Person Recognition with HGR Maximal Correlation on Multimodal data

Yihua Liang(1), Fei Ma(1), Yang Li(1), Shao-Lun Huang(1)

(1) Tsinghua-Berkeley Shenzhen Institute , Tsinghua University

TBSI Tsinghua-Berkeley Shenzhen Institute

Abstract

In video analysis and public surveillance, information from multiple modalities are used to jointly determine the identity of a person. We propose a correlation-based multimodal person recognition framework that is relatively simple but can efficaciously learn supervised information in multimodal data fusion and resist noise.

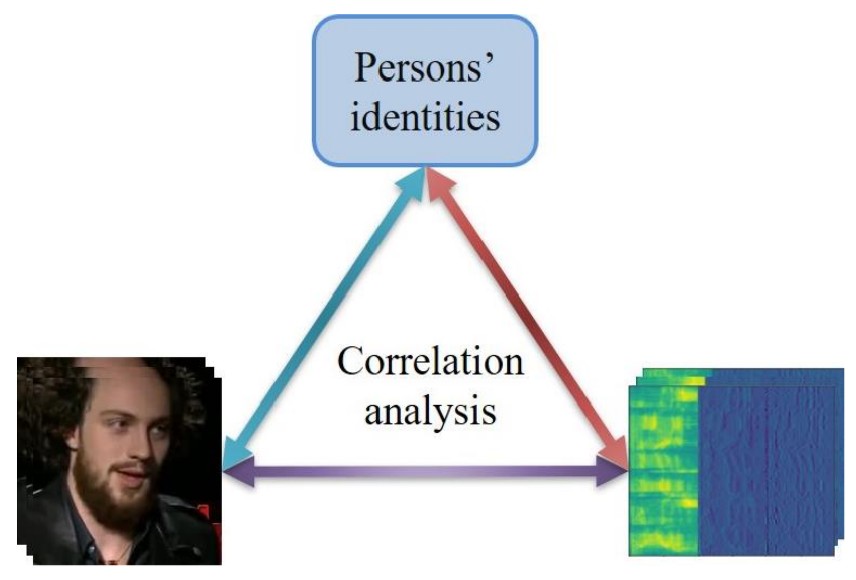
Introduction

Challenges

- Learn person's identity while merging multimodal data for person recognition.
- Hold robustness to noise.

Brief introduction

- Analyze correlation among visual and audio input and identities.



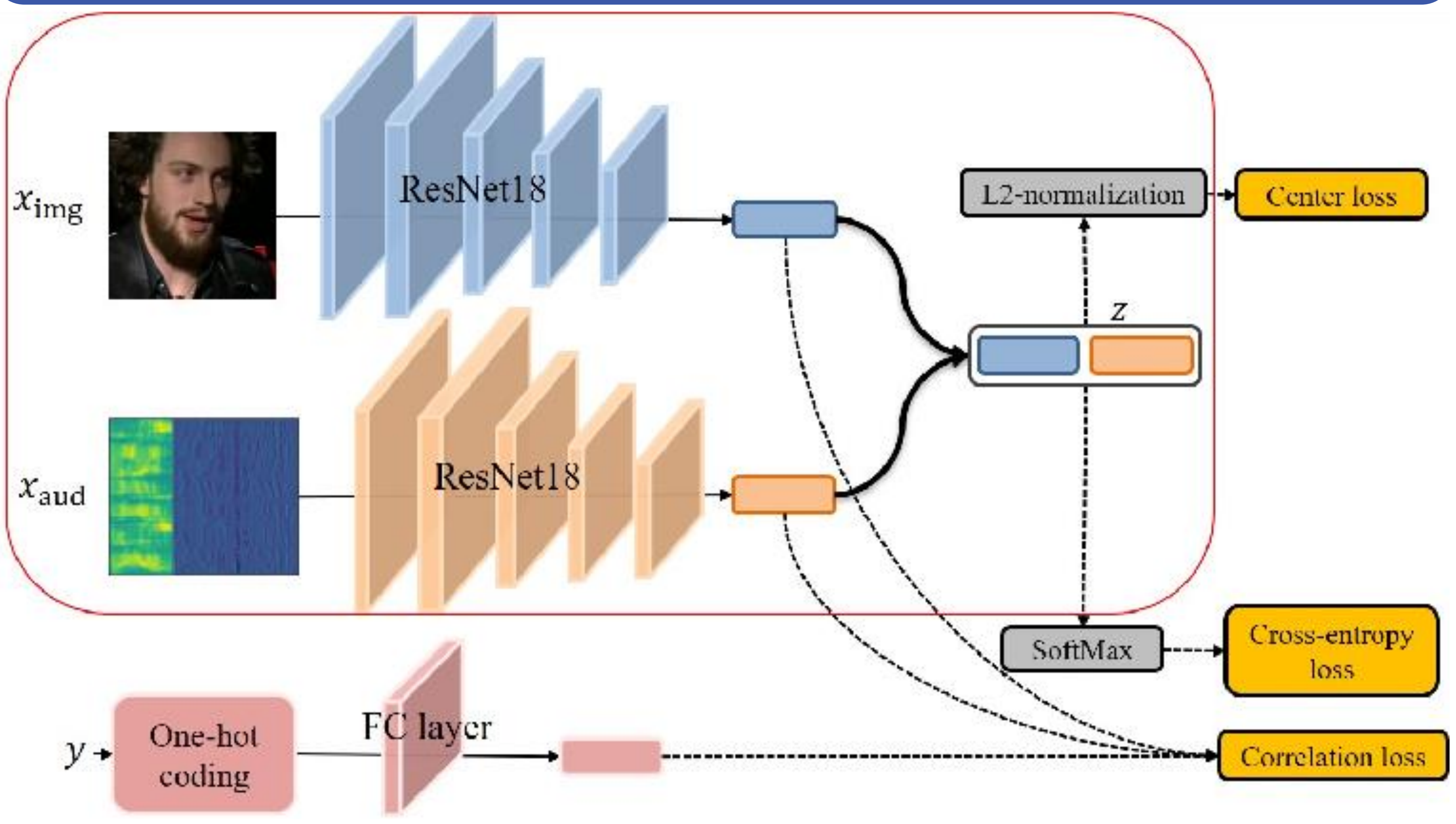
Contributions

- Proposed objective merge multimodal data and learn discriminative embeddings more effectively.
- By maximizing the HGR maximal correlation between labels and input, the embedding robustness under noise is improved.

Existing methods

- Uni-modal methods do not fully utilize multimodal information.
- Correlation based methods leave the extraction of multimodal input's relationship with identity information to downstream tasks.
- Multimodal methods does not consider real world noise.

Method



Training model

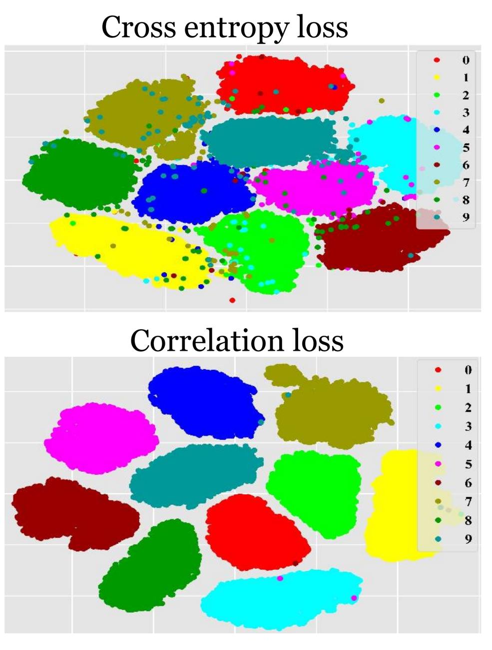
- In training stage, visual feature $f(x_{\text{img}})$ and audio feature $g(x_{\text{aud}})$ are extracted by ResNet18.
- Embedding z is generated by concatenation; Meanwhile, the identity is converted to an one-hot vector and then mapped to feature $h(y)$ by a fully connecting layer.
- In the end, framework is jointly optimized by three loss functions.
- During validation and test, only x_{img} and x_{aud} are taken by the framework

Learning objectives

- Cross entropy loss: basic classification.
- Center loss $\mathcal{L}_{\text{ctr}}$: separates different identities in embedding space.
- Correlation loss $\mathcal{L}_{\text{corr}}$:
  - An adoption of HGR maximal correlation which holds good theoretical interpretation.
  - Effective merge multimodal data.
  - Robustness to noise: lead to larger inter-class margin and less falsely classified points.

$$\mathcal{L}_{\text{ctr}} = \frac{1}{2}\sum_{i=1}^{m} \| z^{(i)} - c^{y^{(i)}} \|_2^2$$

$$\mathcal{L}_{\text{corr}} = -\sum_{l \neq k}^{d} (\mathbb{E}[f_l{}^T f_k] - \frac{1}{2}\text{tr}(\text{cov}(f_l)\text{cov}(f_k)))$$

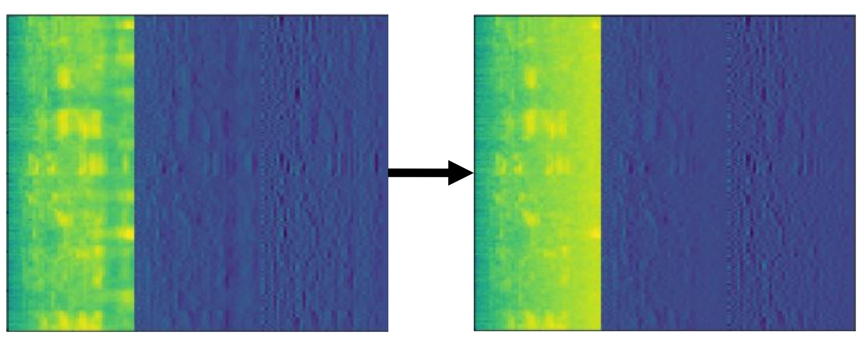


Toy example: Classification results on MNIST

Experiments & Results

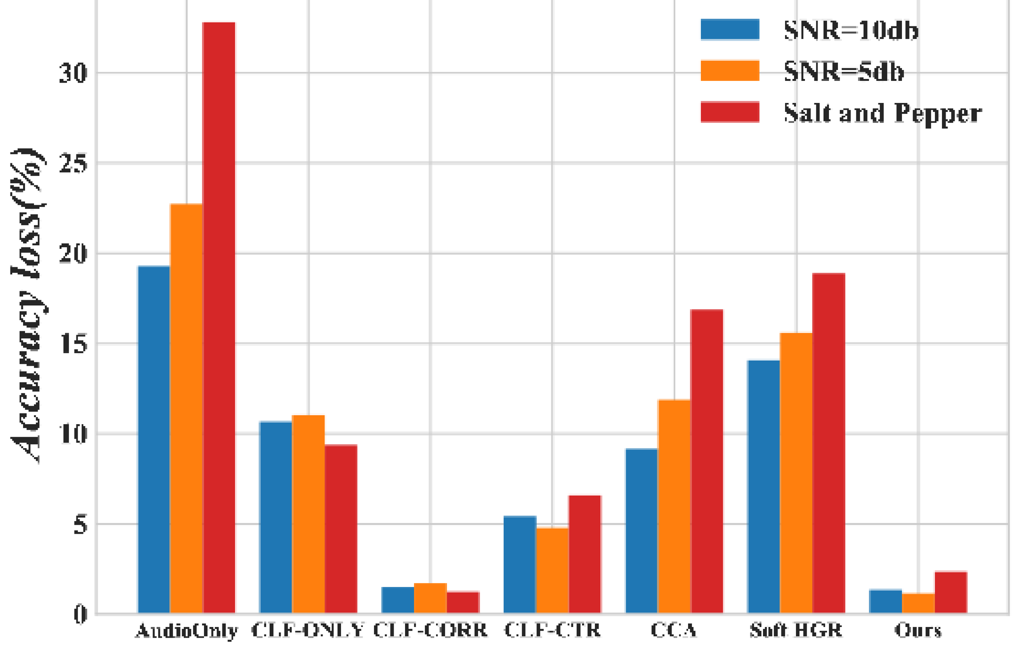
Methods	Accuracy(%)
Product [1]	91.86 ± 0.9
wProduct [1]	90.93 ± 1.7
MMA [2]	91.43 ± 0.7
MSE [3]	90.00 ± 1.0
Imd [4]	91.89 ± 1.3
Ours	**97.56** ± 0.6

Performance comparision



Zero noise mel-spectrum And its derivatives

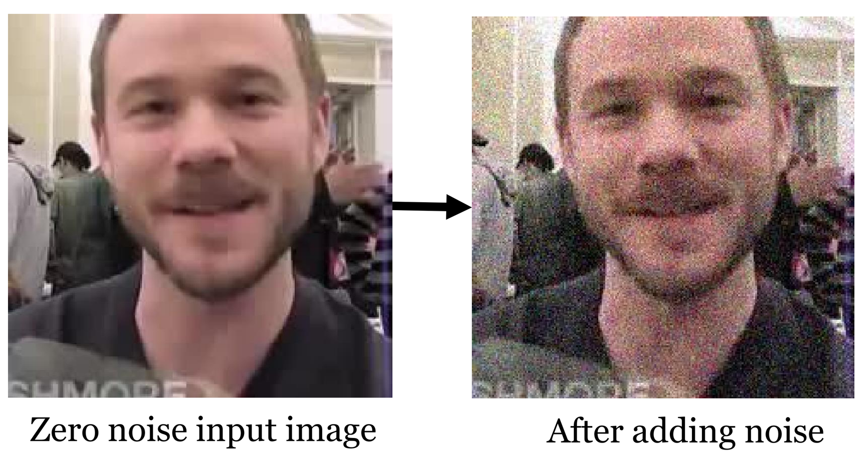
After adding noise



- Ours methods lead to large improvement in accuracy.
- Ours methods effectively reduces the accuracy loss on noisy data.
- It is proved that these improvement are contributed by our application of correlation analysis.

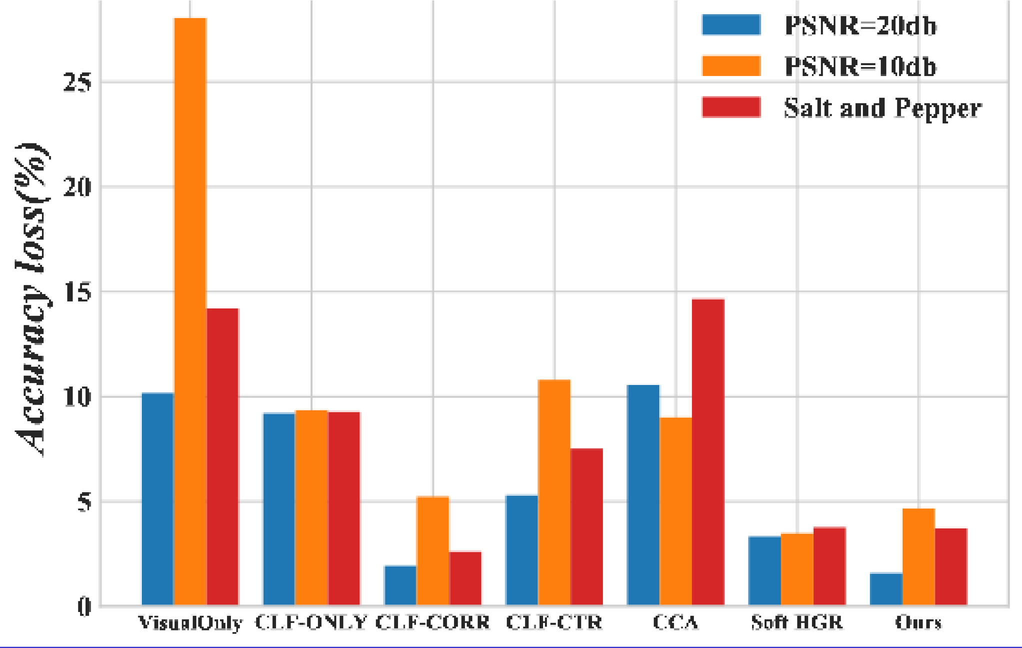
Methods	Accuracy(%)
AudioOnly [5]	83.75 ± 1.1
VisualOnly [6]	88.19 ± 1.6
CLF-CTR	89.54 ± 0.3
CLF-ONLY	89.75 ± 0.6
CLF-CORR	**97.84** ± 0.5
CCA [7]	86.37 ± 0.7
Soft HGR [8]	87.74 ± 1.1
Ours	97.56 ± 0.6

Ablation study



Zero noise input image

After adding noise



Conclusion

The proposed objective not only make the framework acquire more sufficient guidance to supervised target in training but improve its robustness to noise. too. Thus the framework effectually solves the challenges about combining multimodal data and resisting different types of noise.

References

[1] Chowdhury, Y. Atoum, L. Tran, X. Liu, and A. Ross, "Msu-avis dataset: Fusing face and voice modalities for biometric recognition in indoor surveillance videos," in2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018, pp. 3567–3573.

[2] Y. Liu, P. Shi, B. Peng, H. Yan, Y. Zhou, B. Han, Y. Zheng, C. Lin,J. Jiang, Y. Fanet al., "iqiyi-vid: A large dataset for multi-modal person identification,"arXiv preprint arXiv:1811.07548, 20

[3] Chen, S. Wang, and S. Chen, "Deep multimodal network for multi-label classification," in2017 IEEE International Conference on Multi-media and Expo (ICME). IEEE, 2017, pp. 955–960.

[4] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," inProceedings of the 25th ACM international conference on Multimedia. ACM, 2017, pp. 154–162

[5] Li, D. Tuo, D. Su, Z. Li, D. Yu, and A. Tencent, "Deep discriminative embeddings for duration robust speaker verification." in Interspeech,2018, pp. 2262–2266.

[6] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in ECCV. Springer, 2016, pp.499–515.

[7] X. Chang, T. Xiang, and T. M. Hospedales, "Scalable and effective deepcca via soft decorrelation," inCVPR, 2018, pp. 1488–1497.

[8] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang,"An efficient approach to informative feature extraction from multimodaldata," inAAAI, vol. 33, 2019, pp. 5281–5288.

Future Work

- Take more different types of noise into consideration.
- Try to combine attention mechanism with correlation learning.
- Adapt this framework to similar tasks which take multi-modal input and rely on labels.

Acknowledgement

The research of Shao-Lun Huang was supported in part by the Natural Science Foundation of China under Grant 618 07021, in part by the Shenzhen Science and Technology Pro gram under Grant KQTD20170810150821146, and in part b y the Innovation and Entrepreneurship Project for Overseas High-Level Talents of Shenzhen under Grant KQJSCX20180327144037831.

TBSI 清華大學 Tsinghua University Berkeley UNIVERSITY OF CALIFORNIA

Contact: Yihua Liang, liangyh18@mails.tsinghua.edu.cn