ISSN: 2664-9640

Front-View Chinese Sign Recognition in a Continuous-Like Setting: A Reproducible Baseline and Evaluation Protocol

Mingze Han*

International Division, Beijing 101 High School, Beijing 100091, China

*alexhan0722@outlook.com

Abstract

This paper introduces a reproducible front-view benchmark setting for Chinese sign recognition in a continuous-like single-sign scenario. Building on the NationalCSL-DP resource, only red, green and blue (RGB) frames from the front camera are used as input, and this paper investigates the performance of a frame-based isolated sign recognition pipeline when exposed to untrimmed sign sequences instead of action-centric samples. To support reuse, the original per-sequence folders are converted into ImageFolder-style training data, a fixed 30-frame sampling policy is used, and preprocessing metadata is saved together with the checkpoints used for inference. The baseline is a two-tier pipeline consisting of a binary activity model and a ResNet-18 gloss classifier. Among the 6,707 front-view sequences, the resulting continuous-like diagnostic evaluation achieved a Top-1 accuracy of 0.2646, a Top-5 accuracy of 0.5184, an average prediction confidence of 0.4777, and zero no-segment failures under the selected segmentation policy. Further ablation studies have been conducted on frame budget, fusion strategy and segmentation thresholds, and the gap between action-centric frame classification and sequence-level continuous-like evaluation has also been analysed. Based on the above, boundary noise, single-view ambiguity and the mismatch between isolated training and sequence-level inference are all still problems. This study will serve as a reusable baseline and evaluation protocol for future research in front-view Chinese sign language resources, but it will not be a continuous state-of-the-art sign language recognition system.

Keywords

Chinese Sign Language; Language Resources; Reproducibility; Benchmark; Front-view RGB; Isolated-to-continuous Evaluation.

1. Introduction

Sign language technology is relatively difficult to study because there are fewer well-documented resources, opaque evaluation scenarios and non-reproducible baselines for other scholars. Chinese sign language is a typical case, and the available public resources are relatively scarce compared with those for English. Although the coverage and recording quality of recent datasets for sign vocabularies have been improving and expanding, there is still a practical gap between isolated sign recognition (ISLR) and real-world applications; that is, systems in practice often need to deal with longer sequences of input that contain rest poses, transitions and boundary uncertainties.

The deficit is more pronounced in front-view acquisition. A single front-facing camera is simple to install in classrooms, public service windows, and light-weight assistive apps, but it also lacks complementary viewpoint information and makes it difficult to distinguish depth-related hand movements. A resource-oriented work problem has thus emerged: how much performance loss

ISSN: 2664-9640

occurs when a model trained on isolated, action-centric samples is applied to a more realistic sequence-level setting with only front-view input?

To address this problem, a front-view-only experimental setup has been derived from the NationalCSL-DP dataset [1], and a continuous-like single-sign evaluation protocol has been established. Rather than presenting an elaborate temporal structure, this study will use a relatively basic and understandable one as the baseline. The pipeline finds candidate signing intervals first, then uses an isolated gloss classifier on frames in this segment. The design can serve the purposes of resource-oriented evaluation by having individual stages that are independently inspectable, reproducible and replaceable.

Therefore, this paper's contribution is not the establishment of a new high-performance continuous sign language recognition system. Instead, it is a resource-oriented baseline study that documents the behaviour of a front-view Chinese sign recognition pipeline in a continuous-evaluation mode and what errors appear under this condition.

The main contributions of this paper are as follows:

- A front-view-only, continuous-like evaluation set has been created based on the NationalCSL-DP resource, and a practical single-camera scenario for Chinese sign recognition has been introduced.
- A reproducible baseline pipeline is saved to retain frame conversion, fixed sampling, checkpoint metadata and sequence-level inference rules for reuse in future benchmark expansions.
- Protocol-oriented ablations of frame budget, fusion strategy and segmentation thresholds are reported to show how design choices affect sequence-level performance.
- An error analysis of the isolated-to-continuous performance gap is presented here, including boundary noise, single-view ambiguity and class-level variation.

2. Related Work

2.1. Sign Language Resources and Benchmark Settings

Together with the progress in sign language recognition, annotated resources have also been developed. The early work used sensor-based input, such as data gloves [2], but later a vision-based corpus was employed as the main foundation for model development. The RWTH-PHOENIX-Weather-2014 corpus [3] established a well-known benchmark for continuous German Sign Language with gloss-level sentence annotations, and later supported neural sign language translation [4]. WLASL [5] and MS-ASL [6] have extended the vocabulary size and signer diversity of American Sign Language for isolated sign recognition. Huang et al. [7] have published a video-based Chinese Sign Language resource for sentence-level recognition in China that does not require explicit temporal segmentation.

More recently, Jin et al. [1] have introduced the NationalCSL-DP dataset, a large Chinese National Sign Language resource with synchronized front and left views. This resource has a relatively wide range and can be selected from various places in an experiment. Only the front-view part of this dataset will be used in the current study to conduct a low-cost deployment investigation and to document a baseline evaluation setup that other researchers can reproduce.

2.2. Isolated Recognition to Continuous Evaluation

ISLR usually assumes that a video clip has already been cut to a single gloss. CNN backbones such as ResNet [8] are often used in this context, and they have generally been pre-trained on ImageNet [9]. I3D [10] is a video model, and more recent Transformer-style architectures and multimodal systems that combine RGB input with pose or other cues have further improved

ISSN: 2664-9640

recognition accuracy [11, 12]. Kvanchiani et al. [13] have shown that the training strategy and sampling policy are also directly related to the results of ISLR.

Continuous sign language recognition (CSLR), on the other hand, is an unsegmented stream that also needs to address alignment and classification. Hidden Markov Models (HMMs) [2, 3], hybrid CNN-HMM systems [14], and connectionist temporal classification (CTC)-based models [15-17] are all used to solve this alignment problem in different ways. For instance, Pu and others [18] have shown that iterative alignment can reduce the discrepancy between isolated and continuous settings. This paper will not be directly compared with such systems. It examines what occurs when an ISLR-oriented pipeline is shifted to a front-view, sequence-level mode that is more realistic than isolated frame classification but simpler than full sentence transcription.

2.3. Single-View Constraints and Reproducible Baselines

Single-view sign recognition is relatively easy to collect and apply, but it is also ambiguous. Front-view capture may obscure hand depth and self-occlusion, and therefore, it will be more challenging to observe subtle motion trajectories at this time. Therefore, single-view experiments are not only modelling problems but also evaluation problems; they help indicate how much information is preserved under actual capture conditions.

A resource- and evaluation-based reproducible baseline is more likely to be used in the early stages of task definition. A relatively simple, yet transparent, pipeline can be used as a reference point for the future development and assessment of resources. This baseline-oriented approach will be used in the current study.

3. Dataset, Evaluation Protocol and Baseline Pipeline

3.1. Front-View Benchmark Setting Based on NationalCSL-DP

The front-view RGB frames used in this work are from NationalCSL-DP [1]. The present study uses a 500-gloss subset of 6,707 front-view sequences for operation, and the official gloss labels are retained in preprocessing and evaluation to prevent manual relabeling drift. For frame-level baseline training, the data are organised in an ImageFolder-style layout, and sampled frames from each sequence can be reused consistently in all experiments. A fixed 30-frame sampling policy is used per sequence during training to reduce sequence-level imbalance and to keep the preprocessing consistent with the later sequence-level inference.

Table 1 shows the front-view setup employed in this paper. The train/validation split below is a frame-level baseline training configuration. The corpus-wide continuous-like analysis that was reported later is presented here as a diagnostic assessment of the available front-view sequences, and not as a standardised held-out CSLR benchmark.

Table 1. Front-view dataset setting for the baseline study.

Property	Value
Total gloss classes	500
Total front-view sequences	6,707
Frames per sequence (mean)	~85
Frame-level train / val split ratio	80% / 20%
Approx. train sequences	~5,366
Approx. val sequences	~1,341
Original frame resolution	512 x 512 px
Visual input used in this paper	Front view

ISSN: 2664-9640

For the model to be reproducible, each checkpoint also saves the class list, image size, normalisation parameters and model type in addition to the learned weights. The selected method will allow the inference process to be repeated without needing to rebuild the hidden preprocessing assumptions in external scripts.

3.2. Baseline Models

The gloss classifier is a ResNet-18 backbone [8] pre-trained on ImageNet [9]. Replace the output layer with a 500-way classification head. Input frames are resized to 224x224 and normalised according to the standard ImageNet statistics. Training uses stochastic gradient descent (SGD) with a momentum of 0.9, an initial learning rate of 0.01, cosine annealing, a batch size of 64, cross-entropy loss, and 15 epochs. Data augmentation includes random horizontal flipping, random rotation (±10°), and colour jitter.

An activity detector is a binary ResNet-18 baseline and predicts either "signing" or "not_signing". Use the same optimisation settings and have two output classes. Frames in the middle 60% of each sequence are considered proxy signing frames, and frames in the first and last 20% are regarded as proxy non-signing frames. This is a heuristic labelling strategy rather than a manually annotated boundary resource, and it is explicitly stated because it will affect the interpretation of the later sequence-level results.

3.3. Segment Construction and Sequence-Level Prediction

For each sequence, the activity model outputs a per-frame probability p(signing). A moving average is applied to the resulting curve for smoothing, and then it is thresholded to obtain binary activity flags. Positive runs are merged into segments, and only short negative gaps within a segment are allowed; very short segments are discarded. Table 2 shows the baseline hyperparameters.

Table 2. Segmentation Hyperparameters of the Baseline Pipeline.

Parameter	Value	Rationale
SMOOTH_WIN	5 frames	Reduces single-frame noise while preserving short sign intervals
THRESHOLD	0.5	Default decision boundary for the binary activity classifier
GAP	3 frames	Allows brief dips in p(signing) within one candidate segment
MIN_LEN	5 frames	Removes micro-segments that are unlikely to contain a meaningful sign interval

For each detected segment, up to 30 frames with the highest smoothed activity probabilities are retained, and then the gloss classifier is applied to these frames. Segment-level prediction is a weighted average of frame logits. If multiple candidate segments are found in a sequence, the one with the highest confidence in the best label is kept. This way will be relatively simple and serves as a reasonable reference for subsequent comparisons.

4. Experimental Results and Evaluation Analysis

4.1. Action-Oriented Reference Conditions

First, a baseline classifier is used as a reference to evaluate it in a static action-centric manner, with evaluation frames selected from the informative regions of the sign sequence. The validation accuracy in some representative runs reached about 0.986 very early in training. This number is reported as an aligned reference point and is not the main contribution of this paper. The first purpose is to show that front-view RGB frames can contain strong lexical cues when the temporal boundaries are favourable, and this is in line with the previous ISLR findings [13].

ISSN: 2664-9640

4.2. Corpus-wide Continuous-Like Baseline

Then, the entire baseline pipeline is applied to all 6,707 available front-view sequences in a continuous-like single-sign mode. Table 3 shows the corresponding diagnostic data.

Table 3. Corpus-wide diagnostic results of the front-view baseline pipeline.

Metric	Value
Total sequences analyzed	6,707
Sequences with at least one detected segment	6,707 (100%)
Sequences with no detected segment	0
Top-1 gloss accuracy	0.2646
Top-5 gloss accuracy	0.5184
Average confidence of the predicted label	0.4777
Mean detected segments per sequence	1.73

At least one candidate segment for each sequence is produced by the pipeline according to the selected thresholding policy, indicating good sequence coverage. This statistic is not a boundary recall value because the current protocol does not use manually annotated segment boundaries. Therefore, the baseline will still be available for the whole front-view corpus. A large gap between the Top-1 and Top-5 accuracy suggests that although the segmentation boundaries and selected frames may be incorrect, the true gloss is ranked relatively high in both rankings.

4.3. Protocol Sensitivity Studies

Three groups of ablations are carried out to explore the impact of design on the baseline.

Frame Sampling Budget. The number of retained frames in a candidate segment varies. Increase the budget from 10 to 30 to improve Top-1 accuracy; however, the increase after 30 frames is small, and it is likely that any additional frames will be from transition regions.

Table 4. Effect of per-segment frame budget on Top-1 accuracy.

Frames per segment	Top-1 Accuracy	Avg. Confidence
10	0.2231	0.4412
20	0.2487	0.4608
30 (default)	0.2646	0.4777
50	0.2694	0.4803

Fusion Strategy. Compare single-best-frame prediction, uniform averaging and activity-weighted averaging. Weighted fusion has the strongest Top-1 result, indicating that the activity score is a reasonably good, though not perfect, measure of frame informativeness.

Table 5. Comparison of Segment-level Fusion Strategies.

Fusion Strategy	Top-1 Accuracy
Single best frame	0.2104
Uniform average	0.2453
Activity-weighted average (default)	0.2646

Segmentation Threshold Sensitivity. The range of the activity threshold is 0.3-0.7; therefore, there will be a trade-off between the purity and coverage of the segments. A lower threshold retains more transition frames; otherwise, the sign information will be lost. The original threshold of 0.5 is relatively close to what the current protocol is seeking.

ISSN: 2664-9640

In short, the above ablations can be viewed as protocol validations; they indicate that the baseline is sensitive to temporal selection and that sequence-level performance is significantly affected by how the front-view stream is divided.

4.4. Class-Level Patterns

Class-level behaviour will also be investigated to determine where the baseline is most lacking. The best-performing glosses (accuracy > 0.8) generally have characteristic hand shapes or large-amplitude movements. The worst-performing glosses (accuracy less than 0.05) often fall into two categories: signs that differ in subtle finger position or wrist rotation, and signs with a short effective duration whose informative core is contained in only a few frames. Approximately 23% of the gloss classes have not been correctly predicted in the continuous-like setting, although they were recognised in the action-centric condition. The pattern shows that the two main sources of error in the current baseline are boundary quality and single-view ambiguity.

5. Discussion

5.1. The Isolated-to-Continuous Gap

The first is that the performance under the action-centric reference condition (Top-1 ≈ 0.986) is significantly higher than that in the corpus-wide continuous-like setting (Top-1 = 0.2646). Although the baseline classifier is still reasonably discriminative after selecting the informative frames in advance, most of this advantage is lost due to the addition of segmentation and frame selection in the pipeline. This indicates that, as shown in previous studies on alignment-sensitive recognition [18], performance is highly sensitive to the distribution of frames that have been observed at inference time.

That is to say, the core problem is not classifier capability. The sequence-level protocol adds transition frames, residual rest frames, and imperfect boundaries that are less frequent in isolated training data. As a resource and evaluation paper, this gap itself is a significant result and thus indicates that future improvements to the benchmark are urgently needed here.

5.2. What the Front-View Baseline Reveals

The baseline helps to show the three problems practically. First, boundary noise directly affects gloss recognition. Although weighted fusion is better than uniform averaging, the recognition stage is still relatively sensitive to whether the detected interval contains the key information of the sign. Secondly, a single front view is often inadequate to reveal the differences in depth, direction and fine-grained hand orientation of the signs. Third, there is still a large difference between classes, and some glosses are significantly more robust to front-view capture than others.

The above data do not indicate that the current baseline is optimal; instead, they show that the problem of front-view resource distribution is relatively more serious. Therefore, it is used to determine whether there are any problems.

5.3. Reproducibility and Benchmark Applications

The first purpose of this study is to set up an easy-to-examine and reproducible experiment. All of the above have been reported: the dataset conversion procedure, fixed frame budget, checkpoint metadata, and explicit sequence-level decision rules, because otherwise, hidden preprocessing choices would make the results of sign recognition difficult to compare. Therefore, the present protocol should be taken as the beginning for future research on front-view Chinese sign recognition.

At the same time, this evaluation should not be considered a complete standardised CSLR benchmark. The task here is a continuous-like single-sign prediction, not sentence-level

ISSN: 2664-9640

transcription, and the corpus-wide diagnostic results aim to describe the pipeline's behaviour under front-view conditions rather than to claim a new state-of-the-art.

6. Limitations and Future Work

This paper has the following defects. First, the gloss classifier considers frames independently and only combines them at the logit level; thus, motion cues are largely lost. Second, the activity labels are heuristic and have not been manually annotated for segment boundaries. Thirdly, continuous-like evaluation can be best seen as a diagnostic study over the entire corpus rather than a fixed held-out CSLR benchmark.

The above deficiencies indicate some upcoming work. A high-quality reference can be used to manually mark the boundaries or a more stringent held-out sequence method could be employed. Lightweight temporal modules, classifier-guided segment refinement or dual-view extensions can be used in the model to reduce the gap between isolated and sequence-level performance. Add more detailed metadata about the variation of the signer, segment length and failure type on the resource side for future evaluation.

7. Conclusion

This paper reformulates a front-view Chinese sign recognition study as a resource-oriented baseline and evaluation protocol for a continuous-like single-sign setting. Using front-view sequences derived from NationalCSL-DP, it is shown that a reproducible two-stage pipeline can be constructed, and a classifier that performs well under action-centric conditions shows a significant drop in performance after the addition of segmentation and frame selection.

The resulting corpus-wide baseline has a Top-1 accuracy of 0.2646, a Top-5 accuracy of 0.5184, an average prediction confidence of 0.4777, and zero no-segment failures under the selected segmentation policy. Based on the ablations and class-level analysis, frame selection, boundary quality and single-view ambiguity are all contributing factors to reduced performance. Therefore, this work can be used as a reproducible starting point for the future design of benchmarks, error analysis and resource development in front-view Chinese sign recognition.

References

[1] Jin, P., et al. (2025). A large-scale dataset of Chinese national sign language for dual-view isolated sign language recognition. *Scientific Data*, *12*, Article 660. https://doi.org/10.1038/s41597-025-04986-x

[2] Gao, W., Fang, G., Zhao, D., & Chen, Y. (2004). A Chinese sign language recognition system based on SOFM/SRN/HMM. *Pattern Recognition*, *37*(12), 2389–2402.

[3] Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, *141*, 108–125.

[4] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7784–7793).

[5] Li, D., Rodriguez, C., Yu, X., & Li, H. (2020). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 1459–1469).

[6] Vaezi Joze, H. R., & Koller, O. (2019). MS-ASL: A large-scale data set and benchmark for understanding American Sign Language. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[7] Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1, pp. 2257–2264).

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

[9] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255).

[10] Carreira, J., & Zisserman, A. (2017). Quo Vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6299–6308).

[11] Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign language recognition: A deep survey. *Expert Systems with Applications*, *164*, 113794.

[12] Koller, O. (2020). Quantitative Survey of the State of the Art in Sign Language Recognition. arXiv preprint arXiv:2008.09918.

[13] Kvanchiani, K., Kraynov, R., Petrova, E., Surovcev, P., Nagaev, A., & Kapitanov, A. (2024). Training strategies for isolated sign language recognition. arXiv preprint arXiv:2412.11553.

[14] Koller, O., Zargaran, S., & Ney, H. (2017). Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4297–4305).

[15] Graves, A., Fernandez, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)* (pp. 369–376).

[16] Cui, R., Liu, H., & Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, *21*(7), 1880–1891.

[17] Cheng, K. L., Yang, Z., Chen, Q., & Tai, Y. W. (2020). Fully convolutional networks for continuous sign language recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 697–714).

[18] Pu, J., Zhou, W., & Li, H. (2019). Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4165–4174).