Facial Performance Sensing Head-Mounted Display

1 Implementation Details

Data Generation and Preparation. As described in the paper, we record each training subject performing a variety of FACS-based emotional expressions and reciting a set of 30 Harvard sentences while being recorded by our mouth camera. We also record subjects performing a variety of eye expressions that are recorded using the internal infrared camera. The FACS sequences and eye sequences for each subject are animated manually by artists. However, we have the Harvard sentences from only one subject animated by artists, and use Dynamic Time Warping on the audio signal from the other subjects' sentences to align them to the reference subject's sentences, and thus obtain the per-frame blendshape parameters for each subject's speech samples.

Animating the entire sequence of lower face FACS expressions (21 lower face actions) performed by each of the 10 subjects (approximately 1900 frames per subject) took an artist an average of 3 hours. The 30 training sentences spoken by the reference subject (approximately 84 frames each) were more complex to animate, taking on average 2 hours and 20 minutes each for one artist to animate. This motivated our use of Dynamic Time Warping on the speech samples to make the data generation process more efficient.

Dynamic Time Warping between the reference subject's animation sequences and the other subjects' videos was performed in Matlab (R2015b) on a 2014 Macbook Pro (2.8 GHz Intel Core i7, 16 GB RAM). Aligning these 270 speech videos to the reference animation took approximately 14 minutes.

This process gives us a total of 44,226 images (25,370 speech samples and 18,856 FACS samples) with corresponding blendshape coefficients for training our mouth expression regressor. We apply data augmentation to these images by applying random translations (-20 to +20 pixels in the x and y dimensions), rotations (-10 to +10 degrees), scalings (by a factor of 0.8 to 1.2), and changes to the image intensity (by a factor of 0.85 to 1.15 for each channel), which makes our data set more robust to variations in camera placement (which is affected by the placement of the HMD on the user's head), user identity, and lighting conditions. We perform this augmentation 20 times per each original image, giving us a total training set of 884,520 images.

Our neutral face network is trained using 18,526 distinct frames of our training subjects recorded by our mouth camera, labeled as either neutral or non-neutral by Amazon Mechanical Turk users. Using data augmentation as described above produced a total training set of 185,260 images, which we found to be sufficient in our tests.

We use the single internal IR camera in our prototype device to animate both eyes' movements by mirroring the expressions (although this process could easily be adapted to provide asymmetric eye expressions, given a system with multiple internal IR cameras). As there is less variation in the eye expressions (and they are thus controlled by a much smaller number of blendshapes, 7 compared to the 50 used for the mouth regressor), we found that a smaller number of images sufficed for our eye expression regressor. We use a total of 3,657 images from the IR camera as the subjects perform the full range of expressions and look in various directions. An artist was able to generate blendshape coefficients for these sequences in roughly 5 hours. Applying data augmentation as described above gives us a total of 73,140 training images for our eye expression regressor.

Runtime Performance. We use an Ubuntu 14.04 system to regress the blendshape coefficients using the Caffe framework and stream them via UDP to a system running Windows for rendering. In our benchmarks, we measure the following average times to produce the output for a single frame:

- Neutral face network: 1.3 ms
- Eye expression network: 1.1 ms
- Mouth expression network: 1.5 ms

Once the blendshape coefficients are received by the rendering system, it takes approximately 1.6 ms to apply these coefficients to obtain the new facial expression. The new mesh vertex positions are then uploaded to the GPU for rendering. Our framework is parallelized, and thus receiving blendshape coeffients via UDP, applying the blendshape coefficients, and rendering happen in parallel (while the regression is performed in parallel on the other system). The renderer is the bottleneck in our framework, as waiting for the rendering thread to complete the commands to draw a new frame results in a framerate of 38 fps when streaming images from a pre-recorded video sequence (images can be read from disk at roughly 200 fps and processed at this speed by the other parts of our pipeline). However, our camera captures frames at 30 fps, so our system performs consistently at this framerate when recording sequences or using the system for live demonstrations.

2 Additional Results

Only data from 10 subjects were included in the final training set, though we collected recordings from additional subjects. In our paper and video we demonstrate the results of our system on users both in and out of the training set (on sequences that were not included in the final training data). Examples of subjects not included in the training set can be seen in the supplementary video (2:50-3:20, 3:35-3:52, 5:35-5:57, 6:18-6:46). We also demonstrate the system in use by a subject with substantial facial hair (2:05-2:34), though no facial hair was included in the training set. A demonstration of the system in use in the presence of extreme background noise is also included (3:52-4:04).

The input to the original implementation of Cao et al. 2014 was the entire original image seen in the left column of Figure 1, while our mouth regressor operated on the highlighted region. As their approach operates on a single video of the full face, it can animate eye movements such as blinks. The mouth regressor can operate on images in which the rest of the face is occluded from the camera, but only produces blendshape coefficients for the lower face region. Thus, in our approach eye movements are animated using the eye regressor operating on images from the internal IR camera on our prototype HMD.

Figure 2 shows example results from further evaluations in which both our system and the modified implementation of Cao et al. 2014 were trained using the same training images and reference animation data for the depicted user, as described in the evaluations section of the paper. 18 mouth contour landmarks were labeled in 2,500 images of the depicted subject by Amazon Mechanical Turk users. These images, with the corresponding landmarks and artist-generated per-frame weights for the 50 blendshapes used for the lower mouth, were used to train the Cao et al. 2014 regressor. We applied data augmentation to the same images to obtain the 50,000 images used to train our networks. The sample images show several significant differences between the results attained using each approach on the same input image. Example video sequences can be seen in the supplementary video.

Figure 3 shows several additional examples from live demonstrations of our system in different lighting conditions. Note that our full system is able to track the movements of the upper and lower face as well as the rigid motion of the head.

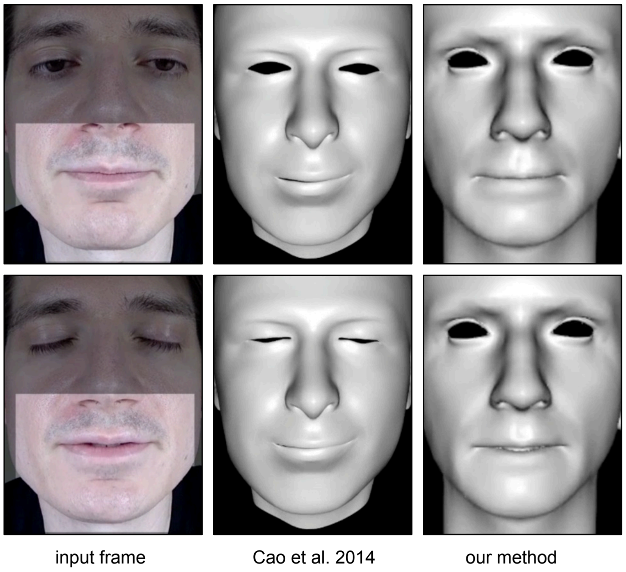


Figure 1: *Further comparison with Cao et al. 2014.*

3 Blendshape Expression Images

Finally, we include renderings of each of the blendshape expressions used in our system. These include images of the 7 FACS-based upper face expressions generated by our eye regressor as well as the 29 speech-based and 21 FACS-based expressions generated by our mouth regressor. (For the blendshapes controlling the tongue we also open the jaw to make the tongue visible.)

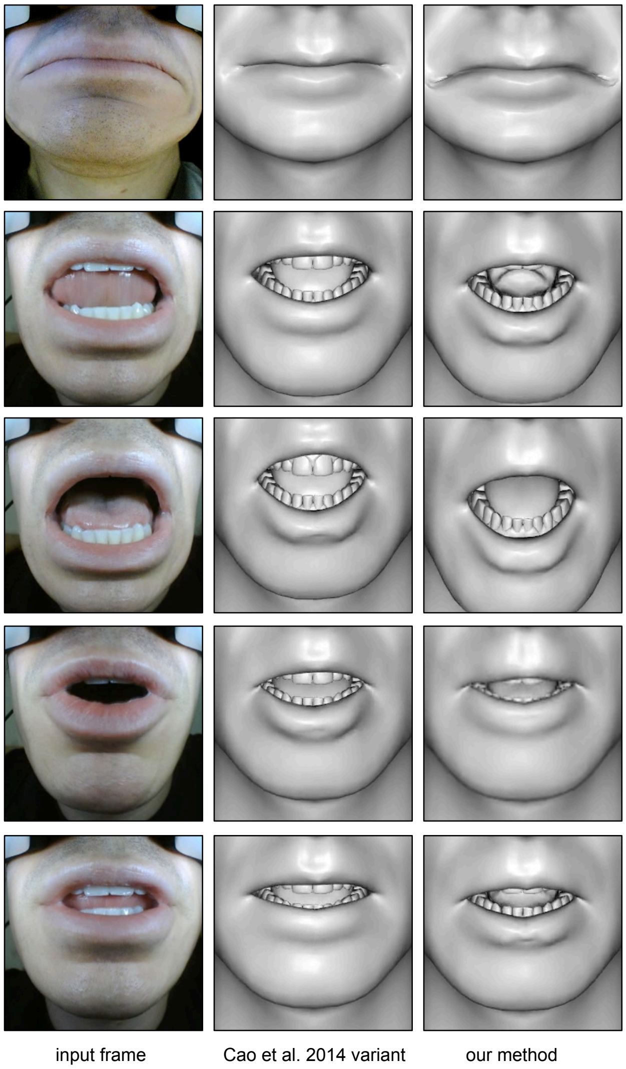


Figure 2: *Comparison with Cao et al. 2014, using the same training images and artist-generated animation data used to train our system.*

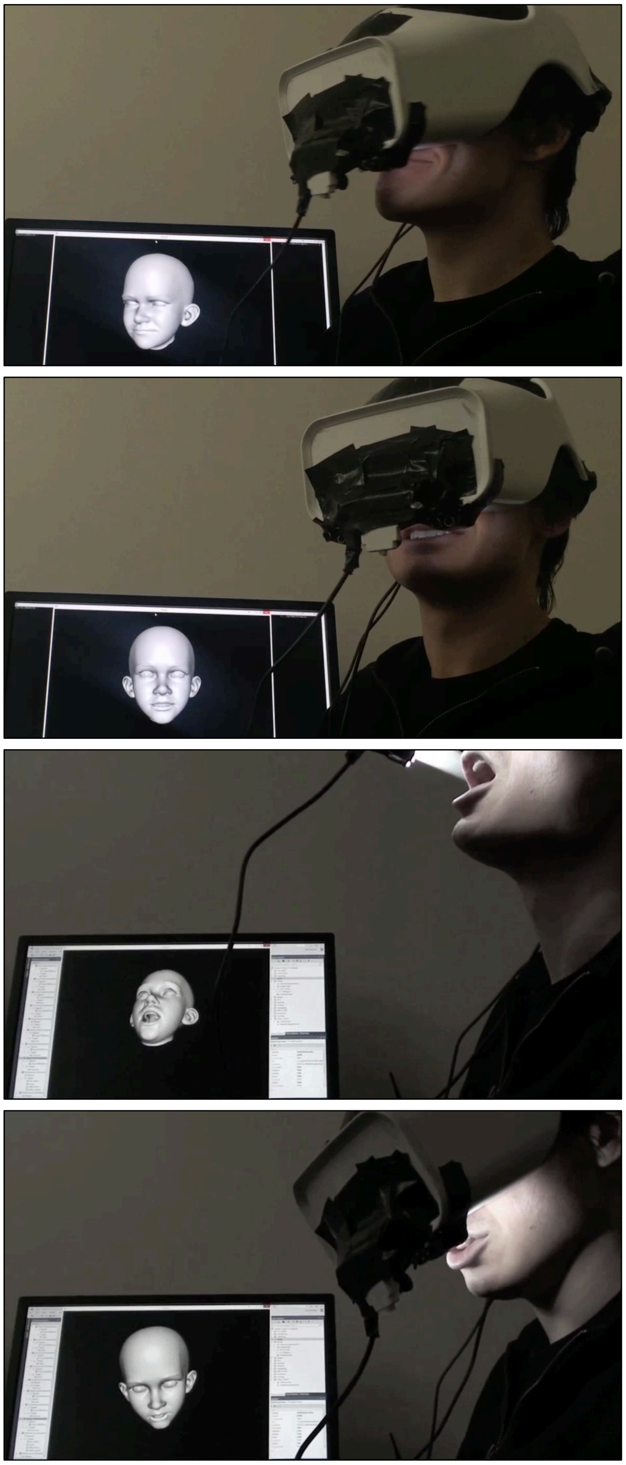
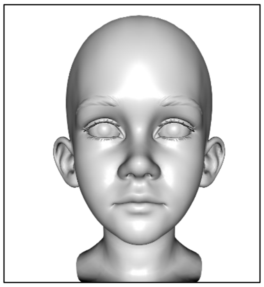
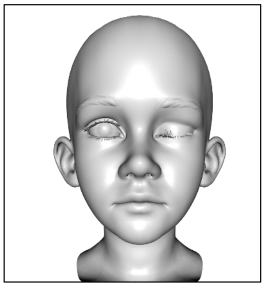


Figure 3: *Images from a live demonstration of our system.*

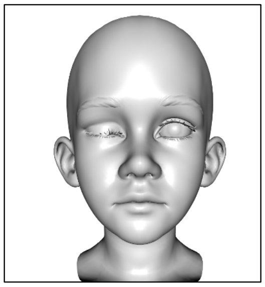


Neutral

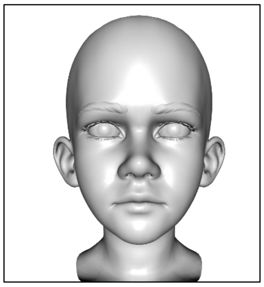
FACS-Based Upper Face Expressions



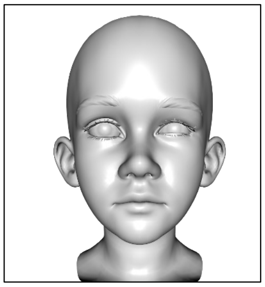
Blink Left



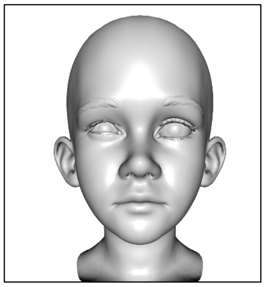
Blink Right



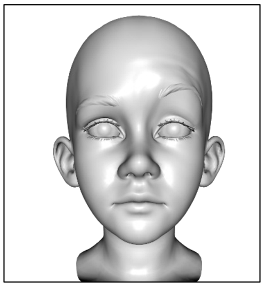
Brow Down



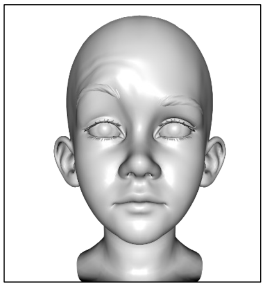
Squint Left



Squint Right

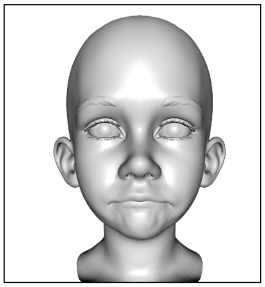


Brow Up Left

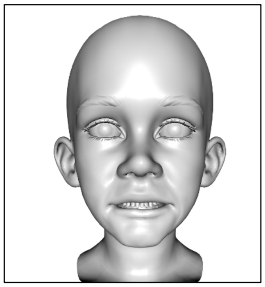


Brow Up Right

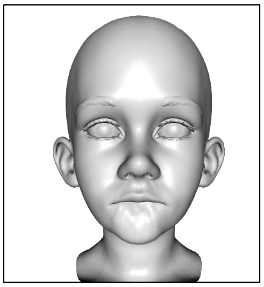
Speech-Based Lower Face Expressions



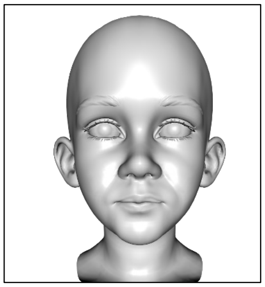
B Sound



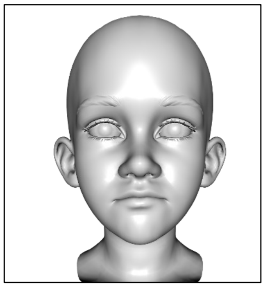
E Sound



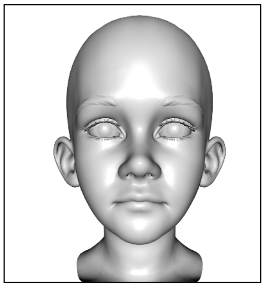
F Sound



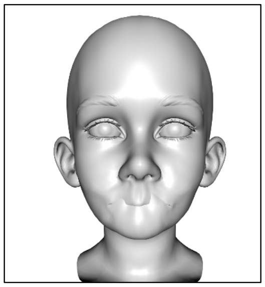
Lip Press



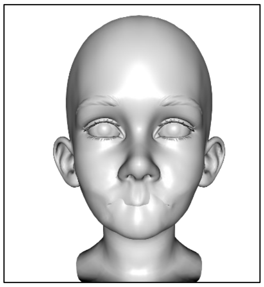
Lip Up



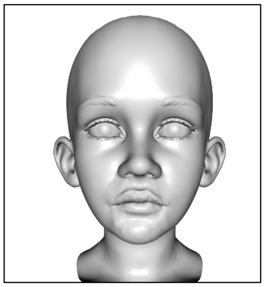
M Sound



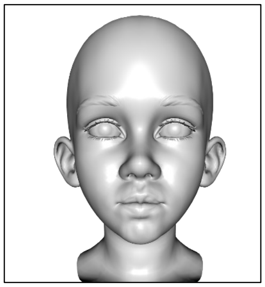
Lip Pull Up Down



Lip Pull Up In

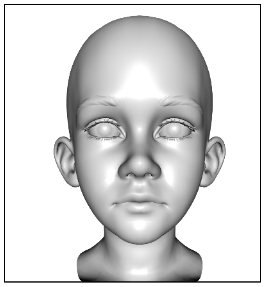


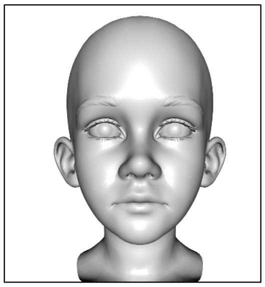
Pucker

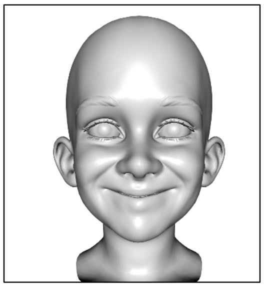


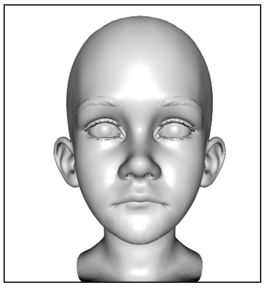
Purse

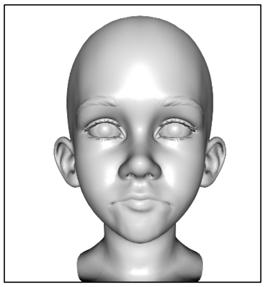
Speech-Based Lower Face Expressions, Continued

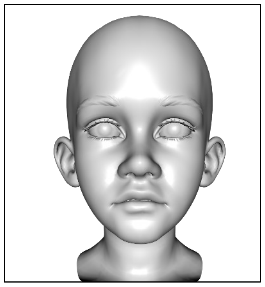

Lower Lip Up In

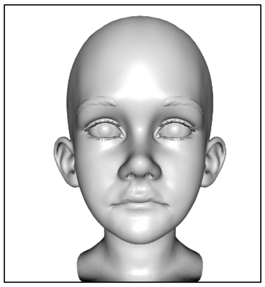

Lower Lip Up

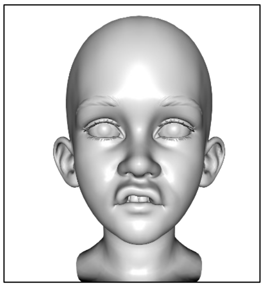

Lip Corners Stretch

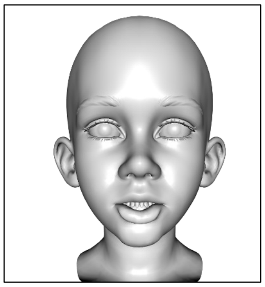

Lower Lip In

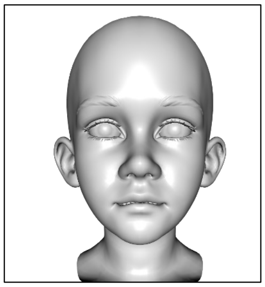

Lower Lip Out

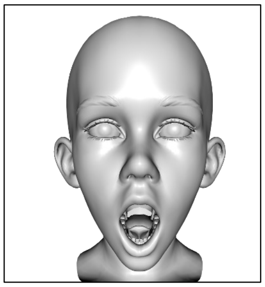

Lips Up Minor

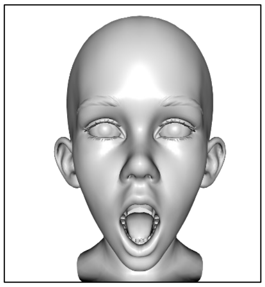

P Sound

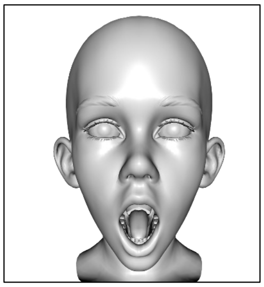

U Sound Upper

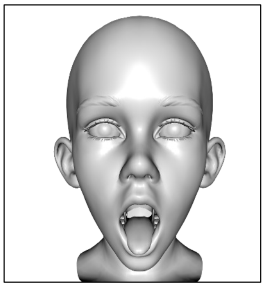

U Sound Lower

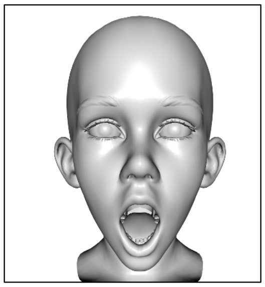

Lip Flatten

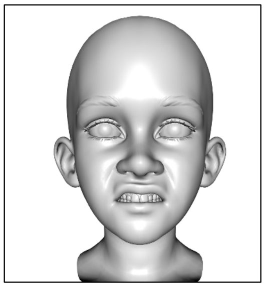

Tongue L Sound

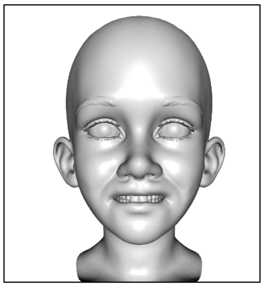

Tongue N Sound

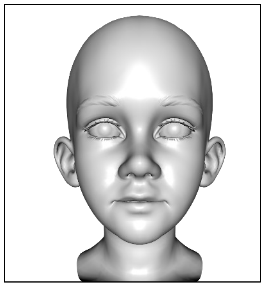

Tongue Narrow

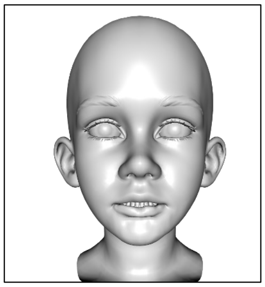

Tongue Out


Tongue Wide

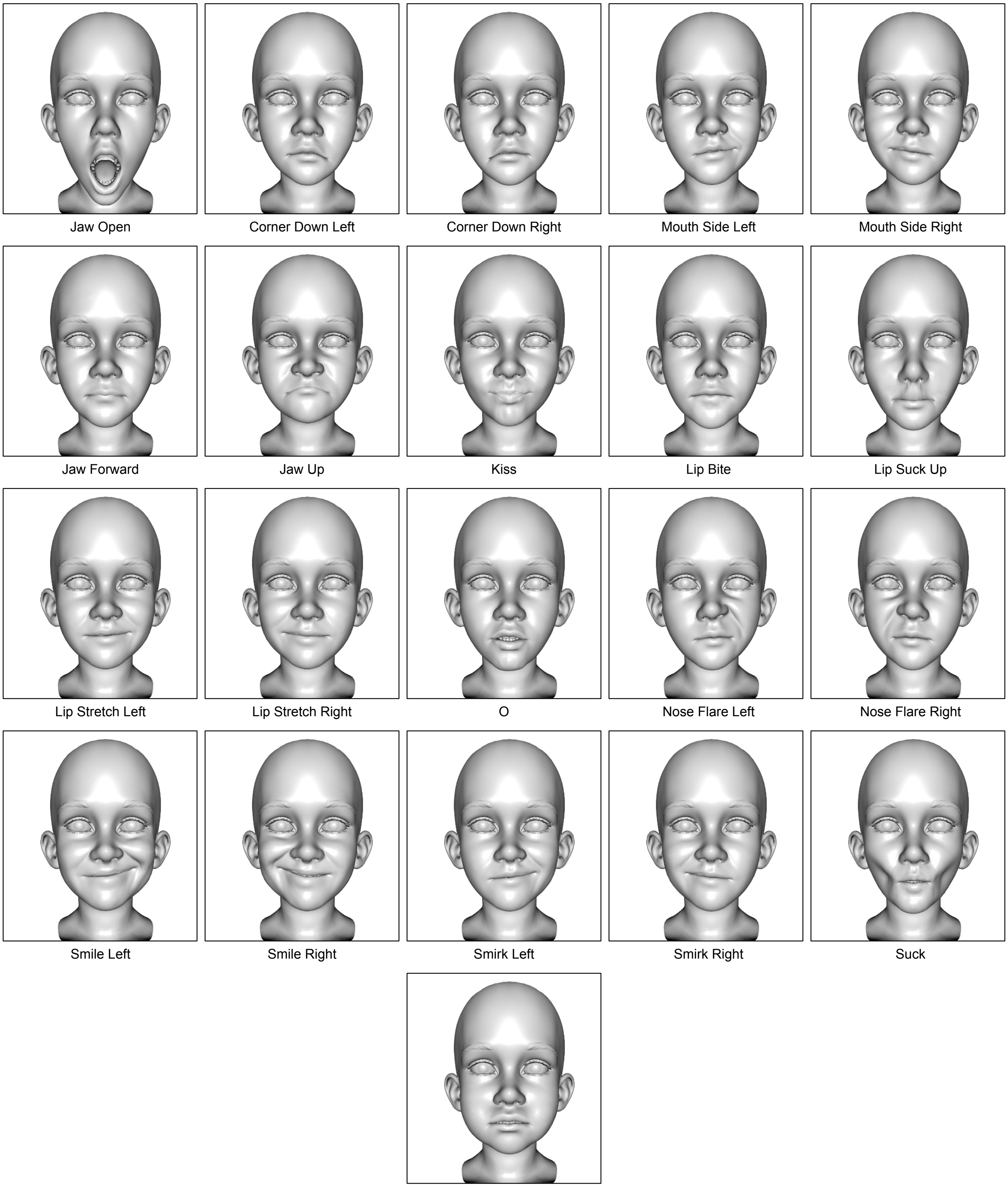

Upper Lip Stretch


Lips Tighten


V Sound


Lower Lip Wide

FACS-Based Lower Face Expressions



Jaw Open

Corner Down Left

Corner Down Right

Mouth Side Left

Mouth Side Right

Jaw Forward

Jaw Up

Kiss

Lip Bite

Lip Suck Up

Lip Stretch Left

Lip Stretch Right

O

Nose Flare Left

Nose Flare Right

Smile Left

Smile Right

Smirk Left

Smirk Right

Suck

Puff