MODEL CARD

MediaPipe Hands

MODEL DETAILS

Two lightweight models, a palm detector (3.7MB size) and a hand landmark model (3.7MB size), to detect palm and predict hand landmarks within an image on a smartphone. Palm detector returns bounding boxes for each palm and hand landmark model predicts keypoints for each hand from the cropped image.

MODEL SPECIFICATIONS

Model Type

- Convolutional Neural Network

Model Architecture

- Palm detector: Adapted SSD with a custom encoder
- Hand landmark model: regression model

Inputs

- Palm detector: A frame of video or an image, represented as a 128 x 128 x 3 tensor. Channels order: RGB with values in [-1.0, 1.0].
- Hand landmark model: 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)

- Palm detector: 1) Predicted offset of predefined anchors represented as a 1 x 896 x 18 tensor. 2) Predicted detection confidence score of each anchor represented as a 1 x 896 tensor.
- Hand landmark model: 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 are 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

Fan Zhang, Google (zhafang@)
Valentin Bazarevsky, Google (bazarevsky@)

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

George Sung, Google (gsung@)

DATE
Unavailable

DOCUMENTATION
Blogpost:
Google AI blog post 19 Aug 2019
Example usage included as part of the open source MediaPipe example documentation hosted at"
http://github.com/google/mediapipe

LICENSED UNDER
Apache License, Version 2.0

Intended Uses

APPLICATION
Detecting prominently displayed hands within images or videos captured by a smartphone camera. This front-facing camera model focuses on relatively large hands.

DOMAIN & USERS
Mobile AR (augmented reality) applications.

OUT-OF-SCOPE APPLICATIONS
Not appropriate for:

- Counting the number of hands in a crowd
- Detecting hands with gloves or occlusions. For example the hand is holding objects or there is decoration on the hand including jewelry, tattoo and henna.
- Detecting hands too far away from the camera (e.g. further than 2 meters)
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology.

Limitations

TRAINING
The models have been trained on limited dataset and are meant for experimental usage.

PERFORMANCE
The models have 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 using a mobile AR application captured with 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

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.

INSTRUMENTATION

- The majority 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.

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 and is taken as 100%. 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 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.

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").

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	14.6	14.6
Central America	14.0	18.9
Caribbean	13.6	21.0
Central Asia	13.6	24.2
Eastern Africa	11.9	20.5
Eastern Asia	12.5	13.8
Europe	13.5	16.8
Middle Africa	8.0	7.1
Northern Africa	13.9	21.0
Northern America	12.2	16.6
Melanesia + Micronesia + Polynesia	**7.7**	6.0
Southern Africa	**15.1**	21.8
South America	12.6	14.8
Southeastern Asia	13.2	15.8
average	12.6	
range	+2.5/-4.9	

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 12.6% +/- 2.2% stdev with a range of [7.7%, 15.1%] 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 9.4% +/- 0.7% stdev with a range of [8.7%, 10.6%] across types.

Evaluation across genders on the validation dataset yields an average performance of 9.4% with a range of [9.3%, 9.6%].

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	9.1
2	9.6	11.0
3	9.0	12.3
4	**8.7**	10.0
5	9.2	11.6
6	**10.6**	12.2
average	9.4	

range	+1.1/-0.7	

Gender	MNAE	Standard deviation
female	9.6	11.1
male	9.3	11.1
average	9.4	
range	+0.2/-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.