# Iterative Alignment Network for Continuous Sign Language Recognition

Junfu Pu  Wengang Zhou  Houqiang Li
University of Science and Technology of China

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

## Introduction

**Brief to SLR**

Continuous sign language recognition (SLR) is a kind of weakly supervised sequence learning task, without rigid annotation of text words to video clips for a complete sign video. The key idea for continuous SLR is to learn the mapping between a sign video and its corresponding annotation of text sentence.

**Contribution**

- A unified deep learning architecture integrating encoder-decoder network and connectionist temporal classification (CTC) for continuous SLR.
- A soft dynamic time warping (soft-DTW) alignment constraint between the LSTM CTC decoders, which indicates the temporal segmentation in sign videos.
- Iterative optimization strategy to train feature extractor and encoder-decoder network alternately with alignment proposals by warping path.

## Iterative Optimization

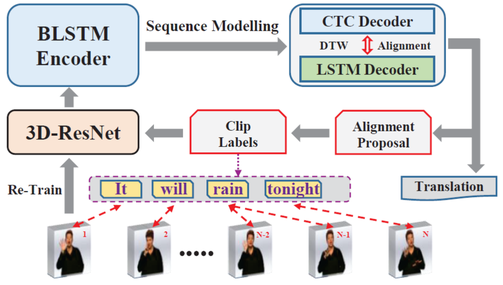
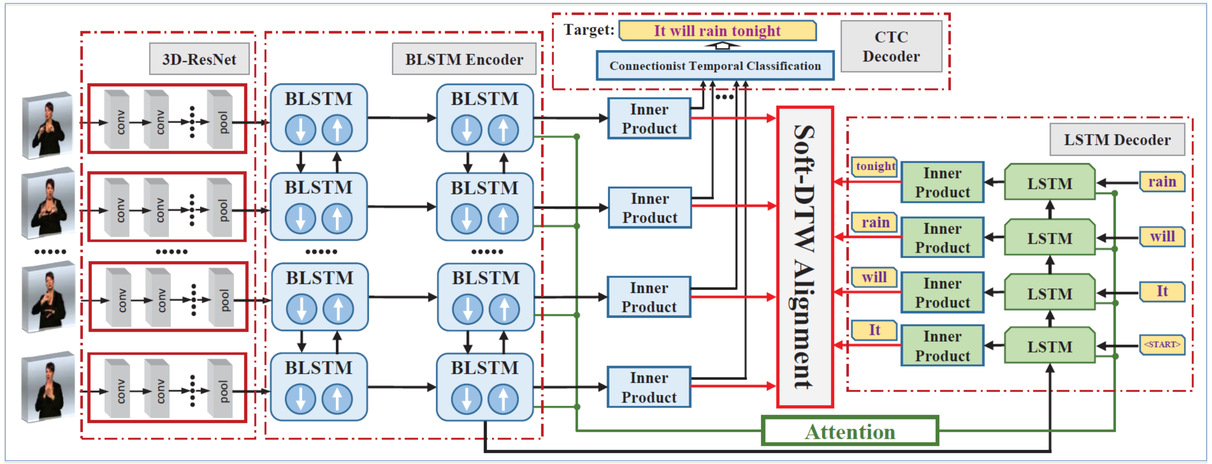



Illustration of iterative training algorithm

- Representative features contribute to good performance.
- When training the network in an end-to-end way, the objective loss has limited contribution to the learning of parameters for low layers of feature extractor due to the chain rules of BP.
- We use soft-DTW alignment proposals as pseudo-labels to learn representative 3D-CNN features, and optimize feature extractor and sequence learning module in EM-like iterations.

1. Extract features from 3D-ResNet for sign videos.
2. Train the encoder-decoder network with CTC by minimizing the total loss $\mathcal{L}$.
3. Get the warping path between the input clips and words provided by soft-DTW.
4. Fine-tune 3D-ResNet with the alignment as supervision.
5. Repeat 1-4. (Optimizing both modules in an EM-like way)

## Framework




### Modules

- Video Representation: $\mathbf{F}^N = (f_1, \cdots, f_N) = \{\mathcal{F}_\Theta(v_t)\}_{t=1}^N$
- Temporal Encoder: $\mathbf{E} = \{e_t\}_{t=1}^N = \mathcal{R}(\{f_t\}_{t=1}^N)$
- LSTM Decoder: $d_k = \mathbf{Decoder}_{lstm}(c_k, s_k, h_{k-1}^d)$

  $z_k = W_{fc2} \cdot d_k + b_{fc2} \quad \boldsymbol{Z} = (Z_{k,l}) = [z_1, z_2, \cdots, z_L]^T$
- CTC Decoder: $y_t = W_{fc1} \cdot e_t + b_{fc1} \quad \boldsymbol{Y} = (Y_{t,l}) = [y_1, y_2, \cdots, y_N]^T$

  $p(\pi|\boldsymbol{V}) = \prod_{t=1}^N p(\pi_t|v_t) = \prod_{t=1}^N Y_{t,\pi_t} \qquad p_{ctc}(\boldsymbol{s}|\boldsymbol{V}) = \sum_{\pi \in \mathcal{M}^{-1}(\boldsymbol{s})} p(\pi|\boldsymbol{V})$

### Optimization and Decoding

- Soft-DTW: $D_{i,j} = d_{i,j} + \min(D_{i-1,j}, D_{i,j-1}, D_{i-1,j-1})$

  $d_{i,j} = ||u_i - v_j||_2 \qquad \min_i^\gamma\{a_i\} = \begin{cases} \min_i\{a_i\}, & \gamma = 0 \\ -\gamma \log \sum_i e^{-a_i/\gamma}, & \gamma > 0 \end{cases}$
- Objective Function:

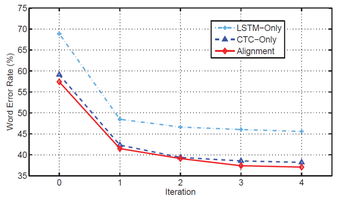
  $\mathcal{L} = \lambda\mathcal{L}_{ctc} + (1-\lambda)\mathcal{L}_{lstm} + \mathcal{L}_{align} + \mu||\omega||^2$

  $\mathcal{L}_{ctc} = -\ln p_{ctc}(\boldsymbol{s}|\boldsymbol{V}) \qquad \mathcal{L}_{lstm} = -\ln p_{lstm}(\boldsymbol{s}|\boldsymbol{V}) \qquad \mathcal{L}_{align} = \mathcal{D}_p(\boldsymbol{Y}, \boldsymbol{Z})$
- Decoding:

  $r(\boldsymbol{s}^i) = \alpha \ln p_{ctc}(\boldsymbol{s}^i|\boldsymbol{V}) + (1-\alpha) \ln p_{lstm}(\boldsymbol{s}^i|\boldsymbol{V}) + \beta \ln L_i$

## Experiments

- RWTH-Phoenix-Weather (German Sign Language):




| Methods | Dev (%) del / ins | WER | Test (%) del / ins | WER |
|---|---|---|---|---|
| 1-Mio-Hands [25, 27] | 16.3 / 4.6 | 47.1 | 15.2 / 4.6 | 45.1 |
| CNN-Hybrid [28] | 12.6 / 5.1 | 38.3 | 11.1 / 5.7 | 38.8 |
| SubUNet [3] | 14.6 / 4.0 | 40.8 | 14.3 / 4.0 | 40.7 |
| Staged-Opt [10] | 13.7 / 7.3 | 39.4 | 12.2 / 7.5 | 38.7 |
| CTF [39] | 12.8 / 5.2 | 37.9 | 11.9 / 5.6 | 37.8 |
| Dilated-SLR [32] | 8.3 / 4.8 | 38.0 | 7.6 / 4.8 | 37.3 |
| LS-HAN [23] | - | - | - | 38.3 |
| Ours (LSTM) | 13.8 / 3.3 | 45.6 | 13.6 / 3.3 | 46.1 |
| Ours (CTC) | 11.4 / 3.8 | 38.2 | 11.9 / 3.5 | 37.9 |
| Ours (Align-end2end) | 12.6 / 2.2 | 69.1 | 22.0 / 2.6 | 69.3 |
| **Ours (Align-iOpt)** | 12.9 / 2.6 | **37.1** | 13.0 / 2.5 | **36.7** |

**Alignment mechanism** **Performance comparison**

**Alignment between video and text annotation**

- CSL-100 (Chinese Sign Language):

| Method | Split I Precision | BLEU-1 | CIDEr | ROUGE-L | METEOR | Split II *Acc-w* | BLEU-1 | CIDEr | ROUGE-L | METEOR | WER |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LSTM&CTC [12, 20] | 0.858 | 0.936 | 8.632 | 0.940 | 0.646 | 0.332 | 0.343 | 0.241 | 0.362 | 0.111 | 0.757 |
| S2VT [38] | 0.897 | 0.902 | 8.512 | 0.904 | 0.642 | 0.457 | 0.466 | 0.479 | 0.461 | 0.189 | 0.670 |
| S2VT (3-layer) [38] | 0.903 | 0.911 | 8.592 | 0.911 | 0.648 | 0.461 | 0.475 | 0.477 | 0.465 | 0.186 | 0.652 |
| HLSTM (SYS sampling) [15] | 0.910 | 0.935 | 8.907 | 0.938 | 0.683 | 0.459 | 0.463 | 0.476 | 0.462 | 0.173 | 0.630 |
| HLSTM [15] | 0.924 | 0.942 | 9.019 | 0.944 | 0.699 | 0.482 | 0.487 | 0.561 | 0.481 | 0.193 | 0.662 |
| HLSTM-attn [15] | 0.929 | 0.948 | 9.084 | 0.951 | 0.703 | 0.506 | 0.508 | 0.605 | 0.503 | 0.205 | 0.641 |
| **Ours** | **0.939** | **0.980** | **9.342** | **0.981** | **0.713** | **0.670** | **0.724** | **3.946** | **0.716** | **0.383** | **0.327** |

**Performance comparison**

## Conclusions

We propose a new deep architecture based on 3D-ResNet and encoder-decoder network with connectionist temporal classification by iterative optimization for continuous SLR.

- Jointly train encoder-decoder network by minimizing CTC loss and cross-entropy loss, with a soft-DTW alignment constraint.
- An iterative training strategy to optimize the 3D-ResNet and sequence learning module based on the warping path.
- Experimental results demonstrate the superiority of our method.