3D Face Reconstruction with Dense Landmarks

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan Garbin, Chirag Raman, Jamie Shotton, Toby Sharp, Ivan Stojiljkovic, Thomas J. Cashman, Julien Valentin

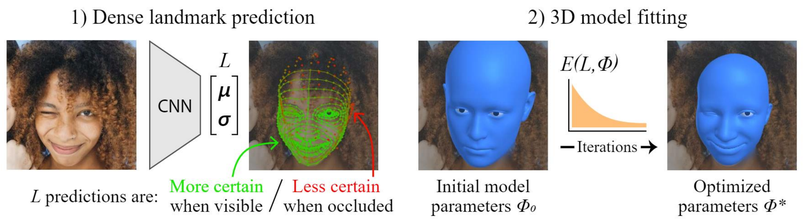
ECCV TEL AVIV 2022

Introduction

We present the first method that accurately predicts ten times as many landmarks as usual, covering the whole head, including the eyes and teeth. This is accomplished using synthetic training data, which guarantees perfect landmark annotations. By fitting a morphable model to these dense landmarks, we achieve state-of-the-art results for monocular 3D face reconstruction in the wild.

We show that dense landmarks are an ideal signal for integrating face shape information across frames by demonstrating accurate and expressive facial performance capture in both monocular and multi-view scenarios. Finally, our method is highly efficient: we can predict dense landmarks and fit our 3D face model at over 150FPS on a single CPU thread.

Method



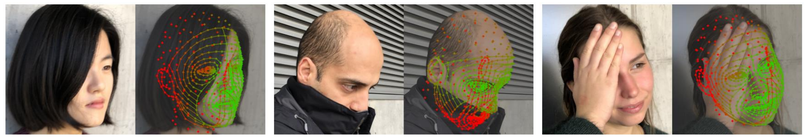
Given an image, we first predict probabilistic dense landmarks L, each with position μ and certainty σ. Then, we fit our 3D face model to L, minimizing an energy E by optimizing model parameters Φ.

While a human might consistently label images with 68 landmarks, manually annotating images with over 700 dense landmarks would be impossible. Instead, we rendered 100,000 synthetic training images using our face synthetics system. Without the perfect annotations provided by synthetic data, dense landmark prediction would not be possible.

We predict each landmark as a random variable with the PDF of a circular 2D Gaussian. This is achieved by training the CNN with a Gaussian negative log likelihood (GNLL) loss

$$\text{Loss}(L) = \sum_{i=1}^{|L|} \lambda_i \left(\underbrace{\log\left(\sigma_i^2\right)}_{\text{Loss}_\sigma} + \underbrace{\frac{\|\mu_i - \mu_i'\|^2}{2\sigma_i^2}}_{\text{Loss}_\mu} \right)$$

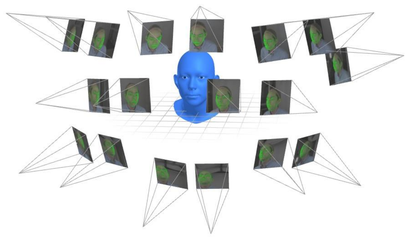
When parts of the face are occluded by e.g. hair or clothing, the corresponding landmarks are predicted with high uncertainty (red), compared to those visible (green).



Given probabilistic dense 2D landmarks L, our goal is to find optimal model parameters $\boldsymbol{\Phi}$ that minimize the following energy E. $E_{landmarks}$ is the only term that encourages the 3D model to explain the observed 2D landmarks. The other terms use prior knowledge to regularize the fit.

$$E(\Phi; L) = \underbrace{E_{\text{landmarks}}}_{\text{Data term}} + \underbrace{E_{\text{identity}} + E_{\text{expression}} + E_{\text{joints}} + E_{\text{temporal}} + E_{\text{intersect}}}_{\text{Regularizers}} \qquad E_{\text{landmarks}} = \sum_{i,j,k}^{F,C,|L|} \frac{\|\mathbf{x}_{ijk} - \boldsymbol{\mu}_{ijk}\|^2}{2\sigma_{ijk}^2}$$

We implemented two versions of our approach: one for processing multi-view recordings offline, and one for real-time facial performance capture on resource-constrained devices. The formulation of our method naturally scales to multiple images and cameras.



In the real-time case, we use the Levenberg-Marquardt algorithm to optimize our model-fitting energy. Camera and identity parameters are only fit occasionally, for the majority of frames we fit pose and expression parameters only. Running on a single CPU thread (i5-11600K),our real-time system spends 6.5ms processing a frame (150FPS), of which 4.1ms is spent predicting dense landmarks and 2.3ms is spent fitting our face model.

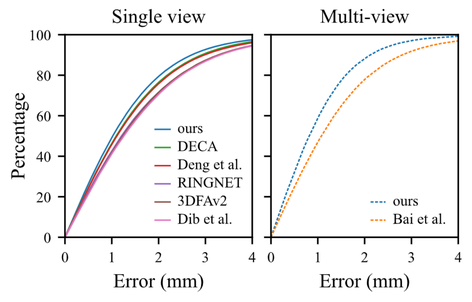
Results

Our method is able to reliably and robustly achieve accurate and expressive results in both multi-view and single-view scenarios.

Method Single view	Error (mm), mean Coop.	 Indoor	 Outdoor
Tran et al.	1.97	2.03	1.93
Genova et al.	1.78	1.78	1.76
Deng et al.	1.66	1.66	1.69
ours	**1.64**	**1.62**	**1.61**

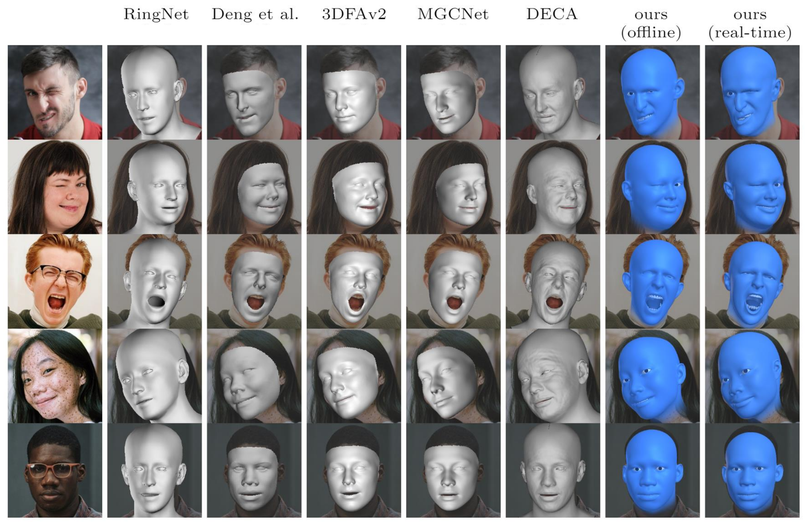
Method Multi-view	Error (mm), mean Coop.	 Indoor	 Outdoor
Piotraschke and Blanz	1.68	1.67	1.72
Deng et al.	1.60	1.61	1.63
ours	**1.43**	**1.42**	**1.42**

We evaluate our method against to MICC dataset in two ways: single view, where we estimate one face shape per frame in a video, and average the resulting face meshes, and multi-view, where we fit a single face model to all frames in a video jointly.



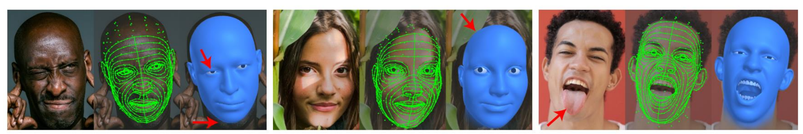
Method Single view	Error (mm) Median	 Mean	 Std
Deng et al.	1.11	1.41	1.21
RingNet	1.21	1.53	1.31
3DFAv2	1.23	1.57	1.39
DECA	1.09	1.38	1.18
Dib et al.	1.26	1.57	1.31
ours	**1.02**	**1.28**	**1.08**
Multi-view			
Bai et al.	1.08	1.35	1.15
ours	**0.81**	**1.01**	**0.84**

We also undertake the NoW challenge for measuring the accuracy and robustness of 3D face reconstruction in the wild in two ways: single view, where we fit our face model to each image separately, and multi-view, where we fit a per-subject face model to all images of a particular subject.



Compared qualitatively to previous recent monocular 3D face reconstruction methods, ours better captures gaze, expressions like winks and sneers, and subtleties of facial identity. In addition, our method can run in real time with only a minor loss of fidelity.

Limitations



Our method depends entirely on accurate landmarks; bad landmarks result in bad fits. Additionally, our face model doesn't include tongue articulation, so we are currently incapable of tracking the tongue. We intend to address these issues by improving our synthetic data and adding tongue articulation.

For more information see our project website by scanning the QR code or going to https://microsoft.github.io/DenseLandmarks/

Work conducted at the Microsoft Mixed Reality & AI Lab Cambridge.

Thanks to Jiaolong Yang and Timo Bolkart for help with evaluation.