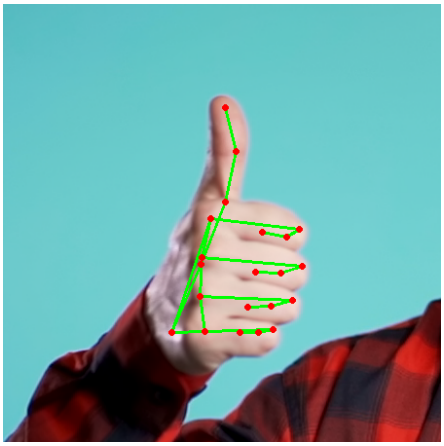
MODEL CARD

MediaPipe Hands (Lite/Full)

SOLUTION DETAILS

Hand tracking neural network pipelines: Lite and Full, to predict 2D and 3D hand landmarks on an image / video sequence. Both pipelines consist of:

- Hand detector model, which locates hand region
- Hand tracking model, which predict 2D keypoints, 3D world keypoints, handedness on a cropped area around hand
- MediaPipe graph, with hand tracking logic.



SOLUTION SPECIFICATIONS

Model Type

- Convolutional Neural Network

Model Architecture

- Two step neural network pipeline with single-shot detector and following regression model running on the cropped region.

Inputs

- A video stream or an image of arbitrary size. Channels order: RGB with values in [0.0, 1.0].

Output(s)

List of detected hands, each containing

- 21 3-dimensional screen landmarks
- A float scalar represents the handedness probability of the predicted hand.
- 21 3-dimensional metric scale world landmarks. Predictions are based on the GHUM hand model.

Landmark screen z-value and 3D metric x, y, z coordinate values estimate is provided using synthetic data, obtained via the GHUM model (articulated 3D human shape model) fitted to 2D point projections.

DETECTOR MODEL SPECIFICATIONS

Model Type

- Convolutional Neural Network

Model Architecture

- Single-shot detector model

Inputs

- A frame of video or an image, represented as a 192 x 192 x 3 tensor. Channels order: RGB with values in [0.0, 1.0].

Output(s)

- A float tensor 2016 x 18 of predicted embeddings representing anchors transformation which are further used in Non Maximum Suppression algorithm.

TRACKER 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)

- A float scalar represents the presence of a hand in the given input image.
- 21 3-dimensional screen 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.
- 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.
- 21 3-dimensional metric scale world landmarks represented as a 1 x 63 tensor. Predictions are based on the GHUM hand model. This output should only be considered valid when the presence score is higher than a threshold.

Landmark screen z-value and 3D metric x, y, z coordinate values estimate is provided using synthetic data, obtained via the GHUM model (articulated 3D human shape model) fitted to 2D point projections.

DOCUMENTATION

Blogpost:
Google AI blog post 2 March 2020
Example usage included as part of MediaPipe February 2021 release

GHUM Paper:
GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6184-6193, 2020

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 is 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.

MNAE

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

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 as the z screen coordinates as well as metric world coordinates are obtained from synthetic data, so for a fair comparison with human annotations, only 2D screen coordinates are employed.

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, the skin tone and gender evaluation end-to-end hand tracking pipeline has been employed via using a hand detector model, hand tracker model and MediaPipe tracking logic.

EVALUATION RESULTS

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

Region	Lite		Full	
	MNAE	Standard deviation	MNAE	Standard deviation
Australia and New Zealand	**13.94**	13.82	10.30	11.86
Central America	12.14	19.82	11.15	18.51
Caribbean	13.42	18.00	11.41	16.98
Central Asia	12.77	20.59	9.55	15.92
Eastern Africa	11.20	18.48	8.23	9.10
Eastern Asia	11.89	15.30	9.90	13.58
Europe	12.81	16.99	10.61	14.44
Middle Africa	**8.43**	8.28	9.19	9.67
Northern Africa	13.13	18.30	**13.00**	20.61
Northern America	11.20	13.04	9.39	10.34
Melanesia + Micronesia + Polynesia	8.91	6.52	**6.10**	5.15
Southern Africa	12.88	14.91	12.18	21.30
South America	12.29	16.35	9.18	10.11
Southeastern Asia	13.20	15.64	11.11	13.19
average	12.02		10.09	
range	+1.93/-3.58		+2.90/-3.99	

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 12.02% +/- 1.6% stdev with a range of [8.43%, 13.42%] across regions for the lite model and an average performance of 10.09% +/- 1.73% stdev with a range of [6.10%, 13.00%] across regions for the full model.

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.
We conjecture that the lower MNAE of this dataset, compared to the geographical dataset, is due to the lower difficulty in the data. The difficulty generally increases with more blurriness, more occlusions, and more image noise.

FAIRNESS RESULTS

Evaluation across 6 skin tone types on the validation dataset yields an average performance of 5.67% +/- 0.94% stdev with a range of [4.88%, 7.25%] across types for lite model and an average performance of 5.08% +/- 0.72% stdev with a range of [4.53%, 6.21%] across types for full model.

Evaluation across genders on the validation dataset yields an average performance of 5.67% with a range of [5.29%, 6.05%] for lite model and an average performance of 5.09% with a range of [4.80%, 5.38%] for full model.

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	Lite		Full	
	MNAE	Standard deviation	MNAE	Standard deviation
1	6.37	10.1	5.76	8.32
2	5.30	4.88	4.67	4.54
3	5.05	7.53	4.60	7.30
4	5.15	4.36	4.69	4.78
5	**4.88**	5.33	**4.53**	5.00
6	**7.25**	9.79	**6.21**	8.66
average	5.67		5.08	
range	+1.58/-0.79		+1.13/-0.55	

Gender	Lite		Full	
	MNAE	Standard deviation	MNAE	Standard deviation
female	**5.29**	6.72	**4.80**	5.40
male	**6.05**	8.11	**5.38**	7.87
average	5.67		5.09	
range	+0.38/-0.38		+0.29/-0.29	

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.

WORLD KEYPOINTS
Hand "world keypoints" or "world landmarks" are (x, y, z) metric scale coordinate locations of hand features. World keypoints 3D metric x, y, z coordinate values estimate is provided using synthetic data, obtained via the GHUM model (articulated 3D human shape model) fitted to 2D point projections.

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

HANDEDNESS
Handedness - flag indicating whether a particular hand is left or right.