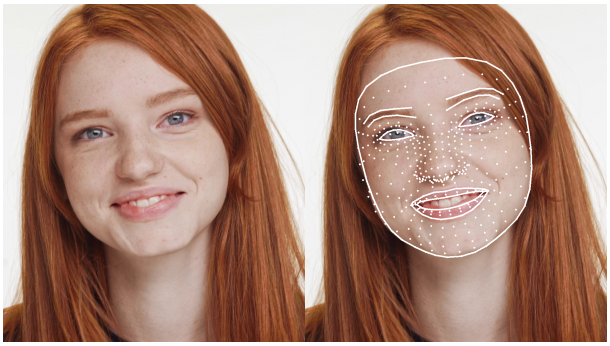
MODEL CARD

MediaPipe Attention Mesh

MODEL DETAILS

A lightweight model for real-time prediction of 3D facial surface landmarks from video captured by a front-facing smartphone camera. Designed for applications like AR makeup, eye tracking and AR puppeteering that rely on highly accurate landmarks for eye (+ iris) and lips regions predicted by the model. Runs at over 50 FPS on Pixel 2 phone.



Left: Input frame. Right: Output face landmarks

MODEL SPECIFICATIONS

Model Type
- Convolutional Neural Network

Model Architecture
- MobileNetV2-like with customized blocks for real-time performance and an attention mechanism to refine lips and eye regions and to predict irises.

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

Outputs
- **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.
- **Lips refined region surface** represented as 80 2D landmarks (inner and outer contours and an intermediate line) flattened into a 1D tensor.
- **Eye with eyebrow refined region surface (x2)** represented as 71 2D landmarks (eye and eyebrow contours with surrounding areas) flattened into a 1D tensor.
- **Iris refined region surface (x2)** represented as 5 2D landmarks (1 for pupil center and 4 for iris contour) flattened into a 1D tensor.
- **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.

MODEL ACCESSIBLE AT

MediaPipe Models and Model Cards - Face Mesh

MODEL DATE

August 28, 2020

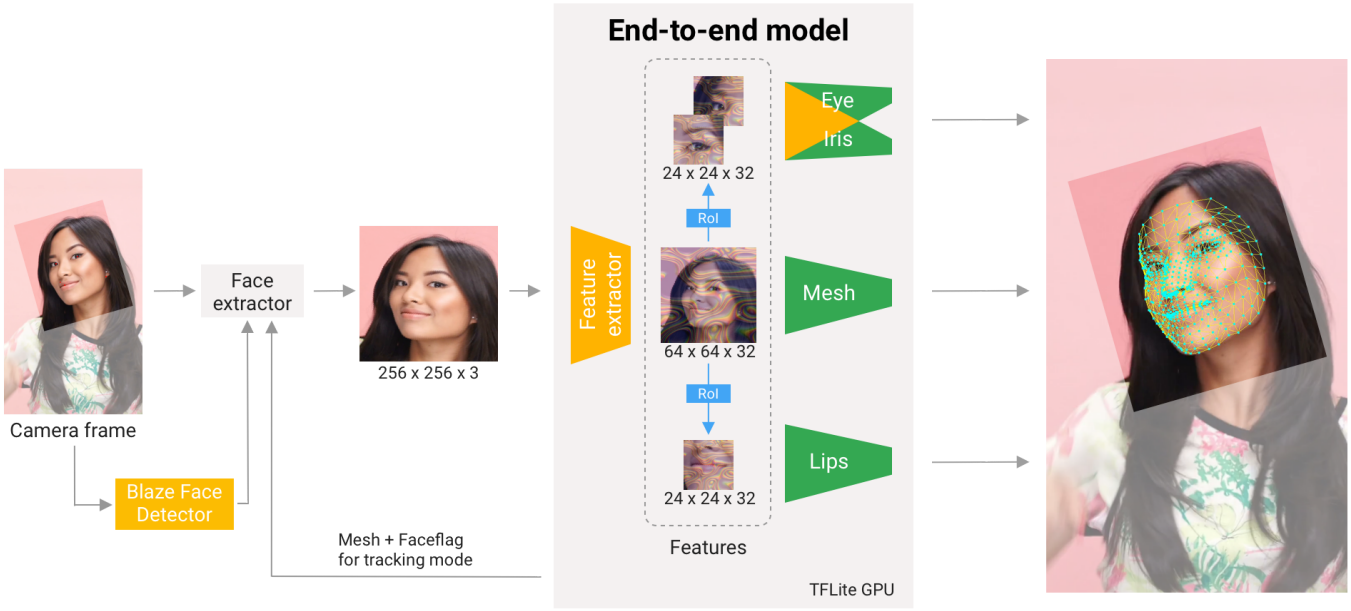
MODEL ARCHITECTURE

The model accepts a 192x192 face image as input. This image is provided by either the face detector or via tracking from a previous frame. After extracting a 48x48 feature map, the model splits into several sub-models (see the figure below). One submodel predicts all 478 face mesh landmarks in 3D and defines crop bounds for each region of interest. The remaining submodels predict region landmarks from the corresponding 16x16 feature maps that are obtained via the attention mechanism.

We concentrate on three facial regions with key contours: the lips and two eyes. Each eye submodel predicts the iris as a separate output after reaching the spatial resolution of 4x4. This allows the reuse of eye features while keeping the dynamic iris independent from the more static eye landmarks.

Individual submodels allow us to control the network capacity dedicated to each region and boost quality where necessary. To further improve the accuracy of the predictions, we apply a set of normalizations to ensure that the eyes and lips are aligned with the horizontal and are of uniform size.

An **attention mechanism** pulls out visual features of a given region of interest by sampling a grid of 2D points in the feature space and extracting the features under the sampled points. This allows to train architectures end-to-end and to enrich the features that are used by the attention mechanism. Specifically, we use a spatial transformer module which is controlled by an affine transformation matrix and allows us to zoom, rotate, translate, and skew the sampled grid of points.



Inference pipeline and Model architecture overview

AUTHORS

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

CITATION

Attention Mesh: High-fidelity Face Mesh Prediction in Real-time, CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, USA, 2020

DOCUMENTATION LINKS

Paper:
https://arxiv.org/abs/2006.10962

LICENSED UNDER

Intended Uses

APPLICATIONS

- Detection of human facial surface landmarks 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 landmarks **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 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 attempt to 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).

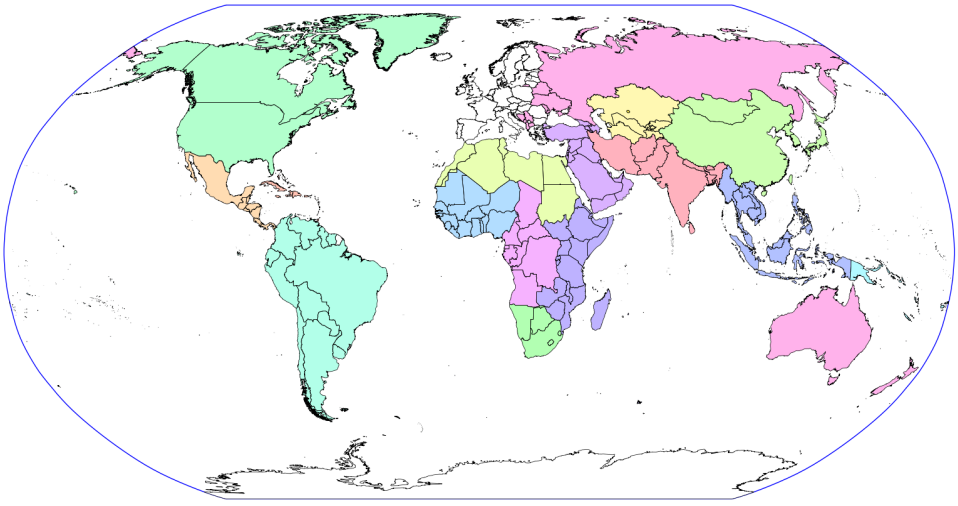
ATTRIBUTES

- The 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 the face bounding box size.
- The image must be rotated in a way that the line connecting the two eye centers becomes horizontal.
- The model is tolerant to certain level of input inaccuracy:
  - 10% shift and scale (taking face width/height as 100% for the corresponding axis)
  - 8° roll

GROUPS

To perform fairness evaluation we group user samples into 17 evenly distributed geographic subregions (based on the 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 IOD

Normalization by interocular distance (IOD) is applied to unify the scale of the samples. IOD is calculated as the distance between the eye centers (which are estimated as the centers of segments connecting eye corners) and is taken as 100%. 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)**.

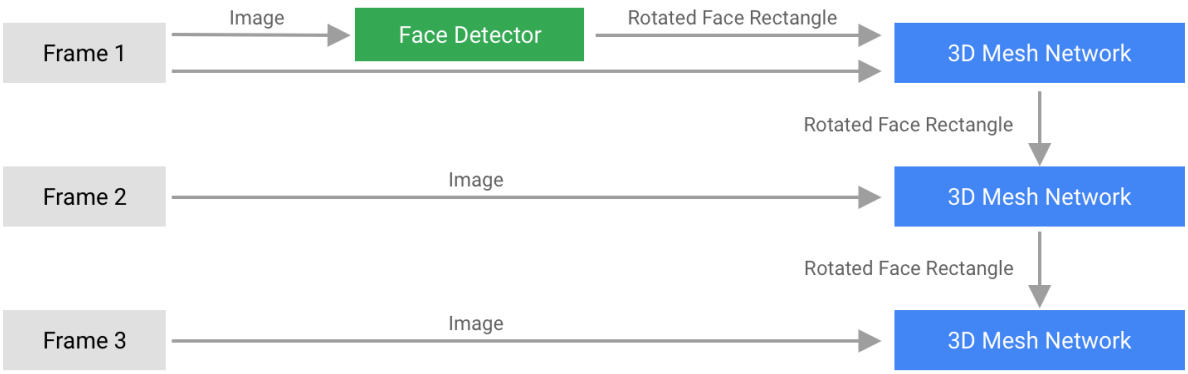
Evaluation Modes

TRACKING MODE

The main mode that takes place most of the time and is based on obtaining a highly accurate face crop from the prediction on the previous frame (frames 2, 3, ... on the image below). Underneath we utilize the MediaPipe Python Solution API for Face Mesh and run the pipeline for several frames on the same image before measuring the tracking accuracy (thus the crop region is determined from the model predictions as in a video stream).

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 where the 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	2.95	0.82	3.26	1.09
Melanesia + Micronesia + Polynesia	**2.73**	0.70	3.05	1.10
Europe	3.85	1.28	3.98	1.35
Central Asia	2.83	0.91	3.06	1.13
Eastern Asia	2.87	0.89	**3.01**	0.96
Southeastern Asia	3.50	1.23	3.94	1.90
Southern Asia	3.44	0.83	3.77	1.04
Western Asia	3.60	1.18	3.87	1.45
Caribbean	3.28	1.21	3.56	1.56
Central America	**3.95**	1.47	**4.28**	2.11
South America	3.74	1.33	4.08	1.86
Northern America	3.77	1.48	4.04	1.68
Northern Africa	2.82	0.70	3.10	0.84
Eastern Africa	2.84	0.85	3.15	1.40
Middle Africa	3.27	1.66	3.58	1.90
Southern Africa	3.13	1.15	3.45	1.39
Western Africa	3.19	1.47	3.34	1.30
Total for all regions	**3.28**	**1.23**	**3.56**	**1.52**

Geographical evaluation

FAIRNESS CRITERIA

We evaluate the model accuracy across representative groups, and compare the error range to the human annotation discrepancy, which has been separately estimated as 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 2.73% to 3.95% (difference of 1.22%)
- Reacquisition mode: from 3.01% to 4.28% (difference of 1.27%)

Comparison with our fairness criteria yields a maximum discrepancy between best and worst performing regions of 1.22% for the thacking mode and 1.27% 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 3.27% to 3.28% (difference of 0.01%)
- Reacquisition mode: from 3.55% to 3.58% (difference of 0.03%)

And across skin tones:

- Tracking mode: from 3.11% to 3.65% (difference of 0.54%)
- Reacquisition mode: from 3.40% to 4.28% (difference of 0.88%)

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	% of dataset	Tracking mode (primary)		Reacquisition mode (on the first frame)	
		Mean absolute error	Standard deviation	Mean absolute error	Standard deviation
Male	46.4%	3.28	1.24	3.58	1.49
Female	53.6%	3.27	1.22	3.55	1.54

Gender evaluation

Skin Tone Type	% of dataset	Tracking mode (primary)		Reacquisition mode (on the first frame)	
		Mean absolute error	Standard deviation	Mean absolute error	Standard deviation
1	1.5%	3.65	1.77	4.28	2.26
2	14.8%	3.41	1.31	3.64	1.46
3	34.5%	3.34	1.18	3.62	1.44
4	28.6%	3.11	1.01	3.40	1.27
5	14.4%	3.16	1.12	3.48	1.74
6	6.3%	3.60	1.95	3.80	2.10

Skin tone evaluation

Release notes

Model updates

The model provides better prediction accuracy of lips and eye regions of a Face Mesh and predicts irises as an additional output. Model increases in size (1.2M -> 2.5M) but keeps real-time performance on a wide variety of devices.

	All	Lips	Eyes
Face Mesh	2.99	3.28	6.66
Attention Mesh	3.11	2.89	6.04

Attention Mesh regions MAE improvement.
Check details in the paper

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.

Attention Mechanism

Attention Mechanism in Deep Learning is a concept of directing model focus to certain areas of processed data in order to obtain higher accuracy for these regions of interest.

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.