IJCAI 2019

合肥工業大學
HEFEI UNIVERSITY OF TECHNOLOGY

Connectionist Temporal Modeling of Video and Language: a Joint Model for Translation and Sign Labeling

Dan Guo, Shengeng Tang and Meng Wang

Lab of Media Computing
Hefei University of Technology

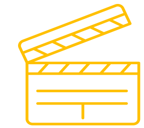
- **Introduction**
- **Method**
- **Experiments**
- **Conclusion**

What is Continuous Sign Language Translation (CSLT)?

Gesture Recognition

Action Detection

Behavior Analysis



Video Understanding

Isolated Sign Language Recognition (ISLR) vs. CSLT

Word:
100

Phrase:
2-3-9-7

Short sentence:
A fly in the soup.

Long sentence:
Even though my mother passed away years ago, I still feel her presence in this home.

Video Classification ⇨ **Video Understanding**

What are the challenges in CSLT?

1 **Potential Semantics**

- Visual hints under sign linguistics are latent and obscure.

2 **Hybrid Tasks**

- CSLT involves hybrid semantics learning under vision understanding, sign recognition, and natural language translation.

3 **Weak Supervision**

- Sign videos have sentence-level annotations, rather than the exact temporal location of each sign action.

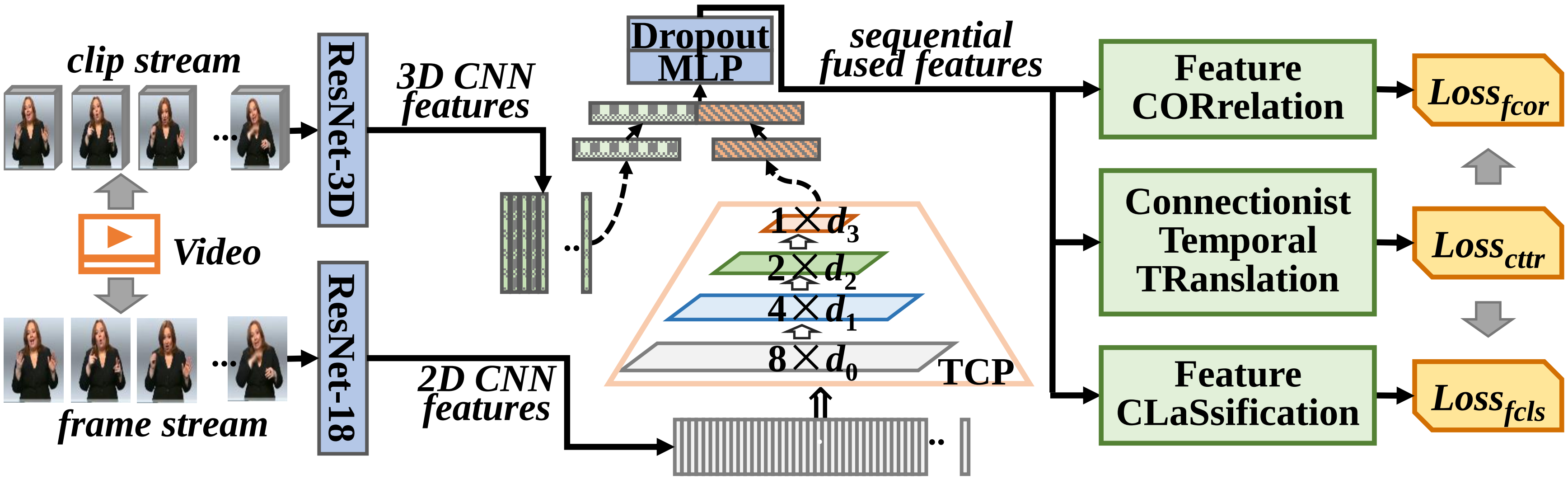
Existing methods and their drawbacks

- **Traditional Temporal Models**: Dynamic Time Warping (DTW) and Hidden Markov Models (HMM).
  - **A lot of time** is spent on training the network.

- **Encoder-decoder Framework**
  - These proposed methods decoded word by word after encoding all visual content. They **do not apply to online CSLT**.

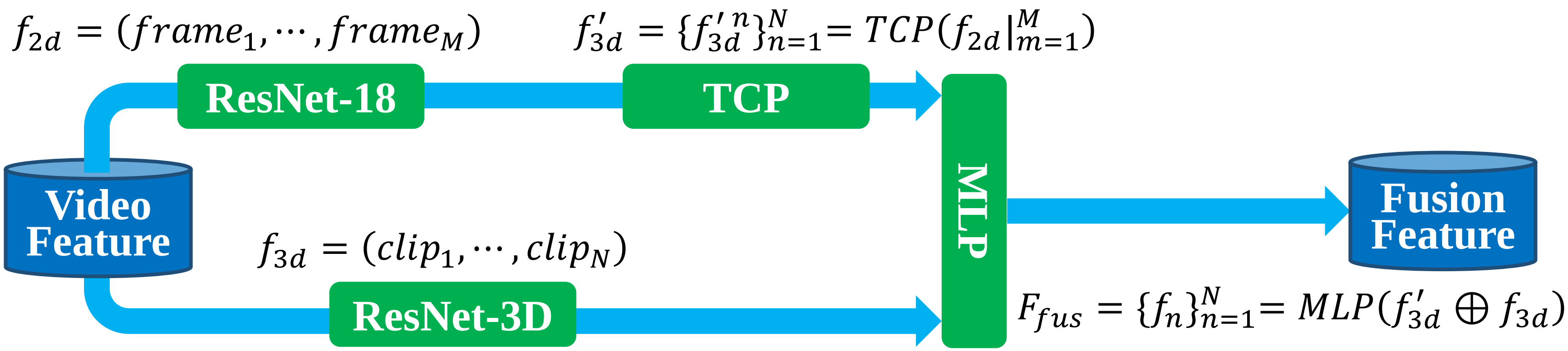
- **Hybrid Models + Offline Optimization**: CNN+RNN+EM
  - Offline iteration **takes a lot of time**, and it is often trained repeatedly **using fixed datasets**, which is not applicable to dataset extension.

Overview: Connectionist Temporal Modeling network

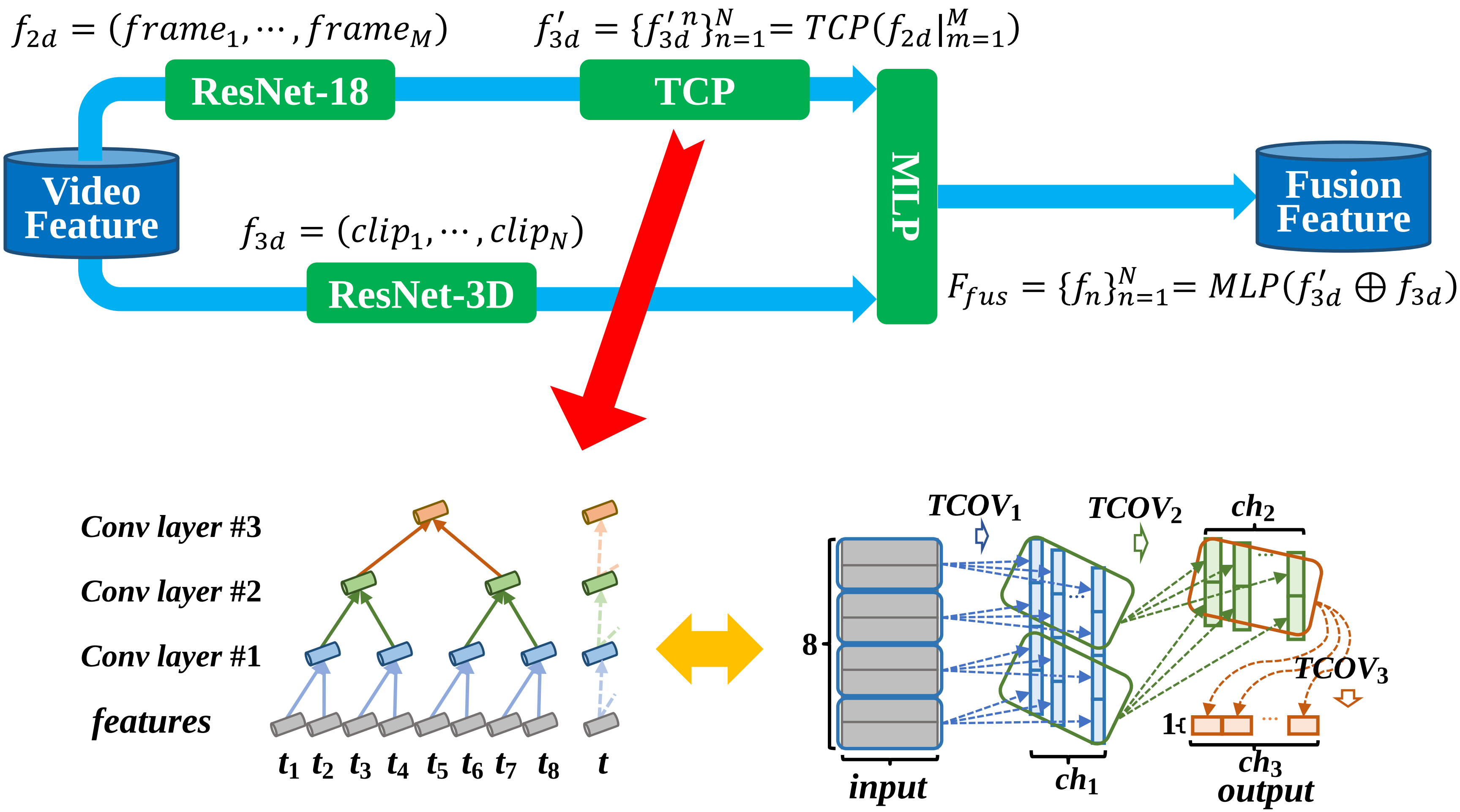


- **Feature Extraction:** 2D frame-level features, 3D clip-level features
- **Temporal Alignment & Fusion:** The TCP module is used to learn the short-term temporal correlation in the 2D features and align them with the 3D features.
- **Joint Loss Optimization:** $Loss_{fcor}$, $Loss_{cttr}$ and $Loss_{fcls}$ is designed to measure feature correlation, sentence decoding, and entropy regularization on sign labeling.

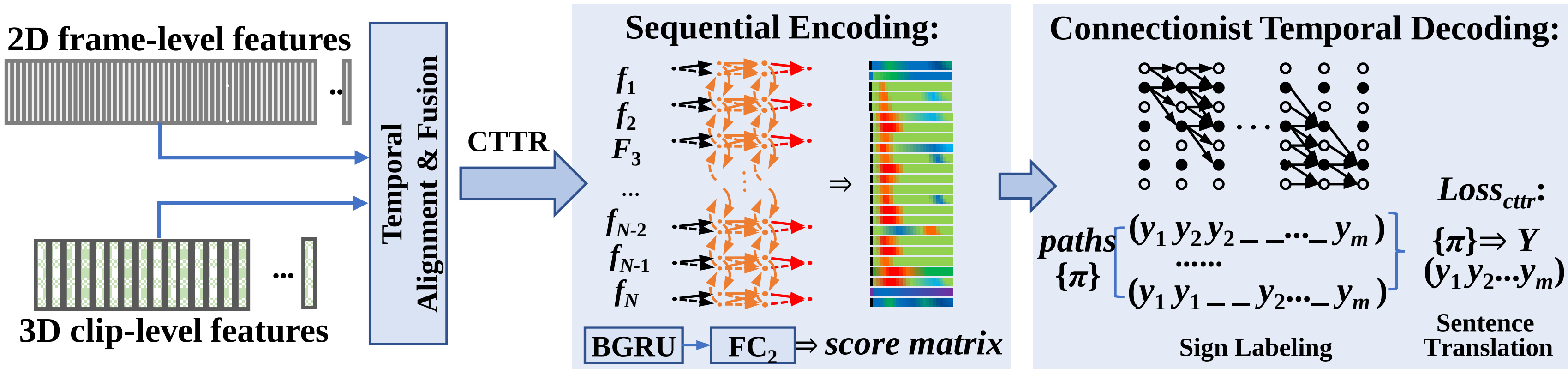
Clip Feature Learning (Temporal Convolution Pyramid)



Clip Feature Learning (Temporal Convolution Pyramid)



Connectionist Temporal TRanslation



Decoding Example: **I _ I have _ a a _ pencil→I I have a a pencil→I have a pencil**

- **Temporal Encoding:**

$$H = \{h_n\}_{n=1}^{N} = \{BGRU(f_n)\}_{n=1}^{N} \qquad P = \{p_n\}_{n=1}^{N} = \varphi_{softmax}\{FC(h_n|_{n=1}^{N})\}$$

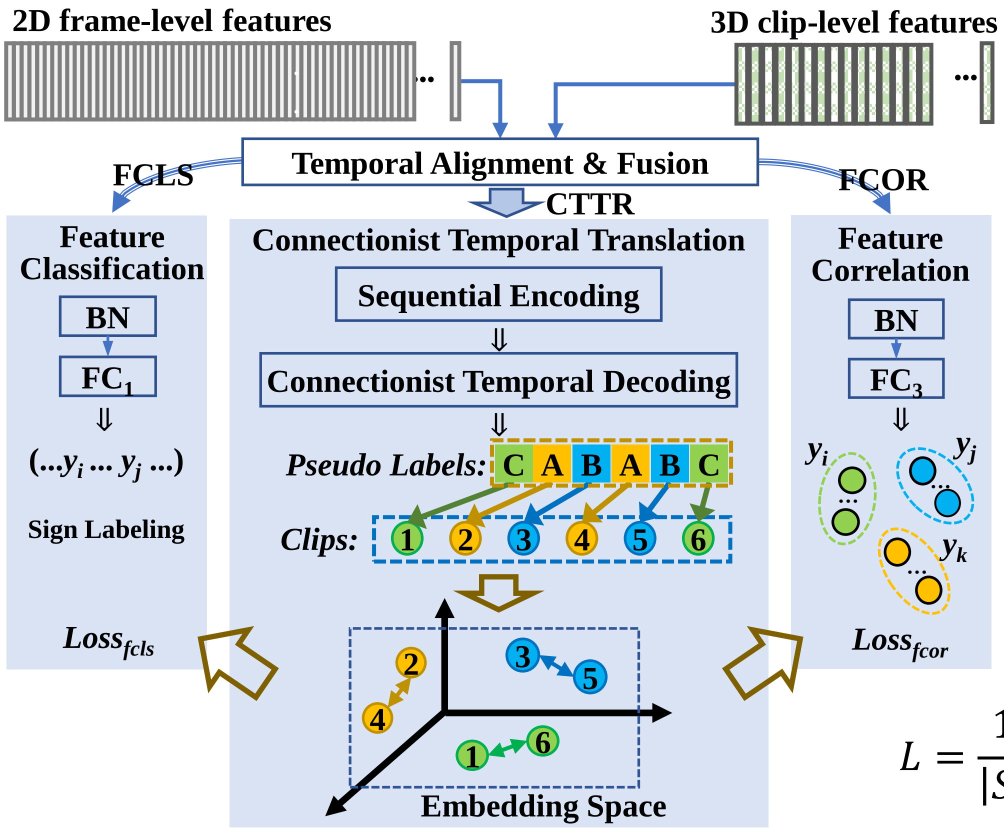
- **Decoding Optimization:**

$$P^{\pi} = Prob(\pi) = \prod_{n=1}^{N} p_n^{\pi_n} \qquad L_{cttr} = \sum_{\pi=B^{-1}(y)} -logP^{\pi} = \sum_{\pi=B^{-1}(y)} \sum_{n=1}^{N} P_n^{\pi_n}$$

Joint Loss Optimization ($Loss_{cttr}$ + $Loss_{fcls}$ + $Loss_{fcor}$)

⇓

Pseudo-supervised Online Learning



- **Cross-entropy Loss:**

$$L_{fcls}(M) = \sum_{m \in M} \sum_{k \in K} y_m^k \log(p_m^k)$$

- **Improved Triplet Loss:**

$$L_{fcor}(T) = \sum_{t \in T} \max\left(s(t_{neg}) - \beta, 0\right) + \sum_{t \in T} \max\left(\beta - s(t_{pos}), 0\right)$$

- **Joint Loss:**

$$L = \frac{1}{|S|}\sum L_{cttr} + \frac{1}{|M|}\sum L_{fcls} + \frac{1}{|T|}\sum L_{fcor}$$

Dataset and Evaluation

- **Dataset1: RWTH-PHOENIX-Weather 2014 (PHOENIX)**
  German Weather Sign Language Dataset. The training, verification and test set do not overlap each other.

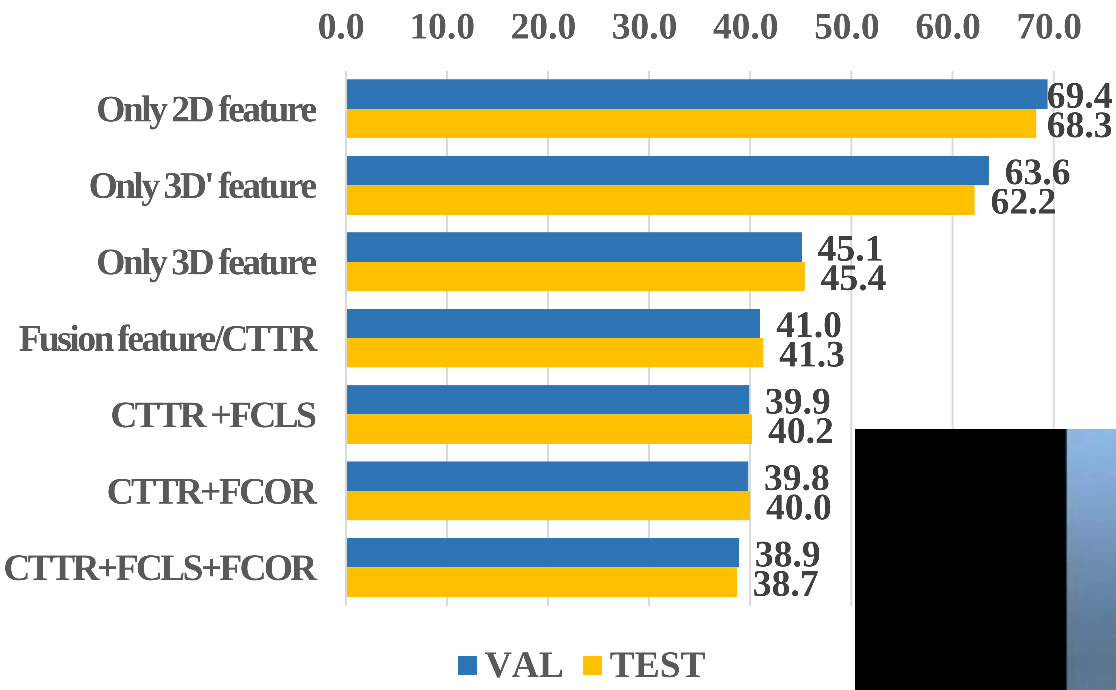
- **Dataset2: USTC Chinese Sign Language (USTC-CSL)**
  Chinese Sign Language Dataset from USTC. The training set and the test set contain different sentences.

Dataset	Split	Signers	Sentences	Videos	Words
PHOENIX	TRAIN	9	5,672	5,672	1,231
	VAL	9	540	540	461
	TEST	9	629	629	497
USTC-CSL	TRAIN	50	94	4,700	178
	TEST	50	6	300	20

- **Evaluation Criterion: Word Error Rate (WER)**

$$WER = \frac{\#ins + \#sub + \#del}{\#num\_words} \times 100\%$$

How much effect does each module have on the results?



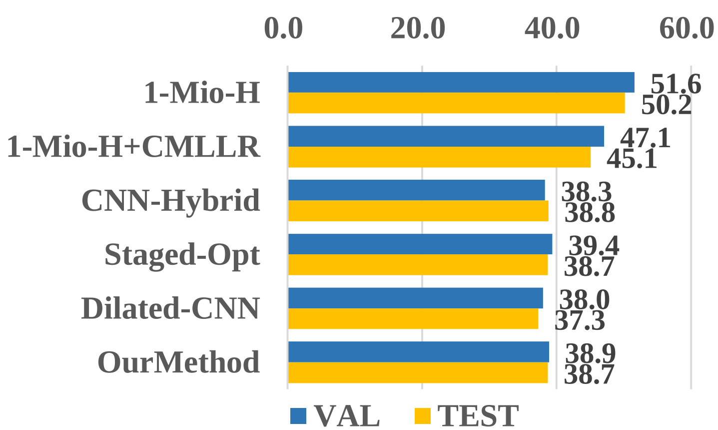
An example of decoding results.

Comparison with existing methods

Methods	Off-line Iterations	Modality			VAL(%)		TEST(%)	
		hand	traj	face	des / ins	WER	des / ins	WER
HOG-3D [Koller *et al.*, 2015]	-	√			25.8 / 4.2	60.9	23.2 / 4.1	58.1
CMLLR [Koller *et al.*, 2015]	-	√	√	√	21.8 / 3.9	55.0	20.3 / 4.5	53.0
1-Mio-H [Koller *et al.*, 2016a]	3	√			19.1 / 4.1	51.6	17.5 / 4.5	50.2
1-Mio-H+CMLLR [Koller *et al.*, 2016a]	3	√	√	√	16.3 / 4.6	47.1	15.2 / 4.6	45.1
CNN-Hybrid [Koller *et al.*, 2016b]	3	√			12.6 / 5.1	38.3	11.1 / 5.7	38.8
Staged-Opt-init [Cui *et al.*, 2017]	-	√			16.3 / 6.7	46.2	15.1 / 7.4	46.9
Staged-Opt [Cui *et al.*, 2017]	3	√			13.7 / 7.3	39.4	12.2 / 7.5	38.7
SubUNets [Camgoz *et al.*, 2017]	-		√		14.6 / 4.0	40.8	14.3 / 4.0	40.7
Dilated-CNN-init [Pu *et al.*, 2018]	-				18.5 / 2.6	60.3	18.1 / 2.8	59.7
Dilated-CNN [Pu *et al.*, 2018]	**5**				**8.3 / 4.8**	**38.0**	**7.6 / 4.8**	**37.3**
Our Method	**-**				**11.6 / 6.3**	**38.9**	**10.9 / 6.4**	**38.7**

Comparison with existing methods

Methods	Off-line Iterations	Modality hand	traj	face	VAL(%) des / ins	WER	TEST(%) des / ins	WER
HOG-3D [Koller *et al.*, 2015]	-	√			25.8 / 4.2	60.9	23.2 / 4.1	58.1
CMLLR [Koller *et al.*, 2015]	-	√	√	√	21.8 / 3.9	55.0	20.3 / 4.5	53.0
1-Mio-H [Koller *et al.*, 2016a]	3	√			19.1 / 4.1	51.6	17.5 / 4.5	50.2
1-Mio-H+CMLLR [Koller *et al.*, 2016a]	3	√	√	√	16.3 / 4.6	47.1	15.2 / 4.6	45.1
CNN-Hybrid [Koller *et al.*, 2016b]	3	√			12.6 / 5.1	38.3	11.1 / 5.7	38.8
Staged-Opt-init [Cui *et al.*, 2017]	-	√			16.3 / 6.7	46.2	15.1 / 7.4	46.9
Staged-Opt [Cui *et al.*, 2017]	3	√			13.7 / 7.3	39.4	12.2 / 7.5	38.7
SubUNets [Camgoz *et al.*, 2017]	-		√		14.6 / 4.0	40.8	14.3 / 4.0	40.7
Dilated-CNN-init [Pu *et al.*, 2018]	-				18.5 / 2.6	60.3	18.1 / 2.8	59.7
Dilated-CNN [Pu *et al.*, 2018]	**5**				**8.3 / 4.8**	**38.0**	**7.6 / 4.8**	**37.3**
Our Method	**-**				**11.6 / 6.3**	**38.9**	**10.9 / 6.4**	**38.7**

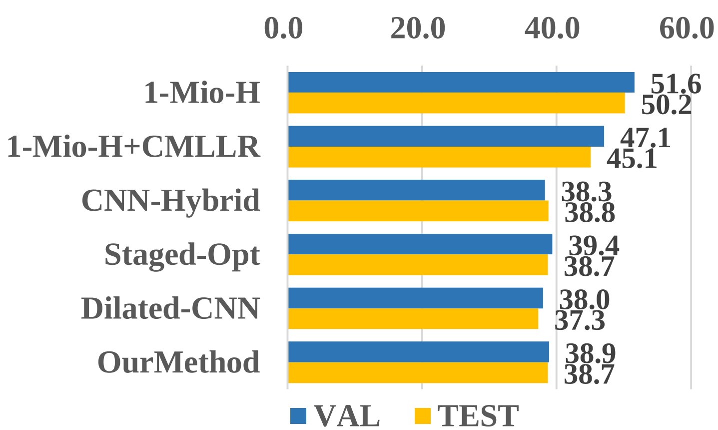


Comparison with Offline Methods

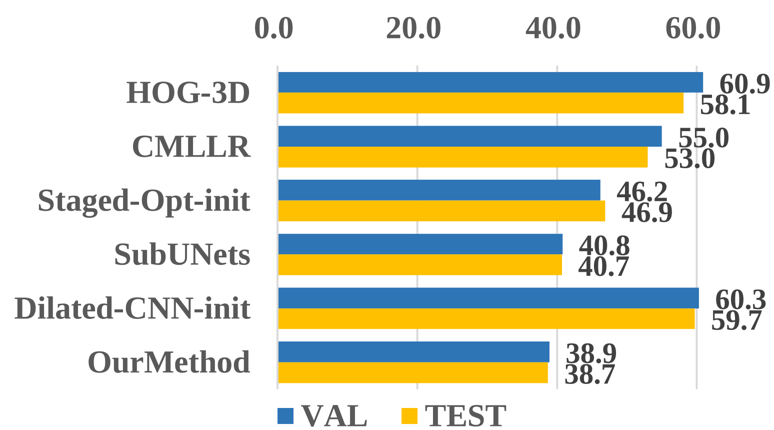
PHOENIX: The effect is close, but the time is greatly reduced(Compared to offline optimization methods).

Comparison with existing methods

Methods	Off-line Iterations	Modality			VAL(%)		TEST(%)	
		hand	traj	face	des / ins	WER	des / ins	WER
HOG-3D [Koller *et al.*, 2015]	-	√			25.8 / 4.2	60.9	23.2 / 4.1	58.1
CMLLR [Koller *et al.*, 2015]	-	√	√	√	21.8 / 3.9	55.0	20.3 / 4.5	53.0
1-Mio-H [Koller *et al.*, 2016a]	3	√			19.1 / 4.1	51.6	17.5 / 4.5	50.2
1-Mio-H+CMLLR [Koller *et al.*, 2016a]	3	√	√	√	16.3 / 4.6	47.1	15.2 / 4.6	45.1
CNN-Hybrid [Koller *et al.*, 2016b]	3	√			12.6 / 5.1	38.3	11.1 / 5.7	38.8
Staged-Opt-init [Cui *et al.*, 2017]	-	√			16.3 / 6.7	46.2	15.1 / 7.4	46.9
Staged-Opt [Cui *et al.*, 2017]	3	√			13.7 / 7.3	39.4	12.2 / 7.5	38.7
SubUNets [Camgoz *et al.*, 2017]	-		√		14.6 / 4.0	40.8	14.3 / 4.0	40.7
Dilated-CNN-init [Pu *et al.*, 2018]	-				18.5 / 2.6	60.3	18.1 / 2.8	59.7
Dilated-CNN [Pu *et al.*, 2018]	**5**				**8.3 / 4.8**	**38.0**	**7.6 / 4.8**	**37.3**
Our Method	**-**				**11.6 / 6.3**	**38.9**	**10.9 / 6.4**	**38.7**



Comparison with Offline Methods

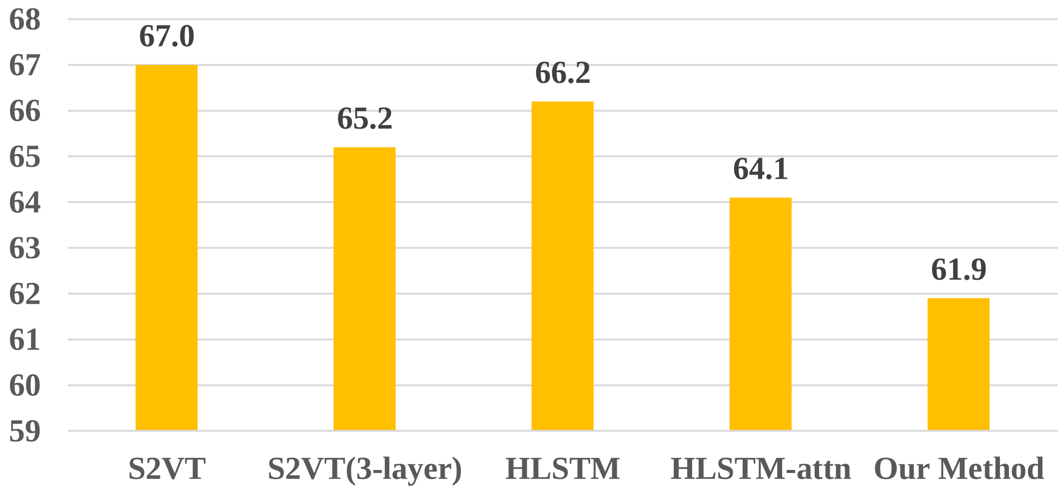


Comparison with Online Methods

PHOENIX: The effect is close, but the time is greatly reduced(Compared to offline optimization methods). BEST(Compared to online optimization methods).

Comparison with existing methods

Methods	TEST WER(%)
S2VT [Venugopalan *et al.*, 2015]	67.0
S2VT(3-layer) [Yao *et al.*, 2015]	65.2
HLSTM [Guo *et al.*, 2018]	66.2
HLSTM-attn [Guo *et al.*, 2018]	64.1
Our Method	**61.9**



USTC-CSL：BEST

The Primary Contributions

- We use the different kind of visual features, and propose the TCP module to learn the short-term association between adjacent frames.

- We propose a connectionist temporal modeling network for long-term sequential learning, where the decoder embeds the dynamic optimization into online learning.

- We design a joint loss function to measure sentence translation, feature correlation, and classification accuracy based on the pseudo labels.

Thanks!

Paper ID: #323

Email: tsg1995@mail.hfut.edu.cn

HEFEI UNIVERSITY OF TECHNOLOGY