MODEL CARD

MediaPipe BlazePose Full

MODEL DETAILS

A lightweight model (15 MB size) to estimate the full body pose of an individual in videos captured by a smartphone or web camera. Runs real-time (~10 FPS) on a Pixel 2 CPU single-core via XNNPack TFLite backend.

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
- Heel Ears (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)
33x3 array corresponding to (x, y, visibility). X, Y coordinates are local to the region of interest and range from [0.0, 255.0]. 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. It does *not* indicate whether the keypoint is occluded by another body part.

AUTHORS

Who created this model?

Valentin Bazarevsky, Google
Ivan Grishchenko, Google
Fan Zhang, Google

DATE
July, 06, 2020

DOCUMENTATION

Paper: https://arxiv.org/abs/2006.10204

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

LICENSED UNDER

Intended Uses

APPLICATION

Full body pose estimation for single-person videos.

DOMAIN AND USERS

- Augmented reality
- Gesture recognition
- Fitness

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
- 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 consented images of people using a mobile AR application captured with smartphone cameras in various “in-the-wild” conditions.

BIAS

This model was trained and evaluated both on visible and hidden points. For cases that the point location is present but hard to define by humans 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.

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

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.

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 face width/height as 100% for corresponding axis)
  - 8° roll

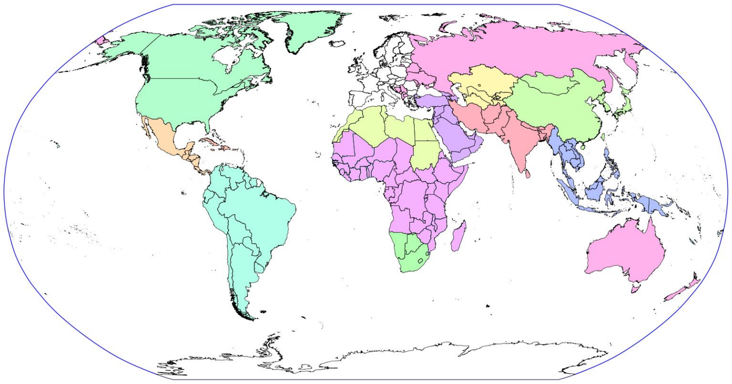
ENVIRONMENTS

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).

GROUPS

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

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

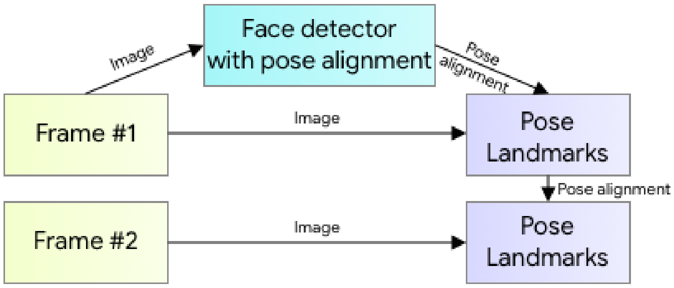


Evaluation modes and metrics

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 Euclidean error between the reference and target keypoint normalized by the torso radius 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.

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 is presented in the table below

Region	**PDJ**	**Standard deviation**
Australia and New Zealand	81.4	16.1
Caribbean	85.4	14.8
Europe	**80.9**	14.9
Northern Africa	85.4	13.5
South America	84.1	14.6
Southeastern Asia	84.1	12.7
Western Asia	84.4	11.8
Central America	84.8	14.7
Central Asia	84.4	14.0
Eastern Asia	81.0	14.3
Middle Africa	85.9	14.4
Northern America	**85.5**	13.9
Southern Africa	83.6	13.5
Southern Asia	86.8	13.2
Total for all regions	**84.0**	**14.2**

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 ~2x the human annotation discrepancy, in our case a total of **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

Comparison with *our fairness goal of **5% PDJ*** discrepancy across 14 regions in tracking mode yields a range from 80.9% to 85.5%.

The model is "fair" based on PDJ parity within a **4.6%** point window.

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

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

VISIBILITY

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