

Pose Estimation – white paper

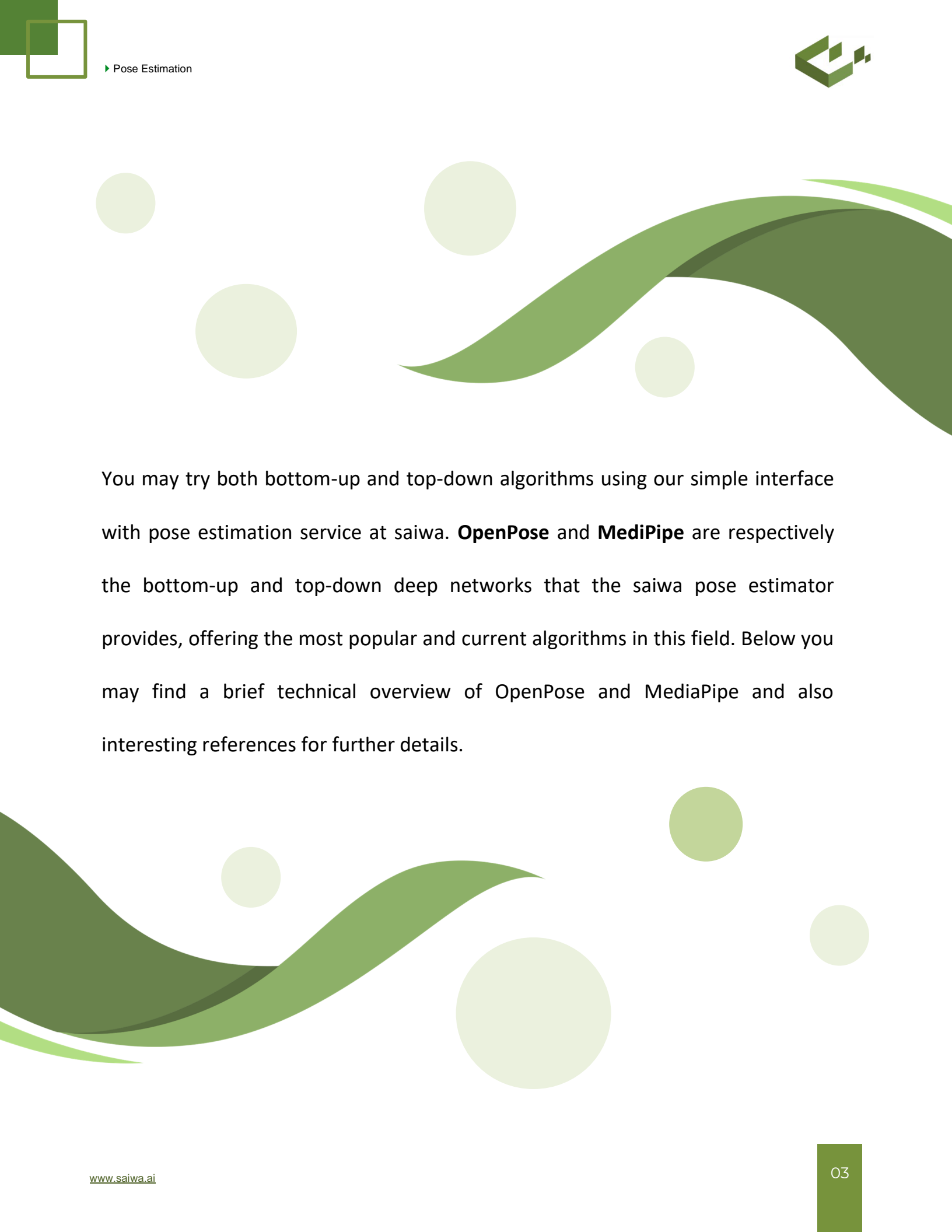
Feb 2022

The logo for Saiwa, featuring the word "saiwa" in a lowercase, white, sans-serif font. A thick green horizontal bar is positioned to the left of the text, extending from the left edge of the page towards the start of the word.

saiwa

simple artificial intelligence web application

Human pose estimation means detecting spatial location of human skeleton joints and classifying them into connections like arm, leg, head, torso and more. The joints and connections between them are known as keypoints and pairs, respectively. Pose estimation is a practical computer vision solution for various problems across a multitude of domains, such as motion analysis, action recognition and sign language recognition. Deep pose estimators have demonstrated remarkable success in estimating keypoints and pairs. Generally, we have two different approaches for pose estimation: **bottom-up** and **top-down**. Bottom-up approaches detect all keypoints of the input image in the first step, and then group them up to form distinct poses. A top-down pose estimator, on the other hand, a human detection method (like detectron2 or yolov5 in saiwa object detection service) is applied first; then, for each detected person (within the detected bounding box), the pose estimation occurs.



You may try both bottom-up and top-down algorithms using our simple interface with pose estimation service at saiwa. **OpenPose** and **MediPipe** are respectively the bottom-up and top-down deep networks that the saiwa pose estimator provides, offering the most popular and current algorithms in this field. Below you may find a brief technical overview of OpenPose and MediaPipe and also interesting references for further details.



OpenPose is a bottom-up multi-person 2D human pose estimator¹ [1]. OpenPose works in real-time invariant to the number of persons in an image and their scales and positions. It detects 135 keypoints in total for each person (as represented in Figure 1) including:

- 15, 18 or 25-keypoints for body and foot
- 2x21-keypoints for the two hands
- 70-keypoints for face

¹ In a single person mode OpenPose is able to reconstruct a 3D pose as well.



OpenPose network follows a four-stage pipeline, shown in Figure 2. It initially extracts input image features using convolution and pooling layers. The extracted features are sent to two parallel branches of CNN layers to jointly predict a set of 18 confidence maps for body part detection, and another set of 38-part affinity fields for parts association. The last two stages are used to clean up the estimations made by the two CNN branches. The parsing step performs a set of bipartite matches to associate body parts candidates. Using the confidence maps, bipartite graphs are made between pairs of parts. Through part affinity field values, the bipartite graphs are pruned. Finally, a greedy parsing stage is run to group keypoints of individual skeletons. For more technical details of the OpenPose pipeline and its corresponding stages, please refer to [1].



saiwa employs an open-source and robust implementation of OpenPose from CMU Perceptual Computing Lab which is a public release, maintained and used across many applications and fields of research [2]. Figure 3 shows a few instances of human pose extracted using OpenPose method in saiwa post estimation service.

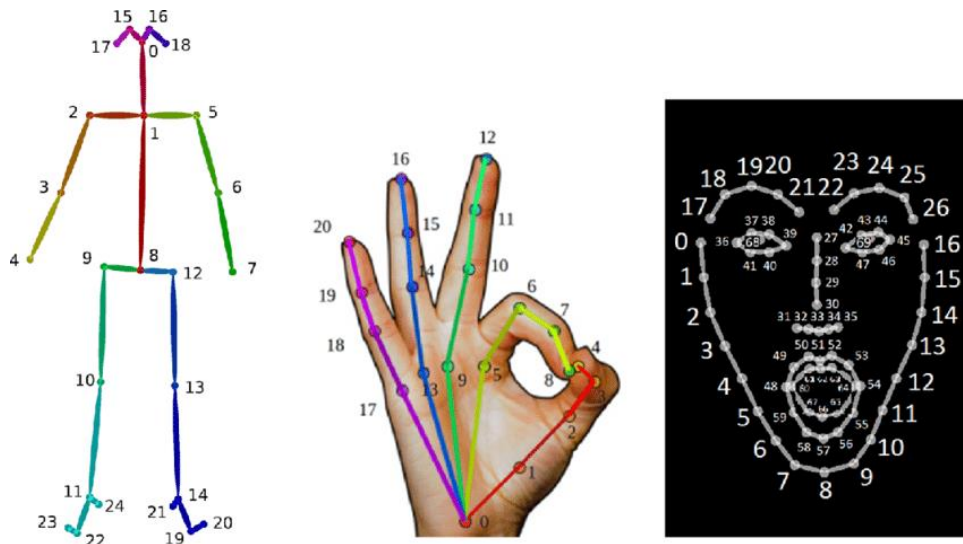


Figure 1. OpenPose 135 keypoints in three main categories; i.e. body and feet, hand and face

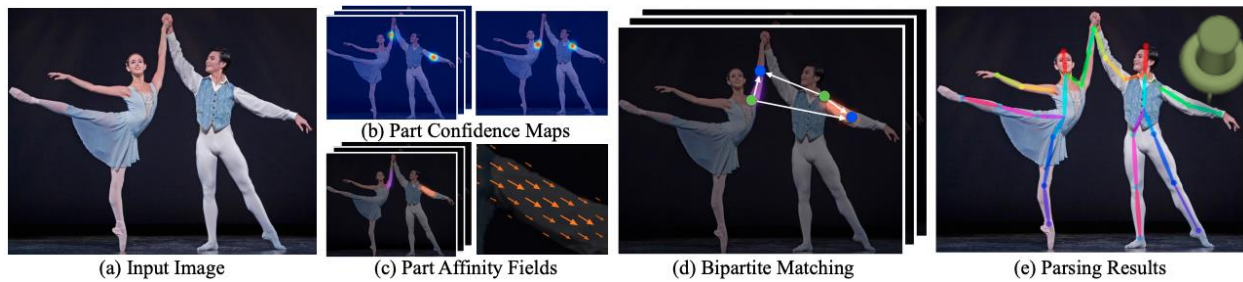


Figure 2. OpenPose pipeline overview (printed from [1])



Figure 3. OpenPose results calculated using saiwa pose estimation online interface.



MediaPipe is Google's open-source cross-platform solution for media processing like face detection, object detection, tracking, etc [3]. One of the interesting applications of MediaPipe is pose estimation using **BlazePose** deep network [4]. BlazePose following a top-down pose estimation approach, detects 33 keypoints for one person as it is shown in

Figure 4.

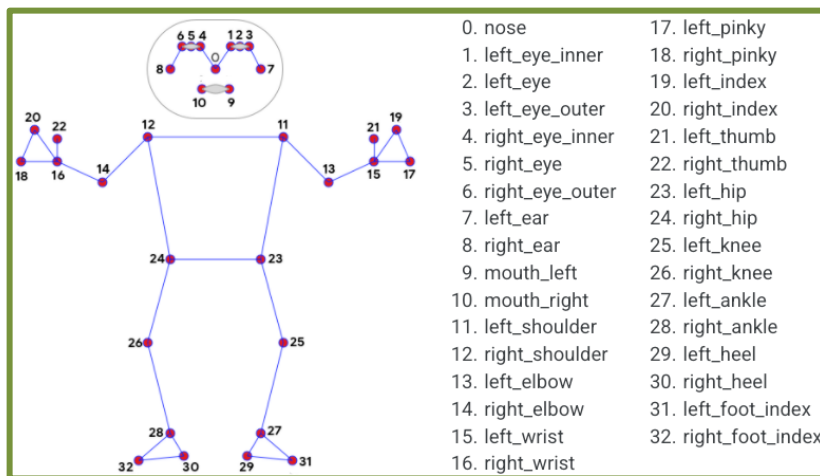


Figure 4 . keypoints as it is extracted using BlasePose (printed from [3])



BlazePose similar to OpenPose has a pipeline architecture. BlazePose pipeline has two stages: detector and tracker. Detector locates the pose region-of-interest (ROI) using an extension of BlazeFace (another MediaPipe solution). Here, BlazePose makes an assumption that the head should be visible for the corresponding person. The tracker subsequently predicts all 33 pose keypoints from this ROI with three degrees of freedom: (x, y location and visibility). Figure 5 shows BlazePose tracker network architecture. By employing a regression approach, the tracker is supervised by a heat map/offset prediction of all keypoints. For more details on BlazePose pipeline please refer to [4].

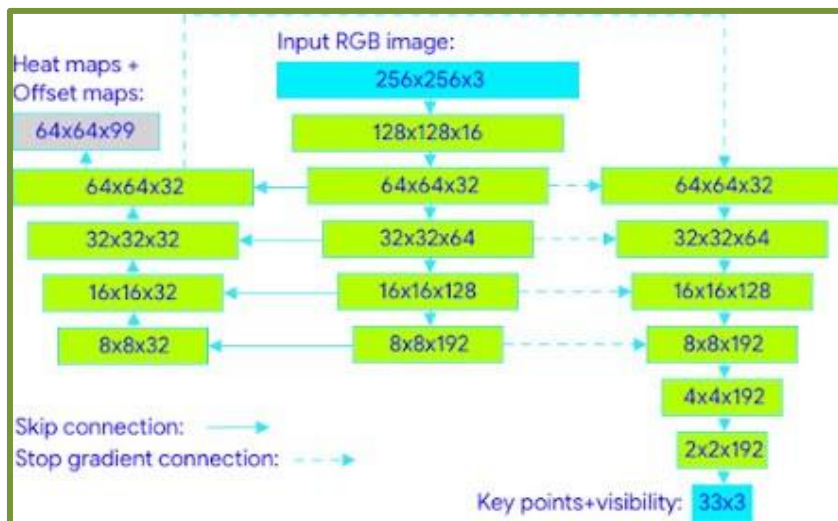


Figure 5. MediaPipe BlazePose tracker architecture (printed from [4])



With the saiwa pose estimation service we use a standard implementation of MediaPipe as it has been released by Google [3]. In Figure 6 a few output results from saiwa using MediPipe are represented.



Figure 6. MediaPipe BlazePose results calculated using saiwa pose estimation online interface.

POSE ESTIMATION



References:

[1] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[2] <https://github.com/CMU-Perceptual-Computing-Lab/openpose>.

[3] <https://google.github.io/mediapipe/>.

[4] Bazarevsky, Valentin, et al. "Blazepose: On-device real-time body pose tracking." arXiv preprint arXiv:2006.10204. 2020.



info@saiwa.ai

720 Guelph Line Burlington, ON L7R 4E2 Building

+15148131809

www.linkedin.com/company/saiwa

www.instagram.com/saiwa.ai