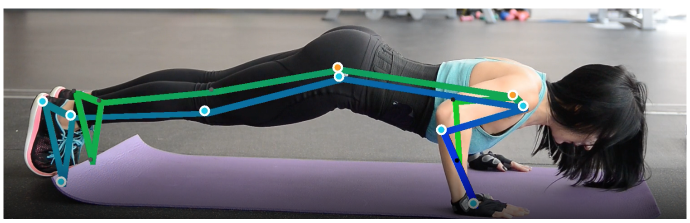
MODEL CARD

MediaPipe BlazePose GHUM 3D

MODEL DETAILS

Lite (3MB size), Full (6 MB size) and Heavy (26 MB size) models, to estimate the **full 3D body pose** of an individual in videos captured by a **smartphone or web camera**. Optimized for **on-device, real-time fitness applications**: Lite model runs ~44 FPS on a CPU via XNNPack TFLite and ~49 FPS via TFLite GPU on a Pixel 3. Full model runs ~18 FPS on a CPU via XNNPack TFLite and ~40 FPS via TFLite GPU on a Pixel 3. Heavy model runs ~4 FPS on a CPU via XNNPack TFLite and ~19 FPS via TFLite GPU on a Pixel 3.



Depth is encoded via gradient from blue (closer) to green (further). Invisible (occluded) keypoints marked as black.

Returns 33 keypoints describing the approximate location of body parts:

- Nose
- Right eye (3 keypoints): Inner, Center, Outer
- Left eye (3 keypoints): Inner, Center, Outer
- Ears (2 keypoints): Right, Left
- Mouth (2 keypoints): Right Corner, Left Corner
- Shoulder (2 keypoints): Right, Left
- Elbow (2 keypoints): Right, Left
- Wrist (2 keypoints): Right, Left
- Pinky knuckle (2 keypoints): Right, Left
- Index knuckle (2 keypoints): Right, Left
- Thumb knuckle (2 keypoints): Right, Left
- Hip (2 keypoints): Right, Left
- Knee (2 keypoints): Right, Left
- Ankle (2 keypoints): Right, Left
- Heels (2 keypoints): Right, Left
- Foot Index (2 keypoints): Right, Left

MODEL SPECIFICATIONS

Model Type
Convolutional Neural Network

Model Architecture
Convolutional Neural Network: MobileNetV2-like with customized blocks for real-time performance.

Input(s)
Regions in the video frames where a person has been detected. Represented as a 256x256x3 array with aligned human full body part, centered by mid-hip in vertical body pose and rotation distortion of (-10, 10). Channels order: RGB with values in [0.0, 1.0].

Output(s)

- 33x5 tensor corresponding to screen projected keypoints (x, y, z, visibility, presence).
- 33x3 tensor corresponding to 3D world metric scale coordinates (world x, world y, world z).
- Scalar in range from [0.0, 1.0] corresponding to the presence flag indicating the probability a person is present on a passed image.

For more details about model output ranges and scale consider the “Model outputs detailed specification” section.

Keypoint screen z-value and 3D world 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.

MODEL OUTPUTS DETAILED SPECIFICATION

- X, Y screen projected coordinates are local to the region of interest and range from [0.0, 255.0].
- Z coordinate is measured in "image pixels" like the X and Y screen coordinates and represents the distance relative to the plane of the subject's hips, which is the origin of the Z axis. Negative values are between the hips and the camera; positive values are behind the hips. Z coordinate scale is similar with X, Y scales but has different nature as obtained not via human annotation, by fitting synthetic data ([GHUM model]) to the 2D annotation. Note, that Z is not metric but up to scale.
- Visibility is in the range of [min_float, max_float] and after user-applied sigmoid denotes the probability that a keypoint is located within the frame and not occluded by another bigger body part or another object.
- Presence is in the range of [min_float, max_float] and after user-applied sigmoid denotes the probability that a keypoint is located within the frame.
- World X, Y, Z coordinates, representing keypoint location in space, measured in meters and normalized to center of subject hips and range from [-1.5, 1.5]. Whis coordinates obtained not via human annotation, but by fitting synthetic data (GHUM model) to the 2D annotation, person foreground/background segmentation mask and camera intrinsic parameters.

AUTHORS

Who created this model?

Valentin Bazarevsky, Google
Ivan Grishchenko, Google
Eduard Gabriel Bazavan, Google

DATE

June, 22, 2021

CITATION

How can users cite your model?

BlazePose: On-device Real-time Body Pose tracking, CVPR Workshop on Computer Vision for Augmented and Virtual Reality,
Seattle, WA, USA, 2020

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

DOCUMENTATION

- BlazePose: On-device Real-time Body Pose tracking
- GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models

LICENSED UNDER

Apache License, Version 2.0

Intended Uses

APPLICATION

3D full body pose estimation for single-person videos on mobile, desktop and in browser.

DOMAIN AND USERS

- Augmented reality
- 3D Pose and gesture recognition
- Fitness and repetition counting
- 3D pose measurements (angles / distances)

OUT-OF-SCOPE APPLICATIONS

- Multiple people in an image.
- People too far away from the camera (e.g. further than 14 feet/4 meters)
- Head is not visible
- Applications requiring metric accurate depth
- Any form of surveillance or identity recognition is explicitly out of scope and not enabled by this technology

Limitations

PRESENCE OF ATTRIBUTES

Tracks only one person on scene if multiple 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

When degrading the 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.

PRIVACY

This model was trained and evaluated on images, including consented images (30K), of people using a mobile AR application captured with smartphone cameras in various "in-the-wild" conditions. The majority of training images (85K) capture a wide range of fitness poses.

BIAS

This model was trained and evaluated both on visible and hidden points. For cases where the point location is present but hard to define by a human annotator, it is annotated with a "best guess" and default pose.
Model has been qualitatively evaluated on users with missing limbs and prosthetics and degrades gracefully by predicting average point location.

The model is providing 3D coordinates obtained from synthetic data using the GHUM model (articulated 3D human shape model), fitted via an algorithm to the 2D key point projections.

Training Factors and Subgroups

INSTRUMENTATION

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

ENVIRONMENTS

Model is trained on images with various lighting, noise and motion conditions and with diverse augmentations. However, its quality can degrade in extreme conditions. This may lead to increased "jittering" (inter-frame prediction noise).

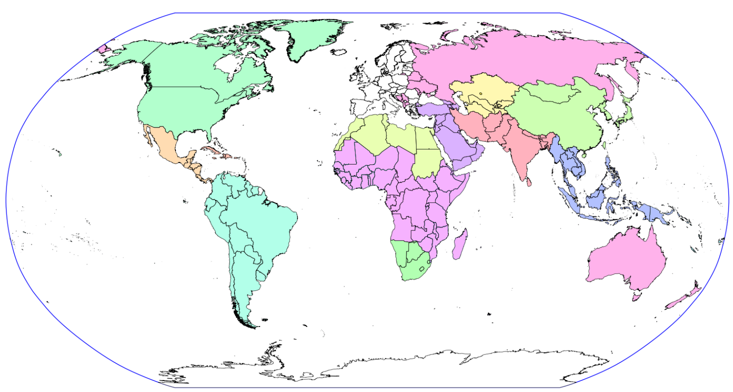
ATTRIBUTES

- Human Full-body cropped from the captured frame should contain a single person placed in the center of the image.
- There should be a margin around the square circumscribing full-body calculated as 25% of size.
- Model is tolerant to certain level of input inaccuracy:
  - 10% shift and scale (taking body width/height as 100% for corresponding axis)
  - 8° roll

GROUPS

To perform fairness evaluation we group user samples into 14 evenly distributed geographic subregions (based on [United Nations geoscheme]() with merges):

Central America	Caribbean
Southern America	Northern America
Central Asia	Northern Africa
Eastern Asia	Middle Africa
Southeastern Asia	Southern Africa
Southern Asia	Australia and New Zealand
Western Asia	Europe (excluding EU)

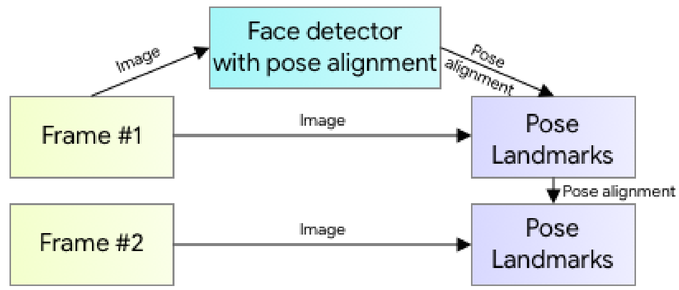


Evaluation modes and metrics

Evaluation Modes

TRACKING MODE

Main mode that takes place most of the time and is based on obtaining a highly accurate full-body crop from the prediction on the previous frame (frames 2, 3, ... on the image)



Model Performance Measures

PDJ, Average percentage of detected joints
(Also known as PCK@0.2 - Percent of Correct Keypoints)

https://github.com/cbsudux/Human-Pose-Estimation-101

We consider a keypoint to be correctly detected if predicted visibility for it matches ground truth and the absolute 2D Euclidean error between the reference and target keypoint normalized by the 2D torso diameter projection is smaller than 20%. This value was determined during development as the maximum value that does not degrade accuracy in classifying pose / asana based solely on the key points without perceiving the original RGB image.

The model is providing 3D coordinates, but the screen z-coordinate, as well as world 3D coordinates obtained from synthetic data, so for a fair comparison with human annotations, only 2D screen coordinates are employed.

Evaluation results

Geographical Evaluation Results

DATA

- **Contains 1400 samples evenly distributed across 14 geographical subregions** (see specification in Section "Factors and Subgroups"). Each region contains 100 images.
- All samples are picked from the same source as training samples and are characterized as smartphone back-facing camera photos taken in real-world environments (see specification in "Factors and Subgroups - Instrumentation").

EVALUATION RESULTS

Detailed evaluation for the tracking modes across 14 geographical subregions, gender and skin tones is presented in the table below

Region	Lite model		Full model		Heavy model	
	PDJ	Standard deviation	PDJ	Standard deviation	PDJ	Standard deviation
Australia and New Zealand	94.1	8.3	95.4	7.6	98.0	3.9
Caribbean	94.8	8.3	97.7	4.9	98.9	2.9
Europe	90.3	12.6	95.1	8.4	97.8	4.8
Northern Africa	94.7	8.2	97.7	5.2	98.9	3.3
South America	95.0	8.3	97.8	4.8	99.0	3.0
Southeastern Asia	94.5	8.1	97.1	5.1	98.5	3.9
Western Asia	94.9	7.8	97.7	5.6	99.0	2.9
Central America	95.4	6.5	97.6	4.5	98.6	2.9
Central Asia	94.3	8.7	96.8	5.4	97.9	4.7
Eastern Asia	91.3	12.5	94.6	8.3	97.4	5.1
Middle Africa	94.6	9.5	96.6	6.6	98.3	4.6
Northern America	93.4	8.6	96.2	6.3	98.7	3.3
Southern Africa	92.9	10.0	95.1	6.9	97.0	6.3
Southern Asia	93.4	9.0	96.8	6.2	98.2	4.5
Average	**93.8**		**96.6**		**98.3**	
Range	**5.1**		**3.2**		**2.0**	

Geographical Fairness Evaluation Results

FAIRNESS CRITERIA

We consider a model to be performing unfairly across representative groups if the error range on them spans more than ~3x the human annotation discrepancy, in our case a total of **7.5% PDJ.**

FAIRNESS METRICS & BASELINE

We asked two annotators to re-annotate the Pose Validation dataset, yielding a PDJ of **97.5%**
This is a high inter-annotator agreement, suggesting that the PDJ metric is a strong indicator of precise matches between predicted keypoints and ground truth keypoints.

FAIRNESS RESULTS

Evaluation across 14 regions of heavy, full and lite models on smartphone back-facing camera photos dataset results an average performance of 98.3% +/- 0.6% stdev with a range of [97.0%, 99.0%] across regions for the heavy model, an average performance of 96.6% +/- 1.3% stdev with a range of [94.6%, 97.8%] across regions for the full model and an average performance of 93.8% +/-1.5% stdev with a range of [90.3%, 95.4%] across regions for the lite model.
Comparison with our fairness criteria yields a maximum discrepancy between average and worst performing regions of 2.0% for the heavy, 3.2% for the full and 5.1% for the light model.

Skin Tone and Gender Evaluation Results

DATA

- **1400 images, 100 images from each of 14 the geographical subregions** were annotated with perceived gender and skin tone (from 1 to 6) based on the Fitzpatrick scale.

EVALUATION RESULTS

Evaluation on smartphone back-facing camera photos dataset results in an average performance of 98.2% with a range of [97.7%, 98.8%] across all skin tones for the heavy model, an average performance of 96.3% with a range of [94.7%, 97.1%] across all skin tones for the full model and an average performance of 94.2% with a range of [91.4%, 96.3%] across regions for the lite model. The maximum discrepancy between worst and best performing categories is 1.1% for the heavy model, 2.5% for the full model and 4.9% for the lite model.

Evaluation across gender yields an average performance of 98.9% with a range of [97.9%, 98.9%] for the heavy model, an average performance of 96.7% with a range of [96.0%, 97.3%] for the full model, and an average of 93.9% with a range of [93.1%, 94.7%] for the lite model. The maximum discrepancy is 1.0% for the heavy model, 1.3% for the full model and 1.6% for the lite model.

Skin tone type	**% of dataset**	**Lite model**		**Full model**		**Heavy model**	
		PDJ	**Standard deviation**	**PDJ**	**Standard deviation**	**PDJ**	**Standard deviation**
1	1.3	96.3	2.5	95.1	5.5	98.8	1.4
2	9.5	91.4	10.1	94.7	7.7	97.7	4.2
3	34.3	93.9	9.3	96.7	6.1	98.2	4.4
4	36.2	94.3	9.0	97.0	6.2	98.6	3.9
5	14.2	94.5	9.3	97.2	5.5	98.5	3.9
6	4.5	96.1	7.1	97.1	5.8	98.7	3.7
Average		**94.2**		**96.3**		**98.2**	
Range		**4.9**		**2.5**		**1.1**	

Gender	**% of dataset**	**Lite model**		**Full model**		**Heavy model**	
		PDJ	**Standard deviation**	**PDJ**	**Standard deviation**	**PDJ**	**Standard deviation**
Male	45.9	94.7	9.0	97.3	5.67	98.9	3.5
Female	54.1	93.1	9.5	96.0	6.80	97.9	4.7
Average		**93.9**		**96.7**		**98.4**	
Range		**1.6**		**1.3**		**1.0**	

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.

PERSPECTIVE PROJECTION
Perspective projection or perspective transformation is a linear projection where three dimensional objects are projected on a picture plane.

KEYPOINTS
"Keypoints" or "landmarks" are (x, y, z) coordinate locations of body parts.

In this model we separate key points into two groups:

- Screen landmarks - coordinate locations of body parts projected to the users screen.
- World landmarks - coordinate locations in the real world 3D space

VISIBILITY
Visibility denotes the probability that a keypoint is located within the frame and not occluded either by other body parts or other objects.

PRESENCE
Presence denotes the probability that a keypoint is located within the frame. It does not indicate whether the keypoint is occluded by another body part.