A lightweight model to predict 52 facial blendshapes from facial landmarks in real time. This model is intended for use with the FaceMesh model (model card). Facial landmarks used by this model first need to be produced by the FaceMesh model that runs on monocular video captured by a front-facing camera.

**MODEL SPECIFICATIONS**

**Model Type**
- Convolutional Neural Network

**Model Architecture**
- MLP-Mixer ([Keras](https://keras.io))

**Inputs**
- 146 landmarks, a subset of the 478 landmarks predicted by FaceMesh (model card). Listed in Appendix.

**Output(s)**
- 52 facial blendshape coefficients as float values in [0, 1] range. Predicted blendshapes are listed in Appendix.

**LICENSED UNDER**

Apache License, Version 2.0

**AUTHORS**
- Ivan Grishchenko, Google
- Geng Yan, Google
- Andrei Zanfir, Google
- Eduard Gabriel Bazavan, Google

**MODEL DATE**
November 11, 2022

---

1 Based on Model Cards for Model Reporting, In Proceedings of FAT* Conference (FAT*2019). ACM, New York, NY, USA
Intended Uses

APPLICATIONS
- Detection of human facial expressions from monocular video.
- Optimized for videos captured on front-facing cameras of smartphones.
- Well suitable for mobile AR (augmented reality) applications, especially for controlling an avatar’s face.

DOMAIN & USERS
- The primary intended application is AR entertainment.
- Intended users are people who use augmented reality for entertainment purposes.

OUT-OF-SCOPE APPLICATIONS
Not appropriate for:
- This model is not intended for human life-critical decisions.
- Predicted facial blendshapes do not provide facial recognition or identification and do not store any unique face representation.

Limitations

PRESENCE OF ATTRIBUTES
The model is intended to be used along with the FaceMesh model using facial landmarks produced by it as input. Thus many limitations come from the underlying FaceMesh model.

INPUTS
Videos used by the underlying FaceMesh model to predict face landmarks should be captured in "selfie" mode. As such, it's not suitable for detecting faces:
- looking away from the camera (more than 80°),
- only partially visible (less than 50% of the face),
- located too far away from the camera (cropped face can’t be rescaled to the upstream facemesh model input of 256x256 without quality degradation and loss of accuracy in landmarks and blendshapes prediction).

ENVIRONMENT
When degrading the environment light, noise, motion or face overlapping conditions on FaceMesh input video one can expect degradation of predicted landmarks and as a result inaccuracy and increase of “jittering” in blendshapes estimation (although the model tries to cover such cases during training by introducing noise augmentation to input landmarks).

TRADE-OFFS
The model is optimized for real-time performance on a wide variety of mobile devices, but is sensitive to noise in input facial landmarks that lay outside of FaceMesh distribution.

Factors and Subgroups
The Blendshapes model works together with the FaceMesh model. For factors and subgroups of the upstream model, check the FaceMesh model card.

INSTRUMENTATION
- All dataset samples were captured in a controlled environment in the lab. Each sample contains aligned multi-view images with reconstructed unified 3D GHUM Mesh of a human subject. Each human

ENVIRONMENTS
- Model is trained on millions of samples generated with diverse identities (i.e. unique subjects), various blendshape combinations representing both common and random expressions and transformations (e.g. rotations).
subject performed an extensive set of predefined facial expressions recorded over time.

- All subjects have 52 individual blendshapes automatically transferred from the canonical Mesh with consideration of given subject expressions distribution. Blendshapes for the canonical GHUM Mesh were designed by an artist.

- To make sure that the model trained on GHUM Mesh is compatible with FaceMesh predictions we use accurate mapping with landmark augmentations. However quality may degrade in extreme conditions.

**ATTRIBUTES**

- Model expects 146 2D facial landmarks as input. These landmarks should be produced by the Face Mesh model.
- No additional facial landmark transformations are required, the model encapsulates all necessary scale, rotation and translation normalizations.

**Metrics**

**Model Performance Measures**

Mean Absolute Deviation (MAD)

Mean Absolute Deviation quantifies the bias of the model for each subgroup. We calculate this metric by computing the mean activation across all samples for given expressions, then absolute deviation for each sample in the subgroup and finally we average across all samples and expressions in the subgroup.

Standard Deviation (STDEV)

Standard deviation allows us to better understand the distribution of predicted blendshape coefficients within a given subgroup relative to the mean of all samples. To compute this metric we use the same absolute deviations and mean activations across all samples as in Mean Absolute Deviation.

**Evaluation, Datasets and Results**

**Skin Tone and Gender Evaluation**

**DATASET**

Contains 511 samples (i.e. multi-view facial images and 3D mesh reconstructions), captured at the lab, which were annotated with perceived gender (male and female) and skin tone (from 1 to 10) based on the monk scale.

**FAIRNESS CRITERIA**

The SDM of each subgroup should be within one standard deviation from those values of the entire dataset to be considered fair.

Due to a high ambiguity of blendshapes annotation, we compare blendshape predictions with average

Across Gender:

- MAD: worst case 0.203, difference is 0.004, within the standard deviation of 0.237.

Across skin tone:

- MAD: worst case 0.207, difference is 0.008, within the standard deviation of 0.237

Observed discrepancy across different genders and skin tones is less than one defined in our fairness criteria. We therefore consider the model performing well across groups.
blendshape activations across all subjects for one given expression (as recorded by each subject). We also only compare blendshapes with high activation values (greater than 0.25) to avoid noise in statistics.

### EVALUATION RESULTS
Detailed evaluation across genders and skin tones is presented in the tables below.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Test subset items and %</th>
<th>MAD</th>
<th>STDEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>7,479</td>
<td>41.10%</td>
<td>0.203</td>
</tr>
<tr>
<td>Female</td>
<td>10,720</td>
<td>58.90%</td>
<td>0.196</td>
</tr>
<tr>
<td>Total</td>
<td>18,199</td>
<td>100%</td>
<td>0.199</td>
</tr>
</tbody>
</table>

*Gender evaluation, Signed Deviation*

<table>
<thead>
<tr>
<th>Skintone</th>
<th>Test subset items and %</th>
<th>MAD</th>
<th>STDEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone_1 + Tone_2</td>
<td>6,543</td>
<td>35.95%</td>
<td>0.192</td>
</tr>
<tr>
<td>Tone_3</td>
<td>4,620</td>
<td>25.39%</td>
<td>0.200</td>
</tr>
<tr>
<td>Tone_4</td>
<td>2,411</td>
<td>13.25%</td>
<td>0.199</td>
</tr>
<tr>
<td>Tone_5</td>
<td>1,412</td>
<td>7.76%</td>
<td>0.202</td>
</tr>
<tr>
<td>Tone_6</td>
<td>809</td>
<td>4.45%</td>
<td>0.201</td>
</tr>
<tr>
<td>Tone_7</td>
<td>546</td>
<td>3.00%</td>
<td>0.200</td>
</tr>
<tr>
<td>Tone_8</td>
<td>697</td>
<td>3.83%</td>
<td>0.210</td>
</tr>
<tr>
<td>Tone_9</td>
<td>443</td>
<td>2.43%</td>
<td>0.206</td>
</tr>
<tr>
<td>Tone_10</td>
<td>718</td>
<td>3.95%</td>
<td>0.207</td>
</tr>
<tr>
<td>Total</td>
<td>18,199</td>
<td>100.00%</td>
<td>0.199</td>
</tr>
</tbody>
</table>

*Skin tone evaluation, Signed Deviation*

### Release notes

**Model updates**

This model is the first model we introduce to predict 52 facial blendshape coefficients in real time, on-device, from a single RGB image based on FaceMesh landmarks estimation.
Definitions

Blendshape

The blendshape model is represented as a linear weighted sum of the target faces, which exemplify user-defined facial expressions or approximate facial muscle actions. Blendshapes are therefore quite popular because of their simplicity, expressiveness, and interpretability. Blendshapes are therefore quite popular because of their simplicity, expressiveness, and interpretability.

Augmented Reality (AR)

Augmented reality, a technology that superimposes a computer-generated image on a user’s view of the real world, thus providing a composite view.

Appendix

List of predicted blendshapes

1 - browDownLeft  2 - browDownRight  3 - browInnerUp  4 - browOuterUpLeft  5 - browOuterUpRight  6 - cheekPuff

Subset of 146 landmarks from 478 landmarks

