MODEL CARD

MediaPipe BlazeFace

FULL RANGE

MODEL DETAILS

A relatively lightweight model (1.1MB in size) for detecting one or multiple faces within an image captured by a smartphone camera, primarily targeting back-facing camera images. Runs super-real-time (~90FPS on Pixel 3 GPU, ~165FPS on Pixel 4 GPU, ~50FPS on Pixel 3 single-core CPU with XNNPACK inference, ~145FPS on Pixel 4 CPU with XNNPACK).

For each detected face, returns:

- Facial bounding box coordinates
- 6 approximate facial keypoint coordinates:
    - Left eye (from the observer's point of view)
    - Right eye
    - Nose tip
    - Mouth
    - Left eye tragion
    - Right eye tragion
- Detection confidence score

MODEL SPECIFICATIONS

Model Type
Convolutional Neural Network

Model Architecture
Convolutional Neural Network: CenterNet-like with a custom encoder.

Input(s)
RGB image (possibly a video frame) resized to 160x192 pixels, represented as a 160x192x3 array of float values in the range [-1.0, 1.0].

Output(s)
Tensor of predicted embeddings representing anchors transformation which are further used in Non Maximum Suppression algorithm.

AUTHORS

Who created this model?
Valentin Bazarevsky, Google

Who provided the model card?
Yury Kartynnik, Google

DATE

January 19, 2021

DOCUMENTATION

Paper: https://arxiv.org/abs/2006.10204

CITATION

How can users cite your model?
V. Bazarevsky et al. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Long Beach, CA, USA, 2019.

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

APPLICATION

Detecting prominently displayed faces within images or videos captured by a smartphone camera.

DOMAIN AND USERS

- Live perception pipelines
- Mobile AR (augmented reality) applications
- User interface enhancements (e.g. Google Meet auto-zoom, AutoFlip)

OUT-OF-SCOPE APPLICATIONS

- Counting the number of people in a crowd
- Detecting faces looking away from the camera, significantly inclined from the vertical orientation, or individuals' back of the head
- Detecting people too far away from the camera (e.g. **further than 5 meters**)
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology

Limitations

PRESENCE OF ATTRIBUTES

Produces only up to a given limit (e.g. 10) of detections even if more people are present.

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.

ENVIRONMENT

In presence of degrading 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).

Ethical Considerations

HUMAN LIFE

The model is not intended for human life-critical decisions. The primary intended application is entertainment and assistive technologies.

PRIVACY

This model was trained and evaluated on consented images of people using a mobile AR application captured with smartphone cameras in various “in the wild” conditions.

BIAS

The model has been trained on images captured with smartphone cameras. While these images have significant variability, please consider whether your use case is significantly different from this domain.

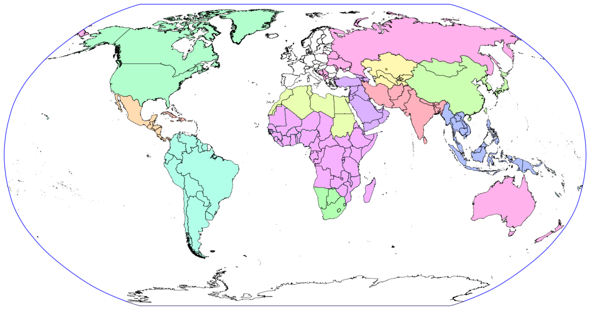
Factors and Subgroups

INSTRUMENTATION

- All dataset images were captured on a diverse set of front- and back-facing smartphone cameras. The dataset consists of 100+K images.
- All images were captured in a real-world environment with different light, noise and motion conditions via an AR (Augmented Reality) application.

ATTRIBUTES

- Face roll and pitch (tilt) angles should be not more than 45 degrees away from the straight orientation. The yaw (pan) angle should not exceed 90 degrees.
- Face bounding box sides should be at least 5% of the corresponding image sides.
- At least 70% of the face bounding box should lie inside the input image.



ENVIRONMENTS

The model is trained on images with various lighting, noise and motion conditions and with diverse augmentations. The test data is coming from the same distribution but without the augmentations applied.

GROUPS

To perform fairness evaluation, we group user samples into slices by several criteria.

The geographical diversity of the samples is ensured by even representation of the 17 *geographic subregions* (based on United Nations geoscheme with mergers):

Australia and New Zealand
Melanesia, Micronesia, Polynesia
Europe*
Central Asia
Eastern Asia
Southeastern Asia
Southern Asia
Western Asia
Caribbean
Central America
South America
Northern America
Northern Africa
Eastern Africa
Middle Africa
Southern Africa
Western Africa

** Excluding the EU*

Additionally, disaggregation is performed based on two *annotator-perceived* characteristics of persons on a subset of images containing only a single face: *perceived gender presentation* (whether a person appears *feminine* or *masculine)* and *skin tone*.

We annotate skin tone using five groups roughly aligned with Fitzpatrick skin type categories. We collapse Fitzpatrick types I and II which annotators have difficulty distinguishing between based on visual images alone. Each group corresponds to a range of actual skin tones.

Evaluation metrics

MODEL PERFORMANCE MEASURES

Average Precision, AP
Area under the interpolated precision-recall curve, obtained by plotting (recall@X, precision@X) points for different values of the decisive confidence threshold X.

True positives @X
Correct face predictions where there are faces (when thresholded by confidence >= X).
False positives @X
Incorrect face predictions where there are no faces (when thresholded by confidence >= X).
False negatives @X
Missed faces (when thresholded by confidence >= X).

Precision @X
True positive rate among all face predictions (when thresholded by confidence >= X).
Recall @X
True positive rate among all ground truth faces (when thresholded by confidence >= X).
Precision and recall are represented by point estimates as well as posterior probability distribution characteristics using the model following [1]. The characteristics used are: 95% credible interval, mean, median, and mode.
[1] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation." European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2005.

Median IOD MAE
Median [over the data points] Interocular distance-normalized Mean [over the keypoints of one face] Absolute Error (MAE) on keypoint coordinates.
Interocular distance (IOD) is estimated as the distance between the eye centers computed as midpoints of segments connecting eye corners; MAE is represented as the percentage of the IOD.

Median IOD Jitter
Median Interocular distance-normalized "jittering" estimate.
MAE between backprojected results on slightly shifted images.
Evaluated similarly to Median IOD MAE, but instead of comparing the predictions of the model against the human annotations, they are related to the predictions of the model given a slightly shifted image (with an appropriate opposite shift of the results). Used to quantify the robustness of the model to e.g. camera movements.

Evaluation results

DATA

- **Dataset I.** Contains **1060** samples: **1020** images containing one or more faces, evenly distributed across **17 geographical subregions** (see the specification in Section "Training Factors and Subgroups"), i.e. **60** images **per region**, **plus 40 images not containing faces** (the same set of no-face images is used in each region).
- **Dataset II.** Contains **1000** samples of images containing exactly one face, **500** of which are perceived by the annotators as feminine and **500** as masculine, respectively.
- **Dataset III.** Contains **500** images with exactly one face each, **100** images per skin tone category as perceived by the annotators.

Note: The datasets I–III are not disjoint.

All samples are picked from the same source as the training samples and are characterized as smartphone front- and back-facing camera photos taken in real-world environments (see specification in "Factors and Subgroups - Instrumentation"). See the face size and angle distribution among the dataset images on the following page.

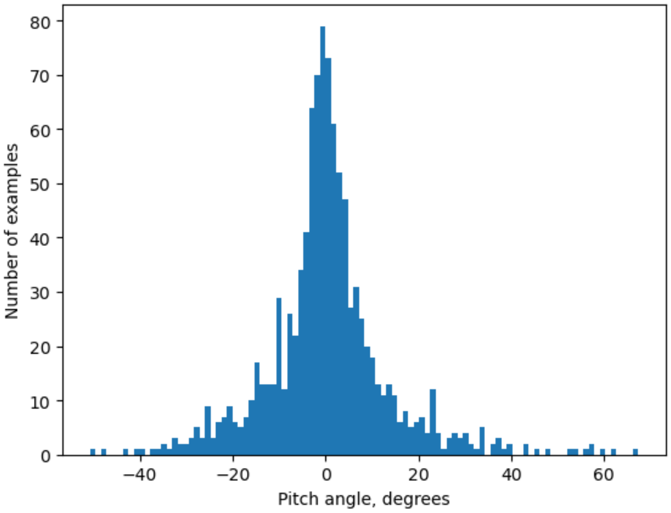
FAIRNESS EVALUATION RESULTS

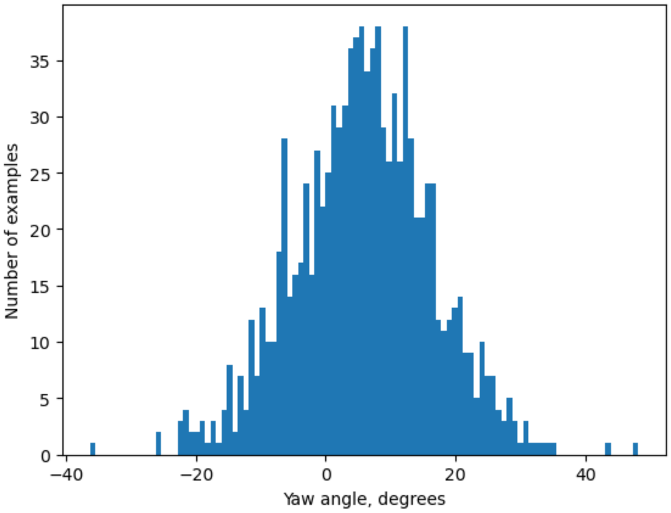
Detailed evaluation for the models across 17 geographical subregions, as well as annotator-perceived genders and 5 skin tone categories is presented in the accompanying spreadsheet.

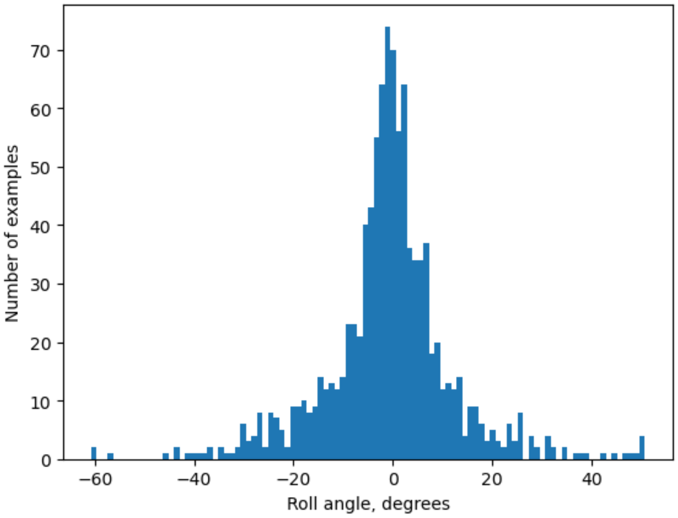
Average **recall** across perceived genders is 98.9% (99.6% vs. 98.2%); across skin tones it is 99.2%, varying from 98% to 100%; and across subregions the recall varies from 96.7% to 100% with the global average of 99.3%.
Average **precision** is consistently 99.8% for both genders, varies between 99% to 100% across skin tones with the combined average of 99.6%, and between 88.1% and 89.6% across subregions with the global average of 89.3%.

FACE ORIENTATION ANGLE AND SIZE DISTRIBUTION

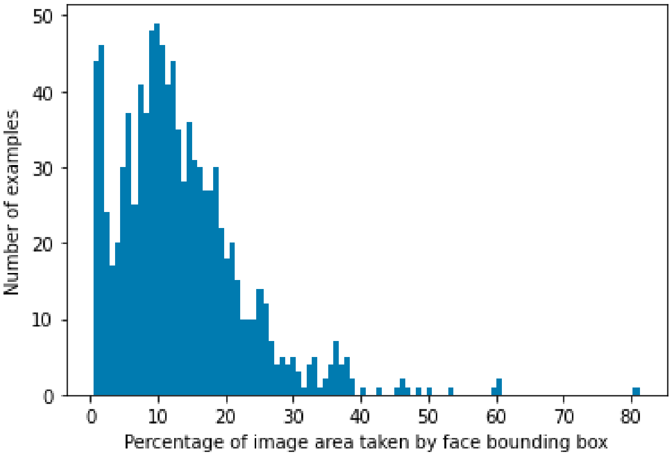
The angles were derived automatically from a fitted 3D facial mesh via MediaPipe Face Geometry Module.







FACE SIZES



Definitions

BOUNDING BOX

A bounding box is an axis-aligned rectangle containing the object of interest (a face in our case).

KEYPOINTS

"Keypoints" or "landmarks" are prominent facial locations. The models represent them with (x, y) coordinates.

AUGMENTED REALITY (AR)

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