Impact of Pose Estimation Models for Landmark-based Sign Language Recognition

Cristian Lazo-Quispe
Universidad Nacional de Ingeniería
Lima, Perú
clazoq@uni.pe

Joe Huamani-Malca
Pontificia Univeresidad Católica del Perú
Lima, Perú
huamani.jn@pucp.edu.pe

Manuel Stev Harold Huamán-Ramos
Pontificia Univeresidad Católica del Perú
Lima, Perú
mhuamanr@pucp.edu.pe

Gissella Bejarano
Baylor University
Texas, USA
gissella_bejaranonic@baylor.edu

Pablo Rivas
Baylor University
Texas, USA
pablo_rivas@baylor.edu

Tomas Cerny
Baylor University
Texas, USA
tomas_cerny@baylor.edu

Abstract

Sign Language Recognition (SLR) models rely heavily on advances reached by Human Action Recognition (HAR). One of the simplest and most dimensional-reduced modalities is the skeleton joints and limbs represented with key-point landmarks and edges connecting these landmarks. These skeletons can be obtained by pose estimation, depth maps, or motion capture. For HAR, models are usually interested in less granularity of pose estimation compared to SLR, where the landmark estimation of not only the pose and body but the facial gestures, hands, and fingers is crucial. In this work, we compare three whole-body estimation libraries/models that are gaining attraction in the SLR task. We first find their relation by identifying common keypoints in their landmark structure and analyzing their quality. Then, we complement this analysis by comparing their annotations in three sign language datasets with videos of different quality, backgrounds, and regions (Peru and USA). Finally, we test a sign language recognition model to compare the quality of the annotations provided by these libraries/models.

1 Introduction

Most recent work in Sign Language Processing (SLP), such as sign language recognition, sign language translation, and production, relies heavily on advances reached by Human Action Recognition (HAR). More specifically, the skeleton modality produced by pose estimation models perfectly suits SLP tasks that should be invariant to the clothing or background of the subjects. For example, the skeleton modality is a simple and dimensional-reduced modality that has helped to recognize pedestrian activities in urban mobility scenarios [11], player movement individually and in a group for sports [10], or monitor elderly patients [17, 23]. In these tasks, landmarks at a low granularity have been enough to achieve significant performance. However, for tasks related to sign language processing, greater granularity is essential because of the importance of hands and gestures in sign meaning.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Most SLP tasks using the skeleton modality rely on two pose-estimation models available through open-source libraries, either *MediaPipe* or *OpenPose*. The latter is among the most mentioned/adopted in the linguistic and sign language research areas, i.e., at least 20% of papers accepted in the Sign Language Workshop at LREC 2022 mention *OpenPose*. However, some other models might be more suitable and accurate for hands and face landmarks estimation or a combination of them. In this work, we compare three pose-estimation models, MediaPipe, OpenPose, and HRNet-based, for sign language datasets. First, we identify the common and correspondent landmarks produced by the different pose-estimation libraries and compare their quality through the percentage of in-range (inside frame boundaries), out-of-range and missing values. We compare these landmarks for three video-based sign language datasets with different quality (resolutions), background (one-color background and in a classroom), and region (Peru and USA). Then, we complement this analysis by testing the performance of two sign language recognition (SLR) models using the landmarks produced by the pose-estimation libraries. Finally, we compare two groups of landmarks from body, face, and hands to analyze their impact and show that a prior feature selection might help when training sign language recognition models.

2 Related Work

In this section, we summarize the state-of-the-art for 2D pose estimation that has been usually trained on body-joint datasets. Few works are starting working on whole-body joint estimations, and even fewer have been trained on sign language datasets. Most of the advances in this task consist of models with several stages and managing high and low-resolution of images.

There are two other categories to analyze the state-of-the-art, whether they perform a top-down or bottom-up approach or a regression [8, 30] or heatmap approach, which is the most used lately. Top-down methods estimate a bounding box through a person detector and, later, the body-joint locations [8, 13]. In contrast, bottom-up approaches locate all the keypoint landmarks and, later, group them into different individuals. As mentioned in [18], bottom-up might be more suitable for urban mobility applications where several individual poses should be recognized in one image. Similarly, most recent work uses a heatmap rather than the regression approach.

One of the first deep learning-based approaches, but pose only, was DeepPose [30], followed by ConvNet [29] and Convolutional Pose Machine (CPM) [31] and the Stacked Hourglass approach [22]. Later, other works attend all keypoints but are trained with different datasets for each section of the body, like the hands and face. Some of the previous work in pose-only estimations were used in works such as [27, 6], which cascades several hourglass networks and use CPM, respectively. Another important previous work is [28] that presents HRNet, which also follows approaches based on [22, 32, 9] but with the difference of a parallel component maintaining the high-resolution input instead of restoring it from the reduced resolution process. However, few works are trained to recognize the whole body keypoints at the time. For example, [15] introduces a whole-body dataset by manually extending the annotation of COCO. Additionally, this work trains ZoomNet in a bottom-up manner that zooms in on all more specific areas to determine finer keypoint locations. Another work that tests the whole-body dataset is [33], which defines a weighted method to consider the importance of different parts of keypoints based on OPenPifPaf [18]. OpenPifPaf is a ResNet-based two-head network that predicts precise location (Part Intensity Field) on one side and association (Part Association Field) on the other. Finally, some other recent methods intend pose estimation in a whole stage and consider the video's temporal dimension [1].

Before 2016, the most used metrics were: Probability of Correct Pose or Percentage of Correct Parts (PCP), which measures the distance between predicted locations of a limb compared to the ground truth; Percentage of Correct Key-points (PCK), which considers a correct estimation if this is within a certain threshold distance of the true joint; or Percentage of Detected Joints (PDJ) which is similar to PCK with the difference that the threshold is a percentage of the torso. Most recent works focus more on Average Precision (AP), mean Average Precision (mAP), and mean Average Recall (mAR), and still, some report PCK.

3 Methodology

We aim to provide more insights on the suitability of three pose-estimation libraries or models to annotate landmarks used in a SLP task such as sign language recognition. First, we introduce technical details about these three libraries. Then, we pseudo-label three sign language datasets with these libraries and compare the quality of their annotations. Finally, we complement this analysis by using the annotated datasets to perform sign language recognition tasks with two SLR models in the section 4. Our analysis scripts, adaptation of pose-estimation libraries and sign language recognition models are publicly available[1].

3.1 Pose-estimation Libraries

In this section we describe the models used in the MediaPipe and OpenPose library, as well as the wholePose model.

MediaPipe (Holistic) integrates in one architecture different machine learning models for pose [2], face detection [3], face recognition [16], hands [20] and uses an appropriate image resolution for each region. It works with a total of 543 landmarks (33 pose landmarks, 468 face landmarks, and 21 hand landmarks per hand) [21]. For pose estimation, Mediapipe[2] adopts a combination of a heatmap and a regression approach. Besides, [2] reports training with two datasets of individuals performing fitness activities.

OpenPose is an open-source real-time system[3] that is built in a bottom-up approach [6]. To estimate the pose, the model relies on subcomponents that first transform the input image into a feature map for body part detection through a CNN . Then, the model produces confidence maps and part affinity fields for association (PAFs) that represent the degree of associations. This network is based on the architecture defined in [31] and an extension and variation work presented in [7]. To estimate the hands, the model uses the work in [26] centered in hand pose detection, and in the same fashion the facial keypoint estimation. Some disadvantages of this library are mentioned in [12, 25].

WholePose is what we call the model used in [14] which is based in an HRNet [28]. They report the use of HRNet whole-body pose estimator provided by MMPose[4] and tested in 27 selected keypoints. Original HRNet is trained in COCO and MPII Pose Human datasets estimating many less keypoints. This approach ranked first at the 2021 Looking at People Large Scale Signer Independent Isolated SLR CVPR Challenge - RGB Track.

3.2 Analysis

The first dataset we analyze is from two interpreters of the Educational Peruvian TV Show *Aprendo en Casa* (AEC) and presented in [4] in a controlled white background. The second dataset is a corpus collected in [24] with deaf Peruvian signers and with a noisy background, usually a classroom (PUCP-DGI156, but called in this paper as PUCP). The third dataset is a sample of the most used American Sign Language dataset [19] (WLASL).

We analyze the quality of 71 landmarks of four sections: pose, face, left hand and right hand. We also analyze the estimation in the following three categories **in-range:** when landmarks are within the dimension of the frame; **missing points**: when the keypoint estimation model cannot define the landmarks section of a certain part of the body; **out of range**: which are points that are not exactly seen but estimated out of the frame. As we can see in Figure 1, MediaPipe identifies more values of missing and out-of-range in the three datasets compared to openPose and wholePose. WholePose does not provide explicitly labeled missing keypoints estimation; MediaPipe recognizes more out-of-range keypoints in the three datasets; OpenPose recognizes more in left and right hands for the three datasets. In Figure 2, we show the quality of one example of frame from each dataset.

We also analyze the general quality of videos by calculating percentage of frames with all landmarks in-range. MediaPipe and OpenPose have less percentage of frames identifying the right wrist and specially the left wrist compared to WholePose. This holds for the three datasets and in Figure 3,

[1] https://github.com/gissemari/impactOfPoseLibrariesInSLR

[2] https://google.github.io/mediapipe/solutions/holistic

[3] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[4] https://github.com/open-mmlab/mmpose

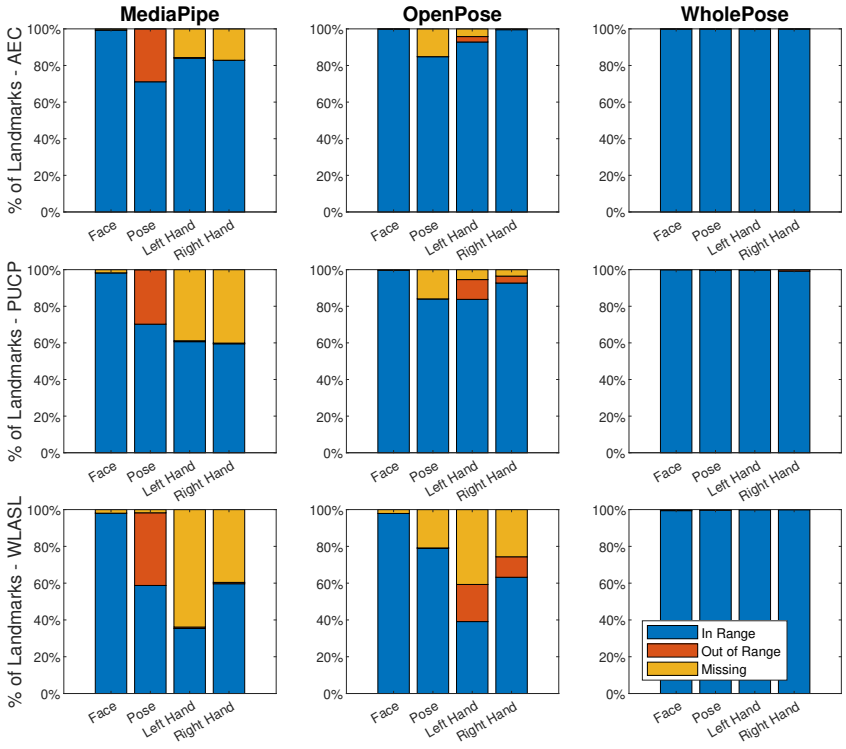

Figure 1: Percentage of Model Failure Cases in the three datasets

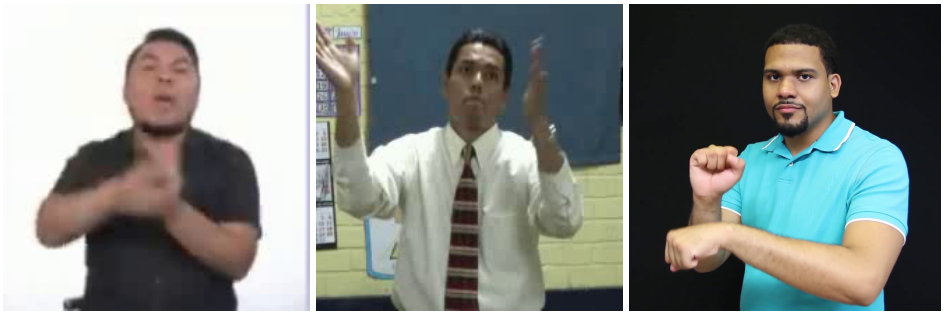

Figure 2: Example of visual quality for videos in left: AEC; center: PUCP; and right: WLASL

we show the particular case for WLASL where those percentages drop to around 70% and 50% for right and left hand respectively. On the top right plot, we analyze all the landmarks for left hand and they oscillate between a 38% and 55% for WLASL too. In the case of PUCP (bottom plots), only MediaPipe drops to around 60% for right and left hand keypoints, and MediaPipe and WholePose maintain a percentage of more than 90% and 80% for right and left hand respectively.

4 Experiments

The quality of the pose-estimation libraries is assessed by using their landmarks prediction on three datasets to test two sign language recognition models. In particular, the three sign language video-based datasets (AEC, PUCP and WLASL) are fed to each of the three pose estimation models. From the total amount of landmarks or pose estimated, we select one group of 29 and another of 71, creating other two landmark datasets for each sign language dataset, i.e. the AEC set of 29 landmarks are

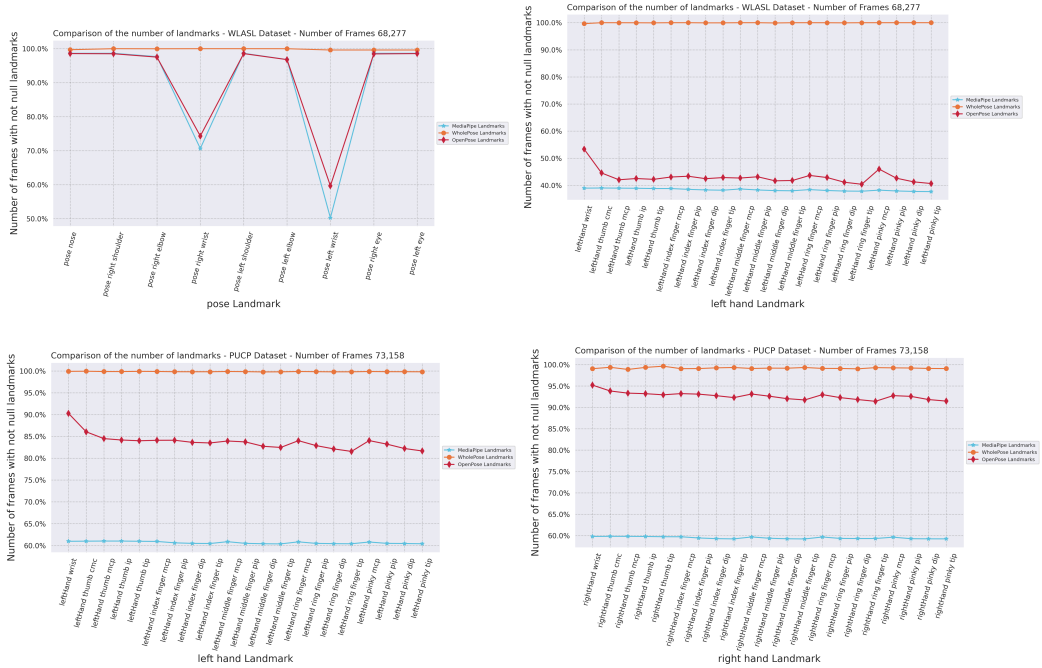

Figure 3: Percentage of in-range landmarks in body section Left Hand, Right Hand and Face

three (MediaPipe-29, OpenPose-29, and WholePose-29). We try to also test if more landmarks imply better performance in the SLR models. Therefore, we report our results considering 29 and 71 keypoint landmarks in Top-1 and Top-5 (if the ground truth corresponds to one of the most probable 5 predicted classes).

A gloss is a token to describe the class or label of an specific sign. Our sign language recognition models were trained for a sample of glosses that have more than 15 instances, excluding classes that perform very bad in initial experiments, and the ones that imply pointing which are very similar ("I","YOU","THERE", etc.). These conditions resulted in 28 classes of AEC, 36 of PUCP, and 101 of WLASL.

4.1 SLR Models

The first sign language recognition model consists of the landmark-based and graph-based component of [14] adapted to work with 71 keypoint landmarks, as they originally consider only 27 keypoints. The second one is a transformer-based model [5] tested in WLASL dataset [19]. It processes the keypoint estimation through the Vision API[5]. For simplicity, we identify these two models as SmileLab and Spoter respectively. Both models train between 3 and 8 millions of parameters depending on the size of landmarks.

4.2 Results and Discussion

We divide each dataset in 80% for the training and 20% for the testing. For hyperparameter tuning, we selected one landmark dataset (a certain group of landmarks estimated with one of the pose-estiation models for a SL dataset) and tune the learning rate (LR) for an specific SLR model. Then, we use the same LR for the landmark datasets produced by the other two pose-estimation library on that same SL dataset and report the results.

To guarantee robustness of the comparisons, we run each experiment five times and report the average accuracy along with the standard deviation, as show in Table 1 for results in Top-1 and 2 for results in Top-5.

[5]https://developer.apple.com/documentation/vision

As it can be seen in Table 1 and 2, none of the libraries works the best in all the datasets and settings. WholePose is more stable because it ranks first or second for the two SLR models for AEC and PUCP. MediaPipe works better in all the settings for WLASL. This seems contradictory to the Analysis Subsection where the percentage of frames with all landmarks in-range for MediaPipe in this dataset is low. Moreover, from the three datasets, WLASL is the one with better resolution quality which make us think that MediaPipe work better when the video is from a good quality. OpenPose works better in PUCP dataset for both sets of landmarks and both SLR models [5, 14].

The greater the number of landmarks we use in training the model, the better the top 1 accuracy we obtain, but this only happens in Spoter model. The opposite result is found in SmileLab, only two out of nine models have higher accuracy using 71 landmarks than 29 landmarks for Top-1. For Top-5, the same rate two out of nine remains, as it can be seen in Table 1. Therefore, there is no enough evidence to conclude that it is better to use a higher number of landmarks for the SLR task.

SLR Model	Library	Top-1					
		AEC		PUCP		WLASL	
		29	71	29	71	29	71
Spoter	MediaPipe	**0.649** ± 0.017	**0.665** ± 0.022	0.366 ± 0.021	0.390 ± 0.025	**0.634** ± 0.011	**0.701** ± 0.018
	OpenPose	0.528 ± 0.010	0.544 ± 0.031	**0.467** ± 0.020	**0.505** ± 0.009	0.473 ± 0.007	0.576 ± 0.011
	WholePose	0.613 ± 0.028	0.627 ± 0.018	0.442 ± 0.032	0.453 ± 0.011	0.418 ± 0.028	0.502 ± 0.004
SmileLab	MediaPipe	0.573 ± 0.018	0.571 ± 0.025	0.277 ± 0.014	0.265 ± 0.017	**0.677** ± 0.023	**0.533** ± 0.025
	OpenPose	0.590 ± 0.019	0.611 ± 0.021	**0.421** ± 0.021	**0.405** ± 0.018	0.562 ± 0.026	0.445 ± 0.018
	WholePose	**0.646** ± 0.022	**0.675** ± 0.008	0.390 ± 0.021	0.380 ± 0.026	0.584 ± 0.021	0.518 ± 0.018

Table 1: Spoter [5]and SmileLab [14] Top-1 results for groups of keypoints:29 and 71

SLR Model	Library	Top-5					
		AEC		PUCP		WLASL	
		29	71	29	71	29	71
Spoter	MediaPipe	0.869 ± 0.0125	0.882 ± 0.0134	0.751 ± 0.032	0.735 ± 0.015	**0.893** ± 0.010	**0.919** ± 0.008
	OpenPose	0.879 ± 0.010	0.869 ± 0.0.024	**0.784** ± 0.024	**0.777** ± 0.013	0.797 ± 0.007	0.838 ± 0.012
	WholePose	**0.8854** ± 0.014	**0.905** ± 0.015	0.781 ± 0.016	0.758 ± 0.016	0.764 ± 0.030	0.801 ± 0.015
SmileLab	MediaPipe	0.813 ± 0.007	0.803 ± 0.021	0.608 ± 0.040	0.613 ± 0.028	**0.876** ± 0.012	**0.796** ± 0.022
	OpenPose	0.842 ± 0.011	0.853 ± 0.015	**0.700** ± 0.027	**0.697** ± 0.030	0.798 ± 0.019	0.734 ± 0.018
	WholePose	**0.866** ± 0.007	**0.854** ± 0.015	0.690 ± 0.044	0.689 ± 0.019	0.812 ± 0024026	0.767 ± 0.015

Table 2: Spoter [5]and SmileLab [14] Top-5 results for groups of keypoints:29 and 71

5 Conclusions and Future Work

This work presents an analysis of three whole body pose-estimation libraries, i.e., MediaPipe, OpenPose, and WholePose and compare them using basic statistics and through the additional task of sign language recognition. We found that WholePose library shows less number of bad-quality landmarks and performs better most of the time in the two sign language recognition models we tested. These findings show that, contrary to the most-used pose estimation library being OpenPose, sign language researchers might want to start using more other models such as WholePose and MediaPipe.

Most of the state-of-the-art and analyzed libraries are being trained to look at the whole body. However, we believe pose-estimation models considering the special context of sign language, mostly focused on the torso, hands, and facial gestures, can help improve these models for this type of task.

For future work, we would like to fill the missing landmarks of one pose-estimation library with the in-range annotations from other pose-estimation library. In that way, we can get a combined landmark dataset and measure it is worth to have a more complete dataset and the accuracy of an SLR model improves with respecto to annotations done by individual pose-estimation libraries.

Finally, the training of these pose-estimation models relies on iterative supervision or metrics of most common people size. However, there are other scenarios, like people of smaller size which have to be considered when using these models to deploy products and end-user solutions. Testing in these context can help improve pose-estimation models.

Acknowledgments and Disclosure of Funding

P. Rivas contributed to this work while funded by the National Science Foundation under grants CNS-2136961 and CNS-2210091.

References

[1] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7035–7044, 2020.

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

[3] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

[4] Gissella Maria Bejarano, Joe Huamani-Malca, Francisco Cerna-Herrera, Fernando Alva-Manchego, and Pablo Rivas. PeruSIL: A framework to build a continuous Peruvian Sign Language interpretation dataset. In Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, Johanna Mesch, and Marc Schulder, editors, *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 1–8, Marseille, France, jun 2022. European Language Resources Association (ELRA).

[5] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 182–191, January 2022.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2021.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[8] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[10] Moritz Einfalt, Dan Zecha, and Rainer Lienhart. Activity-conditioned continuous human pose estimation for performance analysis of athletes using the example of swimming. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 446–455, 2018.

[11] Zhijie Fang and Antonio M. López. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4773–4783, 2020.

[12] Manolis Fragkiadakis, Victoria Nyst, and Peter van der Putten. Signing as input for a dictionary query: Matching signs based on joint positions of the dominant hand. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 69–74, Marseille, France, May 2020. European Language Resources Association (ELRA).

[13] Thomas Golda, Tobias Kalb, Arne Schumann, and Jürgen Beyerer. Human pose estimation for real-world crowded scenarios. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[14] Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3413–3423, June 2021.

[15] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, page 196–214, Berlin, Heidelberg, 2020. Springer-Verlag.

[16] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.

[17] Jyothsna Kondragunta, Ankit Jaiswal, and Gangolf Hirtz. Estimation of gait parameters from 3d pose for elderly care. In *Proceedings of the 2019 6th International Conference on Biomedical and Bioinformatics Engineering*, pages 66–72, 2019.

[18] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11969–11978, 2019.

[19] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020.

[20] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[21] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[22] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.

[23] Yoon Jung Park, Hyocheol Ro, Nam Kyu Lee, and Tack-Don Han. Deep-care: Projection-based home care augmented reality system with deep learning for elderly. *Applied Sciences*, 9(18), 2019.

[24] Miguel Rodriguez Mondoñedo. Lengua de Señas Peruana (LSP) PUCP-DGI156, 2022.

[25] Pascal Schneider, Raphael Memmesheimer, Ivanna Kramer, and Dietrich Paulus. Gesture recognition in rgb videos using human body keypoints and dynamic time warping. In Stephan Chalup, Tim Niemueller, Jackrit Suthakorn, and Mary-Anne Williams, editors, *RoboCup 2019: Robot World Cup XXIII*, pages 281–293, Cham, 2019. Springer International Publishing.

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

[27] Zhihui Su, Ming Ye, Guohui Zhang, Lei Dai, and Jianda Sheng. Cascade feature aggregation for human pose estimation. *arXiv preprint arXiv:1902.07837*, 2019.

[28] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019.

[29] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 648–656, 2015.

[30] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[31] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[32] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[33] Duncan Zauss, Sven Kreiss, and Alexandre Alahi. Keypoint communities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11057–11066, 2021.