Critical Analysis and Recommendations

This study explored and compared two different approaches in sign language recognition. Method 1 utilize extracted MediaPipe 3D Landmark whereas method 2 utilize video frame sequences with Mediapipe Landmark. Since both methods uses different number of epochs for model training due to memory constraints, the model performance at epoch 20 were extracted for comparison.

Table 1 Method 1 model performance on training and validation set at epoch 20.

Method 1 At epoch 20	**MediaPipe 3D Landmark**			
	Model 1 (Conv1D-LSTM)		**Model 2 (Conv1D-GRU)**	
	LSA64	**WLASL**	**LSA64**	**WLASL**
Training accuracy	94.10%	67.67%	91.31%	67.43%
Training loss	0.3233	1.0701	0.3995	1.0746
Validation accuracy	94.18%	38.68%	92.70%	35.83%
Validation loss	0.2880	2.0694	0.3478	1.9236

Table 1 Method 2 model performance on training and validation set at epoch 20.

Method 2 At epoch 20	**Video Frames Sequence with MediaPipe**			
	Model 1 (Conv3D-LSTM)		**Model 2 (Conv3D-GRU)**	
	LSA64	**WLASL**	**LSA64**	**WLASL**
Training accuracy	97.01%	94.92%	99.37%	99.80%
Training loss	0.1626	0.4427	0.07	0.0153
Validation accuracy	78.23%	17.57%	77.19%	19.03%
Validation loss	0.9523	3.6310	1.2999	8.0969

It was observed that when adopting frame-sequence method, both models were able to capture the features from training data and learned very quickly, achieved more than 90% training accuracy and less than 0.5 training loss even on WLASL dataset. However, based on the performance on validation data, it shows that by using this method, the models were not able to generalize on unseen data so quickly. Hence, overfitting was observed in all models under this method. Conversely, by using MediaPipe 3D Landmark, although the model performance on training set was not as good as in method 2, all models were able to generalize on unseen data faster. Therefore, in comparison to method 2, it was noticed that when applying method 1

on both datasets, both validation accuracy and loss showed better results. However, this observation may not be solely contributed by the different method, the difference in model architecture may have also caused this observation. Besides that, the frame-sequence method requires a much longer time to train especially on data with high diversity but small in size. For example, with LSA64 dataset and conv-LSTM model, MediaPipe 3D-Landmark approach requires only 310ms/step while frame-sequence approach requires 10s/step in training.

In each method, two different models were trained and validated using two datasets LSA64 and WLASL dataset.

Table 2 Method 1 models performance comparison.

Method 1 200 epochs	**MediaPipe 3D Landmark**			
	Model 1 (Conv1D-LSTM)		**Model 2 (Conv1D-GRU)**	
	LSA64 (87ms/step)	**WLASL (106ms/step)**	**LSA64 (75ms/step)**	**WLASL (82ms/step)**
Test accuracy	98%	60%	97.33%	66.67%
Test loss	0.0799	5.2784	0.2267	2.8769
F1-score	0.98	0.53	0.97	0.62
precision	0.98	0.5	0.98	0.60
recall	0.98	0.6	0.97	0.67
Model Fitness	Good Fit	Overfit	Good Fit	Overfit

Table 3 Method 2 models performance comparison.

Method 2 20 epochs	**Video Frames Sequence with MediaPipe**			
	Model 1 (Conv3D-LSTM)		**Model 2 (Conv3D-GRU)**	
	LSA64 (26s/step)	**WLASL (236ms/step)**	**LSA64 (35s/step)**	**WLASL (147ms/step)**
Test accuracy	86%	20%	92.67%	13.3%
Test loss	0.4753	4.2496	0.2995	7.2542
F1-score	0.89	0.17	0.81	0.13
precision	0.92	0.16	0.83	0.13
recall	0.91	0.20	0.81	0.13
Model Fitness	Overfit	Overfit	Overfit	Overfit

The results showed that all models performed better when using LSA64 dataset than the WLASL dataset. The reason behind is that, despite some videos were recorded in an outdoor environment, the videos in LSA64 were still in a more controlled settings and consistent. Moreover, the distance between the camera and signer are similar across all videos. LSA64 contains more data than WLASL, each word has 50 videos while all videos show the same sign. Hence, the model can learn the features in a shorter time. Conversely, WLASL dataset is more

diverse as it was gathered from various sources. The videos were recorded in different backgrounds by different signers. Hence, WLASL is closer to the real-life situation. Overall, the results shows that LSTM model outperforms GRU in all experiments by 6% to 7% except in method 1 when using WLASL dataset.

There were several challenges encounters when conducting this study. Mediapipe Landmarker failed to track the hand landmarks in part of the videos. Therefore, when adopting method 1, normalization with minimum-maximum value of the hand is not feasible as this will yield Nan value. Therefore, the distance values were normalized with respect to the signer's shoulder width which is always present throughout the video.

In addition to that, as the second method takes in video frame sequences as data input, it is more computation expensive. Due to the hardware memory constraints, the training sessions for method 2 was conducted using Google Colab Notebook. However, as there is also an absolute timeout of 24 hours in Google Colab. Due to the time limit, the number class prediction and training epochs was restricted to 15 words and 20 epochs respectively. Hence, to keep the variables consistent in this experiment, only 15 words were used from each dataset in the study. To further reduce the computation cost, all frames were transformed to grayscale and image size was reduced to 100*100. In addition to that, the number of frames were also restricted to 8 frames for all videos. By skipping the first 500msec of the video, this can ensure that the complete sign sequence was preserved.

Another challenge faced when applying method 2 on WLASL dataset is that the distances between the signers and camera varies. In some videos, the landmarkers appear to be obscure which affects the model performance. This issue was mitigated by cropping the video frames with bounding box to focus on the signer. Besides that, as some videos were no longer available from the source, the reduction in data size causes some words having more than one sign due to dialect variation. To alleviate the effect, each word was limit to one sign and data augmentation was performed on the training set including rotation, horizontal flip, vertical flip, cropping and the combination of these techniques.

In fact, the performance can be further improved by stabilizing the hand landmarks. As sign language recognition depends heavily on the hand orientation and movements, without proper detection of hand landmarks causes majority of the features to be missing. As a result, the model fails to learn different features across the words. During the study, it was observed that MediaPipe hands tracking solution was not able to track the hands all the time. To improve this

condition, several preprocessing steps may help include applying bounding box on hands to enable more accurate hand tracking.