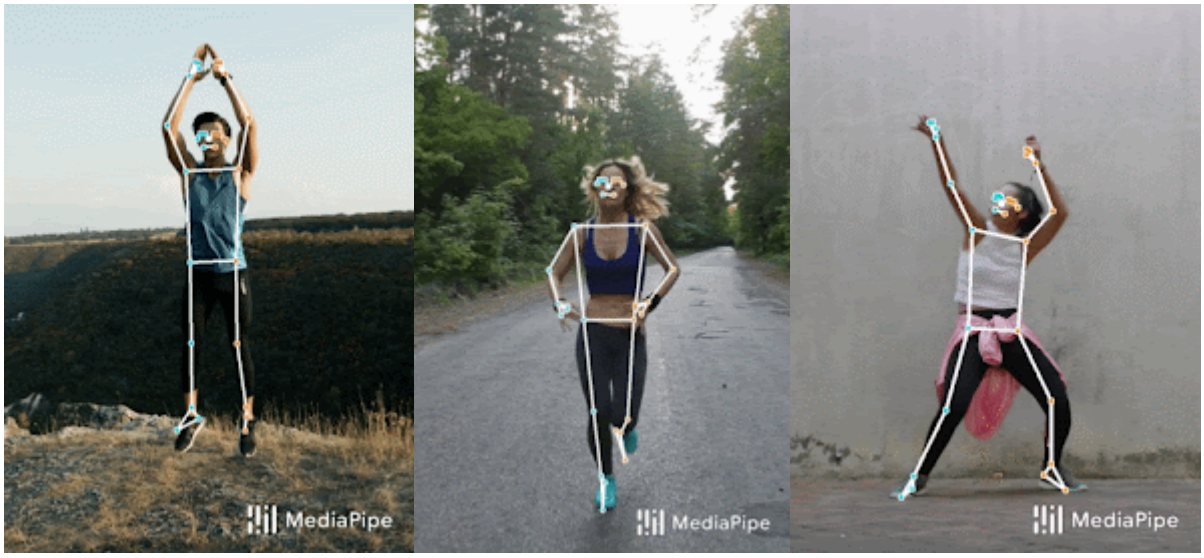
On-device, Real-time Body Pose Tracking with MediaPipe BlazePose

Thursday, August 13, 2020

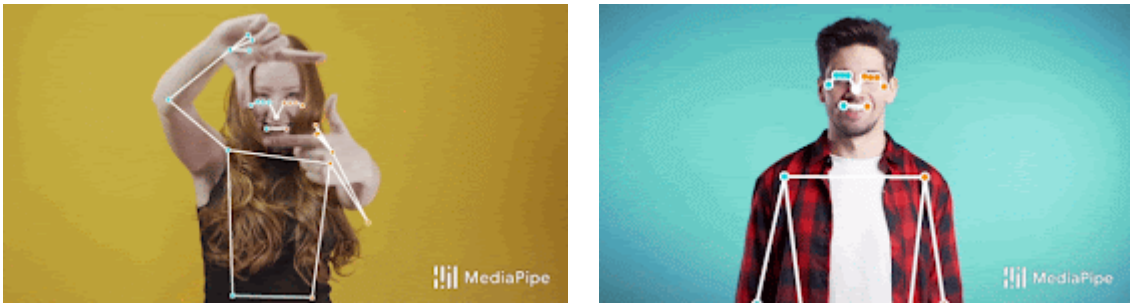
Posted by Valentin Bazarevsky and Ivan Grishchenko, Research Engineers, Google Research

Pose estimation from video plays a critical role enabling the overlay of digital content and information on top of the physical world in augmented reality, sign language recognition, full-body gesture control, and even quantifying physical exercises, where it can form the basis for yoga, dance, and fitness applications. Pose estimation for fitness applications is particularly challenging due to the wide variety of possible poses (e.g., hundreds of yoga asanas), numerous degrees of freedom, occlusions (e.g. the body or other objects occlude limbs as seen from the camera), and a variety of appearances or outfits.



BlazePose results on fitness and dance use-cases.

Today we are announcing the release of a new approach to human body pose perception, BlazePose, which we presented at the CV4ARVR workshop at CVPR 2020. Our approach provides human pose tracking by employing machine learning (ML) to infer 33, 2D landmarks of a body from a single frame. In contrast to current pose models based on the standard COCO topology, BlazePose accurately localizes more keypoints, making it uniquely suited for fitness applications. In addition, current state-of-the-art approaches rely primarily on powerful desktop environments for inference, whereas our method achieves real-time performance on mobile phones with CPU inference. If one leverages GPU inference, BlazePose achieves super-real-time performance, enabling it to run subsequent ML models, like face or hand tracking.

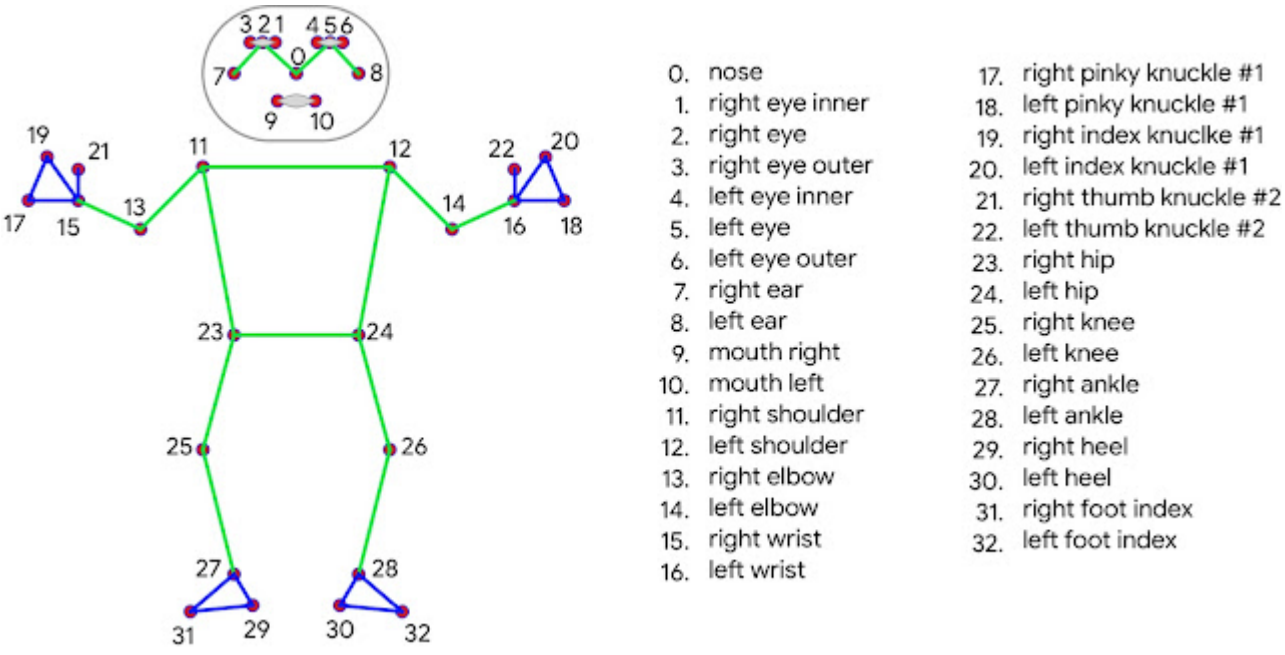


Upper-body BlazePose model in MediaPipe

Topology

The current standard for human body pose is the COCO topology, which consists of 17 landmarks across the torso, arms, legs, and face. However, the COCO keypoints only localize to the ankle and wrist points, lacking scale and orientation information for hands and feet, which is vital for practical applications like fitness and dance. The inclusion of more keypoints is crucial for the subsequent application of domain-specific pose estimation models, like those for hands, face, or feet.

With BlazePose, we present a new topology of 33 human body keypoints, which is a superset of COCO, BlazeFace and BlazePalm topologies. This allows us to determine body semantics from pose prediction alone that is consistent with face and hand models.

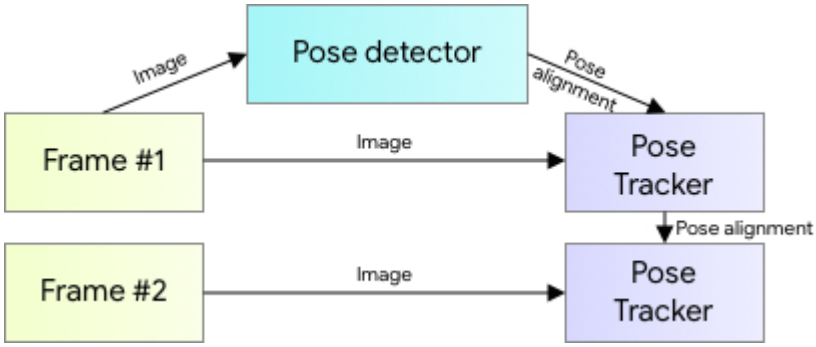


BlazePose 33 keypoint topology as COCO (colored with green) superset

Overview: An ML Pipeline for Pose Tracking

For pose estimation, we utilize our proven two-step detector-tracker ML pipeline. Using a detector, this pipeline first locates the pose region-of-interest (ROI) within the frame. The tracker subsequently predicts all 33 pose keypoints from this ROI. Note that for video use cases, the

detector is run only on the first frame. For subsequent frames we derive the ROI from the previous frame's pose keypoints as discussed below.

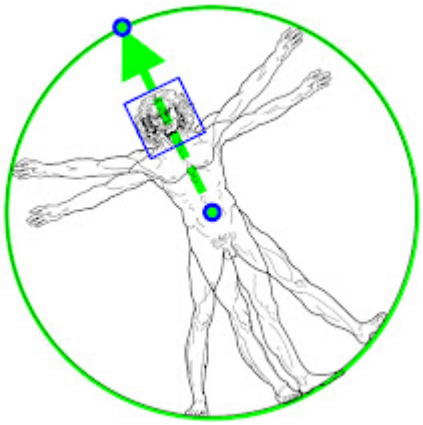


Human pose estimation pipeline overview.

Pose Detection by extending BlazeFace

For real-time performance of the full ML pipeline consisting of pose detection and tracking models, each component must be very fast, using only a few milliseconds per frame. To accomplish this, we observe that the strongest signal to the neural network about the position of the torso is the person's face (due to its high-contrast features and comparably small variations in appearance). Therefore, we achieve a fast and lightweight pose detector by making the strong (yet for many mobile and web applications valid) assumption that the head should be visible for our single-person use case.

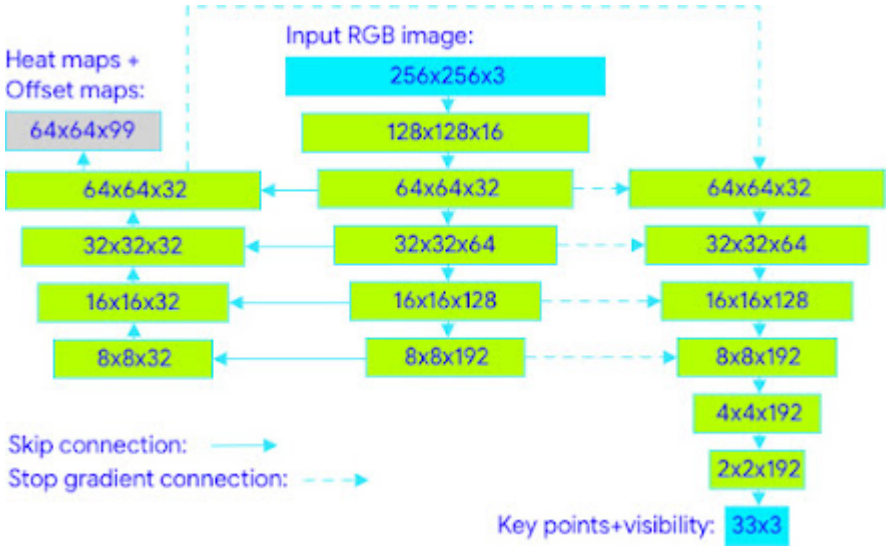
Consequently, we trained a face detector, inspired by our sub-millisecond BlazeFace model, as a proxy for a pose detector. Note, this model only detects the location of a person within the frame and can not be used to identify individuals. In contrast to the Face Mesh and MediaPipe Hand tracking pipelines, where we derive the ROI from predicted keypoints, for the human pose tracking we explicitly predict two additional *virtual* keypoints that firmly describe the human body center, rotation and scale as a circle. Inspired by Leonardo's Vitruvian man, we predict the midpoint of a person's hips, the radius of a circle circumscribing the whole person, and the incline angle of the line connecting the shoulder and hip midpoints. This results in consistent tracking even for very complicated cases, like specific yoga asanas. The figure below illustrates the approach.



Vitruvian man aligned via two virtual keypoints predicted by our BlazePose detector in addition to the face bounding box

Tracking Model

The pose estimation component of the pipeline predicts the location of all 33 person keypoints with three degrees of freedom each (*x, y* location and visibility) plus the two virtual alignment keypoints described above. Unlike current approaches that employ compute-intensive heatmap prediction, our model uses a regression approach that is *supervised* by a combined heat map/offset prediction of all keypoints, as shown below.



Tracking network architecture: regression with heatmap supervision

Specifically, during training we first employ a heatmap and offset loss to train the center and left tower of the network. We then remove the heatmap output and train the regression encoder (right tower), thus, effectively using the heatmap to supervise a lightweight embedding.

The table below shows an ablation study of the model quality resulting from different training strategies. As an evaluation metric, we use the Percent of Correct Points with 20% tolerance (PCK@0.2) (where we assume the point to be detected correctly if the 2D Euclidean error is smaller than 20% of the corresponding person's torso size). To obtain a human baseline, we asked annotators to annotate several samples redundantly and obtained an average PCK@0.2 of 97.2. The training and validation have been done on a geo-diverse dataset of various poses, sampled uniformly.

	Mean 2D Euclidean error, normalized by torso size	*PCK@0.2*
Heatmaps	16.2	83.6
Regression without Heatmaps loss	15.9	79.9
Regression with heatmap regularization	**14.4**	**84.1**

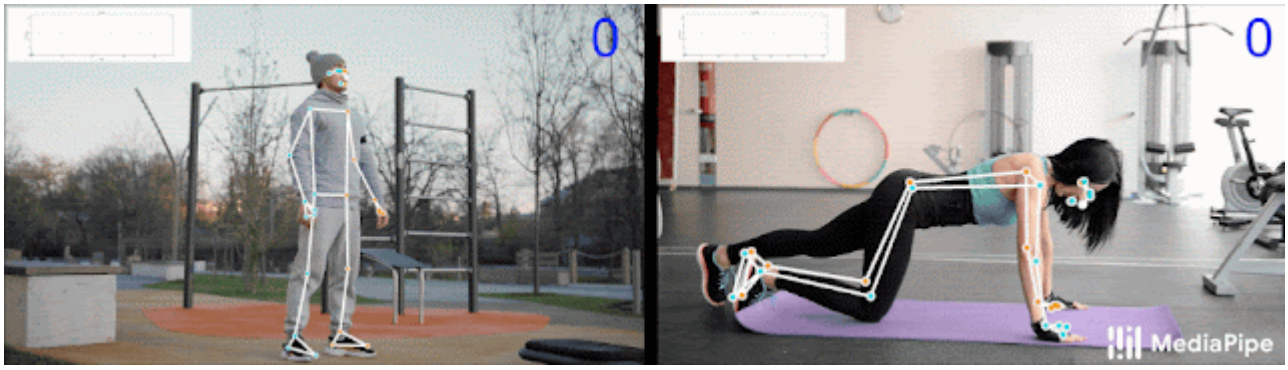
To cover a wide range of customer hardware, we present two pose tracking models: lite and full, which are differentiated in the balance of speed versus quality. For performance evaluation on

CPU, we use XNNPACK; for mobile GPUs, we use the TFLite GPU backend.

	Pixel 3 CPU, FPS using XNNPack	*Pixel 3 GPU, FPS using TFLite GPU*	*Browser, MBP '17, 3.5 GHz i7, FPS MediaPipe WASM*
BlazePose lite	44	112	13.5
BlazePose full	18	69	6.2

Applications

Based on human pose, we can build a variety of applications, like fitness or yoga trackers. As an example, we present squats and push up counters, which can automatically count user statistics, or verify the quality of performed exercises. Such use cases can be implemented either using an additional classifier network or even with a simple joint pairwise distance lookup algorithm, which matches the closest pose in normalized pose space.

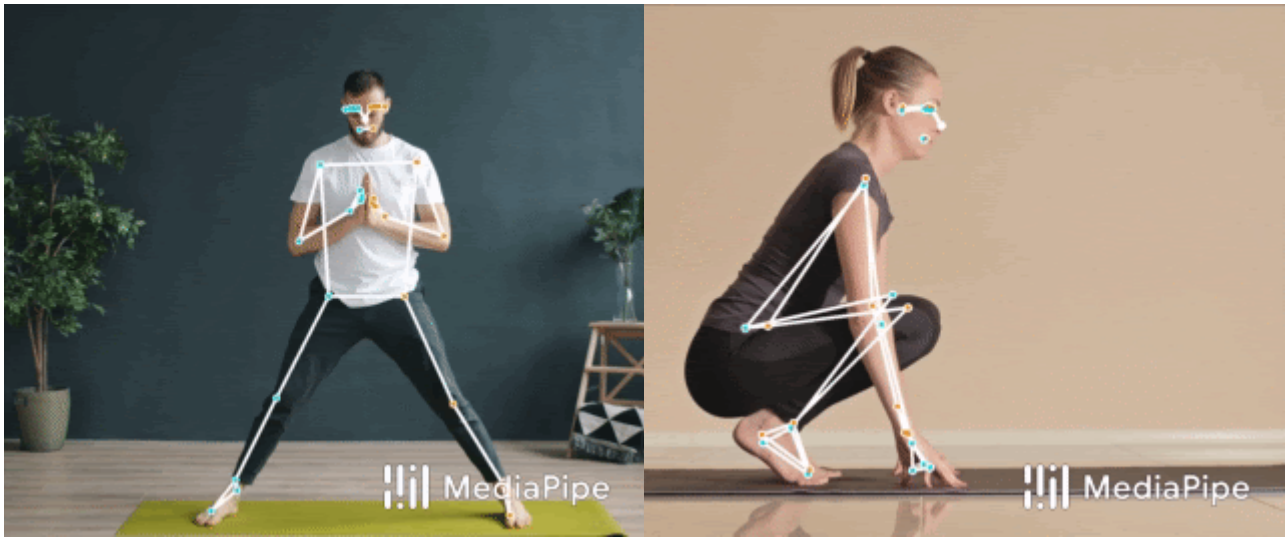


The number of performed exercises counter based on detected body pose. Left: Squats; Right: Push-Ups

Conclusion

We have released a version of BlazePose targeting upper body use cases in MediaPipe running on Android, iOS and Python. BlazePose will also be made available to the broader mobile developer community via the Pose detection API in the upcoming release of ML Kit. Apart from the mobile domain, we preview our web-based in-browser version as well. We hope that providing this human pose perception functionality to the broader research and development community will result in an emergence of creative use cases, stimulating new applications, and new research avenues.

We plan to extend this technology with more robust and stable tracking to an even larger variety of human poses and activities. In the accompanying Model Card, we detail the intended uses, limitations and model fairness to ensure that use of these models aligns with Google's AI Principles. We believe that publishing this technology can provide an impulse to new creative ideas and applications by the members of the research and developer community at large. We are excited to see what you can build with it!



BlazePose results on yoga use-cases

Acknowledgments

Special thanks to all our team members who worked on the tech with us: Fan Zhang, Artsiom Ablavatski, Yury Kartynnik, Tyler Zhu, Karthik Raveendran, Andrei Vakunov, Andrei Tkachenka, Marat Dukhan, Raman Sarokin, Tyler Mullen, Gregory Karpiak, Suril Shah, Buck Bourdon, Jiuqiang Tang, Ming Guang Yong, Chuo-Ling Chang, Juhyun Lee, Michael Hays, Camillo Lugaresi, Esha Uboweja, Siarhei Kazakou, Andrei Kulik, Matsvei Zhdanovich, and Matthias Grundmann.