. The first video is an Inertial Measurement Unit (IMU) calibration video, while the second video is similar in structure

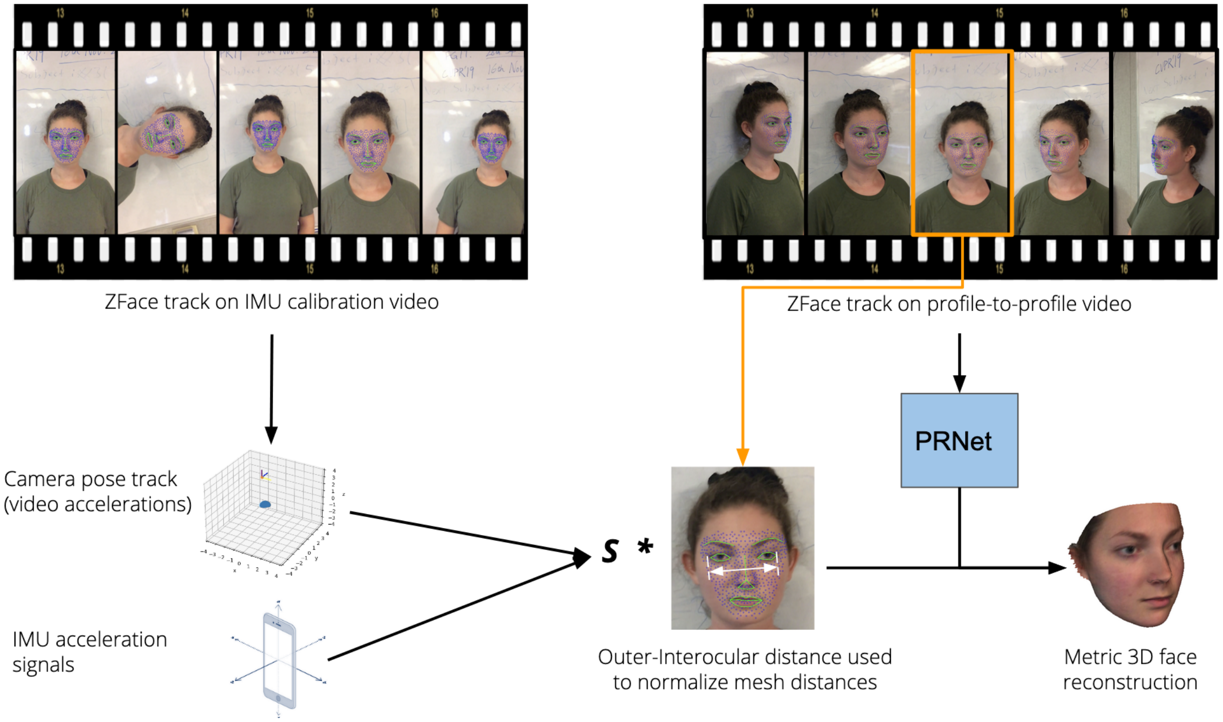

ZFace track on IMU calibration video
ZFace track on profile-to-profile video
PRNet
Camera pose track (video accelerations)
IMU acceleration signals
s *
Outer-Interocular distance used to normalize mesh distances
Metric 3D face reconstruction

Figure 6.1: Overview of pipeline for 3D metric face reconstruction from video. The left half provides the overview of the metric scale estimation, while the right half combines the metric scale with the 3D face reconstruction for the final result.

to the 3DFAW-Video iPhone videos as it has uncontrolled camera motion and captures the full range of the face by a profile to profile camera sweep around the subject.

For any 2D monocular facial task, face tracking is an essential procedure. For this task we extensively use the ZFace tracker [27] ported to the python language, for use in mobile and personal devices. On the profile to profile video, collected at a native 30FPS, we run the ZFace tracker to estimate the headpose of the subject, which provides us the ability to choose specific frames for reconstruction. For this project, the frontal frame as shown in Figure 6.1 was selected to be used as it had the least self occlusions. These frames were then run through PRNet model to get its single-image reconstruction, as shown in Figure6.1. These results were the state-of the art at the time and hence augmenting the results with additional frames were not done. Furthermore, a smaller model, such as one based on MobileNet [24]based single-image reconstruction would be ideal to run on personal devices.

6.1.2 Metric Scale Estimation

The IMU calibration video that is collected using the mobile application is crucial for the metric scale estimation. The IMU calibration video is also structured, as it is includes a rapid rotation in the x,y plane, and followed by rapidly moving back and forth from the subject along the z axis, as shown in Figure 6.2. The video itself is captured at a high frame rate of approximately

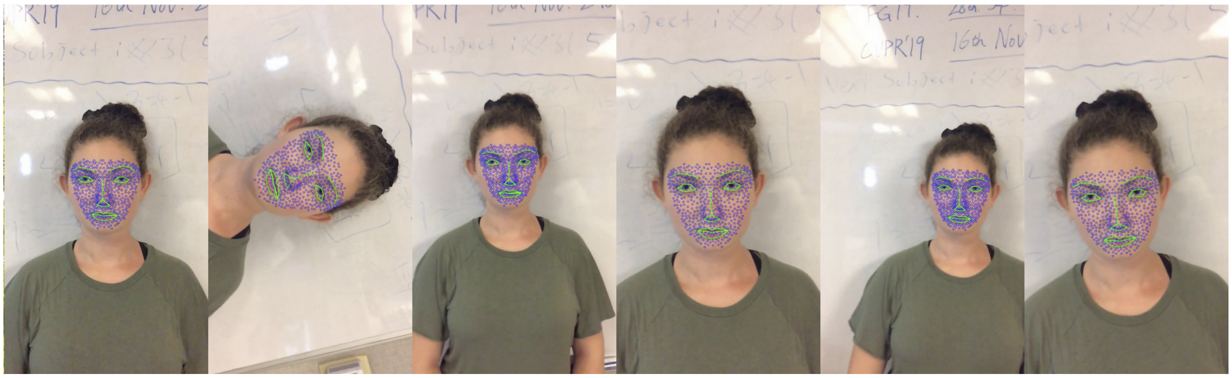


Figure 6.2: The IMU calibration video is collected using the data collection application that also records the IMU data. The IMU video contains quick in-plane rotations followed by back and forward movements to ensure that the signal alignment is accurate, and as shown above tracked with ZFace.

240FPS. This allows for smooth tracking of the subject over the video, using ZFace. Assuming that the subject is stationary, and we have the camera intrinsic matrix, the 3D face tracks are used to calculate the camera pose by solving the Perspective-n-Point problem. We then additionally correct the camera tracks for degenerate cases, and calculate the camera's accelerations.

The metric scale estimation from the captured high frame video and the accompanying IMU reading are based on [22]. Although the mobile application is setup to record the accelerometers and the video in tandem, due to device limitations these are never fully synchronized. Furthermore, the IMU data frequency is much higher than the video accelerations, and this makes the the temporal alignment bit more challenging. The IMU linear accelerations signal and the camera accelerations are aligned using Normalized Cross Correlation (NCC), but can have errors with smaller motion due to noise and hence gravity is introduced into the camera accelerations which allows for more accurate alignment. Then [22] provides the following objective function formulation:

$$\underset{s,\mathbf{b},\mathbf{g}}{\operatorname{argmin}}\, \eta\{s\hat{\mathbf{A}}_V + 1 \otimes \mathbf{b}^T + \hat{\mathbf{G}} - \mathbf{D}\mathbf{A}_I\mathbf{R}_I\} \tag{6.1}$$

Where $s, \mathbf{b}, \mathbf{g}$ are the constant parameters being estimated which are the scale, bias of the IMU sensor, and the gravity component respectively. $\mathbf{D}\mathbf{A}_I\mathbf{R}_I$ is the down-sampled, body centric IMU accelerations, $\hat{\mathbf{G}}$ is the gravity component estimated from the camera rotations, $\otimes$ represents the kronecker product operator, and finally $\hat{\mathbf{A}}_V$ is the accelerations from the camera video. η is a penalty function, which [22] empirically shows works best when selected to be the $L_{2,1}$ norm penalty, which is the sum of L_2 norms of the objective over all the frames of the video. The expression 6.1 can be solved by using a linear programming solver, by alternatively solving between the $\{s, \mathbf{b}\}$, and $\mathbf{g}$. We use the alternating direction method of multipliers (ADMM) solver as implemented by [22] to solve the linear program. The know gravity component, $\mathbf{g} = 9.81ms^{-2}$, is used to normalize the $\{s, \mathbf{b}\}$ values, and iteratively solved till the scale and bias converge.

Once the scale values are estimated using this method, it was however, empirically found to be not fully within the accuracy desired. Hence, we additionally apply a linear regressor, learned on 3DFAW-Video dataset 3D scans, to correct the scale estimates from this method to be within the acceptable error range.

The 3D reconstruction and the scale estimation are done separately, and the final output 3D reconstruction from PRNet is then scaled to metric scale using the corrected scale factor by using outer inter-ocular(OIC) distance from the ZFace track as normalization factor (see Figure 6.1). The generated mesh's preliminary analysis of metric reconstructions were also done by calculating the outer inter-ocular(OIC) distances and comparing against the ground truth to ensure that the method does not produce anomalous results. These 3D face reconstructions were used with predetermined anthropomorphic features on the face to predict a fit, which was then validated in simulation using CFD Research Corporation's (CFDRC) computational fluid dynamics respiratory model.

6.1.3 Results

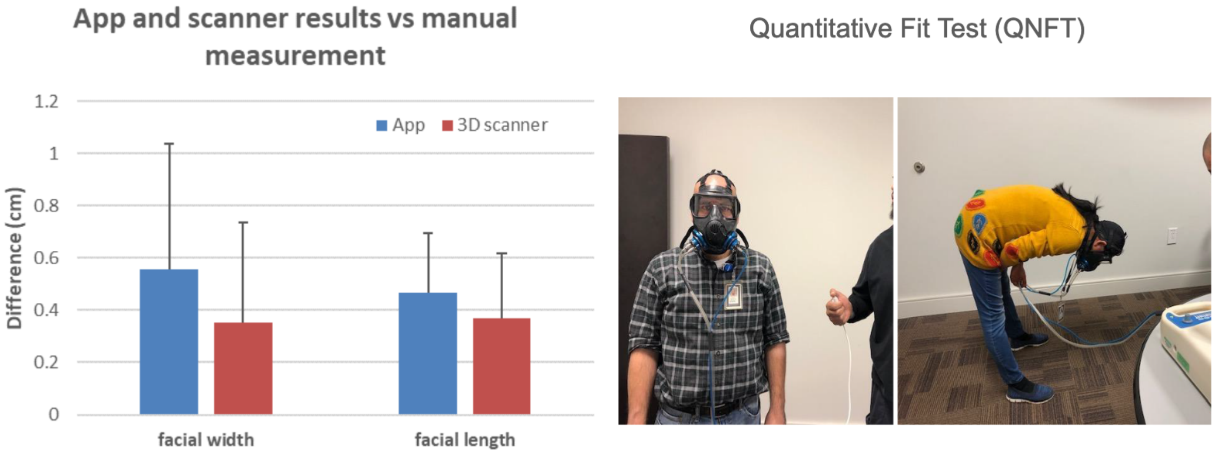


Figure 6.3: (Left) The comparison of our metric 3D face reconstruction application against 3D scanner on the face width and face height estimates. (Right) Participants in the QNFT, checking fit and leaks during large motions. Both the results shown above are courtesy of the CFDRC experiments as reported in [39].

All the experiments where conducted by our collaborators at CFDRC. As reported in [39], our metric face reconstructions were tested on 20 individuals on whom first manual measurements where taken. Then a 3D scanner was used to capture the facial structure of the participants, and followed by the metric 3D face reconstruction application we built. For the comparison, the manual sizing was taken to be the ground truth. The results of the quantitative comparison of the face dimensions using the scanner and our application can be found in Figure 6.3(Left). The application is able to estimate the dimensions to a very close accuracy to the 3D scanner, with the difference between the two methods estimated less than 0.3cm for face width, and 0.1cm for face length. Our technique, when compared against ground truth measurements on the total set of facial features for mask-sizing, also reportedly produced a mean error below 4mm and a mean percentage error below 5.4%.

Moreover, 13 subject took the Quantitative fit test (QNFT) in which the participant physically tries on the mask with the size predicted by our application. Then the participants' mask fit was tested during various body motions as shown in Figure 6.3(Right). All the masks predicted by our application passed the QNFT, showcasing its effectiveness as a fast, portable, and non-contact based replacement for manual mask sizing techniques.

6.2 Facial Video based Physiological Signal Estimation

While the first application was more directly involving the result of 3D face reconstruction, here we introduce an auxiliary task using 3D face reconstructions on video data. While most of the single-image reconstruction methods might ignore multi-view geometry for creating their reconstructions, for a task that requires quick 3D face alignment, it is not an issue. Hence we discuss the use of 3D face reconstructions as a normalization technique made possible by single image reconstruction methods.

6.2.1 3D Face Reconstruction as Normalization Technique

The task of normalizing data is an extremely powerful technique for a wide variety of applications further down stream. Hence, when engineering large data pipelines it is almost always useful to remove as many factors contributing to the variation in the data as possible, especially when it obstructs the main task. For instance [26] showed that normalizing datasets using a spatial transformer network allowed the network to learn invariance to rigid transformations and certain types of warping, which greatly improved their accuracy on classification tasks.

The more recently released Automated Facial Affect Recognition (AFAR) toolbox [35] extends this normalization to faces, and provides a very powerful facial video normalization procedure, which removes in-plane orientation changes on the face. While appearance variation due to headpose can only be partially addressed using the AFAR toolbox, as it provides a 2D normalization, it still greatly simplifies the following tasks such as automated facial Action Unit (AU) Detection. Furthermore, in the field of psychology where human experts are still common as a source of ground truth annotations, a 2D normalized video can make the annotation task much easier and more accurate.

However, with the use of 3D face reconstruction one can go a step further. One such method was introduced in [4], which uses a UV space based mapping to embed all facial features in a

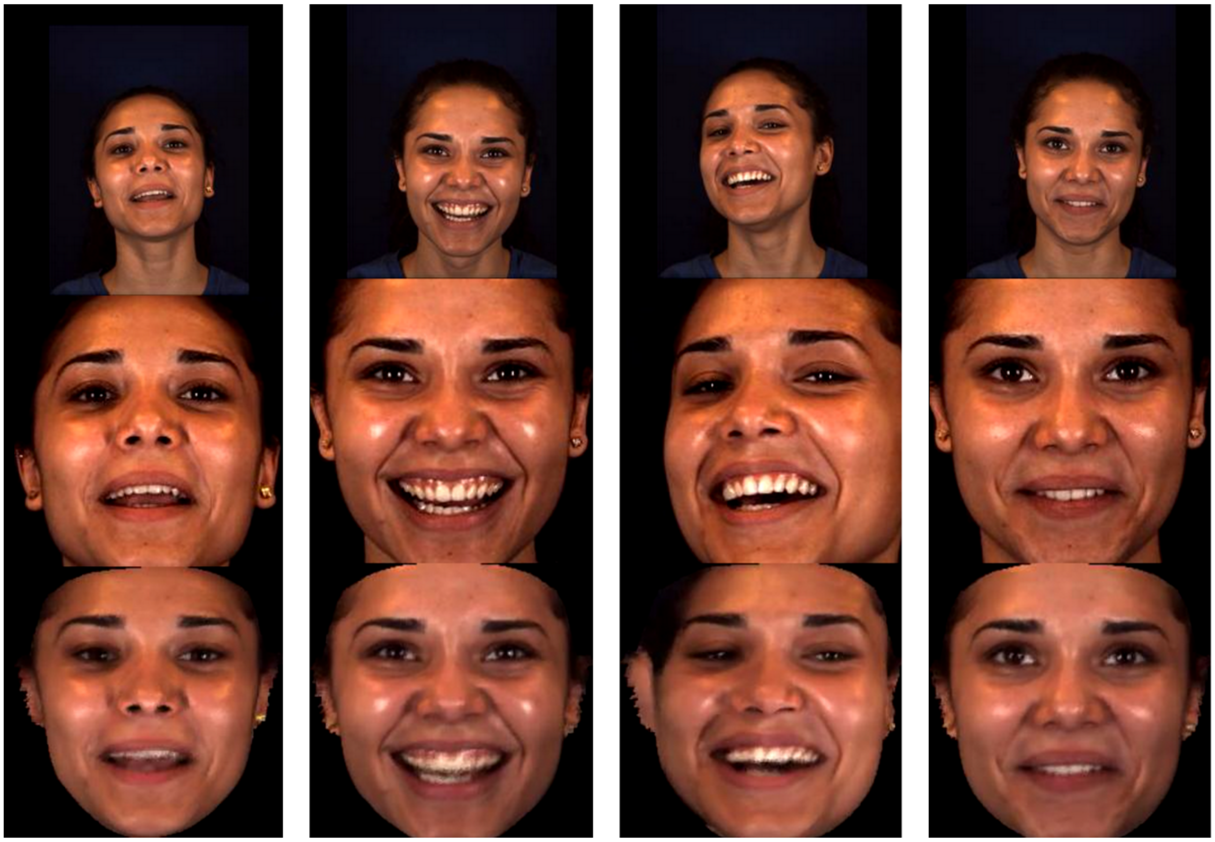


Figure 6.4: The 2 different normalization, techniques and their edge case behaviours can be seen in this montage. The top row of images are from the original unaltered video for reference, while the second row contains frames from the 2D normalized AFAR toolbox output. The bottom row contains our implementation of 3D face normalization using a re-rendering of single-image 3D face reconstructions.

2D manifold. Since the 2D manifold is canonical and fixed, it provides a seamless correspondence between individuals. We use a similar approach, but instead focus on using a 3D rendering based technique so as to still provide a human friendly 3D normalized representation. In order to produce a 3D normalized facial video, first every frame has to 3D reconstructed separately. Hence, single image reconstruction methods such PRNet becomes extremely useful. Once all the frame are reconstructed, they are re-rendered using the 3D reconstructions at each frame, but from a fixed frontal view. By ensuring all the reconstructions are frontalized and aligned, a smooth frontal view only video will be generated.

Figure 6.4 shows the differences between the 2 different types of facial normalizations. The top, middle, and bottom rows refer to the original frames, 2D normalized frames, and our 3D normalized frames respectively. The vertical set of frames are selected to show the boundary case behaviours of the different normalizations. With the 2D normalized video, when there are large out-of-plane rotations such as in the first and third columns, a major portion of the face such as the forehead temple regions or the cheeks become self-occluded, and their appearance changes. However, even with these headpose changes, the 3D normalization roughly maintains its appearance, due to the re-rendering trick. In fact, while the 2D normalization only removes in-plane rotations, scale, and 2D translations, which amount to 4 degrees of freedom removed, the more powerful 3D face normalization removes all 6 rigid degrees of freedom (all head orientations and translations). Hence, 3D normalizations therefore only leaves the non-rigid deformations due to expression changes to be handled.

6.2.2 Heart Rate Estimation from Video

Tele-health in recent years have become a very popular field in both research and industry. The COVID-19 pandemic, has also piqued the interest of many towards tele-health approaches, and hastened those who are already in the field. While tele-health is still growing there is a push towards automating vital sign detection through solely the commonly available sensors suite of a smartphone. Such physiological measurements can be used for final disease state estimation as a tool for health care workers, and also help improve access to healthcare due to the reduced cost, training, and its ability to capture this data in a remote, non-contact manner.

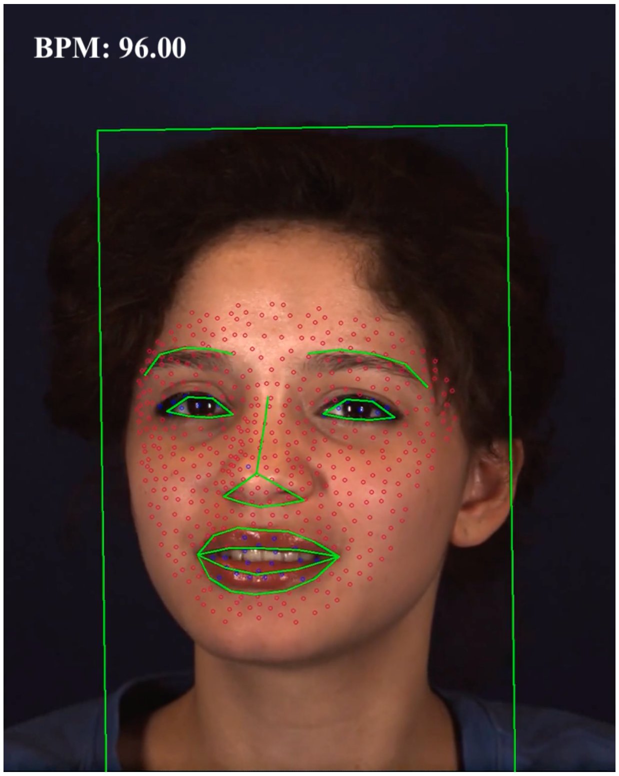


Figure 6.5: Heart-rate estimation from video using the ZFace-BKF method.

Recent years have seen multiple vision based approaches towards tasks such as heart rate estimation. Approaches known as remote-PhotoPlethysmography (rPPG) try to estimate the heart rate of a person from analysis of the light reflected off the skin's surface. These methods specifically try to separate out the pulsatile blood volume signal, which is then cleaned to derive the heart rate estimate. For rPPG methods, a robust face tracking method is necessary for accurate results, as the same points in the face have to be examined over the temporal dimension to infer the pulsatile signal. For example,the Bounded Kalman Filter (BKF) method for heart rate estimation [36], uses a bounded Kalman filter to track a set of 5 vertical points in each of the 3 empirically selected regions of interest such as the cheeks and the forehead. While the method shows promising results and has been deployed to work on in-the-wild dataset, it suffers in low lighting, and videos with high face motions. Hence, we augment the method by replacing the existing frame by frame landmark estimation method with the ZFace tracker[27].

This modified ZFace-BKF method, also allows for the heart rate estimation from a larger number of tracked 3D points, as opposed to just 15 points over the whole face. The ZFace tracker along with a skin segmentation method [10], recognizes the 3D tracked points which correspond to facial hair and ignores them for the heart rate estimation. Once the BKF method has calibrated itself using a fixed number of frames, the average hue of all the valid 3D points is calculated by projecting the pixel values to the HSV color space, and stored in an array. These aggregated values, are selected by a moving window of a fixed size (typically 30s long), normalized and a fast Independent Component Analysis (ICA) conducted on it to separate the signals in the normalized values. This signal is then de-trended, a butter-worth bandpass filter applied , and it's max power spectrum value calculated using a fast fourier transform (FFT). This is then output as the heart rate estimate for the frame, as shown in Figure 6.5. By a sliding window technique, a per frame heart-rate can be calculated after the initial number of calibration frames.

When using the original videos, the headpose and expression changes along with lighting is a major factor in the quality of tracking and therefore the derived heart rate estimate. With the use of normalization techniques such as 2D normalization implemented by AFAR toolbox, or our PRNet based 3D normalization, the tracking is much less complex due to the reduced degrees of freedom.

6.2.3 MMSE-HR Dataset

In order to compare the 2 normalization techniques against each other, we use the MMSE-HR dataset which is subset of the much larger BP4D+ dataset[53] containing both physiological signals, as well as 3D scans and thermal video of the subjects. The MMSE-HR dataset has become widely used for benchmarking heart rate estimation methods. The dataset, contains ground truth physiological signals such as heart rate, respiration rate and blood pressure collected using a Biopac MP150 data acquisition system. It contains a total of 40 subjects divided into 2 subsets, called the 'first10' and the 'rest30' which are aptly named after the number of subjects they contain. Each of the subsets are subject independent.

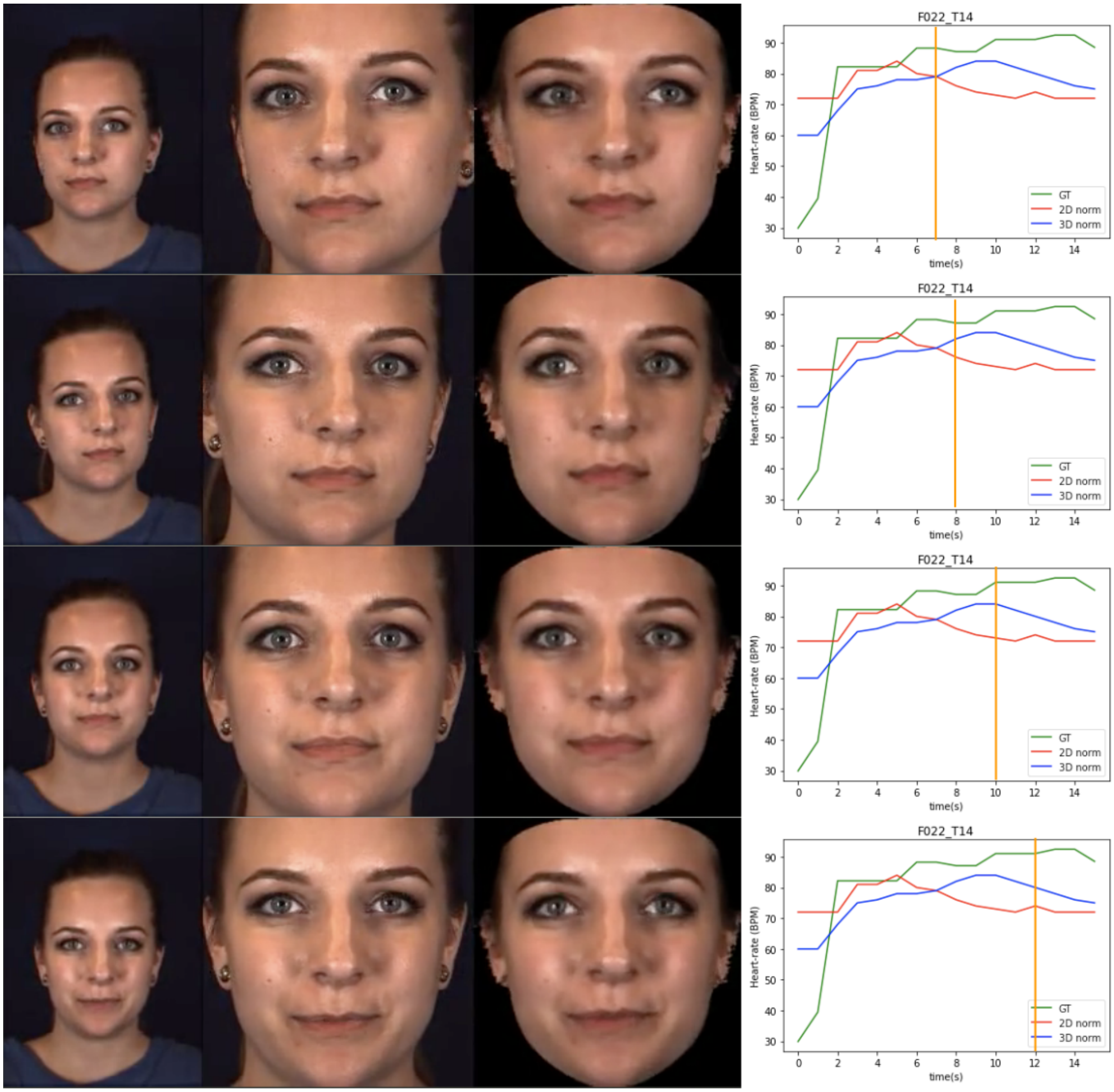


Figure 6.6: The original video frame, 2D normalized video frames, the 3D normalized (ours) video frame are shown in the left, center, and right image columns respectively. The graph on the right shows the heart rate estimate from the ZFace-BKF method for both the normalization techniques and the ground truth at the time the frames are captured given by the vertical yellow bar.

Normalization Technique	mean RMSE with ZFace-BKF[36]	mean RMSE with POS[47]
2D normalization [35]	26.03	33.13
3D normalization w/ PRNet	23.73	35.52

Table 6.1: The mean RMSE scores of the different normalization techniques on the 'first10' subset of the MMSE-HR dataset using 2 different heart-rate estimation methods.

6.2.4 Results

To evaluate the two normalization techniques, we proceed by using 2 different heart-rate estimation methods. For this experiment, both the normalization techniques are applied to the videos in the 'first10' subset of MMSE-HR, containing 40 videos. Then the ZFace-BKF method[36], as well as the older, signal processing based POS method[47] are used for evaluation, and the results compared with the ground truth heart rates from the dataset. Note that there is no additional pre-processing of the heart-rate ground truth than that which is provided. The results of the experiment are shown in Table 6.1.

The 3D normalization seems to work well with the heart-rate estimation method and does well to reduce the loss much more than the 2D normalized data, when compared to the ground truth using the ZFace-BKF method. This shows that 3D normalization based on 3D face reconstruction is a viable alternative to existing 2D normalization techniques. However on older techniques like the POS[47], the effect of lighting and resolution still leaves much to be improved in the quality of the 3D face normalization as it does not outperform 2D normalization. It is also difficult to separate the factors that are affecting these heart rate estimation methods in their implementation which leads to the variations in the results displayed, as each method makes different assumptions. Nevertheless, 3D face normalization will be a great boon for more qualitative tasks that use face shape and expression, and are less effected by re-rendering noise etc.

An example of the two normalization techniques with their synchronized heart rates estimates from ZFace-BKF can be seen in Figure 6.6. The heart rate estimate at the beginning of the video is more accurate with the 2D normalization than with the 3D normalization. At this stage in the video (0-7s), the subject is mostly stationary, and hence should have been the easy case for the heart-rate estimation. The difference between the 2D and 3D normalization during this period is most likely due to the higher resolution of the 2D normalized video as compared to the 3D normalized video. Increasing the resolution of the 3D normalized video should improve the heart-rate estimates in the stationary case.

However, at 7s into the video, the subject shakes their head from side to side and followed by tilting the head up and down as shown by the 4 frames picked at the peak of each movements (Figure 6.6). It can be seen that even though the 3D normalized video with the lower resolution was not performing as well as the 2D normalized video in the stationary case, when the subject starts introducing the head poses changes at 7s, the 2D normalized videos starts to negatively correlate with the ground truth. However, even with the reduced resolution, the 3D normalized video follows the ground truth closely as the head pose changes have significantly less effect on the 3D normalized video. Furthermore, the 3D normalized video only starts to diverge from the ground truth when the subject starts undergoing large facial expression changes (at about 11s), which the 3D normalization cannot yet control. Additionally, while 3D normalizations as shown in the Figure 6.6 can remove head orientation changes from the video, the illumination on the face (which is a function of the head pose and lighting) will still change, and cannot be controlled by 3D normalization. In fact, techniques to integrate these additional factors to 3D face normalization, provide good avenues for further research on the topic.

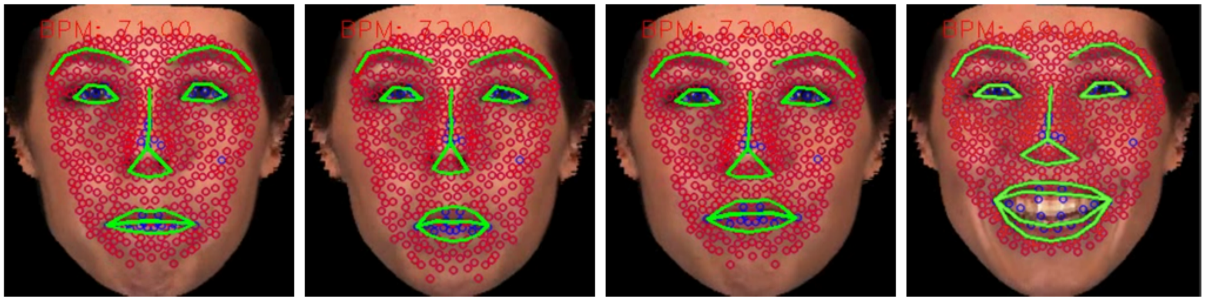


Figure 6.7: Selected frames from heart rate estimation on 3D normalized video, showing the ability of ZFace to track changes in facial expression. The circles show the the points on the skin being tracked in 3D, and only the red colored points are used for the heart-rate estimation procedure.

One of the other limitations of 3D face normalization is the large amount of computation that is required to run 3D reconstruction on every frame of every video in a dataset. While this can be done offline without issues, for tasks such as heart-rate detection which would eventually want to be run online, such a technique is not fast enough as of yet. However, newer single-image reconstruction methods such as 3DFFAv2[21] are able to run at much higher frame rates and closer to real-time. Therefore, in the upcoming years the re-rendering based 3D face normalization can possibly made more computationally feasible.

The combination of a good face tracker and data normalization such 3D face normalization shows promise in estimating more accurate bio-signals from videos, as opposed to letting the data be fully unconstrained. Figure 6.7, shows the ZFace track on the 3D normalizations, where the tracker now only needs to account for the non-rigid deformations of the face due to expression changes. While this standard for 2D normalized videos, note that it's redundant for the 3D normalized videos as the PRNet reconstructions provide the ability to track the deformations of the mesh over time. This also has the advantage of a denser 3D track than ZFace. Hence, a two pronged approach, with 3D face normalization and mesh deformation tracking would allow for possible improvements in tasks such as heart-rate estimation and action unit detection.

Chapter 7

Conclusion

The field of 3D face reconstruction is always evolving due to new advancements in deep learning architectures, and more efficient methods created to run under constrained resources etc. The smartphone and mobile applications have pushed human-centric computer vision tasks such as 3D face reconstructions in the recent past. With the world slowly embracing the virtual space as a real dimension of their lives, many tasks such as online shopping, or consulting a doctor will continue to move further into the virtual space and thereby introduce new applications for human-centric vision.

In this work we introduced the cosine fusion method that utilizes multiple single-image reconstructions to estimate a more accurate multi-frame consistent reconstruction. While the method outperforms the base single-image reconstruction method it is based on, we plan to take the idea further by jointly regressing a displacement map to modify the reconstruction geometry to recover the local details that are currently missing. We also released the 3DFAW-Video benchmarking dataset based on 3D ground truth data to help accurate evaluation of 3D face reconstruction methods. However, more importantly we applied the technique of 3D face reconstructions to real in-the-wild applications such as mask sizing and provided an application for 3D reconstruction based normalization, more directly in the tele-health space. In future work, we aim to improve the quality of our reconstructions by introducing deep learning for fine detail estimation, as well as look into new applications of 3D reconstruction for deployed, computer vision based systems in-the-wild.

Bibliography

[1] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2020. (document), 2.3.1, 4.4, 4.5

[2] Shubham Agrawal, Anuj Pahuja, and Simon Lucey. High accuracy face geometry capture using a smartphone video. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 81–90, 2020. 2.2.1

[3] Andrew D Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2d/3d hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80, 2011. **??**, 3.1

[4] Anil Bas, Patrik Huber, William AP Smith, Muhammad Awais, and Josef Kittler. 3d morphable models as spatial transformer networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 904–912, 2017. 6.2.1

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2, 2.1

[6] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 2.1, **??**, 3.1

[7] Zhang Chen, Yu Ji, Mingyuan Zhou, Sing Bing Kang, and Jingyi Yu. 3d face reconstruction using color photometric stereo with uncalibrated near point lights. In *2020 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2020. 2.2.2

[8] Nikolai Chinaev, Alexander Chigorin, and Ivan Laptev. Mobileface: 3d face reconstruction with efficient cnn regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2.3.1

[9] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, volume 2008, pages 129–136, 2008. 3.2.2

[10] Djamila Dahmani, Mehdi Cheref, and Slimane Larabi. Zero-sum game theory model for segmenting skin regions. *Image and Vision Computing*, page 103925, 2020. 6.2.2

[11] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d

face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3.2.3, 5.1

[12] Pengfei Dou and Ioannis A Kakadiaris. Multi-view 3d face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80:80–91, 2018. 2.3.2

[13] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, 2017. 2.3.1

[14] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 2.1

[15] Nathan Faggian, Andrew Paplinski, and Jamie Sherrah. 3d morphable model fitting from multiple views. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6. IEEE, 2008. 2.3.2

[16] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi 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. 2.3.1, 3.3, 4.2

[17] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 780–786. IEEE, 2018. **??**, 3.1

[18] Douglas Fidaleo and Gérard Medioni. Model-assisted 3d face reconstruction from video. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 124–138. Springer, 2007. 2.2.1

[19] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2.1

[20] Paulo FU Gotardo, Tomas Simon, Yaser Sheikh, and Iain Matthews. Photogeometric scene flow for high-detail dynamic 3d reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 846–854, 2015. 2.2.1

[21] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. *arXiv preprint arXiv:2009.09960*, 2020. 2.3.1, 6.2.4

[22] Christopher Ham, Simon Lucey, and Surya Singh. Hand waving away scale. In *European conference on computer vision*, pages 279–293. Springer, 2014. 6.1.2, 6.1.2

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30:6626–6637, 2017. 5.3

[24] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 6.1.1

[25] *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. IEEE. 1.2.2, 2.1

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 6.2.1

[27] László A Jeni, Jeffrey F Cohn, and Takeo Kanade. Dense 3d face alignment from 2d videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 1.4, 2.3.1, 6.1.1, 6.2.2

[28] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016. 3.1

[29] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):29, 2013. 3.2.2

[30] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 3.2.3

[31] Rohith Krishnan Pillai, Laszlo Attila Jeni, Huiyuan Yang, Zheng Zhang, Lijun Yin, and Jeffrey F Cohn. The 2nd 3d face alignment in the wild challenge (3dfaw-video): Dense reconstruction from video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. **??**, 3.3

[32] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL `https://doi.org/10.1145/3130800.3130813`. 2.1

[33] Shu Liang, Linda G Shapiro, and Ira Kemelmacher-Shlizerman. Head reconstruction from internet photos. In *European Conference on Computer Vision*, pages 360–374. Springer, 2016. 2.2.2

[34] Yuping Lin, Gérard Medioni, and Jongmoo Choi. Accurate 3d face reconstruction from weakly calibrated wide baseline images with profile contours. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1490–1497. IEEE, 2010. 2.2.1

[35] Itir Onal Ertugrul, László A Jeni, Wanqiao Ding, and Jeffrey F Cohn. Afar: A deep learning based tool for automated facial affect recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019. 6.2.1, **??**

[36] Sakthi Kumar Arul Prakash and Conrad S Tucker. Bounded kalman filter method for motion-robust, non-contact heart rate estimation. *Biomedical optics express*, 9(2):873–897, 2018. 6.2.2, **??**, 6.2.4

[37] Chengchao Qu, Eduardo Monari, Tobias Schuchert, and Jürgen Beyerer. Fast, robust and automatic 3d face model reconstruction from videos. In *2014 11th IEEE International*

Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 113–118. IEEE, 2014. 2.3.2

[38] Eduard Ramon, Janna Escur, and Xavier Giró-i Nieto. Multi-view 3d face reconstruction in the wild using siamese networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3.3, **??**

[39] Paulien Roos, Katherine Marschner, Laszlo Jeni, Rohith Krishnan Pillai, and Vincent Harrand. A mobile phone app for automated and accurate sizing of respirator masks. In *58th Annual SAFE Symposium 2020*, pages 45–46. SAFE Association, 2020. (document), 6.3, 6.1.3

[40] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. **??**, 3.1

[41] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018. 2.3.1

[42] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. *arXiv preprint arXiv:2007.12494*, 2020. 2.3.1

[43] Xiaohu Shao, Jiangjing Lyu, Junliang Xing, Lijun Zhang, Xiaobo Li, Xiangdong Zhou, and Yu Shi. 3d face shape regression from 2d videos with multi-reconstruction and mesh retrieval. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3.3, **??**

[44] Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In *Face and Gesture 2011*, pages 611–618. IEEE, 2011. **??**, 3.1

[45] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 2.3.1

[46] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019. 2.3.2

[47] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64(7):1479–1491, 2016. **??**, 6.2.4

[48] Xueying Wang, Yudong Guo, Bailin Deng, and Juyong Zhang. Lightweight photometric stereo for facial details recovery. In *IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR), 2020. 2.3.2

[49] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3d face morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2019. 2.3.2

[50] Stefanos Zafeiriou, Grigorios G Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. The 3d menpo facial landmark tracking challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2503–2511, 2017. 3.1

[51] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2315–2324, 2019. 2.3.1

[52] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 3.1

[53] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. **??**, 3.1, 6.2.3

[54] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 1.2.3, 2.3.1, 3.1, 3.3

[55] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 78–92, 2017. 3.1