Alignment Constraints for Video-based Sign Language Understanding

Yuecong Min
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)
Institute of Computing Technology, CAS, Beijing, 100190, China
yuecong.min@vipl.ict.ac.cn

Abstract

Video-based sign language understanding aims to recognize gloss sequences and translate a spoken sentence from the corresponding sign video. Compared to other sequence processing tasks (e.g., automatic speech recognition and neural language translation), sign language understanding datasets have limited samples and only provide sentence-wise annotations, which brings great difficulties for effective designs. My doctoral research focuses mostly on the alignment constraint design with limited supervised samples on sign language understanding tasks, such as selecting suitable modalities and extraction modules, designing proper visual alignment constraints to relieve overfitting, and exploring the effects of supervision signals. Although the proposed methods improve the performance on current sign understanding datasets, they still face difficulties in real-world scenarios. Recent developments in foundation models provide a great opportunity to make human-environment interaction techniques applicable, and my research plan mainly focuses on human-object interaction modeling and the data closed-loop platform design.

Personal Informations

My Ph.D. advisor is Prof. Xilin Chen. I started the PhD journey on September 2017 and I will defend it on December 2023. The foreseeable thesis title is “Alignment Constraints for Video-based Sign Language Understanding”.

1. Research Progress

Sign Language is a complete and natural language that conveys information through both manual components (hand/arm gestures) and non-manual components (facial expressions, head movements, and body postures) [16] with its own grammar and lexicon [20]. Vision-based Sign Language Understanding (SLU) aims to automatically recognize gloss sequences[1] (Sign Language Recognition, SLR) or translate a spoken sentence (Sign Language Translation, SLT) from a sign video, which can bridge the communication gap between the Deaf and hearing people. It also provides more non-intrusive communication channels for sign language users.

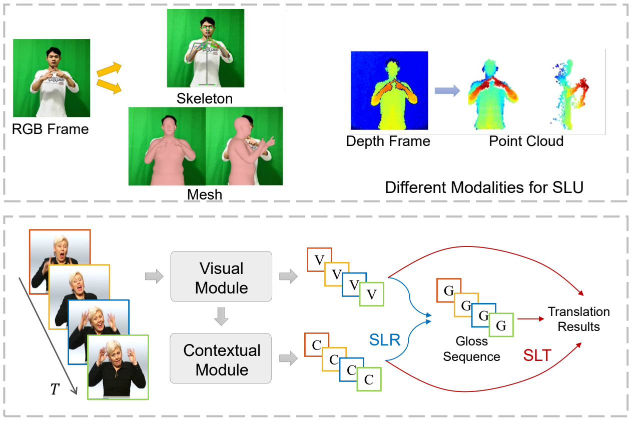


Figure 1. Illustration of commonly used modalities for SLU (the upper) and the relationship between different SLU tasks (the lower). We leverage the monotonically aligned nature of source features to design alignment constraints for SLU tasks.

1.1. SLU from Different Modalities

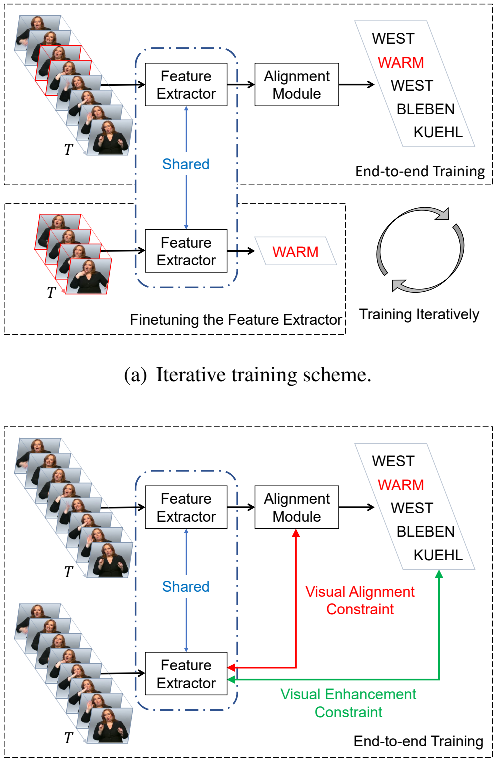
With the success of deep neural networks, SLU attracts more attention in recent years [9]. A key difference between SLU and other sequential tasks is its visual nature: SLU models should localize and identify fine-grained patterns (*e.g.* hand movements and postures) from video data, which contain much redundant spatio-temporal information and are easily affected by illumination and background changes. To guide the learning of visual features, previous works adopt an iterative training scheme [3] and pose-guide module designs [23]. However, they greatly increase the training time and make the network design complicated. As shown in Fig. 1, there are multiple types of modalities for SLU. We first explore the choice of input modality and provide solutions for different situations.

[1] Gloss is the written approximation of a sign.

SLR from Point Clouds in Close-range Interaction. Compared to the color camera, the depth sensor is more sensitive to distance changes and is more robust to illumination and background changes, which can also provide satisfactory segmentation results for close-range interaction. In our preliminary work [12], we first sample point clouds of signers from depth videos, and then extend the pioneering PointNet++ [18] to make it can process point cloud sequences. Specifically, we find there is no need to keep the high density of point clouds in gesture recognition and leverage the complementary information from neighboring frames can further reduce the density of point clouds. Therefore, we randomly sampled 128 points from each depth frame, and propose a spatio-temporal grouping layer to gather information from spatio-temporal local neighbors.

We further propose a PointLSTM module (code is available [2]) to explore the long-term spatio-temporal information from point cloud sequences [15]. The orderless of point clouds is the main challenge to applying LSTM on point cloud sequences. Therefore, we maintain the hidden and cell states for each point and update its states based on its spatio-temporal neighbors. We also proposed an approach with point-shared states to reduce the computation costs. Experimental results show that the proposed PointLSTM can achieve better performance than RGB and depth video-based methods and approach human recognition performance. The proposed method also surpasses the skeleton-based methods on gesture datasets when the estimated results are imperfect.

Visual Alignment Constraints Makes RGB-based CSLR Networks End-to-end Trainable. Although point clouds based approaches achieve satisfactory performance on gesture recognition, RGB sensors can better capture visual signals (e.g., facial expression) of sign languages, and most sign videos on the Internet are recorded by RGB sensors. The iterative training scheme [3] is widely used in RGB-based SLR approaches to capture correct visual signals, but it greatly increases the training time. To understand the difficulties behind the end-to-end training, we revisit the iterative training scheme in SLR and attribute the problem to the overfitting of the alignment module [13]. To make SLR models end-to-end trainable, we propose two kinds of visual alignment constraints (code is available [3]) as shown in Fig. 2(b). The visual enhancement constraint enforces the feature extractor to make predictions based on visual features only, and the visual alignment constraint aligns the short-term visual and long-term contextual predictions. With the combination of the two constraints, the proposed SLR model is end-to-end trainable and achieves superior performance to the latest methods on SLR datasets.



(a) Iterative training scheme.

(b) End-to-end training with the proposed VCs.

Figure 2. Training process comparison on SLR.

Skeleton-based Solutions can Achieve Comparable Performance with Video-based Solutions. Different to general actions, the information conveyed in sign language is totally independent of environments (*e.g.*, backgrounds), which indicates that conducting SLU on skeleton data is lossless if the skeleton data are perfect. Besides, skeleton data also lay a good foundation for the co-occurrence signals (*e.g.*, hand shape, facial expression, and lip pattern) exploration in CSLR. We propose a simple yet effective GCN-based approach named CoSign [6] to incorporate co-occurrence signals in SLR and explore the potential of skeleton data. Specifically, we propose a group-specific GCN to better exploit the knowledge of each signal and a complementary regularization to prevent complex co-adaptation across signals. Furthermore, we propose a two-stream framework that gradually fuses both static and dynamic information in skeleton data. Experimental results on three public datasets show that the proposed CoSign achieves competitive performance with recent video-based approaches while reducing the computation cost during training.

1.2. Weakly Supervised Sign Localization

To better understand the information conveyed in sign videos, we need to know not only “what” but also “where”

[2] `https://github.com/ycmin95/pointlstm-gesture-recognition-pytorch`

[3] `https://github.com/ycmin95/VAC_CSLR`

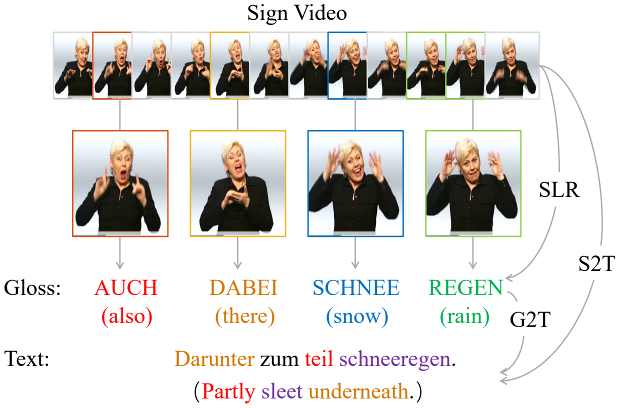


Figure 3. An example from Phoenix14T [1]. The goal of SLR is to recognize a gloss sequence, which is monotonically aligned with sign clips, from the sign video. S2T and G2T aim to translate sign videos and gloss sequences into spoken language sentences, and G2T is often regarded as the 'upper bound' of S2T.

the signs occurred. CTC loss is a popular objective function in sequence recognition tasks, which provides supervision for unsegmented sequence data through aligning sequence and its corresponding labeling iteratively. However, networks trained with CTC will conservatively predict a series of spikes, and they are hardly to provide accurate boundaries for occurred signs.

RadialCTC Controls the Peaky Behavior of CTC by Modifying the Logit of the Blank Class. To explore the alignment between gloss sequences and video clips, we revisit the iterative alignment of CTC. The blank class of CTC plays a crucial role in the alignment process and is often considered responsible for the peaky behavior of CTC. We propose an objective function named RadialCTC that constrains sequence features on a hypersphere while retaining the iterative alignment mechanism of CTC [14]. The learned features of each non-blank class are distributed on a radial arc from the center of the blank class, which provides a clear geometric interpretation and makes the alignment process more efficient. Besides, RadialCTC can control the peaky behavior by simply modifying the logit of the blank class. Experimental results of recognition and localization demonstrate the effectiveness of RadialCTC on both SLR and scene text recognition.

1.3. Utilization of Visual Signals in SLT

As a typical visual language, sign language has its own grammar and lexicon, which means that SLR results do not correspond to the linguistic habit of the hearing. SLT is an essential step to further bridge the communication gap between the Deaf and hearing people. Similar to SLR, SLT also faces the data scarcity problem, but it can leverage powerful large language models. A pioneering work [2] shows the potential of mBart [11] on SLT, which significantly improves the translation quality on public datasets but inevitably faces the hallucination problem [5]. In other words, the SLT models tend to generate fluent but inadequate translation, and we attribute this problem to the lack of faithfulness (*i.e.* the SLT models fail to capture correct visual signals). Our recent work focuses on the utilization of visual signals in SLT.

Alignment Constraints can Improve Faithfulness of SLT Networks. Gloss sequences play a critical role in both SLR and SLT. As shown in Fig. 3, the monotonous alignment between the gloss sequence and sign clips makes it possible to leverage CTC to provide supervision for SLR. On the other hand, Gloss sequences are widely used as the input of Gloss2Text (G2T) task to estimate the upper bound of Sign2Text (S2T) task. However, as a visual language, sign language conveys information through multiple visual signals and glosses are imprecise representations. We explore the association among different SLT subtasks and integrate them into a unified framework. We further propose two kinds of constraints: the alignment constraint aligns the visual and linguistic embeddings and the consistency constraint integrates the advantages of subtasks. Experimental results show that the proposed method is competitive against previous SLT methods by increasing the utilization of visual signals, especially when glosses are imprecise.

2. Ongoing and Future Directions

To summarize, my research in computer vision and pattern recognition focuses on algorithms that learn to understand sequential data with weak supervision, especially for sign videos, and by leveraging the alignment of spatial and temporal cues in videos. My previous efforts mainly focus on the perception of human behavior, especially through sign language, and my final goal is to build an AI system that has the capabilities of perception, reaction, and interaction. There is still a long way to go before reaching this goal, and below I outline four topics that I plan to pursue.

- **Pracitcal Sign Language Understanding System.** With the rapid development in recent years, a SLU model can produce accurate recognition ($< 20\%$ WER) and translation ($> 28\%$ BLEU-4) results on public datasets. However, there is still a large gap between sign language datasets and realistic scenarios. Current datasets are often collected under constrained conditions and it is worth further studying the performance of SLU models in wild scenarios with diverse signers and conversational signings. The first step is to build a data engine that can collect high-quality sign videos from the Internet automatically and then design a suitable benchmark to evaluate the generalization ability of SLU models. The final goal is to leverage the data engine to design a unified visual tokenization with minimal information loss and the corresponding large language model, which can integrate SLU into NLP systems seamlessly [21]. The whole process is challenging and collaboration with the Deaf communities is essential to make the final system practical.

- **Fine-grained Human-Object Interaction (HOI)**. Similar to SLU, hands play a dominant role in humans' interactions with objects. How to understand state changes caused by human manipulation is essential for HOI [4]. Recent hand-centric low-level tasks (*e.g*. hand pose estimation and HOI) and high-level tasks (*e.g*. gesture and sign language understanding) are nearly independent, but the final goal of both kinds of tasks is consistent: build an efficient human environment interaction system. Previous research shows that the ability of infants to imitate [7] is the emergent product of a system of social, cognitive, and motor components. I believe one potential way is to combine different supervision signals from both low levels (*e.g*., hand and object keypoints) and high levels (*e.g*., gesture and interaction type) with a unified framework. Intuitively, the high-level interaction is easier to predict when low-level predictions are accurate, and the low-level predictions are more reasonable with proper high-level guidance. Besides, the physical constraints are essential.

- **Fine-grained Human Motion Generation.** Generating human action from text is a challenging yet fascinating research goal. Recent work [22] focuses on body motion generation, but hand motion generation is a more challenging but essential task. According to my experience of SLU, the set of gestures is finite and unified across the world, it is essential to build a large language model for gesture understanding and generation. I believe the discrete nature of language is essential to the success of recent foundation models [19, 17], which provides a stable mapping for learning and greatly reduces the search space. Compared to images, skeleton data are more suitable to be tokenized as a universal representation for different downstream tasks, including both recognition and generation. How to achieve this goal is challenging yet fascinating.

- **Data Closed-loop Platform.** Compared to using expensive mocap data, the massive interactive data on the Internet can be leveraged in a self-supervised manner, which provides a good supplement to real-world data and improves the generalization and robustness of HOI approaches. With the rapid development of vision techniques [8, 10], it becomes feasible to build an automatic data annotation engine to collect data from the Internet. Moreover, it is feasible to build a data closed-loop platform if the first two topics go well. From the generation side, the skeleton sequence can be generated by first building an action-style embedding space, then sampling an action style, and finally generating the corresponding sequence. The generated sequences can be used to improve the generalization of basic models for HOI for the interaction side.

References

[1] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pages 7784–7793, 2018.

[2] Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *CVPR*, pages 5120–5130, 2022.

[3] Runpeng Cui, Hu Liu, and Changshui Zhang. A deep neural framework for continuous sign language recognition by iterative training. *TMM*, 21(7):1880–1891, 2019.

[4] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023.

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM CSUR*, 55(12):1–38, 2023.

[6] Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. Cosign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition. In *ICCV*, 2023.

[7] Susan S Jones. The development of imitation in infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528):2325–2335, 2009.

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

[9] Oscar Koller. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*, 2020.

[10] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *CVPR*, 2023.

[11] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742, 2020.

[12] Yuecong Min, Xiujuan Chai, Lei Zhao, and Xilin Chen. Flickernet: Adaptive 3d gesture recognition from sparse point clouds. In *BMVC*, volume 2, page 5, 2019.

[13] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *ICCV*, pages 11542–11551, 2021.

[14] Yuecong Min, Peiqi Jiao, Yanan Li, Xiaotao Wang, Lei Lei, Xiujuan Chai, and Xilin Chen. Deep radial embedding for vi-

sual sequence learning. In *ECCV*, pages 240–256. Springer, 2022.

[15] Yuecong Min, Yanxiao Zhang, Xiujuan Chai, and Xilin Chen. An efficient pointlstm for point clouds based gesture recognition. In *CVPR*, pages 5761–5770, 2020.

[16] Sylvie CW Ong and Surendra Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *T-PAMI*, 27(06):873–891, 2005.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022.

[18] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[20] Wendy Sandler and Diane Lillo-Martin. *Sign language and linguistic universals*. Cambridge University Press, 2006.

[21] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. In *ACL*, pages 7347–7360, 2021.

[22] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *CVPR*, 2023.

[23] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI*, volume 34, pages 13009–13016, 2020.