ICPR2020 Poster Presentation

Audio-Visual Speech Recognition Using A Two-Step Feature Fusion Strategy

Presenter: Wanlu Xu

Author: Hong Liu, Wanlu Xu, Bing Yang

School: Peking University, China

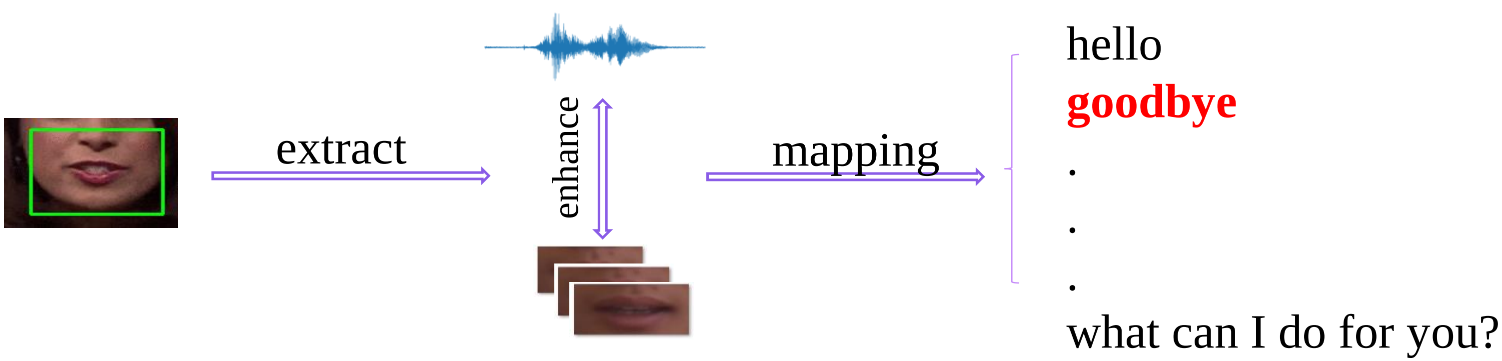
Date: Dec 5th, 2020

Outline

- Introduction
- The Proposed Method
- Experiments and Discussion
- Conclusion

What is the task?

- **Audio-visual speech recognition (AVSR)**



S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, “LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,” in IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, pp. 1–8.

Applications and Challenges

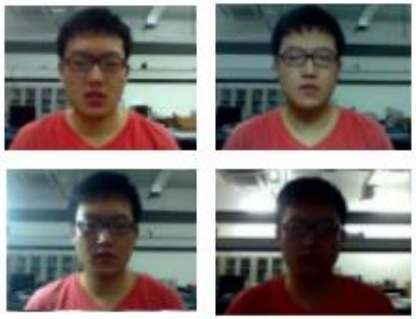
Intelligent monitoring

Human-robot interaction

Target tracking

Intra-class variations

Inter-class similarities



various lighting

complex backgrounds

Motivation

- **Lip-reading method**
  - The visual information is particularly important when the audio information is contaminated severely in a noisy environment.

How to capture the long-range dependencies of sequential data?

- **Fusion method**
  - Only consider the fusion in a single stage of the network, which may not be able to balance the integrity and representativeness of audio and visual information.

How to design a fusion method to better integrate the two features?

- **Research contents**
  - A non-local block is inserted into the visual branch to capture long-range features of lip frames.
  - A two-step feature fusion strategy is proposed to combine audio and visual information in the diverse stages.

Outline

- Introduction
- **The Proposed Method**
- Experiments and Discussion
- Conclusion

Two-Step Feature Fusion Network

- **Pipeline of the network**

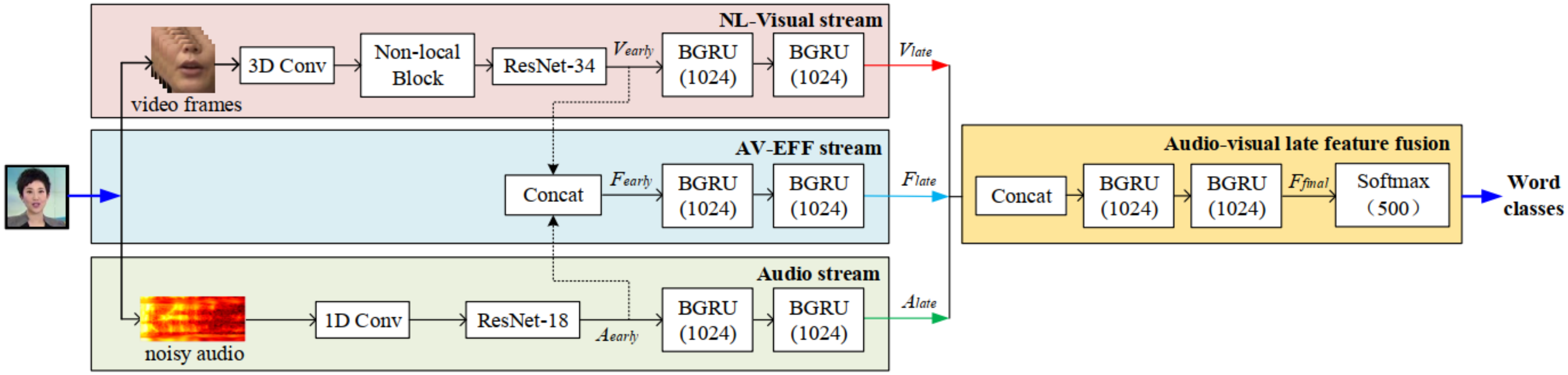
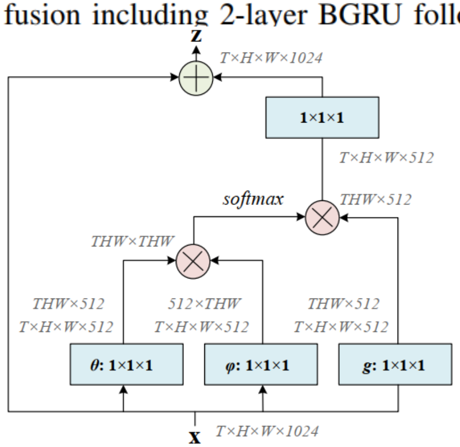
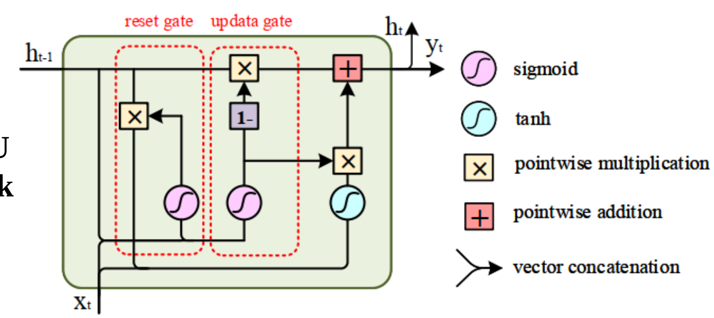


Fig. 1: Overall framework of the two-step feature fusion network, which consists of two parts. The first part has three streams including NL-Visual stream, audio stream, and audio-visual early feature fusion (AV-EFF) stream. The second part is audio-visual late feature fusion including 2-layer BGRU followed by a softmax layer that is connected with the output word label.

A space-time non-local block



GRU block



X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

The Proposed Method

NL-Visual Stream

- **3D Conv:**

$$v_{ij}^{xyz} = tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i\;1}\sum_{q=0}^{Q_i\;1}\sum_{r=0}^{R_i\;1} w_{ijm}^{pqr} v_{(i\;1)m}^{(x+p)(y+q)(z+r)}\right)$$

- **non-local operation:**

$$y_i = \frac{1}{\mathcal{C}(x)}\sum_{\forall j} f(x_i, x_j)g(x_j), \quad f(x_i, x_j) = e^{\theta(x_i)^T\phi(x_j)},$$

$$z_i = W_z y_i + x_i,$$

$$Out_{nl} = W_z \frac{1}{\sum_{\forall j} f(x_i, x_j)} \sum_{\forall j} e^{W_\theta v_i^T W_\phi v_j} W_g v_j + v_i.$$

- **early and late visual feature:**

$$V_{early} = ResNet34(Out_{nl}),$$
$$V_{late} = BGRU(V_{early}),$$

Audio Stream

- **pre-processing:**

$$x(k) = \sum_{n=0}^{N-1} x(n)e^{\frac{j2\pi nk}{N}}, \quad 0 \le n, K \le n-1.$$

$$a_{ij}^{x} = tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1} w_{ijm}^{p} a_{(i-1)m}^{(x+p)}\right),$$

- **early and late audio feature:**

$$A_{early} = ResNet18(a),$$
$$A_{late} = BGRU(A_{early}),$$

Audio-Visual Late Feature Fusion

- **concatenation:**

$$F_{final} = BGRU(Concat(V_{late}, A_{late}, F_{late})),$$

- **classification result:**

$$\begin{aligned} L_{final} &= \arg\max(softmax(F_{final})) \\ &= \arg\max_{j \in 1,\dots,K}\left(\frac{e^{F_{final}^{j}}}{\sum_{k=1}^{K} e^{F_{final}^{k}}}\right), \end{aligned}$$

Loss Function

- **cross-entropy:**

$$L(y, l) = -log\frac{e^{y_l}}{\sum_{i=1}^{C} e^{y_i}},$$

Outline

- Introduction
- The Proposed Method
- **Experiments and Discussion**
- Conclusion

Experiments

- **Datasets**

LRW dataset: Lip Reading in the Wild (LRW) dataset was released in 2016, which is the largest publicly available lipreading dataset in English. The dataset consists of short segments (1.16 seconds) from BBC programs, mainly news and talk shows. It is a very challenging dataset with more than 1000 speakers, 500 words, 538766 samples, and large variation in head pose and illumination.

LRW-1000 dataset: LRW-1000 dataset was released in 2019, which is a more challenging Naturally-Distributed Large-Scale dataset in Mandarin and contains 1000 classes with 718018 samples from more than 2000 individual speakers. Each class corresponds to the syllables of a Mandarin word composed of one or several Chinese characters. It is currently the largest word-level lipreading dataset and also the only public large-scale Mandarin lipreading dataset.

[1] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in Asian Conference on Computer Vision, 2016, pp. 87–103.

[2] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, “LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,” in IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019, pp. 1–8.

Experiments

- **Training Process**

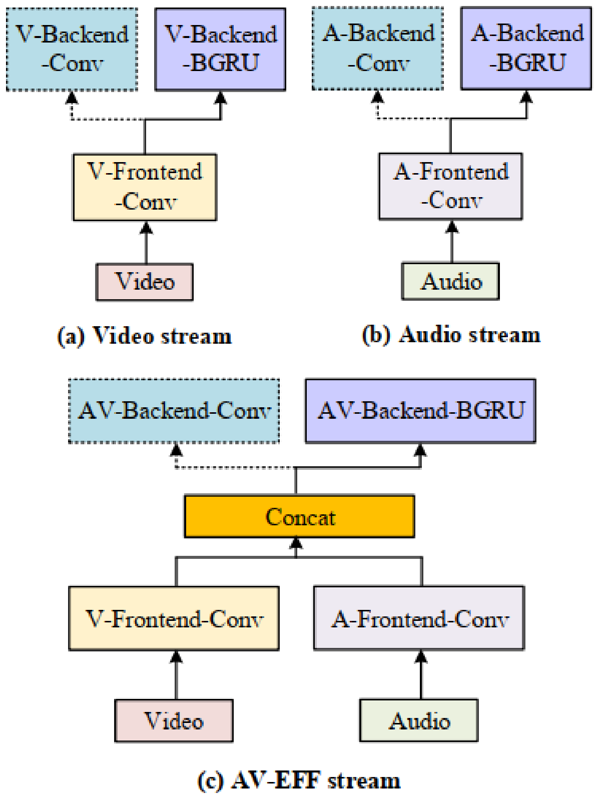


Fig. 5: The network partition of three streams in a two-step feature fusion network. The solid line represents the final classification network structure, and the dotted line represents the feature extraction network structure during the training.

Single stream training:

Firstly, the frontend convolution network is connected to the backend convolution network for pre-training. After that, the backend convolution network is discarded and the backend BGRU layers are added. The backend BGRU layers are firstly trained separately with the remaining parameters fixed, and then trained end-to-end with the frontend convolution network. Early stopping is applied with a delay of 5 epochs.

Muliti stream training:

Once the single stream has been trained, then they are used for initializing the corresponding streams in the multi-stream architecture. Specifically, another 2-layer BGRU is added on top of all streams to fuse the single stream outputs. The top BGRU is trained with the weights of the single-stream fixed firstly. After that, the entire network is fine-tuned end to end. Early stopping is also applied with a delay of 5 epochs.

Experiments

- **Experimental Results**

TABLE II: Comparison of our methods with the state-of-the-art methods on LRW and LRW-1000 datasets. Clean represents in a noiseless environment.

Task	Method	LRW Accuracy(%)	LRW1000 Accuracy(%)
Lip-reading	LSTM-5 [30]	71.50	25.76
	D3D [31]	78.02	34.76
	3D+2D [21]	83.00	**38.19**
	Multi-Grained [32]	83.34	36.91
	ResNet34+BGRU(Baseline) [8]	82.80	36.72
	NL-Visual(Ours)	**83.41**	37.03
AVSR (clean)	MCNN [33]	96.98	39.60
	ETE-AVSR(Baseline) [8]	97.60	37.52
	Two-Step(Ours)	**98.26**	**41.57**

- For lip-reading experiments on the LRW dataset, we can find that the performance of our method is superior to other state-of-the-art methods, and can achieve the best performance among them by adding non-local block to the baseline ResNet34+BGRU model.
- On a more challenging LRW-1000 dataset, our method is better than most state-of-the-art methods, except for 3D+2D method.

Experiments

■ Experimental Results

TABLE III: Ablation experiments of our two-step feature fusion method under different SNR(dB) conditions.

Baseline [8]	NL-Visual	AV-EFF	-5	0	5	10	15	20	clean
✓			86.66	94.13	96.29	96.70	97.00	97.50	97.90
✓	✓		88.21	95.01	97.18	97.22	97.53	97.86	98.10
✓		✓	90.65	95.56	97.28	97.74	98.04	98.08	98.14
✓	✓	✓	**92.10**	**96.19**	**97.35**	**97.86**	**98.08**	**98.15**	**98.26**

TABLE IV: Performance of our three single streams and fusion model under different SNR(dB) conditions.

Modality	Method	-5	0	5	10	15	20	clean
Single	Audio only	71.60	90.55	95.34	96.89	97.32	97.58	97.70
	Visual only	83.41	83.41	83.41	83.41	83.41	83.41	83.41
	AV-EFF only	87.63	94.68	96.19	96.69	96.96	97.02	97.10
Fusion	Two-step(Ours)	**92.10**	**96.19**	**97.35**	**97.86**	**98.08**	**98.15**	**98.26**

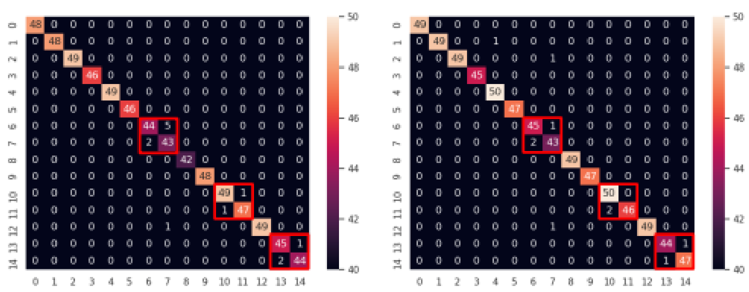
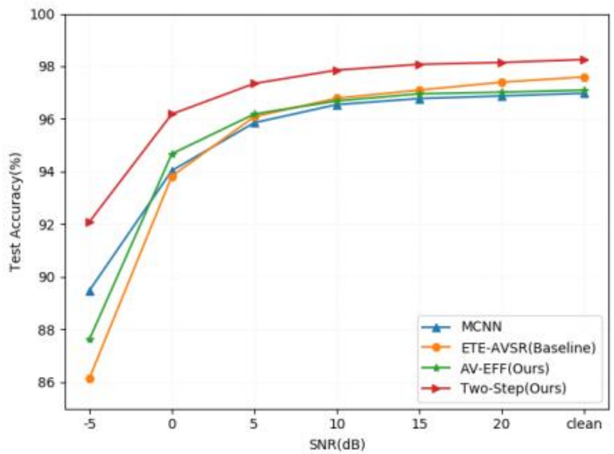


Fig. 6: Confusion matrices of baseline model (left) and our two-step feature fusion network (right) at -5dB SNR.



Outline

- Introduction
- The Proposed Method
- Experiments and Discussion
- **Conclusion**

Conclusion

- A non-local block is inserted in the feature extraction part of the visual stream (NL-Visual) to capture long-range dependencies by calculating the distance of all positions.

- An audio-visual early feature fusion (AV-EFF) stream is added to form a two-step feature fusion strategy that can guarantee integrity and representativeness of features simultaneously.

- The experimental results show that our method can improve the fusion performance in strong noise environment greatly.

Thank You! Q&A