Optimizing Hand Region Detection in MediaPipe Holistic Full-Body Pose Estimation to Improve Accuracy and Avoid Downstream Errors

Amit Moryossef
sign.mt
University of Zurich
amit@sign.mt

Abstract

This paper addresses a critical flaw in MediaPipe Holistic's hand Region of Interest (ROI) prediction, which struggles with non-ideal hand orientations, affecting sign language recognition accuracy. We propose a data-driven approach to enhance ROI estimation, leveraging an enriched feature set including additional hand keypoints and the z-dimension. Our results demonstrate better estimates, with higher Intersection-over-Union compared to the current method.

1 Introduction

In recent years, pose estimation (Pishchulin et al., 2012; Cao et al., 2019; Güler et al., 2018) has emerged as a fundamental component for various applications ranging from action recognition and interactive gaming to the more nuanced field of sign language processing. Among the tools at the forefront of this technological revolution, MediaPipe Holistic (Grishchenko and Bazarevsky, 2020) distinguishes itself by offering real-time processing capabilities across a diverse array of devices, coupled with its adaptability in different runtime environments. This flexibility has facilitated its adoption across both research and practical applications.

However, the MediaPipe Holistic approach exhibits a significant flaw. The heuristic they use for determining the region of interest (ROI) for hands was designed for scenarios where the hand's plane is parallel to the camera——a design choice that does not hold in numerous practical situations. This limitation can lead to inaccuracies in hand ROI prediction, which subsequently affect the detection of hand keypoints, and compromises the overall accuracy of the full-body pose estimation (Moryossef et al., 2021).

The naiveté of the existing hand ROI prediction method in MediaPipe Holistic typically manifests when dealing with non-ideal hand orientations. Given the applications of MediaPipe Holistic in domains such as sign language recognition, enhancing the robustness of hand ROI prediction is needed for accurate downstream solutions.

Concretely, the current approach (Algorithm 1) extract the body pose using BlazePose (Bazarevsky et al., 2020), which includes four hand keypoints per hand - the *Wrist*, the *Index MCP*, the *Pinky MCP*, and the *Thumb*. Then, a rough ROI crop estimate is calculated from the first three points by estimating the position of the center of the hand from the *Index MCP* and *Pinky MCP*, and calculating the hand size as double the distance from the center to the *Wrist*. The center is then shifted, and a bounding box around it is estimated by scaling the hand size using hard-coded values. The 2D hand orientation is then estimated from the angle between the wrist and the center, and a rough crop is produced. This rough crop is fed to a re-croping model which refines the bounding box, to create the hand crop. The hand crop is fed to the hand landmark model, predicting the keypoints for each hand independently.

Algorithm 1 Hand Landmarks from Pose to ROI

```python
def calc_hand_roi(wrist, index, pinky):
    # Estimate middle finger position
    center = (2 * index + pinky) / 3
    # Estimate hand size
    size = 2 * distance(center, wrist)
    # Estimate hand 2D rotation
    rotation = angle(wrist, center) + 90
    # Shift center Y position
    center = shift(center, rotation,
                   by=(0, -0.1 * size))
    # Scale prediction
    size = 2.7 * size
    return center, size, rotation
```

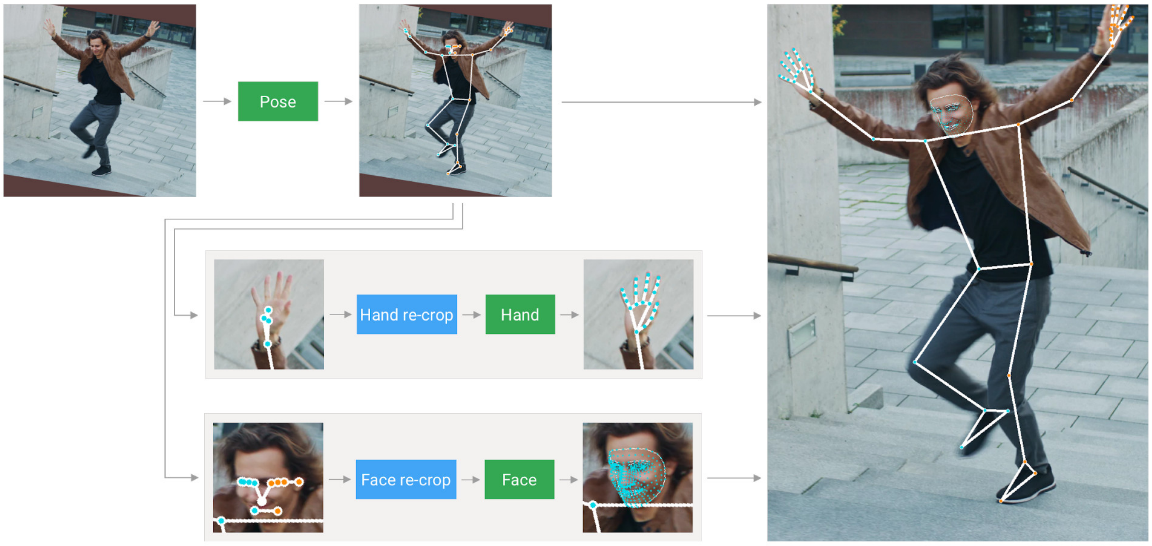


Figure 1: MediaPipe Holistic Pipeline Overview (Grishchenko and Bazarevsky, 2020).

2 Dataset

We utilize the Panoptic Hand DB dataset (Simon et al., 2017)[1], which contains manually annotated 2D hand poses, with 1912 annotations in the training set and 846 in the testing set. Annotations encompass both *right* and *left* hand data, providing a comprehensive basis for analysis.

To estimate the quality of MediaPipe's ROI estimation, we first adopt the MediaPipe framework to define an ROI from the hand keypoints. This process involves bounding all keypoints and adjusting the bounding box by scaling and rotating it based on the angle between the wrist and middle finger. Then, MediaPipe Holistic is employed to predict the body keypoints. We extract the keypoints for the specific hand being analyzed (shoulder, elbow, wrist, thumb, index finger, and pinky), and for consistency, left-hand keypoints are mirrored to simulate right-handed orientation. The ROI is then determined using Algorithm 1.

Table 1 illustrates selected instances from the dataset with varying levels of success in ROI coverage. The 'Worst' and 'Best' examples were automatically identified based on their coverage percentages, at 0.8% and 93.7% respectively, whereas other samples were selected manually.

[1] https://domedb.perception.cs.cmu.edu/

3 Methodology

We explore an enhancement to the hand ROI calculation mechanism within the MediaPipe Holistic framework. We recognize that the current solution only uses three hand points (*wrist*, *index*, and *pinky*) and ignores the *shoulder*, *elbow*, and *thumb* which can also be indicative of the hand region. Furthermore, the current solution only uses the

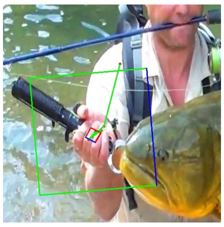
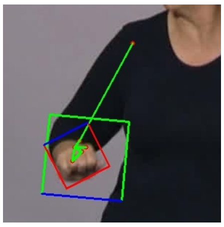
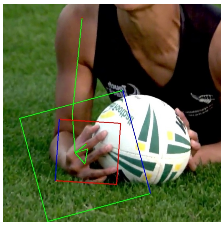
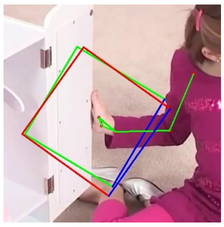
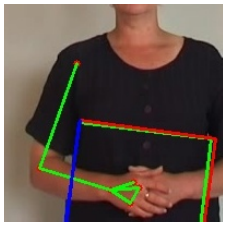
Worst	Bad	OKish	Good	Best
				

Table 1: Selected examples of hand keypoint detections with different ROI coverages. Each image shows hand keypoints in green lines, and two bounding boxes: gold for the ROI and red for the predicted ROI. A blue line indicates the orientation of the bounding box on the bottom edge.

(x, y) coordinates, and ignores the predicted z. As we note that the model seems to perform badly when the hand is perpendicular to the camera, the z dimension can be particularly useful.

We also understand that a good solution to this problem must be very performant, and ideally as interpretable as the current solution. Since we are not members of the MediaPipe project, a complex pull-request with an additional model with obscure weights may not be accepted.

Therefore ideally, we would like to formulate a KAN (Liu et al., 2024), to use all six normalized body right-hand keypoints, alongside the image aspect-ratio (ρ), and predict the ROI parameters - *center*, *size*, and *angle*. After pruning, a KAN can be represented as simple mathematics, allowing us to deliver a solution in code, without loading additional models. We would start the solution by manifesting a KAN from the current mathematical solution. For example, the current *size* estimation as described in Algorithm 1 is mathematically equivalent to the following equation for the *wrist*, *index* and *pinky*:

$$5.4\sqrt{(x_1 - \frac{2x_2 + x_3}{3})^2 + (y_1 - \frac{2y_2 + y_3}{3})^2}$$

However, due to instability in the *pykan* framework, we instead use a standard MLP with a hidden size of 10, to predict each of the ROI parameters separately. This delivers a fast solution of three MLPs with 332 parameters each but lacks the interpretability of a KAN.

4 Evaluation

To evaluate the effectiveness of our proposed methodology, we measure the improvement in hand ROI predictions. We use the following metrics:

- **Center Err** - in percentage based on the image dimensions, how distance is the predicted center from the gold center.
- **Scale Err** - the absolute difference between the two scales, divided by the gold scale.
- **Rotation Err** - the absolute difference between the predicted and gold angles, circularly around 360 degrees.
- **IoU** - intersection-over-union of the predicted and actual hand ROIs.

We anticipate that the model will refine ROI predictions leading to higher precision in hand keypoint detection across a variety of hand orientations and movements. This should be reflected by a higher IoU.

5 Results

We train three separate MLPs on the train set to predict the *center*, *size*, and *angle*. Table 2 shows the results of both the original and our new method on the test set. It shows that the MLP manages to better predict the *center* and *size* of the ROI, but fails to better predict the rotation. We believe that this is due to the simplicity of our network, containing only linear layers with *relu* activations.

Method	IoU ↑	Center ↓	Scale ↓	Rotation ↓
Original	57%	2.51%	30.37%	**32.08**
MLP	**63%**	**2.15%**	**17.91%**	56.96

Table 2: Comparison of our method to the original.

Importantly, on the test set, we find that while the minimum IoU using the original method is 3%, our new method archives a minimum of 16%, indicating that it might work better for edge cases.

Therefore, unless further optimized, we believe the final solution should use the MLP to predict the *center* and *scale*, and use the heuristic to predict the rotation.

6 Limitations

The issues we encountered with *pykan* prevented us from delivering an interpretable solution, which could prove to be better, and more easily accepted by the library maintainers.

While being an improvement on the current methodology, our solution should not be the final one. Users of MediaPipe with more time on their hands could explore additional solutions and validate them on larger amounts of data. Our code is available to ease these future optimizations `https://github.com/sign-language-processing/mediapipe-hand-crop-fix`.

References

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. 2024. Kan: Kolmogorov-arnold networks.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3434–3440.

Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thorm ä hlen, and Bernt Schiele. 2012. Articulated people detection and pose estimation: Reshaping the future. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153.