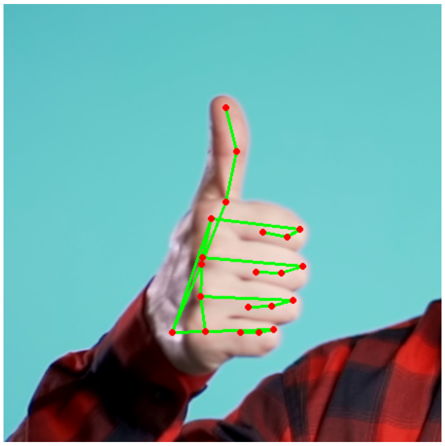
MODEL CARD

MediaPipe Hands (Sparse)

MODEL DETAILS

Lightweight hand landmarks model (3.8MB size), to predict 3D hand landmarks within a cropped region on an image on a smartphone. The hand landmark model predicts keypoints for each hand from the cropped image. The crop comes from a separate hand detector.



MODEL SPECIFICATIONS

Model Type

- Convolutional Neural Network

Model Architecture

- Regression model

Inputs

- A crop of a frame of video or an image, represented as a 224 x 224 x 3 tensor. Channels order: RGB with values in [0.0, 1.0].

Output(s)

- 1) A float scalar represents the presence of a hand in the given input image.
- 2) 21 3-dimensional landmarks represented as a 1 x 63 tensor and normalized by image size. This output should only be considered valid when the presence score is higher than a threshold.
- 3) A float scalar represents the handedness of the predicted hand. This output should only be considered valid when the presence score is higher than a threshold.

AUTHORS
Artsiom Ablavatski, Google (artsiom@)
Fan Zhang, Google (zhafang@)
George Sung, Google (gsung@)

DATE
February 2021

Based on Model Cards for Model Reporting, In Proceedings of FAT* Conference (FAT*2019). ACM, New York, NY, USA, 9 pages

DOCUMENTATION

Blogpost:

Google AI blog post 2 March 2020

Example usage included as part of MediaPipe February 2021 release

LICENSED UNDER

Apache License, Version 2.0

Intended Uses

APPLICATION
Predicting landmarks within the crop of prominently displayed hands in images or videos captured by a smartphone camera.

DOMAIN & USERS
Mobile AR (augmented reality) applications.
Gesture recognition
Hand control

OUT-OF-SCOPE APPLICATIONS
Not appropriate for:

- Counting the number of hands in a crowd
- Predicting hand landmarks with gloves or occlusions. For example when the hand is holding objects or there is decoration on the hand including jewelry, tattoo and henna.
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology.

Limitations

TRAINING
The model has been trained on limited datasets and are meant for experimental usage.

PERFORMANCE
The model has not been tested in "in-the-wild" smartphone camera conditions, including low-end devices, low light, motion blur etc., that can affect performance.

Ethical Considerations

PRIVACY
This model was trained and evaluated on images, including consented images captured using a mobile AR application for smartphone cameras in various "in-the-wild" conditions.

HUMAN LIFE
The model is not intended for human life-critical decisions. The primary intended application is for research and entertainment purposes.

Training Factors and Subgroups

INSTRUMENTATION

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

ENVIRONMENTS
The model is trained on images with various lighting, noise and motion conditions and with diverse augmentations. However, its quality can degrade in extreme conditions.

GROUPS

The 14 groups are based on the United Nations geoscheme with the following amendments: Southern Asia and Western Asia have been united due to their size with Central Asia; Western Africa united with Middle Africa; Europe excludes EU countries.

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

Evaluation metrics

Model Performance Measures

NORMALIZATION BY PALM SIZE

Normalization by palm size is applied to unify the scale of the samples. Palm size is calculated as the distance between the wrist and the first joint (MCP) of the middle finger.

MEAN ABSOLUTE ERROR

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

MNAE

For quality and fairness evaluation, we use MNAE **(Mean of Normalized Absolute Error by palm size)**.

Evaluation results

Geographical Evaluation Results

DATA

- **700 images, 50 images from each of the 14 geographical subregions** (see specification in Section "Factors and Subgroups").
- All samples are picked from the same source as training samples and are characterized as smartphone camera photos taken in real-world environments (see specification in "Factors and Subgroups - Instrumentation").

METHOD

For the geographical evaluation to estimate the cropped region on the image the ground truth information have been used. For the skin tone and gender evaluation end-to-end hand tracking pipeline has been employed via using a hand detector.

EVALUATION RESULTS

Detailed evaluation for hand tracking across 14 geographical subregions is presented in the table below.

Region	MNAE	Standard deviation
Australia and New Zealand	13.3	12.9
Central America	14.3	21.3
Caribbean	13.6	20.6
Central Asia	12.3	19.8
Eastern Africa	9.0	7.6
Eastern Asia	11.7	13.6
Europe	13.7	19.3
Middle Africa	**7.4**	6.4
Northern Africa	13.9	18.9
Northern America	13.0	23.9
Melanesia + Micronesia + Polynesia	7.6	6.7
Southern Africa	**15.5**	21.7
South America	12.0	16.6
Southeastern Asia	13.9	16.6
average	12.2	
range	+3.3/-4.8	

Geographical Fairness Evaluation Results

FAIRNESS METRICS & BASELINE

We asked 5 annotators to re-annotate the validation dataset, yielding an MNAE of **6.0%**
This is a high inter-annotator agreement, suggesting that the MNAE metric is a strong indicator of the hand landmarks.

FAIRNESS RESULTS

Evaluation across 14 regions on the validation dataset yields an average performance of 12.2% +/- 2.5% stdev with a range of [7.4%, 15.5%] across regions.

We found that per-joint MNAE is the smallest at the base of each finger, and gets larger toward the fingertip. We conjecture that the prediction is easier around the palm which is more rigid than the fingers. We also found that the normalized absolute error is larger for blurry or occluded joints. The findings are consistent across all regions. We didn't find any error pattern with regard to the regions.

Skin Tone and Gender

DATA

- **420 images, 35 images from each unique combination of the perceived gender and the skin tone** (from 1 to 6) based on the Fitzpatrick scale.
- All samples are picked from the same source as training samples and are characterized as smartphone camera photos taken in real-world environments (see specification in "Factors and Subgroups - Instrumentation").

FAIRNESS METRICS & BASELINE

We asked 5 annotators to re-annotate the validation dataset, yielding an MNAE of **3.8%**
This is a high inter-annotator agreement, suggesting that the MNAE metric is a strong indicator of the hand landmarks.

FAIRNESS RESULTS

Evaluation across 6 skin tone types on the validation dataset yields an average performance of 8.1% +/- 1.0% stdev with a range of [6.6%, 9.5%] across types.

Evaluation across genders on the validation dataset yields an average performance of 8.09% with a range of [8.04%, 8.15%].

Our findings are the same as in geographical fairness evaluation results above. We didn't find any error pattern with regard to the skin tone types or the gender.

Skin tone type	MNAE	Standard deviation
1	**9.5**	12.0
2	8.0	8.6
3	**6.6**	9.4
4	7.7	8.5
5	8.2	13.2
6	8.7	11.1
average	8.1	
range	+1.4/-1.5	

Gender	MNAE	Standard deviation
female	8.04	10.1
male	8.15	11.4
average	8.09	
range	+0.1/-0.1	

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.

KEYPOINTS
Hand "keypoints" or "landmarks" are (x, y, z) coordinate locations of hand features.