Pose Detection for Word-Level American Sign Language Recognition

Alice Chase[7007281], Kushagra Sharma[7010529], and Dominik Weber[2548553]

Saarland University

Abstract. The progress in Machine Learning translation for sign language progresses slowly. Therefore, we tried to improve the results of a state-of-the-art sign language detection model with the help of pose detection. Related work shows for reasonable classification we need hand and face data. Using the biggest available data set for word-level American Sign Language, we used the top 100 and the top 300 words of this data set and applied pose detection to create additional data to fine-tune existing weights of the recognition model.

Keywords: Sign Language Detection · Pose Detection · Word-Level American Sign Language.

1 Introduction

There is an abundance of machine translation software readily available with decent results, such as Google Translate. However, progress in sign language translation continues to be slow. Sign language presents various unique challenges in creating robust recognition software compared to spoken languages.

Due to these problems in Sign Language Recognition and the importance of accessible technologies, we explored how pose detection could increase the performance of such models. The main premise of our experiment was to add pose detection skeletons to a sign language data set using open source software to see how it affected the accuracy of individual word-level signs with a combined CNN-RNN architecture.

2 Related Work

Before beginning the technical part of the project, it was important to learn more about sign languages and how they are expressed and current challenges in the field. Elliott and Jacob's paper explains that communication via sign language has more factors to it than just hand gestures — facial expressions, lip movements, body language and even objects like tattoos play a part in it's meaning and interpretation.[5] These were important things to consider when choosing a pose detection model. Other research explains how a big challenge is getting large enough data sets due to a relatively small amount of speakers, and additional computer vision variables in data such as lighting, angles, and backgrounds.[8]

Understanding these challenges lead us to decide to use American Sign Language because there is more data publicly available compared to other sign languages. The WLASL (Word-Level American Sign Language) paper which our experiment was centered around is responsible for creating the most comprehensive word-level sign language data set for American sign language.[7] From this we took subsets, we also used one of their given sign language recognition models and compared our accuracies to theirs. The sign language recognition model was a combination CNN-RNN, and there is many other works that have used similar models for sign languages.[3] [9]

Other research has supported pose detection's potential to improve activity recognition in areas other than sign language. [4] [6]

3 Proposed Method

3.1 Data

Since having high quality data is the foundation of supervised learning, we decided to use the WLASL — 'Word-level Deep Sign Language Recognition from Video' data-set.[2] It is the largest video data-set for Word-Level American Sign Language (ASL) recognition, and contains 2000 glosses (words) with 21083 videos featuring 119 different signers. Due to the smaller scope of our project, we used the top-k subsets of 100 and 300 glosses.

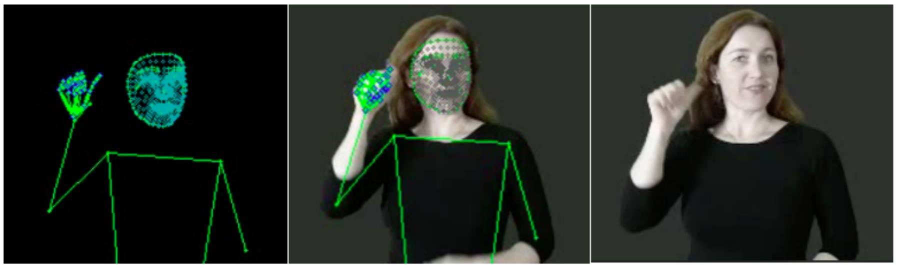
The 300 data set contained 5,117 videos with an average of 17samples per gloss and a total of 109 signers. We used a provided script from WLASL's GitHub to partition it into training, validation, and test data. The 100 subset contains 2038 videos, with an average of 20.4 videos per gloss and 109 different signers.

3.2 Pose Detection

After obtaining the data, we ran it through Media Pipe's holistic pose estimation model.[1] We chose this because it is not merely hand pose estimation but also tracks the upper body and facial landmarks which are important in sign language communication. We modified the code slightly in order to handle our input and get the desired output format since it was originally made for real-time pose detection via webcam but our sign language recognition model require mp4 video inputs. Additionally, we used different colors to highlight the key-points of both hands, so it becomes easier to differentiate between the left and the right hand.

We obtained two outputs from the pose recognition model, one for the pose skeleton overlaid over the original video and the second one for just the pose

Fig. 1. The gloss for 'audiologist' at roughly the same frame for each data type.



skeleton on a plain black background as seen in Figure 1.

The next step is to use the outputs from the pose detection model into the pre-trained sign language recognition model, where we fine-tune the weights with our new augmented data and then run it over the test sets.

3.3 Sign Language Recognition

After reviewing relevant literature, we decided to use 3D Convolutional Networks, as it produced the best results on the original data set.

Specifically, we used the I3D model network as being a 3D Convolution Network it successfully captures the spatial information of the frames as well as the temporal relationship between them in a hierarchical order, which is important for pose detection. The models were pre-trained on ImageNet and fine-tuned on Kinetics before tuned over the original WLASL data sets, which we then further fine-tuned with our pose data.

For each subset of glosses (100 and 300) and data set type (pose only and pose-overlaid) we used different numbers of epochs while training, anywhere between 5-30. Due to memory constraints, we also had to reduce batch size for the 300 subset from 32(used in the original paper) to 6.

4 Experimental Results

We compared our methods of retraining on our augmented dataset(s) to the baseline in the WLASL paper. The italicized numbers are the cases in which we surpassed the accuracy in the original paper. As given in the original paper, we have used accuracy percentage for Top-K classes to report the results, where K is the number of most probable outputs. For instance, if Top-1 accuracy is 65 percent, then the most probable output given by the model is correct 65 percent of the times.

Method		**WLASL-100**			**WLASL-300**	
	Top-1	**Top-5**	**Top-10**	**Top-1**	**Top-5**	**Top-10**
Baseline	65.89	84.11	89.92	56.14	79.94	86.98
Pose-Overlaid	*66.89*	*85.46*	89.5	*57.77*	*80.37*	86.18
Pose-Only	*67.38*	*88.5*	*91.08*	52.21	75.72	82.97
Both	-	-	-	*56.39*	79.52	85.27

5 Conclusions and Future Work

The most significant improvement was observed for the WLASL100 pose-only data. This may be because by just training on the pose skeleton, we remove all the other noise from the video. For Pose overlaid, the model beat the baseline accuracies for Top-1 and Top-5 classes and came close to beating the top-10 class too, for both WLASL100 and WLASL300. While training on both pose only and pose overlaid data together, the model returned similar results to the baseline model. Even though the WLASL300 pose only dataset failed to beat the baseline model, we believe that given a way to counter overfitting and more time to train it can return promising results.

It may be concluded that using the pose skeletons allows the model to make correct guess using fewer number of attempts, as we observed improvement in the Top-1 and Top-5 accuracies for all-most all combinations of methods and data subsets. Although, it is hard to determined exactly if it improves upon the original methods, or it is due to inherent randomization when training models.

A few more methods can be tried to check if we can make further improvements. For instance, it will be interesting to combine the weights from pose laid over training data and pose only training data and use those weights on the test data.

The original WLASL paper also mentions results from pose-based models, which perform human pose estimation by localizing the key points and joints of human body from a single image or video. Since this task should become easier with pose skeleton overlaid on the body using this model also seems promising. However, the data is quite huge, and we were limited by time and resources, so we could not test the result for all possible methods.

We also noticed that the model starts to overfit the data while training, so using some form of regularization might improve the results even further.

References

1. Mediapipe pose detection model, https://github.com/google/mediapipe
2. Wlasl data set and code, https://github.com/dxli94/WLASL
3. Real-Time Sign Language Gesture (Word) Recognition from Video Sequences Using CNN and RNN, Advances in Intelligent Systems and Computing book series, vol. 695. Springer, Singapore (2018)
4. Chéron, G., Laptev, I., Schmid, C.: P-cnn: Pose-based cnn features for action recognition (2015)
5. Elliott, E., Jacobs, A.: Facial expressions, emotions, and sign languages. Frontiers in Psychology **4**(115) (2013). https://doi.org/10.3389/fpsyg.2013.00115
6. Kim, J., Lee, D.: Activity recognition with combination of deeply learned visual attention and pose estimation. Applied Sciences **11**(9) (2021). https://doi.org/10.3390/app11094153, https://www.mdpi.com/2076-3417/11/9/4153
7. Li, D., Opazo, C.R., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison (2020)
8. Nada Bahaa Ibrahim, Hala Zayed, M.S.: Advances, challenges, and opportunities in continuous sign language recognition. Journal of Engineering and Applied **15**(5), 1205–1227 (2019)
9. Pigou L. Dieleman S., Kindermans, P.J..S.B.: Sign language recognition using convolutional neural networks. Lecture Notes in Computer Science p. 572–578 (2015). https://doi.org/10.1007/$978 - 3 - 319 - 16178 - 5_40$