Face Inverse Rendering: Morphable Muscle Models

Cole Sohn,[1] Winnie Lin,[2]

[1]Ron Fedkiw's Lab, CURIS, Stanford University
[2]Ron Fedkiw's Lab, Computer Graphics, Stanford University

Stanford
Computer Graphics

Abstract

The Fedkiw lab is attempting to create 3D facial animation from single–directional video for applications such as face replacements or full 3D animation. Work pioneered by Matthew Cong and Michael Bao lead to detailed differentiable muscle models that can be used to drive surface meshes to match video through inverse rendering in an anatomically accurate way. The main problem with the lab's current pipeline is a bottleneck in speed when deforming muscle models to match artist-defined blendshapes. One solution is to learn muscle volumes and positions from data produced by the simulation to circumvent the need to run the full muscle deformation process. This requires large datasets to run through the simulation and generate training data. This is difficult to achieve through traditional artist-directed means. This leads to the desire to find or generate large amounts of input data to run through the muscle simulation through other means such as existing models that take camera stills as input to generate expressive face meshes.

Stanford
University

Face Inverse Rendering

A goal of the Fedkiw lab through the work of Matthew Cong, Michael Bao, and many others is to simulate accurate facial movement to match recorded video and capture lifelike animation. Cong's work lead to a morphable template muscle model that could fit detailed volumetric muscle and bone meshes to surface face meshes. This could be used to create realistic facial movement based on human anatomy. Cong then worked to make this model more expressive by deforming to artist-defined blendshapes while maintaining physical accuracy using a system of muscle tracks and zero-length springs to drag muscles to artist-desired blendshape locations.

Bao and his team worked to make Cong's morphable muscle model differentiable for optimization and learning as well as driven by blendshapes. This allows the body of work to be compatible with many applications such as inverse rendering to approximate an animated model based on video. The culmination of this work is as follows:

$$
\begin{aligned}
&j, w \rightarrow M(j, w), C(j, w) \\
&\quad \rightarrow X(M, C) \\
&\quad \rightarrow S(X) \\
&\quad \rightarrow I(S, P, R)
\end{aligned}
$$

j and *w* represent jaw skinning and blendshape coefficients

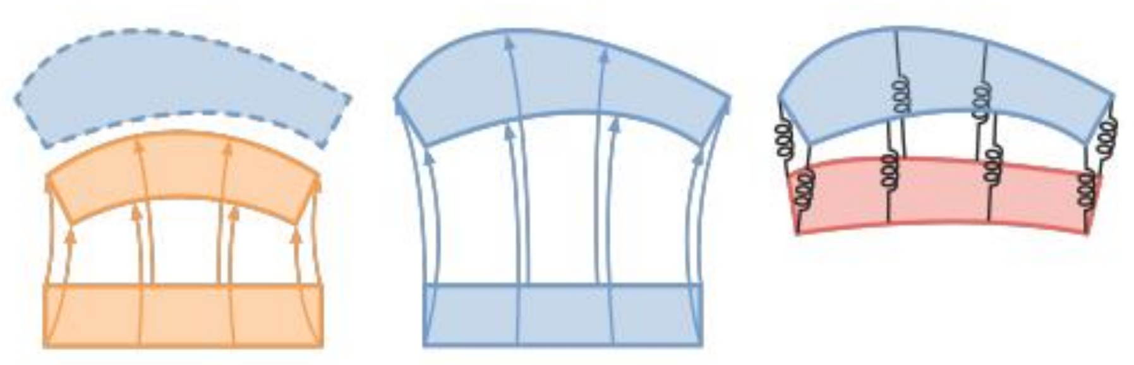
M and *C* are muscle shapes and centerline curves

X and *S* are tet flesh meshes and surfaces

I is a final render, *P* is posing, and *R* is materials, lighting, etc.

Current Problems

The system of simulating muscles *(M, C)* from blendshape coefficients *(j, w)* is complex and slow. A solution is to learn M and C from j and w instead of continuously performing complex spring calculations by generating large amounts of data at various blendshape coefficients (values of w). A subsequent problem that arises is the need for large numbers of models at various blendshape coefficients as input to the simulation for training data generation. Bao's differentiable pipeline requires inputs of blendshapes adhering to the facial action coding system (FACS). Traditionally these would be created by artists, but this is impractical for the required volume of data for learning muscle volumes and positions. Solutions are being looked at for generating large number of sim inputs and many rely on methods of using or generating large input datasets such as the CoMA dataset from Michael Black. Large amounts of input data can be produced are through various published models that predict face shape and expression based on images inputs. When run on images of people producing expressions aligned with FACS, the necessary input data for Cong's simulation can be rapidly generated.



An illustration from Cong et al. illustrating process of "dragging" flesh tet mesh muscle to artist-driven blendshapes with zero length springs

Facial Action Coding System

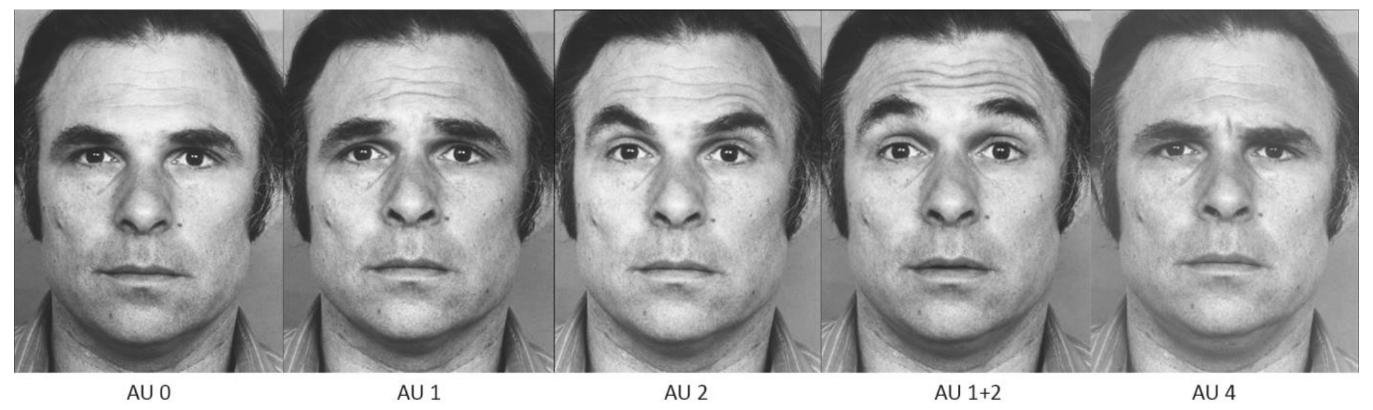
FACS was developed in 1970 by Carl-Herman Hjortsjö and Paul Elkman in 1978 to measure micro-expressions and categorize muscle movement.

Different Action Units (AUs) describe different muscle movements and can be combined to create and match expressions.

FACS is used by 3D artists when rigging and creating blendshape systems for facial animation to create dials and controls to influence overall facial expression.

Facial movement must account for attachment points, muscle volume, compression and tension, etc. that make it difficult to control with a traditional bone rig.

AU 0	Neutral
AU 1	Inner corners of eyebrows lifted
AU 2	Outer eyebrow raised
AU 1+2	Entire eyebrow raised
AU 4	Eyebrows pulled together



FACS can be viewed as a compromise between objectivity and human readability.

ExpressionNet and FaceScape

ExpNet: Bundle of ExpNet with FacePoseNet from Chang et al., and 3DMM face identity shape network from Tran et al.: predicts expression, pose, and shape respectively and morphs a template model.

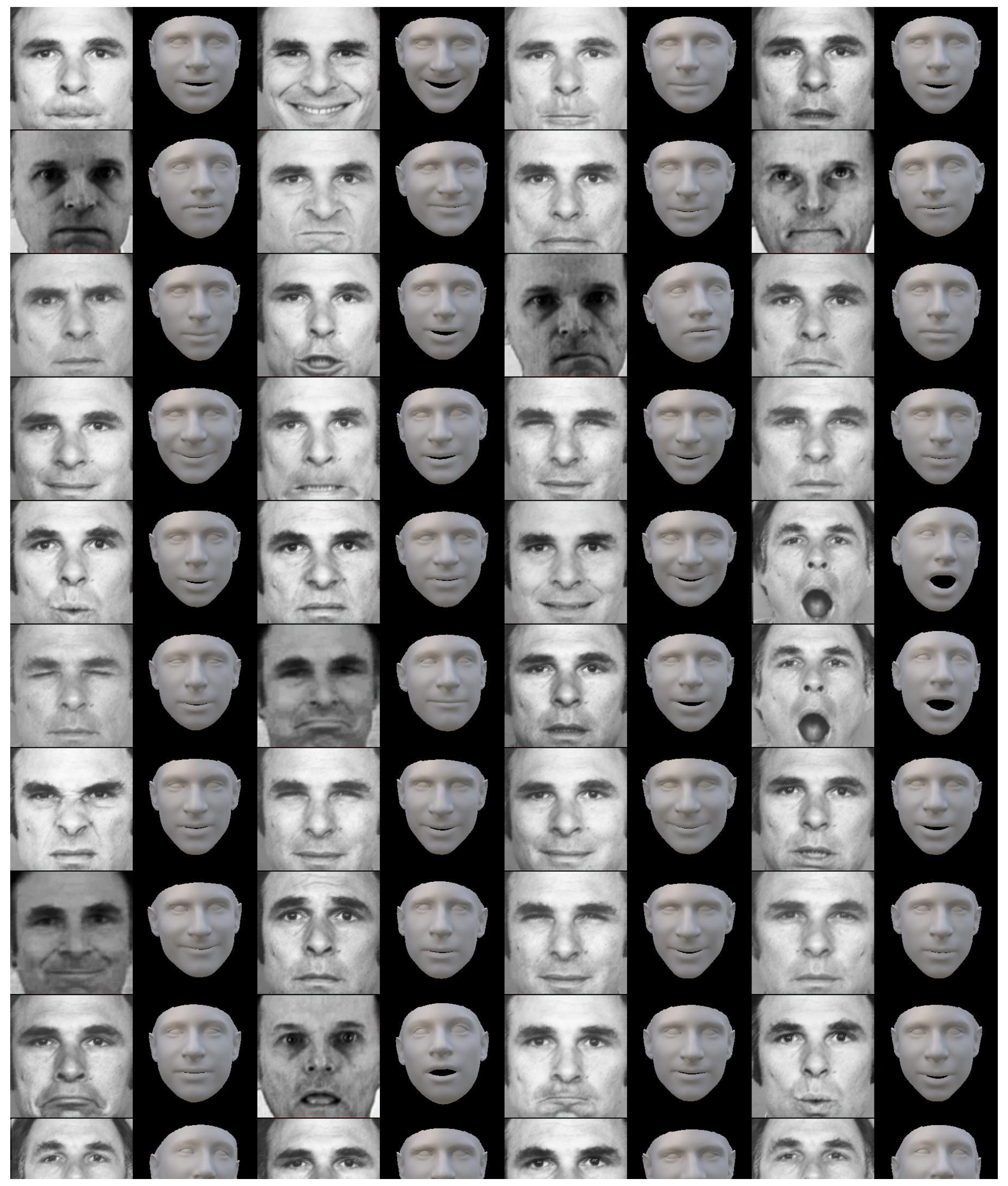


Image Bounding Box Detection: OpenCV
ply->obj: Blender Script
Batch render with Houdini
Image zoom and grid with PILLOW

FaceScape: Requires more complex multi-view data but can simulate highly accurate face shape and expressions. Dataset provides 938 subject faces with FACS based blendshapes. Dataset not yet widely released but can be harnessed for training data for learning improve speed of muscle mesh morphing.