SignVTCL: Multi-Modal Continuous Sign Language Recognition Enhanced by Visual-Textual Contrastive Learning

BMVC 2024 Submission # 335

1 Extraction of Keypoints and Optical Flow.

Keypoints. We adopt the keypoints generated by HRNet [4] trained on COCO-WholeBody [1], which contains 42 hand keypoints, 26 face keypoints, and 11 upper body keypoints, to model the keypoint sequences. To mitigate the sensitivity to noise, we opted to represent the keypoints using heatmaps instead of treating them as a set of coordinate points. Specifically, each coordinate point will be processed by a Gaussian function and become part of the heatmap. We treat each extracted keypoint as an independent channel input to the model. Therefore, the keypoints sequences can be represented as $x^k = \{x_t^k\}_{t=1}^T \in \mathcal{R}^{T\times H\times W\times K}$, K is the keypoints number. The details of this process can be expressed as: $x^k_{(t,i,j,k)} = \exp(-[(i-x_t^k)^2 + (j-y_t^k)^2]/2\sigma^2)$, where (x_t^k, y_t^k) denotes the coordinates of the k-th keypoint in the t-th frame, and σ is a scale controller. In Figure 2, we present some examples of visualizations. For improved visual clarity, we have combined 79 (42+26+11) channels into a single channel.

Optical Flow. Optical flow provides a dense motion representation, which enables the model to capture finer details of the signing motion and facilitates accurate recognition. By leveraging the capabilities of RAFT [7], a deep network architecture explicitly crafted for extracting optical flow, we can proficiently acquire high-quality optical flow information from video data. We store the optical flow as a sequence of images. This format allows us to represent the optical flow information in a manner similar to video data, which can be represented as $x^o = \{x_t^o\}_{t=1}^T \in \mathcal{R}^{T\times H\times W\times 3}$. Some visualizations of optical flow can be seen in Figure 2.

2 Details of Different Fusion Methods.

In this section, we provide the explanation of implementation details for each fusion method. MLP-based fusion method is a simple yet effective method, which contains two linear layers, each layer followed by a GELU activation function except for the last one. The operation of the fusion module can be represented as follows:

$$f^m = \text{MLP}(f^v © f^k © f^o), \quad f^{v'} = f^m + f^v, \tag{1}$$

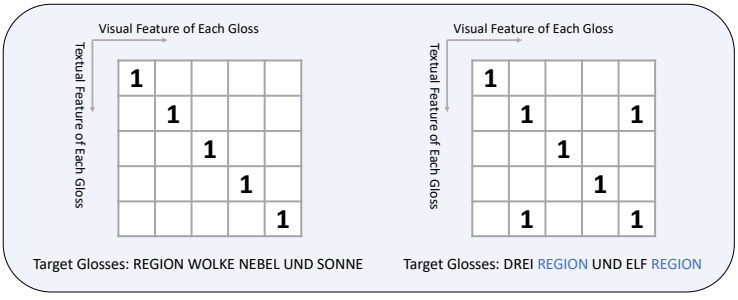

Figure 1: **Examples of Gloss-Level Alignment Ground Truth G.** In G, all positions are 0 except for those equal to 1.

$$f^{k'} = f^m + f^k, \quad f^{o'} = f^m + f^o, \tag{2}$$

where MLP represent the MLP function, f^m is the feature output from our fusion module, f^v, f^k, f^o are features input to the fusion module, and $f^{v'}$, $f^{k'}$, $f^{o'}$ are enhanced features input to the next stage. © denotes channel-wise concatenation. The same meanings of the symbols are below.

For the convolution-based fusion method, we describe the process as:

$$f^m = \text{Conv}(f^v © f^k © f^o), \quad f^{v'} = f^m + f^v, \tag{3}$$

$$f^{k'} = f^m + f^k, \quad f^{o'} = f^m + f^o, \tag{4}$$

where Conv is a $3 \times 3 \times 3$ convolutional layer.

In terms of the attention-based fusion method, the operation of the fusion module can be represented as:

$$f^{v'} = f^v + \text{Attn}(f^v, f^k) + \text{Attn}(f^v, f^o), \tag{5}$$

$$f^{k'} = f^k + \text{Attn}(f^k, f^v) + \text{Attn}(f^k, f^o), \tag{6}$$

$$f^{o'} = f^o + \text{Attn}(f^o, f^v) + \text{Attn}(f^o, f^k), \tag{7}$$

where Attn represents the cross-attention function.

3 Explanation of Gloss-Level Alignment Ground Truth G.

Once the visual representations and their corresponding glosses are identified, we can automatically generate G, which is a square matrix. More specifically, when there are no repeated glosses in the sentence, only the diagonal value of this square matrix is 1, and the value of other positions is 0. When there are repeated glosses, there will be values of 1 in other positions, as shown in Figure 1. Each position corresponds to a ground truth value of the similarity between the visual feature and the textual feature. These two features might originate from the same gloss (positive sample) or from different glosses (negative samples).

Datasets	V	K	O	S1	S2	S3	MLP	AlignG	AlignS	Dev (%)	Test (%)
Phoenix-2014T	✓									20.3	21.0
		✓								25.7	25.9
			✓							37.9	37.4
	✓	✓	✓	✓	✓	✓	✓			18.0	18.6
	✓	✓	✓	✓	✓	✓	✓	✓	✓	**16.7**	**17.7**
CSL-Daily	✓									28.0	27.7
		✓								34.4	33.3
			✓							36.9	35.8
	✓	✓	✓	✓	✓	✓	✓			25.3	24.9
	✓	✓	✓	✓	✓	✓	✓	✓	✓	**24.2**	**24.0**

Table 1: **Ablation study of each component of SignVTCL on Phoenix-2014T and CSL-Daily.** 'V', 'K', 'O' are video, keypoint and optical flow branches, respectively. 'S1' ∼ 'S3' represent three fusion stages in three-branch network. 'MLP' means MLP-based fusion method. 'AlignG' and 'AlignS' denote gloss-level and sentence-level alignments.

4 More Ablation Studies and Visualizations.

We performed individual testing of each branch (using one branch implies utilizing data from only one modality) of our SignVTCL on the Phoenix-2014T and CSL-Daily datasets, as shown in Table 1. Simultaneously, the effectiveness of the fusion stages and method were also validated. After incorporating our visual-textual alignment approach, the SLR error rates can be further significantly reduced. To verify that our gloss-level alignment plays a pivotal role in efficiently aligning visual and textual features, we show more similarity matrices and predicted glosses in Figure 3.

5 Implementation Details.

In both the Phoenix-2014 and Phoenix-2014T datasets, data from three modalities are resized and cropped to dimensions of 224 × 224, while in the CSL-Daily dataset, a crop size of 320 × 320 is applied. During the training phase, data augmentations are applied, consisting of spatial cropping within the range of [0.7-1.0] and frame-rate augmentation spanning [×0.5-×1.5]. For our network training strategy, we implement a cosine annealing schedule spanning 60 epochs. We utilize the Adam optimizer with a weight decay of $1e^{-3}$ and set the initial learning rate to $1e^{-3}$. We train our models on 4 Nvidia A100 GPUs.

References

[1] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020.

[2] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[3] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

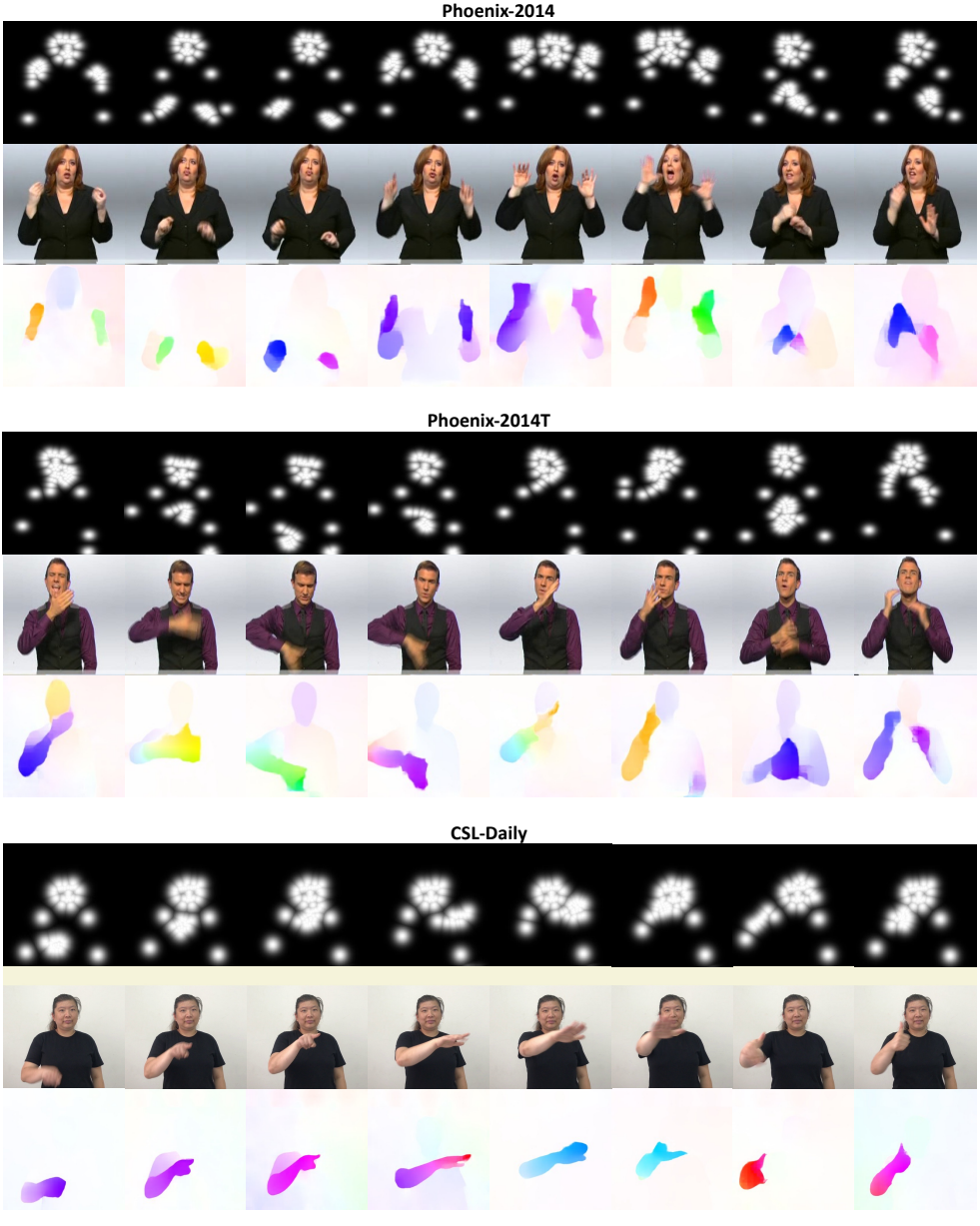

Figure 2: **Visualization of Three Input Modalities on the Phoenix-2014, Phoenix-2014T, and CSL-Daily Datasets.** For each example, the first line is keypoints, the second line is video, and the third line is optical flow.

Figure 3: **Visualization of Gloss-Level Alignment Similarity Matrix and Predicted Glosses on the Phoenix-2014T, Phoenix-2014, and CSL-Daily Datasets.**