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.

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

$$x_{\theta}(R(\theta), t) = R(\theta)x + t.$$

We use OpenBLR [1] to obtain differentially a rendered image $F_{\theta}(x)$, then given the captured image $F$, we can solve both $\theta$ and $t$ through the network $R$ to obtain the deep features. We then define the energy function to be the $L_2$ norm of the difference between the deep fractions,

$$\frac{1}{2} \left\| F_{\theta}(x) - F(x) \right\|^2_2.$$

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 estimate the facial pose and expression for 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 400 where numerous captured images exhibit motion blur. These images are captured at a resolution of 2880 x 2304, but we downsample them to 720 x 540 before feeding them through our pipeline.

Our approach can totally 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.

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: