Impact of Video Length Reduction due to Missing Landmarks on Sign Language Recognition Model

Anonymous CVPR submission

Paper ID 53

Abstract

Sign Language Processing (SLP) has become an increasingly challenging field, particularly in the areas of sign language recognition (SLR), translation, and production. One of the primary challenges in SLP is pose estimation, which can be impacted by missing landmarks due to occlusions or limitations in the model's performance. In this study, we propose a method for evaluating the impact of missing landmarks on the performance of an SLR transformer-based model for the Isolated Sign Language Recognition (ISLR) task. We train and test the Spoter model on two subsets of Peruvian Sign Language datasets, and evaluate its performance using top-1 and top-5 validation accuracy. The study finds that removing frames with missing landmarks did not significantly impact accuracy in most of the cases, which suggests that additional preprocessing steps may not be necessary to deal with missing landmarks in this particular task. These findings contribute to the ongoing research in SLP and highlight potential avenues for improving SLP tasks.

1. Introduction

Sign Language Processing (SLP) has recently received a great deal of attention due to its relevance in sign language recognition (SLR), translation, and production. These advances in SLP have heavily relied on the progress achieved by Human Action Recognition (HAR), particularly in pose estimation models. The skeleton-based modality produced by these models is well-suited for SLP tasks that require invariance to clothing or background of the subjects. Nonetheless, pose estimation still poses a significant challenge to SLP, as missing landmarks can occur due to occlusions or model limitations.

In this study, we propose a method to filter out frames with missing landmarks to investigate their impact on the performance of an SLR model. To achieve this, we extract left and right hand landmarks, which are usually prone to errors in the hand pose estimation model. We employ the Mediapipe Holistic model [12] to address this issue, which initially produces null entries when a landmark is missing from a video frame. Unlike previous studies that replaced null entries with the wrist points from the same pose estimation model [11], we filter out frames with missing landmarks to investigate their impact on the SLR model's performance.

To assess the impact of missing landmarks on the performance of the SLR model, we employ the Spoter transformer-based model to train and test on two subsets (baseline and reduced) of two Peruvian sign language datasets: AEC and PUCP305. We chose the Spoter model due to its state-of-the-art performance on the WLASL dataset, which is widely used to evaluate sign language recognition models [3]. We evaluate the Spoter model's performance using top-1 and top-5 validation accuracy to determine if removing frames with missing landmarks can improve the performance of the SLR model. Our findings provide valuable insights into the impact of missing landmarks on the performance of the SLR model and suggest potential avenues for improving SLP tasks.

2. Related work

Pose estimation is a challenging research area in computer vision with various applications, including sign language recognition and detection. Previous research has utilized different techniques for pose estimation, such as contour-based features, histogram of gradients (HOG) features, and edges [1, 2, 7, 10].

Landmark localization, a crucial component of human pose estimation, involves identifying the precise positions of specific body parts in an image. Reliable landmark estimation is vital for robust vision tools, such as hand tracking, gesture recognition, facial expression recognition, and eye gaze tracking [5, 6, 8].

In Sign Language Recognition, tracking pose, hand, and facial movements is critical. Although several pose estimation models exist, OpenPose and Google's MediaPipe

are the two most commonly used frameworks in the sign language recognition research community due to their ease of use and seamless integration into existing sign language recognition pipelines [4, 12].

After curating the videos, undesired camera movements and jitters may affect the estimated poses' continuity if not corrected. In addition, low-resolution factors, such as blurry video frames or fast hand movements, may cause lost landmarks. Models like MediaPipe may miss landmarks in such situations, and these landmarks will not be included in the landmark group for the video frame. Various studies have proposed several methods, such as linear filters, Kalman filters, particle filters, and interpolation, to address this issue [9, 13, 16].

This study aims to investigate the impact of reducing the dataset by filtering out frames with missing landmarks on the accuracy of an SLR model. By examining the accuracy of pose estimation on four sign language datasets and proposing a methodology to remove frames with missing landmarks, we seek to better understand the significance of missing landmarks in the performance of SLR models.

3. Methodology

The main objective of this study is to evaluate the impact of missing landmarks on the performance of Sign Language Recognition models, specifically focusing on the Spoter model. The importance of filtering out frames containing missing landmarks and its potential influence on the model performance is emphasized.

3.1. Pose-estimation Library

We used Mediapipe library [12] to annotate landmarks used in a Sign Language Processing (SLP) task such as sign language recognition (SLR). The holistic Mediapipe model is used for generating pose, face and hand landmarks in the videos. However, the library is known to generate missing landmarks when it fails to accurately estimate the landmarks positions, particularly the hand pose estimation model, impacting in the SLR model performance.

3.2. Datasets and Data Analysis

To investigate the significance of frames with missing landmarks on hands, two subsets were used for the study:

1. **Baseline Subset:** The baseline subset consists of videos containing pose estimation landmarks, where hand missing landmarks where fixed to the wrist landmark. However, it didn't include those videos with zero frames after processing on the Reduced Subset

2. **Reduced Subset:** The reduced subset is obtained by filtering out such frames by checking if all landmarks within a hand (left or right) are fixed at the same position (wrist), effectively removing frames containing missing landmarks. Those videos with missing landmarks in all their frames where substracted from the whole dataset

We conducted our study on Peruvian Sign Language using two datasets, AEC and PUCP. The PeruSil framework was used to preprocess the datasets and obtain isolated sign videos, and the ConnectingPoints repository was used to extract keypoint landmarks of the signer from each video. For training our sign language recognition models, we carefully selected classes with more than 15 instances and excluded classes that performed poorly in initial experiments or those that involved pointing, resulting in 26 classes for the AEC dataset and 17 classes for the PUCP dataset.

Figure 1 shows the percentage of frames removed for both the baseline and reduced versions of each dataset. We found that the majority of instances had a frame reduction percentage of less than 25% for AEC and 20% for PUCP305. Meanwhile, Figure 2 displays a sequence of frames with and without missing landmarks. We believe that frames without missing landmarks could represent important signs or sign transitions, such as from C to D.

3.3. Model Training and Validation

We divided each dataset (AEC, PUCP) into 80% for training and 20% for testing. The Spoter model was used for SLR, and we used a variable learning rate, with a number of epochs of 250. The number of encoder and decoder layers were 6, and the feedforward dimension was 2048. We used a transformer model with 9 heads for multi-head attention. We trained and tested the Spoter model on both the baseline and reduced subsets of each dataset and compared the performance of the Spoter model to evaluate the effect of frame filtering on SLR performance.

4. Results

In this section, we present the results obtained after evaluating the SLR model in both the Baseline Subset and the Reduced Subset. We conducted an analysis to evaluate the model's performance in each subset under similar conditions. Based on these results, we investigated the impact of missing landmarks on the model and whether their removal would improve its performance. It's worth noting that this analysis was performed under comparable conditions, with the same number of classes and videos, but with a reduction in the number of frames for the Reduced Subset as described earlier.

4.1. Impact of Frame Filtering on Datasets

We examined the impact of filtering frames with missing landmarks, and it resulted in a 14,58% reduction in the

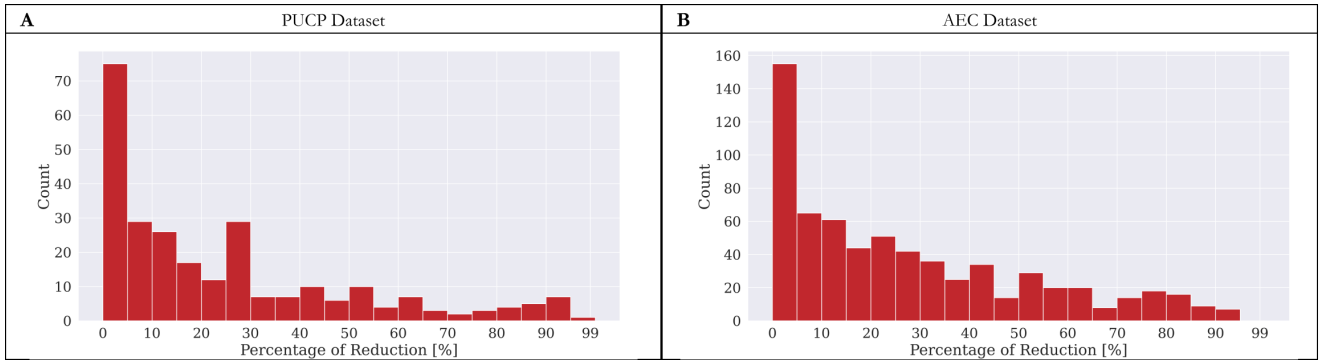

Figure 1. Distribution of Percentage of Reduction in both datasets

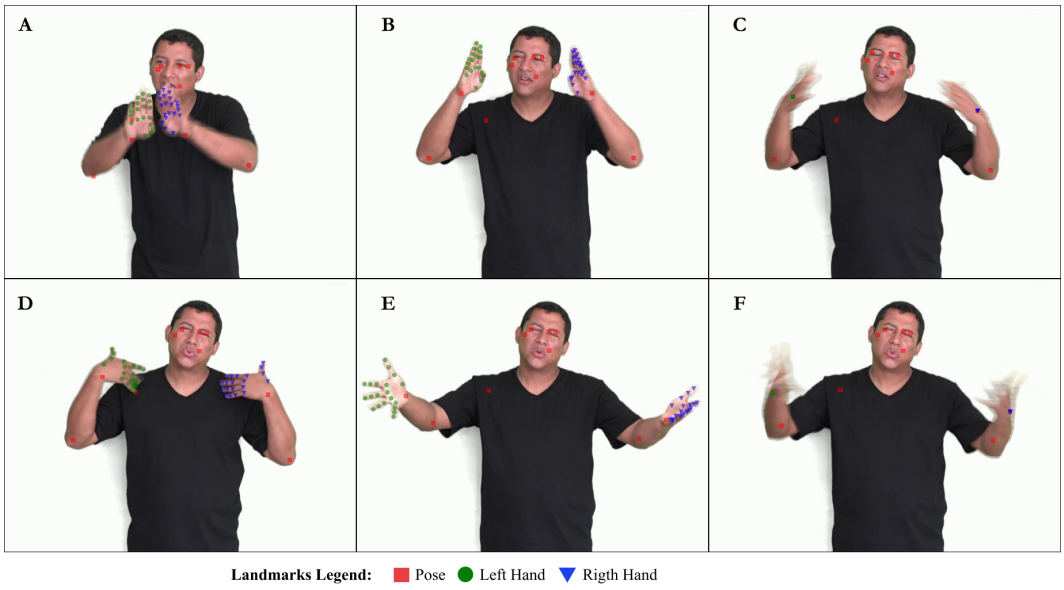

Figure 2. Examples of Landmarks estimation for frames without missing landmarks (A,B,D,E) and with missing landmarks (C,F)

number of videos for the AEC dataset, and a 19,76% of reduction in the PUCP dataset. By reducing the both datasets to the reduced number of videos (videos with zero frames after frame reduction where substracted), a fair comparison could be achieved between the models in the SLR.

4.2. Model Performance

We evaluated the Spoter model's performance on the original and filtered datasets using evaluation metrics such as Top-1 and Top-5 maximum accuracy. The results are presented in Table 1, where we report the Top-1 and Top-5 validation accuracy scores.

We compared the performance of our models on both the baseline and reduced subsets of the AEC and PUCP305 datasets. During five experiments performed under the same conditions, the models achieved an average accuracy of 0.801 $\pm$ 0.017 and 0.751 $\pm$ 0.009 for the AEC baseline and reduced subsets, respectively, in Top-1, and 0.978 $\pm$ 0.000 and 0.970 $\pm$ 0.005 in Top-5. For the PUCP305 dataset, the models achieved an average accuracy of 0.777 $\pm$ 0.022 and 0.777 $\pm$ 0.022 in Top-1 for both the baseline and reduced subsets, respectively, and 0.936 $\pm$ 0.015 and 0.940 $\pm$ 0.014 in Top-5. To assess the statistical significance of our results, we conducted a Kruskall Wallis test using the values obtained from the five experiments. Our analysis revealed a significant difference between the baseline and reduced versions of the AEC dataset for Top-1 accuracy ($p < 0.01$)

5. Conclusions

In this study, we proposed a method to evaluate the impact of missing landmarks on the performance of a transformer-based model for Sign Language Recognition (SLR). We conducted experiments using the Spoter model

Table 1. Accuracy comparison of baseline and reduced models on the AEC and PUCP305 datasets

Dataset	Top-1		Top-5	
	Baseline	**Reduced**	**Baseline**	**Reduced**
AEC	0.801 ± 0.017	0.751 ± 0.009	0.978 ± 0.000	0.970 ± 0.005
PUCP305	0.777 ± 0.022	0.777 ± 0.022	0.936 ± 0.015	0.940 ± 0.014

on two subsets (baseline and reduced) of two Peruvian Sign Language datasets: AEC and PUCP. Based on our key findings, we conclude that missing landmarks may not contribute significantly to the model's learning process, as the Spoter model achieved similar accuracy scores in both Top-1 and Top-5 validation accuracy on the Reduced and Baseline subsets.

However, we found that missing frames can impact slightly on the model's performance by losing representative continuous fractions within multiple videos, partially losing the temporal factor in the frame sequence. Therefore, for future work, we plan to evaluate the impact of reduced frames due to missing landmarks on other datasets and investigate the use of interpolation models on frames with missing landmarks to retain the temporal factor in hand landmark movement. Additionally, we aim to test our approach on larger sign language datasets such as AUTSL [14] and INCLUDE [15] to validate our findings on a broader scale. Overall, our study provides valuable insights into the impact of missing landmarks on SLR performance and suggests potential avenues for improvement in SLP tasks.

References

[1] F Ababsa. A robust method for 3d hand tracking using histogram of gradients and particle filter. In *2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012)*, volume 1, pages 339–342. IEEE, 2012. 1

[2] R Anusha and CD Jaidhar. Human gait recognition based on histogram of oriented gradients and haralick texture descriptor. *Multimedia Tools and Applications*, 79(11-12):8213–8234, 2020. 1

[3] 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*, pages 182–191, 2022. 1

[4] 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. 2

[5] Nasser Dardas, Qing Chen, Nicolas D Georganas, and Emil M Petriu. Hand gesture recognition using bag-of-features and multi-class support vector machine. In *2010 IEEE International Symposium on Haptic Audio Visual Environments and Games*, pages 1–5. IEEE, 2010. 1

[6] Dragos Datcu and Stephan Lukosch. Free-hands interaction in augmented reality. In *Proceedings of the 1st symposium on Spatial user interaction*, pages 33–40, 2013. 1

[7] Xijian Fan and Tardi Tjahjadi. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences. *Pattern Recognition*, 48(11):3407–3416, 2015. 1

[8] Kaoning Hu, Shaun Canavan, and Lijun Yin. Hand pointing estimation for human computer interaction based on two orthogonal-views. In *2010 20th International Conference on Pattern Recognition*, pages 3760–3763. IEEE, 2010. 1

[9] Sowmya Jayaram Iyer, P Saranya, and M Sivaram. Human pose-estimation and low-cost interpolation for text to indian sign language. In *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 130–135. IEEE, 2021. 2

[10] Arjun Jain, Jonathan Tompson, Yann LeCun, and Christoph Bregler. Modeep: A deep learning framework using motion features for human pose estimation. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12*, pages 302–315. Springer, 2015. 1

[11] Cristian Lazo-Quispe, Joe Huamani-Malca, Pontificia Universidad Católica del Perú, Manuel Stev Harold Huamán-Ramos, Gissella Bejarano, Pablo Rivas, and Tomas Cerny. Impact of pose estimation models for landmark-based sign language recognition. 1

[12] 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. 1, 2

[13] Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. *arXiv preprint arXiv:2110.05877*, 2021. 2

[14] Ozge Mercanoglu Sincan and Hacer Yalim Keles. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access*, 8:181340–181355, 2020. 4

[15] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1366–1375, 2020. 4

[16] Kathan Vyas, Rui Ma, Behnaz Rezaei, Shuangjun Liu, Michael Neubauer, Thomas Ploetz, Ronald Oberleitner, and Sarah Ostadabbas. Recognition of atypical behavior in autism diagnosis from video using pose estimation over time. In *2019 IEEE 29th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE, 2019. 2