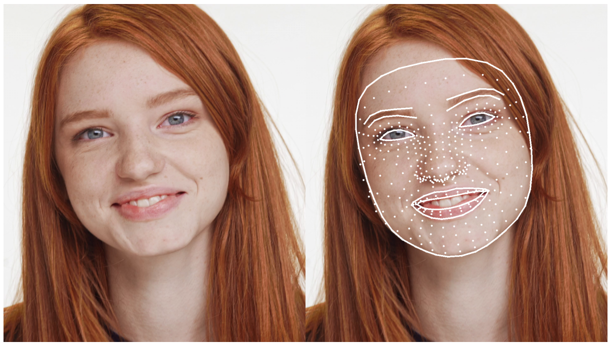
MODEL CARD

MediaPipe Face Mesh

MODEL DETAILS

A lightweight model to predict 3D facial surface geometry from monocular video captured by a front-facing camera on a smartphone in real time.



Left: Input frame. Right: Output face landmarks

MODEL SPECIFICATIONS

Model Type

- Convolutional Neural Network

Model Architecture

- Convolutional Neural Network: MobileNetV2-like with customized blocks for real-time performance.

Inputs

- Image of cropped face with 25% margin on each side and size 192x192 px

Output(s)

- **Facial surface represented** as 468 3D landmarks flattened into a 1D tensor: (x1, y1, z1), (x2, y2, z2), ... x- and y-coordinates follow the image pixel coordinates; z-coordinates are relative to the face center of mass and are scaled proportionally to the face width.
- **Face flag** indicating the likelihood of the face being present in the input image. Used in tracking mode to detect that the face was lost and the face detector should be applied to obtain a new face position. Face probability threshold is set at 0.5 by default and can be adjusted.

AUTHORS

Ivan Grishchenko, Google
Artsiom Ablavatski, Google
Yury Kartynnik, Google

CITATION

Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs, CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Long Beach, CA, USA, 2019

MODEL ACCESSIBLE AT

https://google.github.io/mediapipe/solutions/models.html

MODEL DATE

October 10, 2018

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

DOCUMENTATION LINKS

Paper:
https://arxiv.org/abs/1907.06724
Project Page:
https://sites.google.com/corp/view/perception-cv4arvr/facemesh

LICENSED UNDER

Apache License, Version 2.0

Intended Uses

APPLICATIONS

- Detection of human facial surface geometry from monocular video .
- Optimized for videos captured on front-facing cameras of smartphones.
- Well suitable for mobile AR (augmented reality) applications.

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 face geometry **does not provide facial recognition or identification** and **does not store any unique face representation**.

Limitations

PRESENCE OF ATTRIBUTES

The model is intended to be used primarily in the tracking mode that guarantees certain accuracy of the face location, scale and rotation (see specification in “Attributes").

TRADE-OFFS

The model is optimized for real-time performance on a wide variety of mobile devices, but is sensitive to face position, scale and orientation in the input image.

INPUTS

Videos should be captured in "selfie" mode. As such, it's not suitable for detecting faces:

- looking away from the camera (more than 80°),
- inclined from the vertical orientation (more than 8°),
- only partially visible (less than 50% of the face),
- located too far away from the camera (cropped face can't be rescaled to model input of 192x192 without quality degradation).

ENVIRONMENT

When degrading the environment light, noise, motion or face overlapping conditions one can expect degradation of quality and increase of “jittering” (although we cover such cases during training with real-world samples and augmentations).

Factors and Subgroups

INSTRUMENTATION

- All dataset images were captured on a diverse set of smartphone cameras, both front- and back-facing.
- All images were captured in a real-world environment with different light, noise and motion conditions via an AR (Augmented Reality) application.

ENVIRONMENTS

Model is trained on images with various lighting, noise and motion conditions and with diverse augmentations. However, its quality can degrade in extreme conditions. This may lead to increased "jittering" (inter-frame prediction noise).

ATTRIBUTES

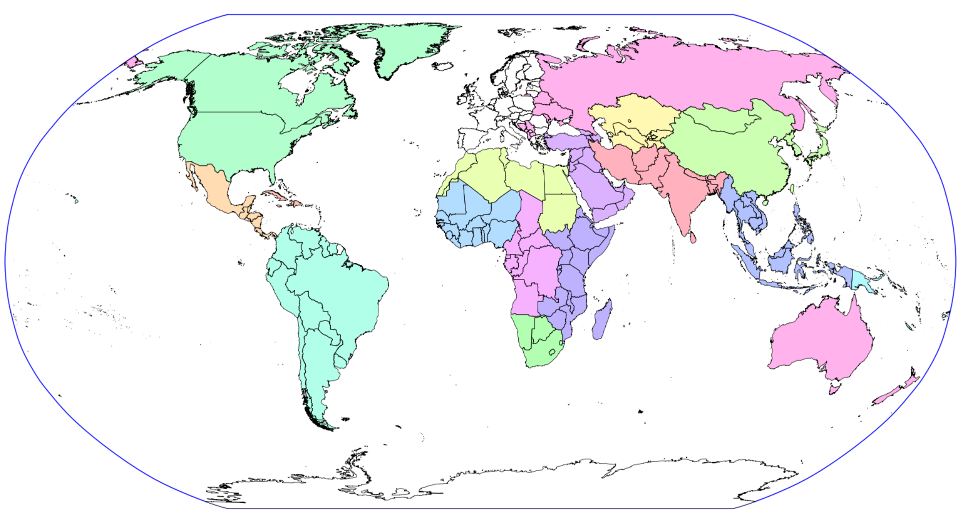
- Face cropped from the captured frame should contain a single face placed in the center of the image.
- There should be a margin around the face calculated as 25% of face size.
- Image must be rotated in a way that a horizontal line can connect the two centers of the eyes.
- Model is tolerant to certain level of input inaccuracy:
    - 10% shift and scale (taking face width/height as 100% for corresponding axis)
    - 8° roll

GROUPS

To perform fairness evaluation we group user samples into 17 evenly distributed geographic subregions (based on [United Nations geoscheme] with merges and no EU countries):

Northern Africa
Eastern Africa
Middle Africa
Southern Africa
Western Africa
Caribbean
Central America
South America
Northern America

Central Asia
Eastern Asia
South-eastern Asia
Southern Asia
Western Asia
Australia and New Zealand
Europe (without EU)
Melanesia, Micronesia,
and Polynesia.



Metrics

Model Performance Measures

NORMALIZATION BY IC

Normalization by interocular distance is applied to unify the scale of the samples and is taken as 100%. IOD is calculated as the distance between the eye centers (which are estimated as the centers of segments connecting eye corners). To accommodate head rotations, 3D IOD from the ground truth is employed.

MEAN ABSOLUTE ERROR

Mean absolute error is calculated as the pixel distance between ground truth and predicted face mesh. The model is providing 3D coordinates, but the z-coordinate is obtained from synthetic data, so for a fair comparison with human annotations, only 2D coordinates are employed.

IOD MAE

For quality and fairness evaluation, we use IOD MAE **(Mean Absolute Error normalized by Interocular Distance)**.

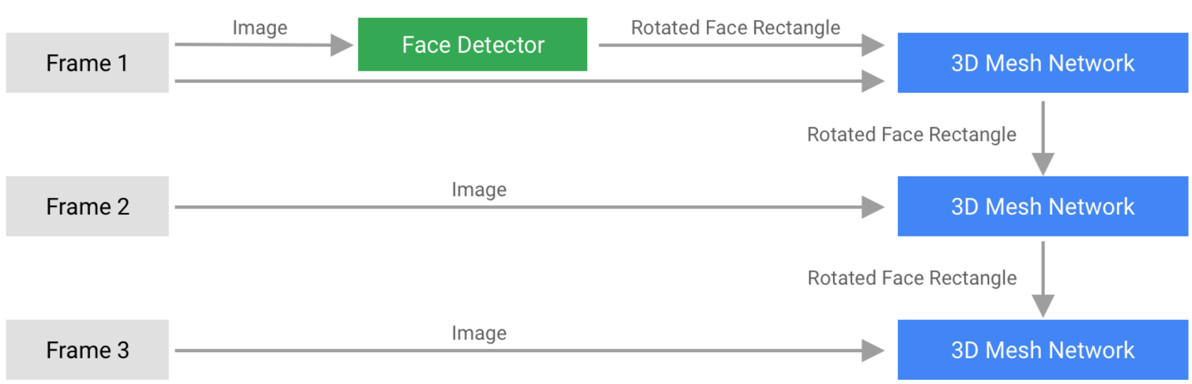
Evaluation Modes

TRACKING MODE

Main mode that takes place most of the time and is based on obtaining highly accurate face crop from the prediction on the previous frame (frames 2, 3, ... on the image below)..

REACQUISITION MODE

Takes place when there is no information about the face from previous frames. It happens either on the first frame (image below) or on the frames when face tracking is lost. In this case, an external face detector is being run over the whole frame. We used BlazeFace Detector for the evaluation of the reacquisition mode.



Evaluation, Datasets and Results

Geographical Evaluation

GEOGRAPHICAL SUBREGIONS DATASET

- **Contains 1700 samples evenly distributed across 17 geographical subregions** (see specification in Section "3a Groups"). Each region contains 100 images.
- All samples are picked from the same source as training samples and are characterized as smartphone front-facing camera selfies taken in real-world environments (see specification in "*Factors and Subgroups - Instrumentation*").

EVALUATION RESULTS

Detailed evaluation for the tracking and reacquisition modes across 17 geographical subregions is presented in the table below.

Region	Tracking mode (primary)		Reacquisition mode (on the first frame)	
	Mean absolute error	**Standard deviation**	**Mean absolute error**	**Standard deviation**
Australia and New Zealand	2.52	0.67	3.87	1.48
Melanesia + Micronesia + Polynesia	**2.40**	0.77	3.82	1.53
Europe	3.16	1.26	4.76	1.87
Central Asia	2.52	0.88	3.76	1.57
Eastern Asia	2.57	0.62	**3.68**	1.01
Southeastern Asia	3.11	0.91	4.38	1.90
Southern Asia	3.05	0.73	4.71	2.27
Western Asia	3.06	0.99	4.66	1.94
Caribbean	2.62	0.77	4.28	1.58
Central America	**3.29**	1.17	**5.24**	3.75
South America	3.09	1.17	4.58	2.08
Northern America	3.21	1.18	4.61	1.97
Northern Africa	2.51	0.74	4.04	2.04
Eastern Africa	2.43	0.63	3.73	1.06
Middle Africa	2.62	0.86	4.35	2.60
Southern Africa	2.48	0.73	3.87	1.50
Western Africa	2.70	0.75	4.44	2.79
Total for all regions	**2.78**	**0.95**	**4.28**	**2.08**

Geographical evaluation

FAIRNESS CRITERIA

We evaluate the model accuracy across representative groups, and compare the error range to the human annotation discrepancy, which is 2.56% IOD MAE. We flag any group that has a higher error rate than this discrepancy.

FAIRNESS METRICS & BASELINE

2.56% IOD MAE was obtained by measuring the discrepancy of the 11 human annotators (same people used for training data annotation) on 58 samples.

FAIRNESS RESULTS

Comparison with fairness goal of 2.56% IOD MAE discrepancy across 17 regions:

- Tracking mode: from 2.40% to 3.29% (difference of 0.90%)
- Reacquisition mode: from 3.68% to 5.24% (difference of 1.56%)

Comparison with our fairness criteria yields a maximum discrepancy between best and worst performing regions of 0.9% for the thacking mode and 1.56% for the reacquisition mode. We therefore consider the models performing well across groups.

Skin Tone and Gender Evaluation

DATASET

Contains 1700 samples, 100 from each of the 17 geographical subregions, which were annotated with perceived gender (male and female) and skin tone (from 1 to 6) based on the Fitzpatrick scale.

EVALUATION RESULTS

Detailed evaluation for the tracking and reacquisition modes across genders and skin tones is presented in the tables below.

FAIRNESS RESULTS

Comparison with fairness goal of 2.56% IOD MAE discrepancy across genders:

- Tracking mode: from 2.77% to 2.80% (difference of 0.02%)
- Reacquisition mode: from 4.04% to 4.14% (difference of 0.1%)

And across skin tones:

- Tracking mode: from 2.69% to 2.92% (difference of 0.24%)
- Reacquisition mode: from 3.87% to 4.99% (difference of 1.12%)

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.

Gender	Tracking mode (primary)			Reacquisition mode (on the first frame)		
	% of dataset	Mean	Standard	% of dataset	Mean	Standard

		absolute error	deviation		absolute error	deviation
Male	46.5%	**2.80**	0.99	46.6%	**4.04**	1.83
Female	53.5%	**2.77**	0.90	53.4%	**4.14**	2.13

Gender evaluation

Skin Tone Type	Tracking mode (primary)			Reacquisition mode (on the first frame)		
	% of dataset	Mean absolute error	Standard deviation	% of dataset	Mean absolute error	Standard deviation
1	1.5%	**2.92**	1.03	1.5%	4.35	2.21
2	14.8%	2.80	0.98	14.7%	3.98	1.77
3	34.5%	2.84	1.04	34.6%	4.10	1.94
4	28.6%	2.73	0.79	28.6%	**3.87**	1.62
5	14.4%	**2.69**	0.91	14.4%	4.22	2.45
6	6.3%	2.88	0.97	6.1%	**4.99**	2.78

Skin tone evaluation

Release notes

`Model updates`

The model provides better prediction accuracy and smaller error discrepancy across geographical subregions while presuming the same size and performance.

Definitions

`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.

`Interocular Distance (IOD)`

An estimate of the distance between the eye centers. To avoid gaze direction dependence, the eye centers are defined as the midpoints of the segments connecting the eye corners.

`Landmarks`

Facial landmarks are 2D (x, y) or 3D (x, y, z) coordinate locations of facial features, such as lips or eyes corners, points on the eyebrows, irises and face contours and intermediate points on cheeks and forehead.

`Mean Absolute Error (MAE)`

Per sample metric calculated as average 2D distance error over all 468 facial landmarks. To normalize scale across samples we divide MAE of every sample by its 3D IOD.