Deep Energies for Estimating Three-Dimensional Facial Pose and Expression

Jane Wu (janehwu@),[1] Xinwei Yao (xinweiyao@)[2]

[1]*CS230: Deep Learning*

[2]*CS229: Machine Learning*

Overview

We tackle the problem of face performance capture.

Inputs:

- Facial performance of an actor captured by a multi-camera rig
- A blendshape rig for the actor where combinations of blendshape weights control facial expressions

Goal: for each captured frame, solve for

- Rigid transformation parameters: translation, rotation
- Weights for jaw/mouth related blendshapes

So that rendered frames of the 3D face model reproduce the captured performance.

Methods

We propose a general strategy that incorporates pre-trained deep neural networks into classical optimization methods. More concretely, we construct an end-to-end differentiable function from the rigid and blendshape parameters to an energy based on deep feature outputs of neural networks.

Let w be the blendshape weights, our 3D model defines a differentiable function x(w) for the triangulated surface of the face. Then given Euler angles θ, its rotation matrix R(θ) and translation vector t, the final vertex positions are

$$x_R(\theta, t, w) = R(\theta)x(w) + t.$$

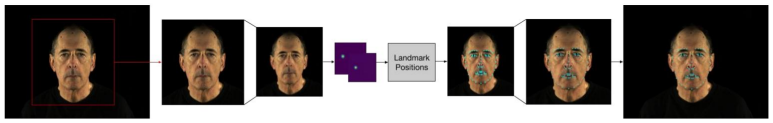
We use OpenDR [1] to obtain differentiably a rendered image $F(x_R)$. Now given the captured image F^*, we can send both F and F^* through the network N to obtain the deep features. We then define the energy function to be the L_2 norm of the difference between the deep features.

$$\|N(F^*) - N(F(x_R(\theta, t, w)))\|_2$$

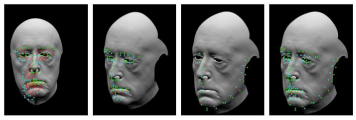
Since the blendshape rig, the renderer and the network are all differentiable functions, we are able to compute the jacobian of the energy with respect to the parameters to solve for, we can now minimize the energy function using classical nonlinear least squares method Dogleg.

We use 3D-FAN [2] for facial landmark detection and Flownet2 [3] for optical flow.

Facial Landmark Detection



Rigid Alignment



Expression Estimation

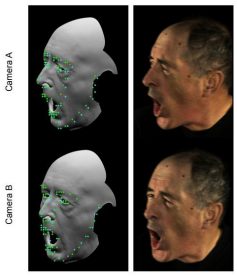
Results

We estimate the facial pose and expression on a moderately challenging performance captured by a single ARRI Alexa XT Studio running at 24 frames-per-second with an 180 degree shutter angle at ISO 800 where numerous captured images exhibit motion blur. These images are captured at a resolution of 2880 x 2160, but we downsample them to 720 x 540 before feeding them through our pipeline.

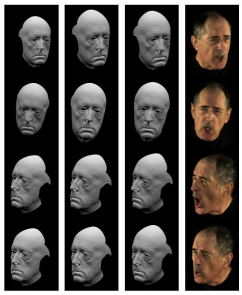
Our approach can trivially be extended to multiple calibrated camera viewpoints as it only entails adding another duplicate set of energy terms to the nonlinear least squares objective function.

We further compare our method to rigid alignment computed manually by a skilled artist.

Stereo Estimation



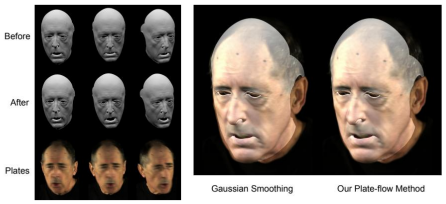
Comparison to Manual Alignment



Optical Flow Refinement/Infill

Refinement

Deep Energy: $\|N(F_2, F_1) - N(F_2^*, F_1^*)\|_2^2 + \|N(F_2, F_3) - N(F_2^*, F_3^*)\|_2^2$



Infill

Fill in frames in the middle with forward pass followed by backward pass

Future Work

While we have only explored using pre-trained facial alignment and optical flow networks, using other types of networks (*e.g.* face segmentation, face recognition, etc.) and using networks trained specifically on the vast repository of data from decades of visual effects work are exciting avenues for future work.

Acknowledgements

We would like to thank Michael Bao and Prof. Ron Fedkiw for their guidance on this project.

References:

[1] M. M. Loper and M. J. Black. Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision*, pages 154–169. Springer, 2014.

[2] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, volume 1, page 4, 2017.

[3] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.