MODEL CARD

MediaPipe Iris

MODEL DETAILS

A 3 megabyte model to predict 2D eye, eyebrow and iris geometry from monocular video captured by a front-facing camera on a smartphone in real time. This model is intended to be used in combination with MediaPipe FaceMesh.

MODEL SPECIFICATIONS

Model Type
- Convolutional Neural Network

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

Input(s)
- Image of proportionally cropped left eye with eyebrow (or horizontally flipped right eye) with a 25% margin on each side and size 64x64

Output(s)
- **Eye with eyebrow region surface** represented as 71 2D landmarks flattened into a 1D tensor: (x1, y1), (x2, y2), ...; x- and y-coordinates follow the image pixel coordinates.
- **Iris surface** represented as 5 2D landmarks (1 for pupil center and 4 for iris contour) also represented as flattened 1D tensor.

AUTHORS

Andrey Vakunov, Google
Ivan Grishchenko , Google
Artsiom Ablavatski, Google
Dmitry Lagun, Google

MODEL ACCESSIBLE AT

https://github.com/google/mediapipe/tree/master/mediapipe/models

DOCUMENTATION LINKS

Not yet available

CITATION

Not yet available

MODEL CREATED ON

May 13, 2020

LICENSED UNDER

Apache License, Version 2.0

_

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

Intended Uses

APPLICATIONS

- Depth estimation using iris from monocular video.
- Optimized for videos captured on front-facing cameras of smartphones.
- Well suitable for mobile AR (augmented reality), computational photography and accessibility applications.

DOMAIN & USERS

- The primary intended applications are in:
  AR entertainment: e.g. iris recoloring and virtual avatars
  Computational photography: e.g. distance estimation and iris reflection, *and*
  Accessibility: e.g. adjusting font size with distance from camera.
- Intended users are people who use augmented reality for entertainment purposes, or computational photography for mobile phone photography.

OUT-OF-SCOPE APPLICATIONS

Not appropriate for:

- This model is not intended for human life-critical decisions.
- Predicted eye and iris geometry **does not provide identity recognition** and **does not store any unique face representation**.
- **Note, that any form of surveillance or identification are explicitly out of scope and not enabled by this technology.**

Limitations

PRESENCE OF ATTRIBUTES

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

INPUTS

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

- Face is directed away from the camera (more than 60° when eye is no longer visible),
- Face is inclined from the vertical orientation (more than 8°),
- Eye is only partially visible (less than 50%) or no iris (eye is closed),
- Eye is located too far away from the camera (cropped eye can't be rescaled to model input of 64x64 without quality degradation).

TRADE-OFFS

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

ENVIRONMENT

When degrading the environment conditions (very dark or noisy camera video, video with a lot of motion or significant eye overlap) 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 real-world environments with different light, noise and motion conditions via an AR (Augmented Reality) application.

ATTRIBUTES

- Eye region cropped from the captured frame should contain a single left eye with eyebrow placed in the center of the image.
- There should be a margin around the eye region calculated as 25% of eye region size. Margin should be applied to a minimal proportional bounding box enclosing eye and eyebrow.
- Image must be rotated in a way that a horizontal line can connect the two corners of the eye.
- Model is tolerant to certain level of input inaccuracy:
  - 10% shift and scale (taking eye region width/height as 100% for corresponding axis)
  - 8° roll

ENVIRONMENTS

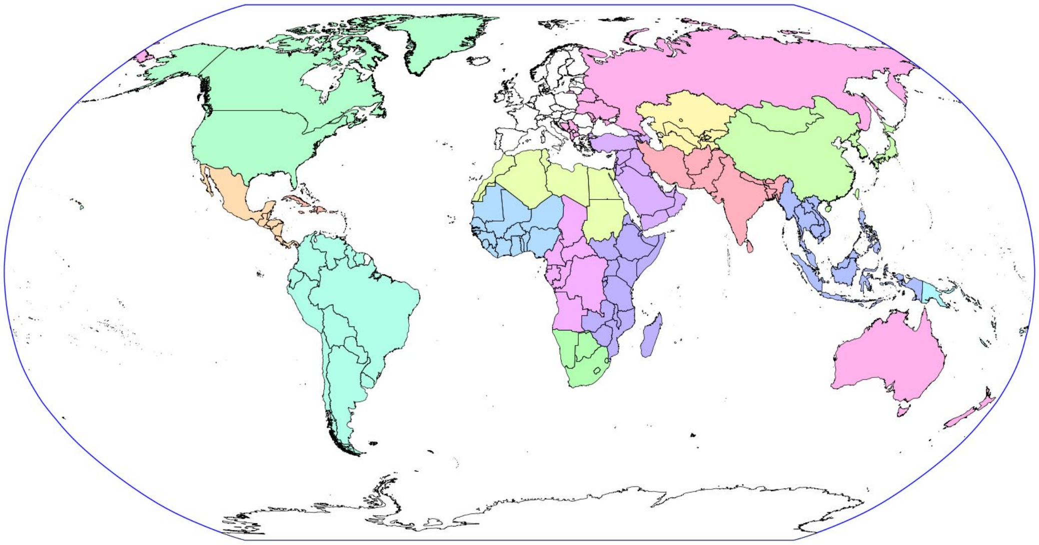
Model is trained on images with various lighting, noise and motion conditions and with diverse augmentations. However, its quality can degrade in extreme conditions (specified in "Limitations-Environment"). This may lead to increased “jittering” (inter-frame prediction noise).

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

IC MAE, WWD MAE

For quality and fairness evaluation, we use IC MAE for eye **(Mean Absolute Error normalized by Intercorner Distance)** and WWD MAE for iris **(Mean Absolute Error normalized by White-to-white diameter).**

MEAN ABSOLUTE ERROR

Mean absolute error is calculated as the pixel distance between ground truth and predicted landmarks. In order to make a fair comparison with human annotators we use 2D coordinates predicted by the model.

NORMALIZATION BY IC

Normalization by intercorner distance is applied to unify the scale of the eyes and is taken as 100%. Unified scale allows to calculate correct average metrics per-region and overall. IC is calculated as the 3D distance between the eye corners taken from the ground truth (3D is used to accommodate for head rotations).

NORMALIZATION BY WWD

Normalization by white-to-white diameter is applied to unify scale of the irises and is taken as 100%. WWD is calculated as the 3D distance between the left and right landmarks on iris contour taken from the ground truth (3D is used to accommodate for head rotations).

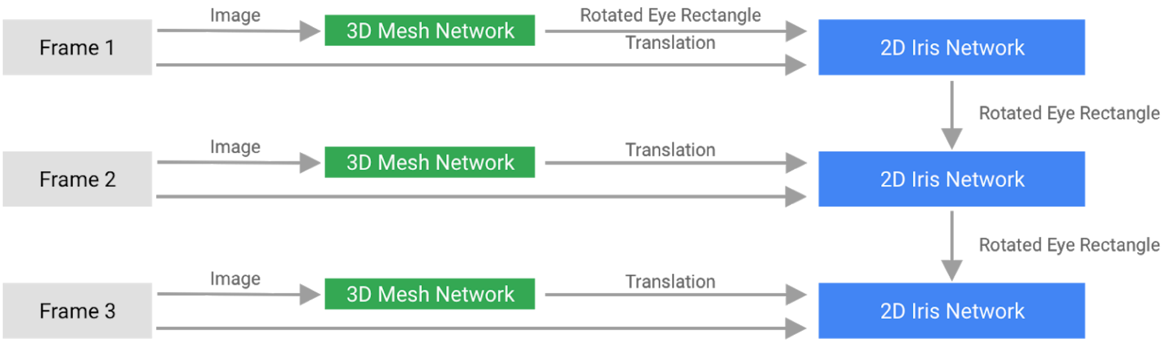
Evaluation Modes

TRACKING MODE

This is the main mode that takes place most of the time. It is based on obtaining a highly accurate eye region crop from the prediction on the previous frame and a translation (crop center) from the face mesh on the current frame (frames 2, 3, ... on the image below).

REACQUISITION MODE

Takes place when there is no information about the eye region from previous frames. It happens either on the first frame (image below) or on the frames when face tracking is lost. In this case, we use the Face Mesh model prediction to obtain eye region crop and 2D translation.



Evaluation, Datasets and Results

Geographical Results

GEOGRAPHICAL SUBREGIONS DATASET

- Contains 1700 samples evenly distributed across 17 geographical subregions (see specification in "*Factors and Subgroups - Groups*"). Each region contains 100 images.
- All samples are picked from the same source as the 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 are available in this sheet.

Fairness Results

FAIRNESS CRITERIA

We consider a model to perform unfairly across representative groups **if the discrepancy in error rates on them spans more than the human annotation discrepancy**.

FAIRNESS METRICS & BASELINE

Discrepancy threshold values were obtained by measuring the discrepancy of the human annotators (same people used for training data annotation) on a variety of samples.

Irises:

- 139 samples by 8 annotators
- 4.01% WWD MAE for irises

Eyes:

- 58 samples by 11 annotators
- 2.56% IOD MAE for eyes

FAIRNESS RESULTS

Comparison with fairness goal discrepancy across 17 regions (i.e. that difference between highest and lowest error rates across the 17 regions is within human annotation discrepancy):

Irises (human discrepancy is 4.01% WWD MAE):

- Tracking mode: from 5.76% to 7.10% (difference of 1.34%)
- Reacquisition mode: from 5.85% to 6.96% (difference of 1.11%)

Eyes (human discrepancy is 2.56% IOD MAE):

- Tracking mode: from 4.68% to 6.81% (difference of 2.13%)
- Reacquisition mode: from 4.61% to 6.53% (difference of 1.91%)

Definitions

Augmented Reality (AR)

Definition: A technology that superimposes a computer-generated image on a user's view of the real world, thus providing a composite view.

Implementation: N/A

Inter-corner Distance (IC)

Definition: Unavailable

Implementation: 3D distance between the eye corners used to normalize all eyes to the same scale.

Mean Absolute Error (MAE)

Definition: An error metric calculated by taking an average of absolute errors. In the context of evaluating a model's accuracy, MAE is the average absolute difference between the expected and predicted values across all training examples.

Implementation: Per sample metric calculated as average 2D distance error over all landmarks of the model output. To normalize scale across samples we divide MAE of every sample by either IC or WWD (for eye and iris metrics correspondingly).

Landmarks

Definition: Synonym for keypoints. The coordinates of particular features in an image. For example, for an image recognition model that distinguishes flower species, keypoints might be the center of each petal, the stem, the stamen, and so on. Commonly used in image recognition models.

Implementation: 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.

White-to-white Diameter (WWD)

Definition: Unavailable

Implementation: 3D distance from from left to right side of the iris used to normalize all irises to the same scale.