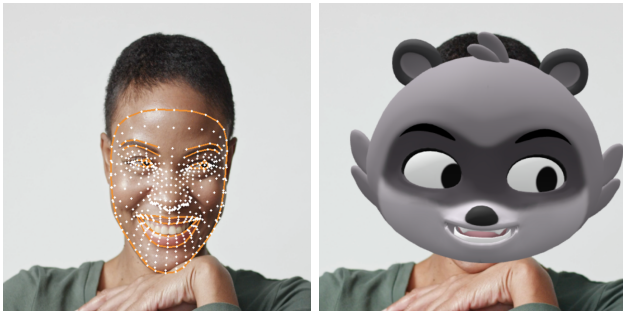
MODEL CARD

Mediapipe Blendshape V2 11/11/2022

MODEL DETAILS[1]

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



Left: Input frame with predicted facial landmarks. Right: 3D avatar controlled with predicted blendshape coefficients

MODEL SPECIFICATIONS

Model Type
- Convolutional Neural Network

Model Architecture
- MLP-Mixer (Keras)

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.

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.

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).

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

FAIRNESS RESULTS

Examine whether subgroups are within one standard deviation away from the metrics of the entire dataset:
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.

Gender	Test subset items and %		MAD	STDEV
Male	7,479	41.10%	0.203	0.242
Female	10,720	58.90%	0.196	0.231
Total	18,199	100%	0.199	0.237

Gender evaluation, Signed Deviation

Skintone	Test subset items and %		MAD	STDEV
Tone_1 + Tone_2	6,543	35.95%	0.192	0.114
Tone_3	4,620	25.39%	0.200	0.235
Tone_4	2,411	13.25%	0.199	0.235
Tone_5	1,412	7.76%	0.202	0.237
Tone_6	809	4.45%	0.201	0.229
Tone_7	546	3.00%	0.200	0.231
Tone_8	697	3.83%	0.210	0.229
Tone_9	443	2.43%	0.206	0.223
Tone_10	718	3.95%	0.207	0.228
Total	18,199	100.00%	0.199	0.237

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. user-defined facial expressions or approximate facial muscle actions. Blendshapes are therefore quite popular because of their simplicity, expressiveness, and interpretability.[2]

`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 (blendshape predicted by the FaceMesh model)
7 - cheekSquintLeft
8 - cheekSquintRight
9 - eyeBlinkLeft
10 - eyeBlinkRight
11 - eyeLookDownLeft
12 - eyeLookDownRight
13 - eyeLookInLeft
14 - eyeLookInRight
15 - eyeLookOutLeft
16 - eyeLookOutRight
17 - eyeLookUpLeft
18 - eyeLookUpRight
19 - eyeSquintLeft
20 - eyeSquintRight
21 - eyeWideLeft
22 - eyeWideRight
23 - jawForward
24 - jawLeft
25 - jawOpen
26 - jawRight
27 - mouthClose
28 - mouthDimpleLeft
29 - mouthDimpleRight
30 - mouthFrownLeft
31 - mouthFrownRight
32 - mouthFunnel
33 - mouthLeft
34 - mouthLowerDownLeft
35 - mouthLowerDownRight
36 - mouthPressLeft
37 - mouthPressRight
38 - mouthPucker
39 - mouthRight
40 - mouthRollLower
41 - mouthRollUpper
42 - mouthShrugLower
43 - mouthShrugUpper
44 - mouthSmileLeft
45 - mouthSmileRight
46 - mouthStretchLeft
47 - mouthStretchRight
48 - mouthUpperUpLeft
49 - mouthUpperUpRight
50 - noseSneerLeft
51 - noseSneerRight
52 - tongueOut (blendshape predicted by the FaceMesh model)

`Subset of 146 landmarks from 478 landmarks`

0, 1, 4, 5, 6, 7, 8, 10, 13, 14, 17, 21, 33, 37, 39, 40, 46, 52, 53, 54, 55, 58, 61, 63, 65, 66, 67, 70, 78, 80, 81, 82, 84, 87, 88, 91, 93, 95, 103, 105, 107, 109, 127, 132, 133, 136, 144, 145, 146, 148, 149, 150, 152, 153, 154, 155, 157, 158, 159, 160, 161, 162, 163, 168, 172, 173, 176, 178, 181, 185, 191, 195, 197, 234, 246, 249, 251, 263, 267, 269, 270, 276, 282, 283, 284, 285, 288, 291, 293, 295, 296, 297, 300, 308, 310, 311, 312, 314, 317, 318, 321, 323, 324, 332, 334, 336, 338, 356,

[2] *Anjyo, K. (2018). Blendshape Facial Animation. In: Handbook of Human Motion. Springer, Cham*

361, 362, 365, 373, 374, 375, 377, 378, 379, 380, 381, 382, 384, 385, 386, 387, 388, 389, 390, 397, 398, 400, 402, 405, 409, 415, 454, 466, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477