Learning Discriminative Joint Embeddings for Efficient Face and Voice Association

Rui Wang
Dep. of CS, Huaqiao University & State Key Lab. of ISN, Xidian U.
wr@stu.hqu.edu.cn

Xin Liu*
Dep. of CS, Huaqiao University & ISN, Xidian U. & Dep. of CS, HKBU
xliu@hqu.edu.cn

Yiu-ming Cheung
Dep. of CS, Hong Kong Baptist University, HK SAR, China
ymc@comp.hkbu.edu.hk

Kai Cheng
Dep. of CS, Huaqiao University
kcheng@stu.hqu.edu.cn

Nannan Wang
State Key Lab. of ISN, Xidian U.
nnwang@xidian.edu.cn

Wentao Fan
Dep. of CS, Huaqiao University
fwt@hqu.edu.cn

ABSTRACT

Many cognitive researches have shown the natural possibility of face-voice association, and such potential association has attracted much attention in biometric cross-modal retrieval domain. Nevertheless, the existing methods often fail to explicitly learn the common embeddings for challenging face-voice association tasks. In this paper, we present to learn discriminative joint embedding for face-voice association, which can seamlessly train the face subnetwork and voice subnetwork to learn their high-level semantic features, while correlating them to be compared directly and efficiently. Within the proposed approach, we introduce bi-directional ranking constraint, identity constraint and center constraint to learn the joint face-voice embedding, and adopt bi-directional training strategy to train the deep correlated face-voice model. Meanwhile, an online hard negative mining technique is utilized to discriminatively construct hard triplets in a mini-batch manner, featuring on speeding up the learning process. Accordingly, the proposed approach is adaptive to benefit various face-voice association tasks, including cross-modal verification, 1:2 matching, 1:N matching, and retrieval scenarios. Extensive experiments have shown its improved performances in comparison with the state-of-the-art ones.

CCS CONCEPTS

• **Information systems** → **Retrieval models and ranking**; *Information retrieval*;

KEYWORDS

Face-voice association; discriminative joint embedding; cross-modal verification; bi-directional ranking constraint

ACM Reference Format:
Rui Wang, Xin Liu*, Yiu-ming Cheung, Kai Cheng, Nannan Wang, and Wentao Fan. 2020. Learning Discriminative Joint Embeddings for Efficient Face and Voice Association. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401302

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401302

1 INTRODUCTION

Many cognitive researches have shown that humans are able to hear voices of known individuals to form mental pictures of their facial appearances, and vice versa. Naturally, face and voice have proven to be the most valuable sources of biometric identity information, which can greatly help identify, search and organize human identities. In the past years, purely face-based or voice-based human recognition methods had been widely studied in the literatures [9, 14], and the above studies lend credence to the hypothesis that it may be possible to find the associations between voices and faces because they both characterize the speaker [2].

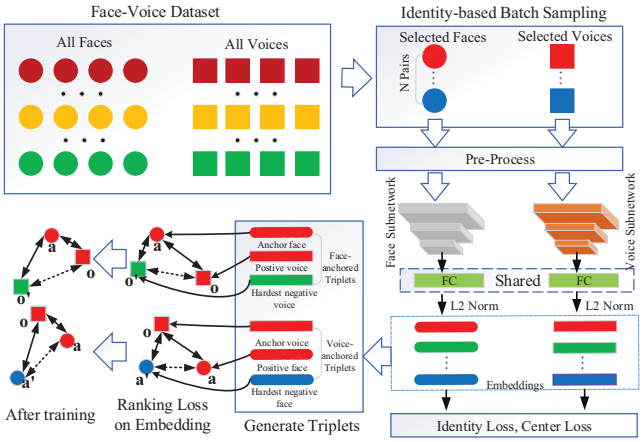

Figure 1: Pipeline of the proposed framework.

With the recent advance of deep learning, there have emerged some studies to mine the correlations between face and voice examples, which can be divided into classification based and metric learning based methods [15]. The former approaches formulate the face-voice correlation task as a classification problem. For instance, SVHF-Net [11] formulates face-voice association problem as a binary selection task. However, this method is designed for a specific matching task, which is not flexible enough for other association tasks. DIMNet [13] utilizes a multi-task classification network to learn the common embeddings, but which does not fully consider the high-level semantic correlations between faces and voices. The latter methods usually construct contrastive loss [10] or triplet loss

[6, 15] to supervise the common representation learning. Along this line, Arsha et al. [10] utilize the contrastive loss to guide the learning, while Kim et al. [6] employ the triplet Loss to supervise the network learning. These two approaches are able to match the human performance on face-voice matching tasks, but their performances are far from the expectation. To the best of our knowledge, automatic face and voice association is still under early study.

In this paper, we address an efficient deep correlated framework to learn discriminative joint embedding for face and voice association, which can semantically correlate face subnetwork and voice subnetwork to benefit various cross-modal matching tasks. The main contributions of this paper are highlighted as follows:

- An end-to-end deep correlated network is proposed to learn discriminative joint embeddings for face-voice associations.
- A bi-directional ranking loss is proposed to enhance the discriminative power of joint embedding, while an online hard negative mining technique is addressed to speed up the training process.
- Extensive experiments have shown its improved performances under various face-voice association and retrieval tasks.

The rest of this paper is structured as follows: Section 2 introduces the proposed model and its implementation details. The extensive experiments and analysis are discussed in Section 3. Finally, we draw a conclusion in Section 4.

2 THE PROPOSED FRAMEWORK

The objective of the proposed deep correlated model is to learn the joint common embeddings to represent both face and voice, which allow them to be semantically correlated. In the following, we first introduce the network architecture, and then describe the loss function, the process of identity-based batch sampling and online hard negative mining scheme.

2.1 Network Architecture

The network architecture is shown in Figure 1, which is a dual-path network consisting of three parts: face subnetwork, voice subnetwork and shared FC structure. Specifically, face subnetwork and voice subnetwork, with independent network parameters, are utilized to learn the high-level correlated and modality-specific features with respect to face and voice. The shared FC structures with the same parameters are employed to jointly learn a shared latent space to bridge the semantic gap between face and voice.

Face Subnetwork. The face subnetwork is implemented using the Inception-ResNet-v1 architecture. The input to the face subnetwork is an RGB image, and the finally fully connected layer of the Inception-ResNet-v1 architecture is reduced to produce a single 512-D embedding for every face input.

Voice Subnetwork. The voice subnetwork is implemented using the DIMNet-voice [13] architecture. In order to produce an efficient voice embedding that is the same size as face embedding, we set the size of last network layer to be 512.

Shared FC Structure. A shared fully connected layer (FC) is utilized to learn a joint discriminative embedding space between face and voice, and their network parameters are shared with each other. Accordingly, this FC layer can project both face and voice examples into 256-D common embedding space.

2.2 Loss Function

A novel loss function, consisting of three constraints, is proposed for discriminative joint embedding learning.

Ranking Constraint. A bi-directional ranking constraint [3] is utilized to ensure the discriminability of the learnt representation. Since it is the hardest negative that determines success or failure as measured by R@1 in retrieval task [7], we focus on hardest negatives for training. Given a l_2 normalized positive pair (a, o), the hardest negatives are given by $a' = \arg\min_{i \neq a} d(i, o)$ and $o' = \arg\min_{j \neq o} d(a, j)$, and this ranking constraint is formulated as:

$$\begin{aligned}\mathcal{L}_r = &\overbrace{\max_{o'}[\alpha_1 + d(a,o) - d(a,o')]_+ + \lambda_3 \max_{o'}[\beta_1 - d(o,o')]_+}^{\text{from face to voice}} \\ &\underbrace{+ \max_{a'}[\alpha_2 + d(o,a) - d(o,a')]_+ + \lambda_4 \max_{a'}[\beta_2 - d(a,a')]_+}_{\text{from voice to face}}\end{aligned} \tag{1}$$

where $d(x)$ is Euclidean distance. Since the ranking constraint is bi-directional and symmetrical, we take the constraint from face to voice as an example to show its functionality. Given a mini-batch, it contains N face embeddings and N voice embeddings. For an anchor face embedding a, the distance of its positive voice embedding o should be smaller than the distance between a and the hardest negative voice embedding o' by a margin α_1:

$$d(a, o) < d(a, o') - \alpha_1 \tag{2}$$

where o and o' belong to different identity, we also want the distance between them is larger than a pre-defined margin β_1:

$$d(o, o') > \beta_1 \tag{3}$$

Accordingly, the distance between anchor sample and the hard negative sample representations is greater than the distance between the anchor and positive representations.

Identity Constraint. In [13], it is found that the attribute of ID also provides strong supervision, and we introduce identity constraint $\mathcal{L}_{id}$ to supervise the joint embedding learning, by using cross-entropy loss function.

Center Constraint. The center constraint is introduced to minimize the intra-class variance, with its definition by:

$$\mathcal{L}_{cen} = \frac{1}{2}\sum_{k=1}^{N} ||x_k - c_{y_k}||_2^2 \tag{4}$$

where N is the total number of training samples, and x_k represents the feature vectors of the k-th training sample (face or voice), c_{y_k} denotes the center of the class and it will be updated every batch during the training process. Accordingly, the features of samples of the same class should be clustered in the same common subspace.

By combining the the ranking constraint, identity constraint and center constraint, the overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_r + \lambda_1 \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{cen} \tag{5}$$

2.3 Online Hard Negative Mining

During training, we randomly select one face image and one voice recording of the same person to construct anchor-positive pairs. To speed up the convergence of our network, we adopt online

hard negative mining scheme for training efficiency. For an anchor-positive pair, the hardest negative example which is the closest to the anchor example is selected to construct a hard triplet.

3 EXPERIMENTS

Dataset. The popular VoxCeleb [12] dataset is selected for face-voice association. The training, validate and test split are depicted in Table 1. In the experiments, the identities between the training and validation set (or testing set) are fully disjoint.

Face data. We extract video frames at a sampling rate of fps = 1, and employ MTCNN to detect facial landmarks from the extracted video frames. Accordingly, the cropped RGB face images of size 224×224×3 are obtained, and followed by preprocessing like [10].

Voice data. We separate the voice data from the original video and utilize voice activity detector (VAD) to eliminate the long silence period in the voice segments. Accordingly, 64-dimensional log mel-spectrograms are generated (window size: 25ms, hop size: 10ms) and followed by mean and variance normalization.

Table 1: Number statistics for the datasets.

	Train	Val	Test
# speaking face-tracks	105,600	12734	30496
# identities	901	100	250

Evaluation metrics. To evaluate the performance, Some standard metrics, including AUC value, accuracy (ACC) and mAP, are selected for quantitative comparison [5].

Implementation details. The algorithm is implemented with Pytorch, with the momentum and weight decay values setting at 0.9 and 0.0005, respectively. Meanwhile, a logarithmically decaying learning rate is initialized to 10^{-3} and decaying value set at 10^{-8}. The trade-off parameters are set at $\lambda_1 = 1, \lambda_2 = 0.001, \lambda_3 = \lambda_4 = 0.1$. The margin values are empirically set at $\alpha_1 = \alpha_2 = 0.6, \beta_1 = \beta_2 = 0.2$. The face subnetwork is initialized with pre-trained weights on the VGGFace2 [1] dataset, and the voice subnetwork is initialized with pre-trained weights on Voxceleb1.

3.1 Performance Analysis and Comparison

To evaluate the model effectiveness, various face-voice association tasks have been extensively tested, elaborated below.

Verification task: There are 5 testing groups based on gender, nationality and age. Similar to [10], **U** represents the unstratified group, **G** denotes that the test set is stratified by gender, and **N** represents that the test set is stratified by nationality. **A** denotes that the test set is stratified by age.

1:2 matching task: We divide the 1:2 matching task into two scenarios, from the face to voice and from voice to face. We name the first scenario as F-V 1:2 matching, where a face image and two audio clips are given, and the purpose is to determine which audio clip corresponds to the face image. Similarly, the second scenario is named as V-F 1:2 matching task, where an audio clip of a voice and two face images are given, and the goal is to determine which face image corresponds to the given voice clip.

1:N matching task: This task is an extension of 1:2 matching, in which the gallery now includes N-1 imposters. Note that, the model just needs to predict the only positive sample from N samples. This task is more challenging with the increase of number N.

Cross-modal retrieval task: This task is an extension of matching task, and one or more instances might match the given anchor.

Table 2 presents the various face-voice association results. Comparing with recent relevant works, it can be found that the proposed joint embedding learning method has yielded the improved verification, 1:2 matching and cross-modal retrieval performances. That is, the learned joint embeddings are discriminative enough to correlate the face and voice modalities. That is, the final outputs of the proposed network model not only can preserve much information about the gender, nationality and age, but also contain valuable information for identity analysis. Consequently, the proposed network model incorporates more capability to learn discriminative joint embedding for challenging face-voice matching task.

Figure 2 shows 1:N matching performance on two cross-modal matching scenarios. It can be found that the proposed network model has achieved the best results on different N values. That is, the joint embedding learning method is powerful to find the associations between faces and voices.

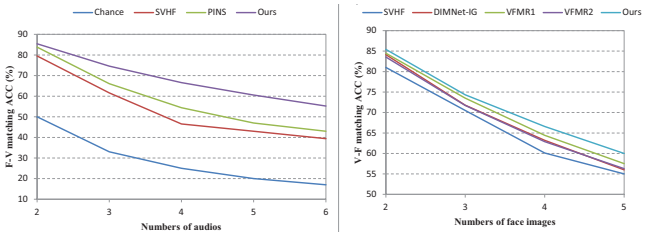

Figure 2: Comparison of 1: N matching performance

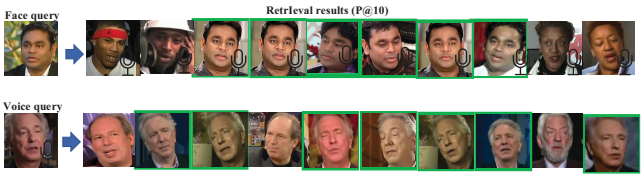

Figure 3: Some cross-modal retrieval examples.

Figure 3 shows some retrieval results with descending order of similarity from left to right, and we highlight the matching samples in green. It can be found that the matching samples account for a large part of top ranked samples, which shows that the proposed model is able to well associate the face and voices.

Further, we utilize t-SNE [8] method to visualize the learned embedding features. As shown in Figure 4, we randomly chose 10 people from the test set and visualize their face embeddings. It can be found that the learned embeddings belonging to the same identity always gather together, while the embeddings belonging to the different identities are far away. It indicates that the proposed network model exhibits high discriminability to learn the joint embeddings, which can well push representations of the same person closer while pulling those of different person away.

3.2 Ablation Study

First, we study the influence of different model configurations. As shown in Table 3, it can be found that the deeper structures, e.g., Inception-ResNet, has achieved better cross-modal verification performance. Within the proposed framework, the fully connected layers of both face and voice branches are shared with each other.

Table 2: Comparison with other models on verification task, 1:2 matching task, retrieval task.

Tasks	Verification (AUC)					F-V 1:2 Matching (ACC)				V-F 1:2 Matching (ACC)				Retrieval (mAP)	
Methods	**U**	**G**	**N**	**A**	**GNA**	**U**	**G**	**N**	**GN**	**U**	**G**	**N**	**GN**	**F-V**	**V-F**
SVHF [11]	-	-	-	-	-	79.50	63.40	-	-	81.00	63.90	-	-	-	-
Horiguchi's [4]	-	-	-	-	-	77.8	60.8	-	-	78.10	61.7	-	-	2.18	1.96
Kim's [6]	-	-	-	-	-	78.60	61.60	-	-	78.20	62.90	-	-	-	-
PINs [10]	78.5	61.1	77.2	74.9	58.8	83.80	-	-	-	-	-	-	-	-	-
DIMNet-I [13]	82.5	71.0	81.1	77.7	62.8	83.52	71.78	82.41	70.90	83.45	70.91	81.97	69.89	4.17	4.25
DIMNet-IG [13]	83.2	71.2	81.9	78.0	62.8	84.03	71.65	82.96	70.78	84.12	71.32	82.65	70.39	4.23	4.42
VFMR3 [15]	-	-	-	-	-	-	-	-	-	71.52	-	-	-	-	5.00
Ours	**85.03**	**73.22**	**84.44**	**79.77**	**65.07**	**85.42**	**73.52**	**84.48**	**71.11**	**85.18**	**74.29**	**83.97**	**70.70**	**6.19**	**6.75**

Table 3: Different configurations on verification task.

Config	Details	Val (AUC %)
Network	VGG-Face+VGG-Vox	78.15
	Inception-Resnet+DIMNet-voice	**85.49**
Last FC	Not shared	73.67
	Shared	**85.49**
Embedding size	128	84.69
	256	**85.49**
	512	85.20
Loss	Only identity loss	66.12
	Only center loss	50.49
	Only ranking loss	83.15
	Ranking loss + identity loss	83.94
	Ranking loss + center loss	83.81
	Full model	**85.49**

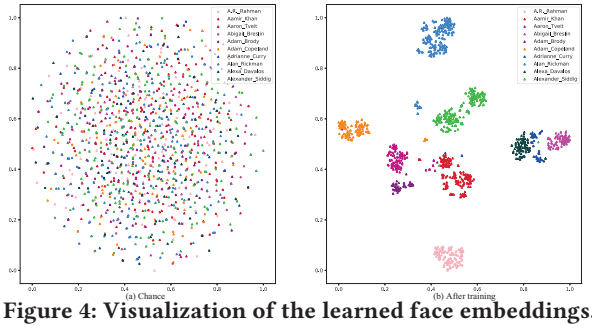

Figure 4: Visualization of the learned face embeddings.

To evaluate its efficiency, we also evaluate the performance without sharing the FC layers. It can be found that the verification performance without sharing scheme drops a lot, which indicates that the shared fully connected layers with the shared parameters are able to well correlate the heterogeneous face and voice.

Next, we train our model with different embedding size (i.e., 128, 256 and 512). It can be found that different embedding sizes have induced a bit different verification performance, and the size of 256 has yielded the better performances.

Moreover, we also verify the effectiveness of the ranking loss, the center loss and the identity loss. It can be observed that the learning model with only the identity loss or the center loss has resulted in a bit poor performance. By integrating the ranking loss, the proposed network model brings significant improvement on verification performance, and the proposed fully model with three constraints has yielded the best performances.

4 CONCLUSION

This paper has presented an efficient deep correlated model to map the face and voice into a shared discriminative embedding space, and the proposed model can well push the representations of the same person closer while pulling those of different person away. Accordingly, the derived cross-modal embeddings are beneficial for various face-voice association and cross-matching tasks, and the extensive experiments have shown its outstanding performances.

ACKNOWLEDGMENTS

The work is supported by National Science Foundation of China (Nos. 61672444, 61673185, 61876142, 61876068, and 61922066), State Key Lab. of ISN of Xidian University (No. ISN20-11), Quanzhou City Science & Technology Program of China (No. 2018C107R), Promotion Program for graduate student in Scientific research and innovation ability of Huaqiao University (No. 18014083018), IG-FNRA of HKBU with Grant: RC-FNRA-IG/18-19/SCI/03, and ITF of ITC of Hong Kong SAR under Project ITS/339/18.

REFERENCES

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *FG*. 67–74.
[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Interspeech*. 1086–1090.
[3] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).
[4] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. 2018. Face-voice matching using cross-modal embeddings. In *ACM MM*. 1011–1019.
[5] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable deep multi-modal learning for cross-modal retrieval. In *ACM SIGIR*. 635–644.
[6] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. 2018. On learning associations of faces and voices. In *ACCV*. 276–292.
[7] X. Liu, Z. Hu, H. Ling, and Y.M. Cheung. 2019. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 99 (2019), 10.1109/TPAMI.2019.2940446.
[8] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008), 2579–2605.
[9] Mitchell McLaren, Luciana Ferrer, Diego Castan, and Aaron Lawson. 2016. The Speakers in the Wild (SITW) speaker recognition database. In *Interspeech*. 818–822.
[10] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Learnable pins: Cross-modal embeddings for person identity. In *ECCV*. 71–88.
[11] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*. 8427–8436.
[12] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).
[13] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, Bhiksha Raj, and Rita Singh. 2019. Disjoint mapping network for cross-modal matching of voices and faces. In *ICLR*.
[14] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*. 499–515.
[15] Chuyuan Xiong, Deyuan Zhang, Tao Liu, and Xiaoyong Du. 2019. Voice-Face Cross-modal Matching and Retrieval: A Benchmark. *arXiv:1911.09338* (2019).