MULTIMODAL LOCALLY ENHANCED TRANSFORMER FOR CONTINUOUS SIGN LANGUAGE RECOGNITION

INTERSPEECH 2023

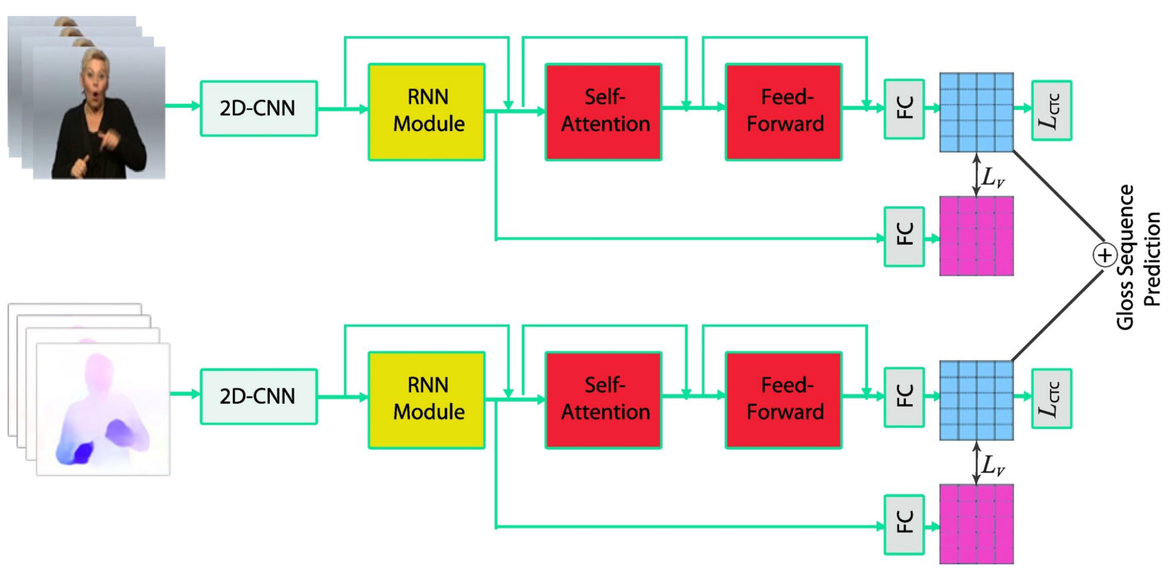
Katerina Papadimitriou Gerasimos Potamianos

Department of Electrical and Computer Engineering, University of Thessaly, Volos 38221, Greece

OVERVIEW

- **Goal:** Continuous sign language recognition from RGB videos.
- **Challenges:**
  - Multitude, complexity, and strong correlation of SL articulators.
  - Absence of gloss-level segmentation.
- **Previous work [1]:**
  - Multiple modalities: signer's pose, shape, appearance, and motion information.
  - Graph convolutional networks with BiLSTMs.
- **Paper contributions:**
  - A multimodal framework: appearance and motion signing streams.
  - A window-based RNN module [2] ➡ local temporal context.
  - A Transformer encoder ➡ both local and global structure modeling.
  - Visual feature and gloss sequence alignment.
- **Results:**
  - Achieves competitive performance on two large-scale German CSLR datasets.

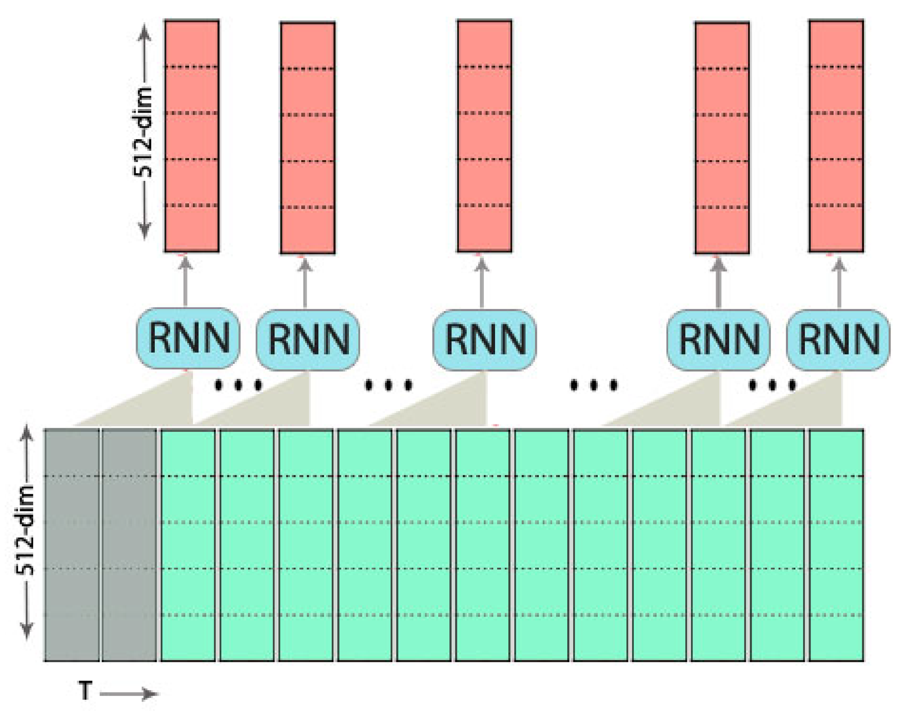
PROPOSED CSLR SYSTEM



- **Visual module:**
  - Two different streams ➡ RGB appearance frames and optical flows.
  - A 2D-CNN based spatial feature learner.
  - A window-based RNN module for local context visual features extraction.
- **Sequence learning model:**
  - Transformer encoder:
    - ➡ Relative position encoding [3] and Gaussian bias [4].
    - ➡ Multi-head attention.
- **Alignment module:**
  - Conjunction of CTC and knowledge distillation loss functions [5].
- **Ensemble module:**
  - Streams alignment through a CTC guiding technique [1] and score fusion.

VISUAL MODULE

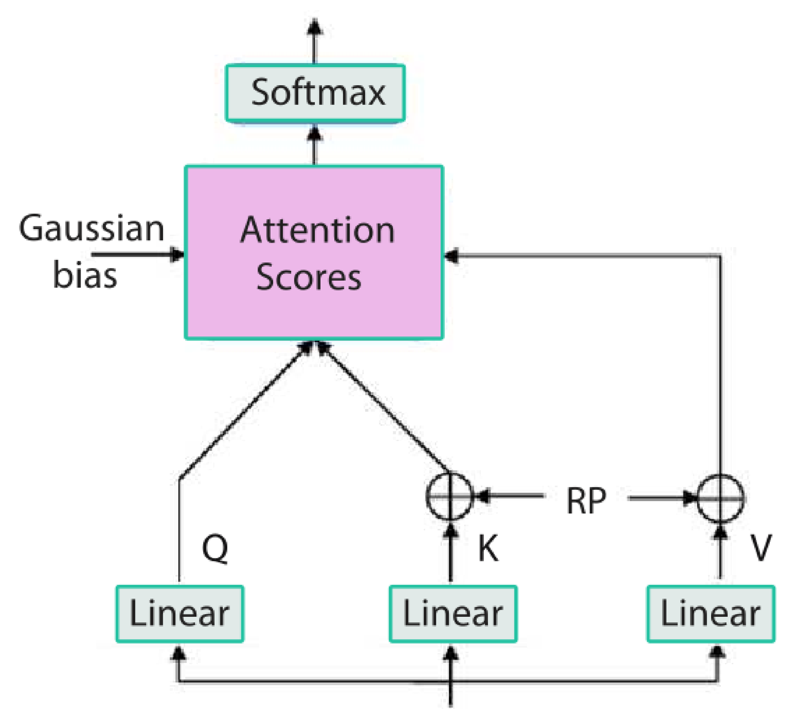
- **Appearance and optical flow features:**
  - Full-frame RGB stream.
  - Motion informative image generation via SpyNet [6].
  - Visual representations based on the VGG11 network [7].
  - 512-dimensional features.



- **Window-based RNN module:**
  - Rearrange the initial frame feature sequence into many short ones.
  - Use a local window of fixed size M for each target frame ➡ local sequences.
  - Local sequences pass through the RNN unit ➡ hidden state representations.
  - RNN module relies on BiLSTM networks.

CSL RECOGNIZER

- **Transformer encoder:**
  - Global long-term dependencies ➡ multi-head attention layer followed by a feed-forward one.
  - Local context dependencies:
    - ➡ Relative representations enhancing neighboring relations.
    - ➡ Gaussian distribution with a fixed window size as additive bias.
- **Alignment module:**
  - Combines the CTC loss with a knowledge distillation loss.
  - Minimizes the distance between the probability distributions of the sequence learning model and the visual module.
- **Ensemble module:**
  - Add RGB and optical flow streams decoding scores through a posterior fusion scheme.
  - Spike timings synchronization via a guiding CTC model.



DATASETS & EXPERIMENTAL SETUP

- **RWTH-PHOENIX Weather 2014 dataset [8]:**
  - 6,841 sentences ➡ 1,232-gloss vocabulary.
  - Multi-signer split ➡ 5,672 training videos, 540 validation, and 629 testing.
- **RWTH-PHOENIX Weather 2014T dataset [9]:**
  - 8,257 sequences ➡ 1,066-gloss vocabulary.
  - Multi-signer setting ➡ 7,096 training videos, 519 validation, and 642 testing.

EXPERIMENTAL RESULTS

- System evaluation on the RWTH-PHOENIX Weather 2014 dataset, against some variations of it, in word error rate (WER, %).

Modalities	RNN	RP	GB	WER (%)
RGB				27.55
	✓			23.25
		✓	✓	24.05
	✓	✓	✓	21.25
Optical Flow				29.18
	✓			26.20
		✓	✓	25.87
	✓	✓	✓	25.07
Both	✓	✓	✓	**20.89**

- Superior performance when all modalities are considered.
- RNN module and relative position (RP) encoding ➡ most robust components.
- Gaussian bias (GB) incorporation benefits system performance.

Proposed Model	WER (%)
w/o $\mathcal{L}_V$	21.75
w/o $\mathcal{L}_G$	24.16
w $\mathcal{L}_V$ & $\mathcal{L}_G$	**20.89**

- Comparison of our proposed model to the literature on the RWTH-PHOENIX Weather 2014 dataset (left) and the RWTH-PHOENIX Weather 2014T dataset (right).
- Outperforms most results in the literature, coming very close to the state-of-the-art.

Model	WER (%)
SubUnet [35]	40.70
SLT [36]	24.59
CNN-LSTM-HMM [37]	24.10
VAC [4]	22.30
SMKD [5]	21.00
STMC [18]	20.70
C2SLR [6]	20.40
STTN [20]	**19.98**
Proposed	**20.89**

Model	WER (%)
Re-Sign [8]	26.60
SFD+SGS+SFL [14]	26.10
Bi-ST-LSTM-A [16]	24.68
SLT [36]	24.59
CrossModal [24]	24.30
CNN-LSTM-HMM [37]	24.10
TDCNN [15]	23.70
SMKD [5]	22.40
ST-GCN [25]	21.34
STMC [18]	21.00
C2SLR [6]	**20.40**
Proposed	**20.73**

CONCLUSIONS

- Proposed a deep learning model for CSLR from RGB videos.
- Investigated the contribution of:
  - A window-based RNN module to capture local temporal context.
  - A Transformer encoder with local context modeling and global structure learning.
  - The design of a multi-modal framework.
  - The conjunction of the CTC loss with a visual alignment loss.
- Achieved competitive performance on two popular German CSLR datasets.

REFERENCES

[1] **Parelli et al.,** "Spatio-temporal graph convolutional networks for continuous sign language recognition," *Proc. ICASSP*, **2022**.
[2] **Zheng et al.,** "Improving end-to-end speech synthesis with local recurrent neural network enhanced transformer ," *Proc. ICASSP*, **2022.**
[3] **Shaw et al.,** "Self-attention with relative position representations," *Proc. NAACL-HLT*, **2018**.
[4] **Yang et al.,** "Modeling localness for self-attention networks," *ArXiv*, **2018**.
[5] **Hinton et al.,** "Distilling the knowledge in a neural network," *ArXiv,*, **2015**.
[6] **Ranjan & Black.,** "Optical flow estimation using a spatial pyramid network," *Proc. CVPR*, **2017**.
[7] **Simonyan and Zisserman,** "Very deep convolutional networks for large-scale image recognition," *Proc. ICLR*, **2015.**
[8] **Koller et al**., "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, **2015**.
[9] **Camgöz et al**., "Sign language transformers: Joint end-to-end sign language recognition and translation," *Proc. ICCV*, **2019**.
[*] See paper for the table citations.

ACKNOWLEDGMENTS

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (HFRI-FM17-2456).