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

Cristian Lazo Quispe, Joe Huamani Malca, Manuel Stev H. Huamán Ramos, Gissella Bejarano, Pablo Rivas, Tomas Cerny

Introduction

We **compare** three **whole-body estimation** libraries/models that are gaining traction in the Sign Language Recognition **(SLR)** task.

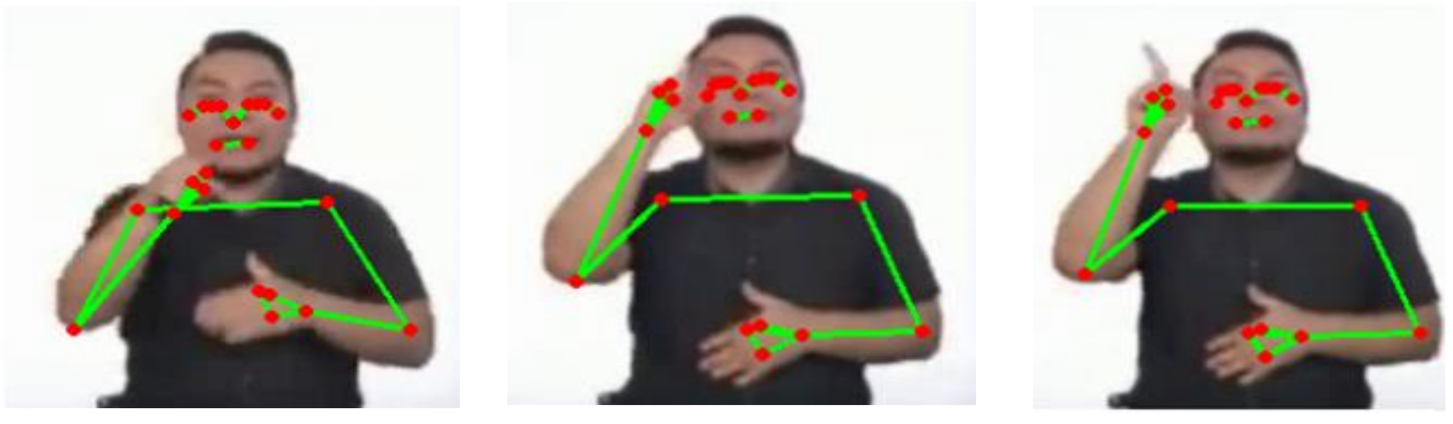


Fig. 1: Example of sequence of frames for the sign “IDEA” of AEC dataset

Data

We use 3 datasets “Aprendo en Casa” (**AEC**), Peruvian signers (**PUCP**), and American Sign Language dataset (**WLASL**).

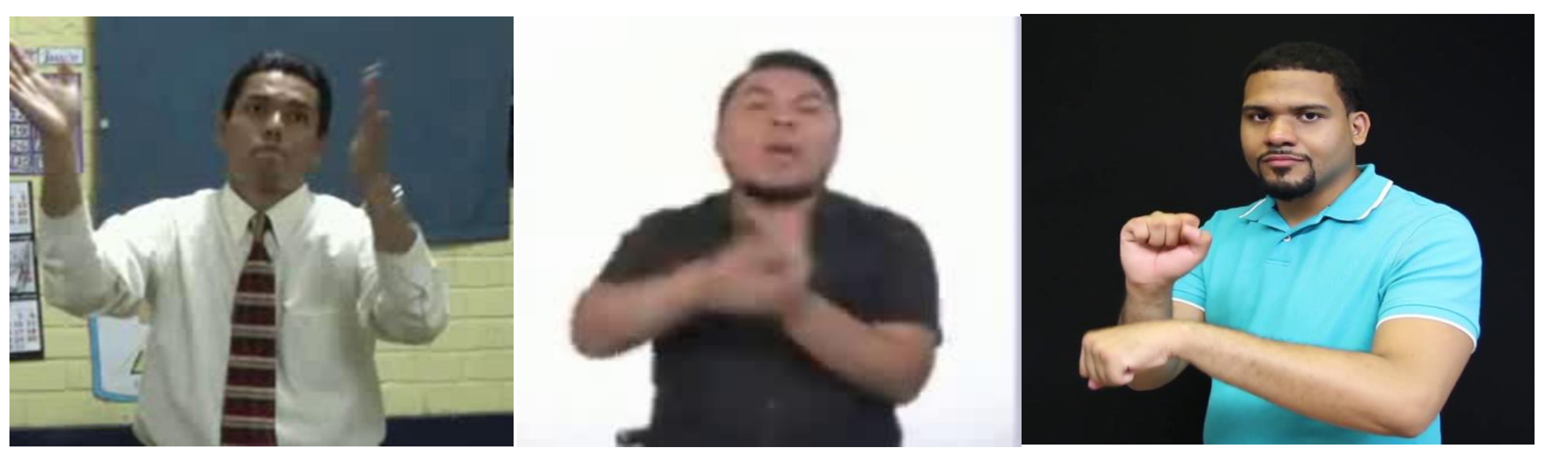


Fig. 2: Example of visual quality for videos AEC, PUCP, and WLASL

Methodology

1. Annotating three datasets with pose estimation libraries (**MediaPipe**, **OpenPose**, and **WholePose**)
2. Comparing the quality of their annotations in four sections: **pose, face, left hand** and **right hand**.
3. Analyzing sign language recognition perform with two SLR models (**SmileLab** and **Spoter** model).

Analysis

We analyze the quality of 71 landmarks of four sections: **pose**, **face**, **left hand** and **right hand**. We also categorized the estimations in **in-range**, **missing points**, and **out of range**, which are points that are not exactly seen but estimated out of the frame.

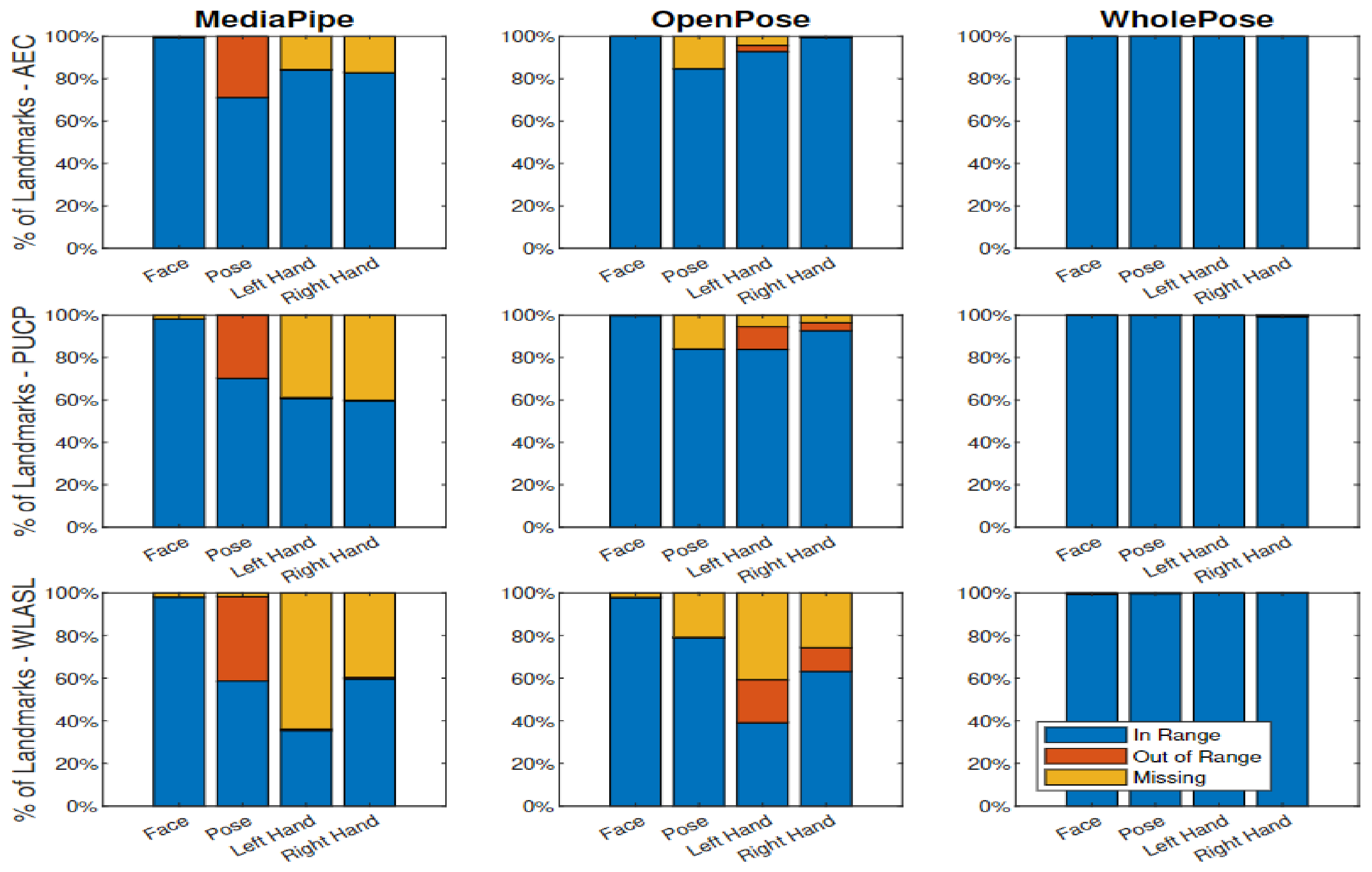
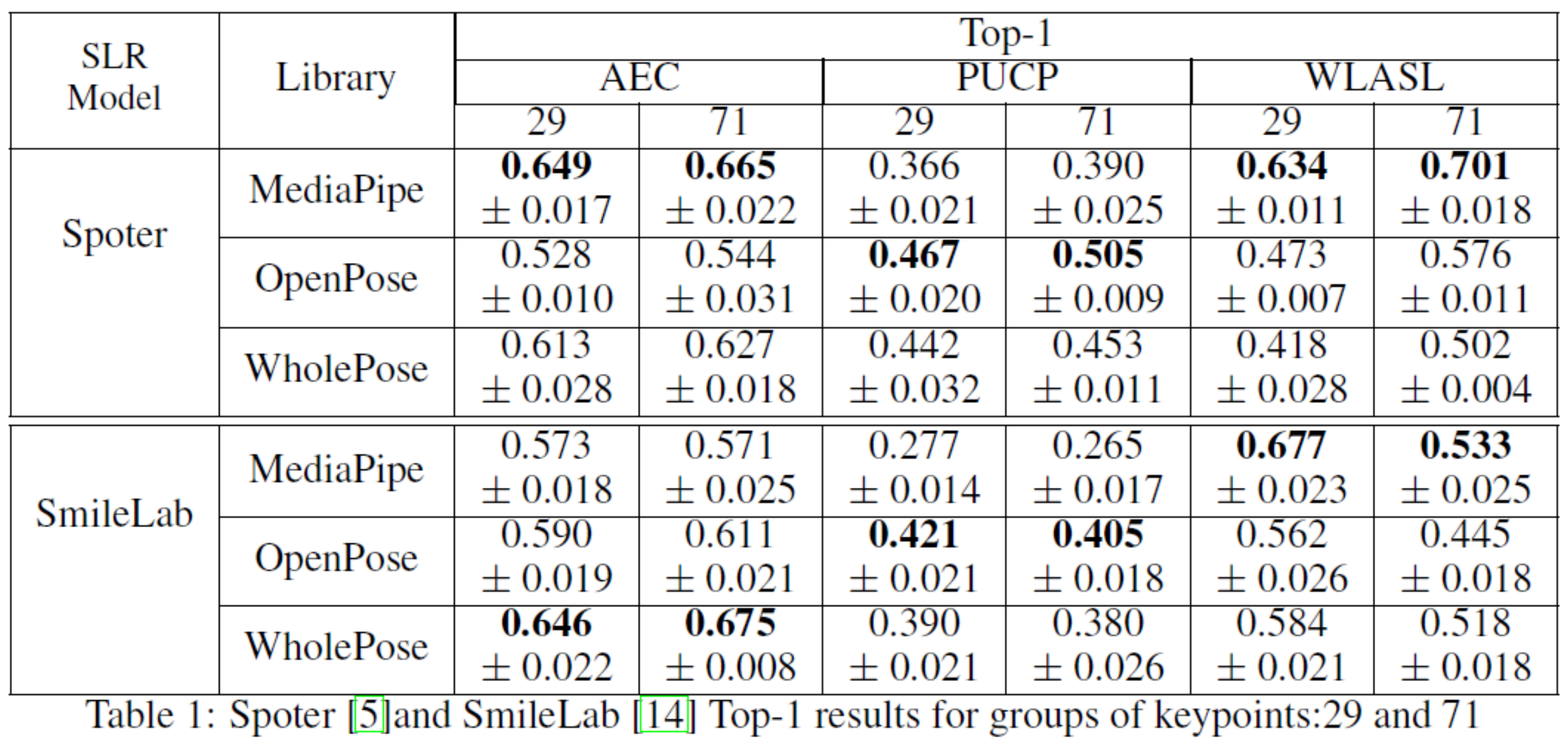


Fig. 3: Percentage of bad quality data in the three datasets

Experiment

We use 28 classes of AEC, 36 of PUCP, and 101 of WLASL. We report our results considering **29** and **71 keypoint** landmarks in Top-1 and Top-5 accuracies (if the ground truth corresponds to one of the most probable 5 predicted classes).

We use two SLR models consist of the landmark-based, **graph-based SmileLab (2021)** and **transformer-based Spoter model (2022).**



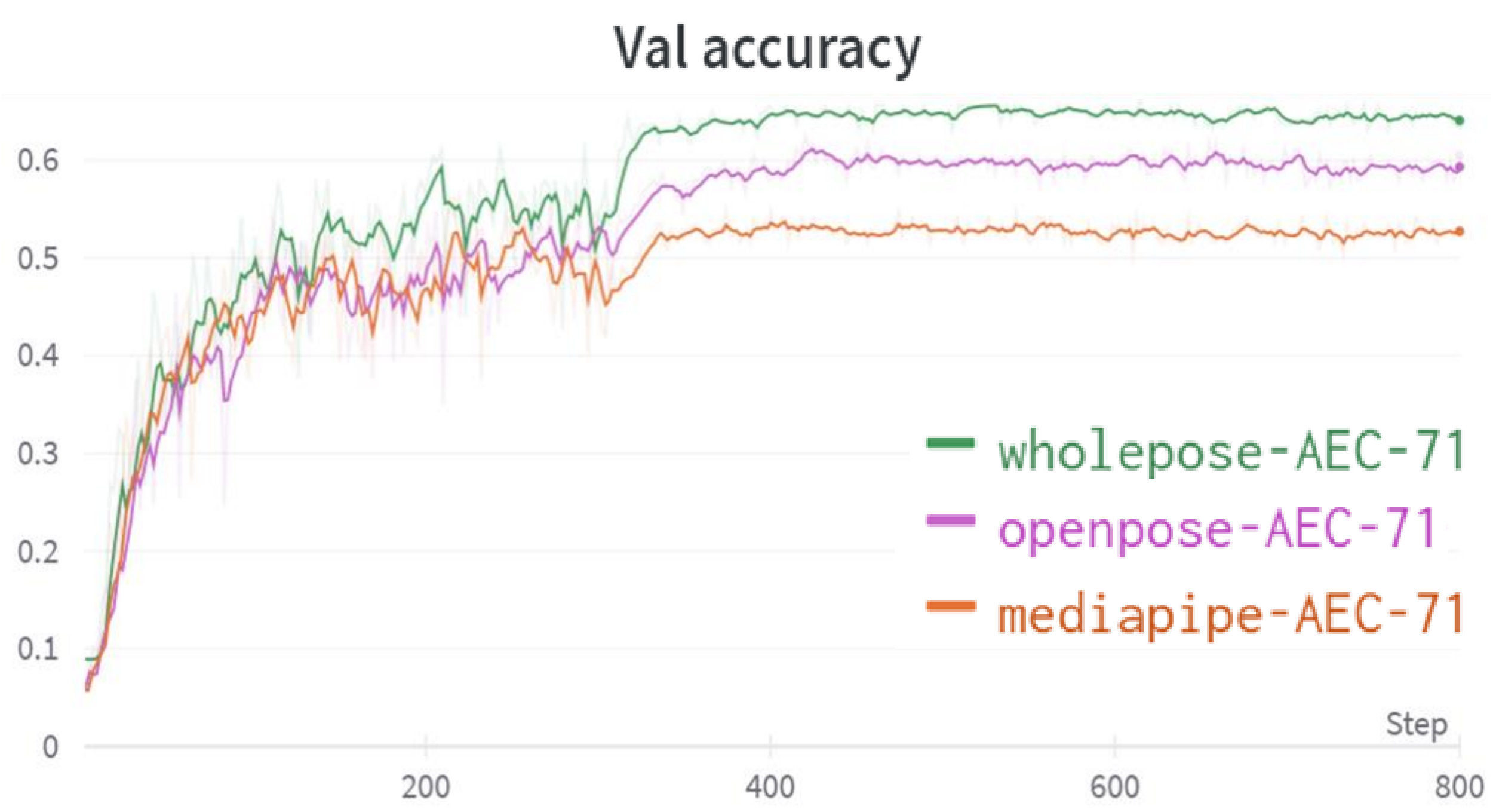
SLR Model	Library	Top-1					
		AEC		PUCP		WLASL	
		29	71	29	71	29	71
Spoter	MediaPipe	**0.649** ± 0.017	**0.665** ± 0.022	0.366 ± 0.021	0.390 ± 0.025	**0.634** ± 0.011	**0.701** ± 0.018
	OpenPose	0.528 ± 0.010	0.544 ± 0.031	**0.467** ± 0.020	**0.505** ± 0.009	0.473 ± 0.007	0.576 ± 0.011
	WholePose	0.613 ± 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

Results and Discussion

We tune the models for each dataset by finding the best learning rate for one pose-library, using the same values for the other two, and training for 400 epochs.

None of the libraries works the best in all the datasets and settings. WholePose works better most of the times compared to the other libraries in the two SL models for AEC and PUCP.



Conclusion

We found that **WholePose** shows **less number of bad-quality** landmarks and **performs better** most of the time in the two SLR models. This 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**. We did not find evidence that more keypoints produce better accuracy in the SL recognition models.

We plan to average several runs of these experiments to provide more robust results.