SELF-SUFFICIENT FRAMEWORK FOR CONTINUOUS SIGN LANGUAGE RECOGNITION

Youngjoon Jang\textsuperscript{1}, Youngtaek Oh\textsuperscript{1}, Jae Won Cho\textsuperscript{1}, Myungchul Kim\textsuperscript{1}, Dong-Jin Kim\textsuperscript{2}, In So Kweon\textsuperscript{1,3}, Joon Son Chung\textsuperscript{1,4}\textsuperscript{†}

\textsuperscript{1}Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
\textsuperscript{2}Hanyang University, Seoul, Republic of Korea

Project page with demo: https://mm.kaist.ac.kr/projects/ssslr

ABSTRACT

The goal of this work is to develop self-sufficient framework for Continuous Sign Language Recognition (CSLR) that addresses key issues of sign language recognition. These include the need for complex multi-scale features such as hands, face, and mouth for understanding, and absence of frame-level annotations. To this end, we propose (1) Divide and Focus Convolution (DFConv) which extracts both manual and non-manual features without the need for additional networks or annotations, and (2) Dense Pseudo-Label Refinement (DPLR) which propagates non-spiky frame-level pseudo-labels by combining the ground truth gloss sequence labels with the predicted sequence. We demonstrate that our model achieves state-of-the-art performance among RGB-based methods on large-scale CSLR benchmarks, PHOENIX-2014 and PHOENIX-2014-T, while showing comparable results with better efficiency when compared to other approaches that use multi-modality or extra annotations.

1. INTRODUCTION

The Continuous Sign Language Recognition (CSLR) task aims to recognize a gloss sequence in a sign language video [2, 3, 4]. To capture the meaning of the sign expressions from a signer, recent works obtain manual and non-manual expressions by fusing RGB with other modalities such as depth [5], infrared maps [6] and optical flow [7], or by explicitly extracting multi-cue features [3, 8, 9, 10] or human keypoints [11] using off-the-shelf detectors. However, using such extra components introduce bottlenecks in both training and inference processes. In addition, most CSLR datasets only have sentence-level gloss labels without frame- or gloss-level labels [3, 12, 13]. To overcome insufficient annotations, the Connectionist Temporal Classification (CTC) [14] loss has been traditionally opted to consider all possible underlying alignments between the input and target sequence. However, using the CTC loss without true frame-level supervision produces temporally spiky attention which can make the model fail to localize important temporal segments [15].

Accordingly, we develop self-sufficient framework for CSLR, which provides meaningful gloss supervision while capturing helpful multi-cue information without additional modalities or annotations. To this end, we propose two novel methods: Divide and Focus Convolution (DFConv) and Dense Pseudo-Label Refinement (DPLR). DFConv is a task-aware convolutional layer which extracts visual multi-cue features by dividing spatial regions to focus on partially specialized features. Note that DFConv is designed to leverage prior knowledge about the structure of human bodies without any additional networks or modalities. In addition, DPLR elaborately refines an initially predicted gloss sequence from the model by referring a ground-truth gloss, and propagates frame-level gloss supervision without additional networks, unlike [7, 16]. We emphasize that DPLR is generally applicable to other CSLR architectures or frameworks [2, 15] to bring performance gain by reducing missing glosses in predictions.

We extensively validate the effectiveness of DFConv and DPLR. We also show that the whole self-sufficient counterpart achieves state-of-the-art results among RGB-based methods and is comparable to other methods that use extra knowledge with better efficiency on two publicly available CSLR benchmarks [12, 13]. To summarize, our main contributions are as follows:

(1) We design a task-specific convolutional layer, named DFConv, that efficiently extracts non-manual and manual features without additional networks or annotations. (2) We also introduce DPLR, a novel pseudo-label generation method, to propagate frame-level supervision by using the combination of the ground truth gloss sequence and the predicted temporal segmentation information. (3) We conduct extensive experiments on two publicly available CSLR benchmarks, showing state-of-the-art performance compared to other RGB-based methods, and competitive results compared to other approaches that use multi-modality or additional knowledge with better efficiency.

2. RELATED WORKS

Multi-cue fusion methods for CSLR task can be categorized into multi-semantic and multi-modal methods. Multi-semantic works [4, 13, 9, 10] utilized hand-crafted or weak-labeled features such as detected hands, trajectories of hands, and body parts, then integrate these features into frames to predict the gloss sequences. On the other hand, multi-modal works [6, 5, 17] use color, depth, and optical flow to extract orthogonal features. [7] proposed a multi-modality integration framework of appearance and motion cues by using both RGB frames and optical flow. Most recently, [11] fused human body keypoints extracted by an off-the-shelf network [18].

\textsuperscript{†}Corresponding authors
\textsuperscript{‡}Glosses are the smallest units having independent meaning in sign language [1].
3. METHOD

CSLR task aims to map a given input video to its corresponding gloss sequence $g = \{g_i\}^N_{i=1}$ with $N$ glosses. As shown in Fig. 2, a sign video is fed into the spatial modeling module consisting of several Divide and Focus Convolution (DFConv) layers, and a multi-cue embedding layer to extract manual and non-manual features. The multi-cue features of all the frames are passed through the temporal modeling module, that is comprised of Bottleneck-based Temporal Attention (BTA), which captures more important information among adjacent frames, and Temporal Multi-Cue (TMC) blocks of [11]. Then, the output of last TMC block is passed through the sequence learning stage, which is composed of a Bi-LSTM layer [26] and FC layer to predict the gloss sequence from the final model output. Finally, the Dense Pseudo-Label Refinement (DPLR) module is introduced to effectively train the latent representations by generating corrected and densified frame-level pseudo-labels.

3.1. Divide and Focus Convolution

We observe from various CSLR datasets [12, 13] that non-manual expressions occur frequently in the upper region of the image, while manual expressions occur mainly in the lower region. As shown in Fig. 1, despite the importance of both non-manual and manual expressions appearing in the entire image area, the conventional 2D convolution layer tends to capture the only one most dominant information (i.e., right hand) over the whole image. To address this issue, we propose a novel Divide and Focus Convolution (DFConv) layer designed to independently capture non-manual features and manual features solely from RGB modality. Unlike other methods that leverage external knowledge, we only design DFConv that captures both manual and non-manual expressions from RGB video, without relying on any additional hand-crafted features or multi-modal data.

In addition, the CSLR task [3, 8, 16, 19] naturally corresponds to a weakly-supervised learning problem due to the lack of frame-level gloss annotations. The challenge lies in the ambiguous semantic boundary of the adjacent glosses from sign videos [3, 13, 20]. To address this issue, some works in CSLR field generate frame-level pseudo-labels from sparse gloss annotations [4, 21], which can be inherently noisy and reliant on the model’s performance. Most recently, the CTC loss [14] is employed to facilitate end-to-end training of a deep learning model [22, 23], and consider all the feasible underlying alignments between the predictions and labels. However, as observed in [15, 24], directly optimizing CTC can cause spiky attention in predictions, favoring more blank glosses. Recent works tackle this issue by balancing the blank output and meaningful glosses [2], and by directly supervising the visual features via visual alignment constraint [15] and mutual knowledge transfer [25]. In contrast, we propose Dense Pseudo-Label Refinement (DPLR) that provides dense and reliable supervision signals obtained by gloss predictions of the model to visual features.

3.2. Dense Pseudo-Label Refinement

Most existing sign language datasets do not have temporally localized gloss labels [3, 12, 13, 28]. Due to the characteristics of the CTC loss used in training CSLR models without frame-level labels, the output sequence predictions of models are naturally induced to be sparse. As

Fig. 2. Overall architecture. In Spatial Modeling, non-manual and manual features are extracted through DFConv followed by a Multi-Cue embedding layer. In Temporal Modeling, temporal features are extracted for gloss sequence prediction by integrating each element. Finally, the gloss probability vectors are obtained through sequence learning.
a result, it is difficult for CSLR models to receive direct and precise alignment supervision for each gloss token. In addition, without alignment supervision, CSLR models learn entire sequences as a whole instead of individual gloss words. This limits the robustness of models severely as they rely on entire sequences. In other words, models can easily confuse similar sequences with slightly different words. In order to mitigate these drawbacks, we introduce an additional training objective called Dense Pseudo-Label Refinement (DPLR) that uses the alignment information predicted by the model to generate Dense Pseudo-Labels (DPL). Then, the model is further refined with these generated pseudo-labels.

In DPLR process, we have two separate cases for generating DPL $\tilde{Q}$ as illustrated in Fig. 3. We first compare the sequence length of non-blank predictions of the model with its corresponding ground truth gloss sequence. If the sequence length is matched, we go to Case 1, where we compare the predicted gloss sequence with the ground truth sequence. If a predicted gloss is wrong, we swap in correct gloss from the ground truth to increase the reliability of the pseudo-labels. As mentioned before, the predictions are sparse due to the nature of the CTC loss, and most of the predictions along the temporal axis are blanks. Here, we create DPL by filling each blank with the nearest predicted glosses. In the case where the predicted sequence length differs from the ground truth by one gloss length, we go to Case 2. Then, we simply densify the pseudo-label using the nearest gloss without swapping any glosses regardless of the correctness of the glosses. In the case that the sequence length differs by more than one gloss, we disregard that sequence as this might cause predictions of the model to degrade, so we do not propagate refinement loss $L_{\text{refine}}$.

Using pseudo-labels only from Case 1 and Case 2, we define the model with Cross Entropy (CE) loss on the latent features similar to [2] as follows:

$$L_{\text{refine}} = CE(\tilde{Q}, \hat{Q}),$$

(1)

where $\tilde{Q}$ is dense pseudo-labels and $\hat{Q}$ is gloss probability acquired from latent features, which is the final output of the inter-cue path (See Fig. 2).

Note that we demonstrate the efficacy of ‘Densify’ and ‘Refine’ processes in Table 4, and show that DPLR is generalizable to other models in our project page.

In addition, the quality of pseudo-labels generated from the model depend heavily on the model’s performance. As the CSLR task aims to translate a sign language video into a gloss sequence by mapping several adjacent frames into one gloss, it is important to extract key frames in the video. Hence, we design the Bottleneck-based Temporal Attention (BTA) module to attend to the temporally salient frames among adjacent frames. BTA consists of a temporal-wise attention map using 1D convolution layers and a max pooling layer to capture the temporally salient frames. The CTC loss is then propagated to the bottleneck, after the max pooled features, hence the name is Bottleneck.

With our additional modules, our final loss function is as follows:

$$L_{\text{total}} = L_{\text{inter}} + \lambda_1 L_{\text{intra}} + \lambda_2 L_{\text{refine}} + \lambda_3 L_{\text{data}},$$

(2)

where, $L_{\text{inter}}, L_{\text{intra}}$ and $L_{\text{data}}$ are all CTC losses. $L_{\text{data}}$ is the average of all the TMC block’s CTC losses and $\lambda_1, \lambda_2, \lambda_3$ are loss weights.

4. EXPERIMENTS

Dataset and Evaluation Metric. We conduct experiments on two publicly available CSLR benchmarks to validate our self-sufficient framework: PHOENIX-2014 [13] and PHOENIX-2014-T [12]. We adopt the Word Error Rate (WER) [13] for evaluation. Furthermore, in our project page, we upload a demo video to visually demonstrate the effectiveness of DFCNN.

<table>
<thead>
<tr>
<th>Method</th>
<th>Extra Annotations</th>
<th>WER (%)</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepHand [19]</td>
<td>Hand</td>
<td>47.1</td>
<td>45.1</td>
<td></td>
</tr>
<tr>
<td>SubNet [3]</td>
<td>Hand</td>
<td>40.8</td>
<td>40.7</td>
<td></td>
</tr>
<tr>
<td>DeepSign [29]</td>
<td>Hand</td>
<td>33.3</td>
<td>38.8</td>
<td></td>
</tr>
<tr>
<td>SignOpt [22]</td>
<td>Hand</td>
<td>39.4</td>
<td>38.7</td>
<td></td>
</tr>
<tr>
<td>LS-HAN [31]</td>
<td>Hand</td>
<td>38.3</td>
<td>38.3</td>
<td></td>
</tr>
<tr>
<td>SF-Net [30]</td>
<td>35.6</td>
<td>34.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPF+TEM [31]</td>
<td>35.6</td>
<td>34.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cm-bilstm-hmm [4]</td>
<td>27.5</td>
<td>28.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Re-sign [23]</td>
<td>26.8</td>
<td>26.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNF [7]</td>
<td>23.8</td>
<td>24.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>-</td>
<td>20.9</td>
<td>20.8</td>
<td></td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Method</th>
<th>Extra Annotations</th>
<th>WER (%)</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMKD [15]</td>
<td>Mouth+Hand</td>
<td>20.8</td>
<td>22.4</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>-</td>
<td>20.3</td>
<td>22.3</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison of performance in WER (%) on PHOENIX-2014-T benchmark. Ours achieves the state-of-the-art performances among RGB-based approaches, while shows comparable performances with the pose-based multi-cue method [11].

4.1. Experimental Results

We compare our framework with recent CSLR methods on both PHOENIX-2014 [13] and PHOENIX-2014-T [12] benchmarks. Tables 1 and 2 show the WER scores, while we specify the type of either extra annotations or modalities used during training for each method.

PHOENIX-2014. Table 1 summarizes the results on Dev and Test splits from PHOENIX-2014 for several CSLR baselines. First, Ours achieves the state-of-the-art performances on Test split among RGB-based approaches. In particular, Ours outperforms the recently proposed FCN [2], fine-grained labeling [23], VAC [15] with alignment supervision to visual features, and CMA [33] with both gloss and video augmentation. Moreover, Ours shows superior performance over several recent methods that explicitly require extra annotations for training [3, 4, 7], and comparable performances to SMKD [25] with algorithmic gloss segmentations and STMC [11] using pose annotations. Note that the proposed method does not require either extra annotations for acquiring the model to detect spatially important regions or additional networks for the refinement of pseudo-labels.

PHOENIX-2014-T. Table 2 shows the results on Dev and Test splits of PHOENIX-2014-T. Ours surpasses cm-bilstm-hmm [4] which is trained with both mouth and hand annotations, and even outperforms SLRT [34] that jointly learns sign recognition and translation task from both sign glosses and sentences. Ours also outperforms SMKD [25], a competing baseline using RGB modality, and shows comparable results to STMC [11].

$$\text{WER} = \frac{\text{(#substitutions} + \text{#deletions} + \text{#insertions})}{\text{(#words in reference)}}$$
When the model predicts correct the number of glosses.

When the model fails to predict correct the number of glosses.

Table 3. Ablation study of DFConv, DPLR, and BTA. All the proposed components of our method gradually improve the performance.

Table 4. Ablation study on the design choice of DPLR. Both ‘Densify’ and ‘Refine’ processes are key in improving performance.

Table 5. Robustness comparison with state-of-the-art methods in simulated real world scenario. We compare the WER on a model that has been trained on a train set without these transformations. Ours denotes a model with r set to 0.35, and Ours1 denotes a model where r is moved along with the transformations during inference (T: vertical translation, S: scale). We note that STMC is our reproduction. We reimplement it as faithfully as possible.

Fig. 4. Gloss predictions in a single sentence sign video from different network architectures (D: deletion, I: insertion, S: substitution). In the fourth and sixth rows, we further visualize two cases of Dense Pseudo-Labels (DPL). Applying the DPLR on the prediction greatly reduces the deletion phenomenon.

Fig. 5. Activation maps from Ours on test-time novel transformations.

5. CONCLUSION

In this paper, we propose two novel methods, DFConv and DPLR, that complement missing annotations in the existing weakly-labeled sign language datasets. To the best of our knowledge, we are the first to propose a method to extract manual and non-manual features individually by designing a task-specific convolution without any additional networks or annotations. In addition, we introduce DPLR module that does not require additional networks during the pseudo-labeling process and demonstrate its effectiveness through various experiments. The experimental results show that our framework achieves state-of-the-art performance on two large-scale benchmarks among RGB-based methods, and also outperforms or is comparable to methods based on multi-modality.