High Accuracy Face Geometry Capture using a Smartphone Video - Supplementary material

A. Fine detail enhancement

Beeler *et al*. [1] proposed using a high-pass filtered version of the texture to emboss a mesh with fine details, such as wrinkled and pores. The recent work of Sela *et al*. [3] proposed using the mesh itself to obtain the high frequency component of the texture, using heat flow to model a low pass filterer version of the texture. In their results, this modification allows them to capture more medium-to-fine scale details, such as the nasolabial folds. However, we observed that their method also tends to distort the mesh, and also pick up on other high frequency noise such as sensor noise that is not desirable. We thus propose to augment our reconstructions as follows:
For a mesh with per vertex texture mapping τ_v, we calculate a low-pass filtered version of the texture as :

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

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

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

$$\mu_v = \tau_{lp} - (M - \Delta t.C)^{-1}.M\tau_v \tag{2}$$

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

$$\vec{\delta_\mu}(v) = ||\mu_v(v) - \mu_m||.\vec{n}(v) \tag{3}$$

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

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

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

Figure 1: Our method naturally generalizes to any face geometry, including deformations caused by expressions.

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

This simplified version of the original approach suggested by [1] allows us to capture details like eye-lids, lip corners and nasolabial-folds in the mesh.

B. Expressions

Our method captures the geometry of the face in a completely data-driven manner, and hence it also generalizes naturally to deformations caused by expressions (Fig 1). Since, is difficult to hold the same expression through a video sequence and then also obtain a corresponding ground truth 3D scan, we skip quantitative evaluation of this. We also note that since there is dense correspondence and consistent topology across meshes, various existing techniques like blendshapes [2] can be applied with our reconstructed meshes to generate animated, expressive face models.

C. Dataset

We found that there is a lack of datasets containing multi-view sequences of faces with consistent geometry in "in-the-wild" settings. To this end, we have constructed our own dataset of 200 sequences of 100 individuals. Each video sequence is shot from an iPhone X, at 1920x1080 resolution and 120fps. Each video sequence is 15-20 seconds long, containing a profile-to-profile sweep of the subject's face. We acquire 2 sequences of the same individual with different background and lighting conditions. For a subset of the dataset, we acquire high accuracy ground truth to serve as validation for testing various methods. For the

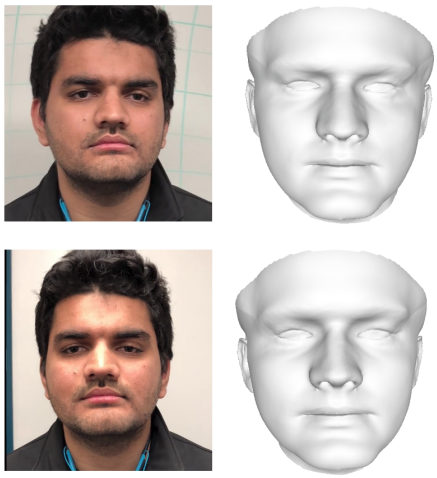

Figure 2: For each subject, we record two video sequences under different lighting and background. For the subject's where ground truth is not available, we self-validate the two reconstructed meshes to be consistent, within a small tolerance.

remaining sequences, we provide our reconstructions as reference, where the meshes are validated to be self-consistent between two sequences of the same subject (Fig 2). While a lot of work has been done on learning-based single view face reconstruction, we wish to also encourage better multi-view methods for face reconstructions with this dataset. In particular, we hope this will encourage self-supervised techniques, where the reconstruction should be consistent across views and sequences.

References

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

[2] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

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