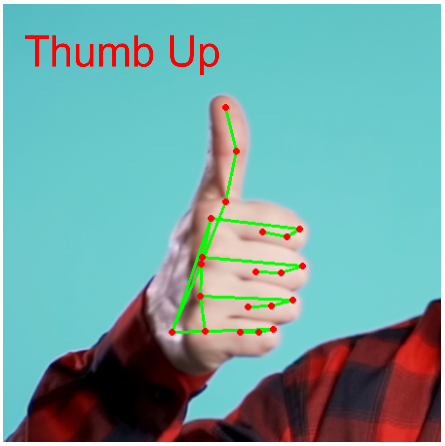
MODEL CARD

MediaPipe Hand Gesture Classification

SOLUTION DETAILS

Hand Gesture Classification model uses hand landmarks produced by MediaPipe Hands Model to classify a hand pose as one of the 8 hand gesture classes, namely,

- Closed Fist
- Open Palm
- Pointing Up
- Thumb Down
- Thumb Up
- Victory
- I Love You
- None of the above gestures



SOLUTION SPECIFICATIONS

Solution Architecture

- Two step neural network pipeline with an embedding model followed by a classification model. This pipeline runs on hand landmarks and related information for a single hand, but does not directly process any images (i.e. RGB pixel data).

Inputs
This pipeline consumes MediaPipe Hands model's outputs:

- 21 3-dimensional screen landmarks represented as a 1 x 63 tensor and normalized by image size.
- A float scalar represents the handedness probability of the predicted hand.
- 21 3-dimensional metric scale world landmarks represented as a 1 x 63 tensor and normalized by image size.
- Refer to this model card for more details.

No image data was directly input into the model.

Output(s)

An 8 element vector that predicts the probability of each of the following classes:

- 0th-element: probability that hand pose is not a known hand gesture to the model
- 1st-8th: probability of hand pose is one of the 7 known gestures.

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

EMBEDDING MODEL SPECIFICATIONS

Model Type

- Fully Connected Neural Network with residual blocks

Model Architecture

- Regression model

Inputs

- 21 3-dimensional screen landmarks represented as a 1 x 63 tensor and normalized by image size.
- A float scalar represents the handedness probability of the predicted hand.
- 21 3-dimensional metric scale world landmarks represented as a 1 x 63 tensor and normalized by image size.

Output(s)

- A float tensor 128x1 embedding tensor of predicted embedding representing the hand landmarks, which is further used in the classification model head, described in the next section.

CLASSIFICATION MODEL SPECIFICATIONS

Model Type

- Fully Connected Neural Network

Model Architecture

- Classification model

Inputs

- A float tensor 128x1 embedding tensor of predicted embedding representing the hand landmarks.

Output(s)

- An 8 element vector that predicts the probability for each of the 8 above mentioned gesture classes.

Intended Uses

APPLICATION

Predict if and what the hand gesture of a given hand's landmark information.

DOMAIN & USERS

Mobile AR (augmented reality) applications
Gesture recognition
Hand control

OUT-OF-SCOPE APPLICATIONS

Not appropriate for:

- Hand gestures involving multiple hands (e.g. two handed heart shape)
- Hand gestures involving motion (e.g. waving goodbye)
- Translate sign language
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology.

Limitations

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.

USER CUSTOMIZATION
Even though our pre-trained hand gesture classification model has not shown bias in perceived gender expression or perceived skin tones, the user trained custom gesture classification models that are based on our model, may introduce unintentional bias, depending on the datasets used to the custom model.

Ethical Considerations

PRIVACY
This model was trained and evaluated from the hand landmarks output data from MediaPipe Hands Model. This classification model does not directly use any images or videos as inputs.

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 hand landmarks from images with various lighting, noise and motion conditions and with diverse augmentations. However, its quality can degrade in extreme conditions.

Evaluation metrics

Model Performance Measures

Specificity-Sensitivity F1 Score (SS F1 Score) is defined as the harmonic mean of sensitivity and specificity, where specificity and sensitivity of the model is defined below.

Sensitivity (True Positive Rate) refers to the probability of a positive test, conditioned on truly being positive.

Specificity (True Negative Rate) refers to the probability of a negative test, conditioned on truly being negative.

Evaluation results

Perceived Skin Tone and Gender Expression

DATA

There are about 2000 images used in perceived Monk skin tone evaluation, and about 8000 images used in perceived gender expression evaluation.

The images are chosen from a diverse variety of perceived gender expressions and skin tones.

We aim to balance the number of images in each perceived skin tone, but since there are fewer people on the two ends of the perceived Monk Skin Tone scale, we are reporting combined statistics for tones 1-2 and 9-10, respectively.

FAIRNESS METRICS & BASELINE

We asked human annotators, who are trained to annotate perceived gender expressions and skin tones, to annotate these images.

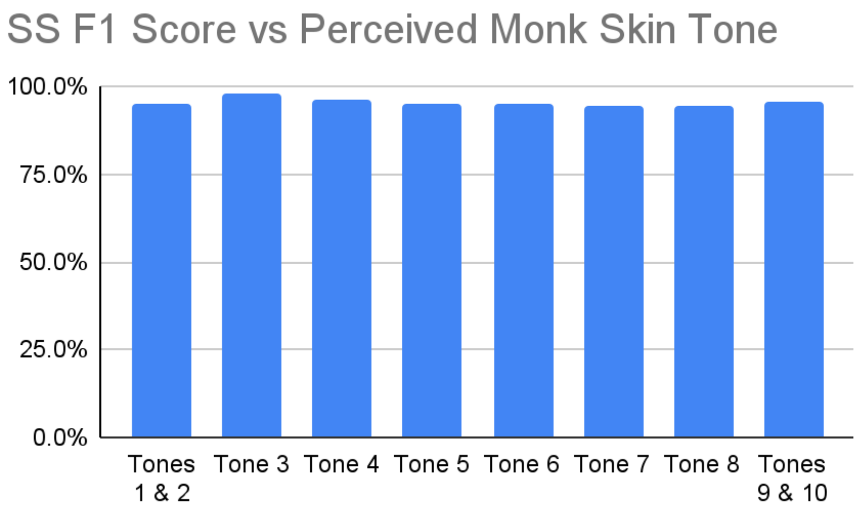
FAIRNESS RESULTS

Evaluation across perceived Monk Skin Tone types on the evaluation dataset yields a weighted average SS F1 score of 95.5% with a range of [94.3%-97.8%] for the model.

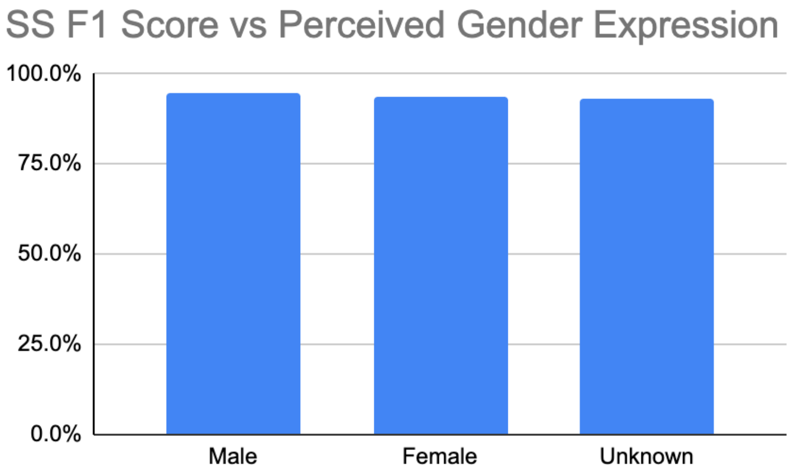
Evaluation across perceived gender expression types on the evaluation dataset yields a weighted average SS F1 score 93.9% with a range of [93.1%-94.4%] for the model.

We didn't find any error pattern regarding the perceived skin tone types or gender expressions, but we acknowledge the limitation of data collection of tones 9-10.

Perceived Monk Skin Tone	SS F1 Score	Δ from Wt Average	% of Dataset
Tones 1 & 2	95.2%	-0.4%	12.3%
Tone 3	97.8%	2.2%	13.4%
Tone 4	96.1%	0.5%	13.4%
Tone 5	95.1%	-0.5%	13.8%
Tone 6	95.4%	-0.2%	13.7%
Tone 7	94.3%	-1.3%	12.7%
Tone 8	94.6%	-1.0%	10.5%
Tones 9 & 10	95.8%	0.3%	10.3%
Wt Average	95.5%		100.0%



Perceived Gender Expression	SS F1 Score	Δ from Wt Average	% of Dataset
Male	94.4%	0.5%	49.6%
Female	93.3%	-0.5%	49.6%
Unknown	93.1%	-0.8%	0.8%
Wt Average	93.9%		100.0%



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.