计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20200547
59(3):694-705,2022

结合双流网络和双向五元组损失的跨人脸-语音匹配

柳 欣[1,2,3] 王 锐[1,3] 钟必能[4] 王楠楠[2]
[1](华侨大学计算机科学与技术学院 福建厦门 361021)
[2](综合业务网理论及关键技术国家重点实验室(西安电子科技大学) 西安 710071)
[3](厦门市计算机视觉与模式识别重点实验室(华侨大学) 福建厦门 361021)
[4](广西师范大学计算机科学与信息工程学院 广西桂林 541004)
(xliu@hqu.edu.cn)

Cross Face-Voice Matching via Double-Stream Networks and Bi-Quintuple Loss

Liu Xin[1,2,3], Wang Rui[1,3], Zhong Bineng[4], and Wang Nannan[2]
[1](*College of Computer Science and Technology, Huaqiao University, Xiamen, Fujian* 361021)
[2](*State Key Laboratory of Integrated Services Networks (Xidian University), Xi'an* 710071)
[3](*Xiamen Key Laboratory of Computer Vision and Pattern Recognition (Huaqiao University), Xiamen, Fujian* 361021)
[4](*School of Computer Science and Information Engineering, Guangxi Normal University, Guilin, Guangxi* 541004)

Abstract Facial information and voice cues are the most natural and flexible ways in human-computer interaction, and some recent researchers are now paying more attention to the intelligent cross-modal perception between the face and voice modalities. Nevertheless, most existing methods often fail to perform well on some challenge cross-modal face-voice matching tasks, mainly due to the complex integration of semantic gap and modality heterogeneity. In this paper, we address an efficient cross-modal face-voice matching network by using double-stream networks and bi-quintuple loss, and the derived feature representations can be well utilized to adapt four challenging cross-modal matching tasks between faces and voices. First, we introduce a novel modality-shared multi-modal weighted residual network to model the face-voice association, by embedding it on the top layer of our double-stream network. Then, a bi-quintuple loss is newly proposed to significantly improve the data utilization, while enhancing the generalization ability of network model. Further, we learn to predict identity (ID) of each person during the training process, which can supervise the discriminative feature learning process. As a result, discriminative cross-modal representations can be well learned for different matching tasks. Within four different cross-modal matching tasks, extensive experiments have shown that the proposed approach performs better than the state-of-the-art methods, by a large margin reaching up to 5%.

Key words face-voice associations; cross-modal perception; double-stream networks; bi-quintuple loss; weighted residual network

收稿日期:2020-07-14;**修回日期**:2021-03-09

基金项目:国家自然科学基金项目(61673185,61922066,61972167);综合业务网理论及关键技术国家重点实验室基金项目(ISN20-11);福建省自然科学基金项目(2020J01084);之江实验室开放课题(2021KH0AB01)
This work was supported by the National Natural Science Foundation of China (61673185, 61922066, 61972167), the Project of State Key Laboratory of Integrated Services Networks (ISN20-11), the Natural Science Foundation of Fujian Province (2020J01084), and the Zhejiang Laboratory (2021KH0AB01).

摘　要　面部视觉信息和语音信息是人机交互过程中最为直接和灵活的方式,从而基于智能方式的人脸和语音跨模态感知吸引了国内外研究学者的广泛关注.然而,由于人脸-语音样本的异质性以及语义鸿沟问题,现有方法并不能很好地解决一些难度比较高的跨人脸-语音匹配任务.提出了一种结合双流网络和双向五元组损失的跨人脸-语音特征学习框架,该框架学到的特征可直接用于4种不同的跨人脸-语音匹配任务.首先,在双流深度网络顶端引入一种新的权重共享的多模态加权残差网络,以挖掘人脸和语音模态间的语义关联;接着,设计了一种融合多种样本对构造策略的双向五元组损失,极大地提高了数据利用率和模型的泛化性能;最后,在模型训练中进行ID分类学习,以保证跨模态表示的可分性.实验结果表明,与现有方法相比,能够在4个不同跨人脸-语音匹配任务上取得效果的全面提升,某些评价指标效果提升近5%.

关键词　人脸-语音关联;跨模态感知;双流网络;双向五元组损失;加权残差网络

中图法分类号　TP18; TP391

语音和视觉信息是人们相互交流的重要载体,也是人机交互过程中最为直接和灵活的方式.心理学中著名的"麦格克效应"(McGurk effect)[1]表明,大脑在感知语音的过程中,人脸信息和语音信息会相互作用.同时,大量神经认知科学的研究表明,人脸信息和语音信息有着相同的神经认知通路[2].在日常生活中,当人们在给好友打语音电话时,虽然只接收到了对方的语音信息,但脑海中会不自觉地浮现出对方的人脸信息,即我们的大脑可以自动地将接收到的语音信息与之前已经存储好的人脸信息进行语义关联.

上述现象和研究表明,个体人脸信息和语音信息之间是存在明显关联特性的.受此启发,人们已逐渐认识到语音特征与视觉特征之间关联的重要性,并进行了多方面的跨模态匹配研究,如跨人脸-语音生物特征匹配、说话人标注以及跨人脸-语音检索等[3-5].因此,有效的人脸-语音相关性挖掘和跨模态匹配研究能够促进认知科学和人工智能技术创新实践的发展,具有重要的现实意义,有着广阔的应用前景.近年来,基于文本和图像的跨媒体检索,受到了国内外研究学者的广泛关注,但基于面部信息与语音信息的跨模态匹配和语义关联挖掘研究较为匮乏.

据文献研究,现有的挖掘人脸信息和语音信息之间关联的方法大致可以概括为2类:1)基于浅层特征相关性学习的方法;2)基于深度兼容性特征学习的方法.具体地,浅层匹配学习方法一般使用子空间相关性学习的方法进行人脸和语音的语义相似性映射挖掘,从而达到缩小面部特征与语音特征间的语义鸿沟的目的.深度特征学习方法旨在通过多层非线性网络结构来实现复杂特征表达能力的逼近,进而实现人脸和语音特征的跨模态语义关联.然而,人脸和语音信息的复杂多样性和非静态性加大了不同模态间潜层语义关联的抽取难度.为满足实际应用需求,现有人脸-语音语义关联方法在相关跨模态匹配方面的效果还需进一步提升.

基于人脸-语音的相关性挖掘和跨模态匹配问题的研究尚为一项新颖的课题,其智能语义关联研究仍处于早期发展阶段[6],并且现有方法或多或少存在一些挑战,包括3方面:1)面部和语音底层特征因维数不同、性质和属性不同,使得彼此之间无法直接参与计算,进而带来了语义表征的差异性和不可比性;2)针对人脸和语音特征的异构性,目前仍缺乏有效方法解决低层特征和高层语义之间存在的语义鸿沟问题;3)现有的异构特征学习和关联性学习结合的不够紧密,从而导致高层一致性语义挖掘的表征学习不够充分.此外,本文通过文献调研发现,大多数跨人脸-语音匹配方法工作只呈现了部分匹配任务评测结果,而其他多样性匹配评测任务还有待挖掘.

针对上述挑战,本文提出了一种基于双向五元组损失的跨人脸-语音特征学习框架,生成的跨模态表示在所有跨模态匹配任务上进行全面了评估测试.首先,本文采用双流架构的网络学习跨人脸-语音特征表示.传统的双流网络采用2条并行且独立的分支处理多模态数据,不同模态之间缺少交互,因而很难学习出高质量的语义特征.为解决特征异构问题,本文在双流网络的顶端引入了一种新的多模态加权残差网络,并采用权重共享策略,以挖掘模态间关联,生成模态不变的跨模态表示;其次,现有的基于距离度量损失的方法中采用的样本对构造策略往往没有挑选出合适样本对,也没有充分利用batch中的数据,使得很多有益于训练的样本未能参与训练,极大地限制了模型的泛化性能.为解决训练样本

不足问题,本文提出了多种有效的样本对构造策略,并基于这些策略提出了多种表现形式的三元组损失,这些三元组损失一起构成一种新的双向五元组损失(bi-quintuple loss, Bi-Q loss).通过优化该损失,可以促使更多有益于训练的人脸样本及语音样本参与训练,进而学到更好的跨模态表示;最后,为了保证人脸特征和语音特征在共享语义空间的可分性,本文在特征层后面引入了一个全连接层进行身份(identity, ID)分类学习,实验表明结合 ID 损失与双向五元组损失可以促进模型的有效收敛,鲁棒性较好.本文工作的贡献主要包括 3 个方面:

1) 提出了一个端到端的跨人脸-语音特征学习框架,该框架在双流网络的顶端引入了一种新的权重共享多模态加权残差网络,可以有效挖掘模态间关联;

2) 设计多种样本对构造策略并提出双向五元组损失,极大地提高了数据利用率和模型泛化性能;

3)本文方法具有较强的扩展性和一般性.相比现有方法,本文学习框架在 4 个不同的跨人脸-语音关联任务上,其跨模态匹配各项指标上几乎取得了全面提升,某些指标上的提升近 5%.

1 相关工作

人类面部视觉信息和语音信息是人机交互过程中最为直接和灵活的方式,从而基于人脸和语音的关联性挖掘及其跨模态协同感知吸引了国内外研究学者的广泛关注.早期针对人脸和语音的相关性挖掘主要是基于浅层特征相关性学习的方法.例如,Hasan 等人[7]通过认知学的角度利用功能性磁共振成像分析了人脸和语音在身份鉴别上的潜在关联特性;类似地,针对人脸和语音 2 种不同模态特征之间存在的"语义鸿沟"问题,Li 等人[8]通过跨模态因子分析(cross-modal factor analysis)来缩小人脸与语音特征间的语义鸿沟,接着利用典型相关分析法(canonical correlation analysis, CCA)进一步关联 2 种模态特征集,从而实现说话人的跨视听媒体数据互标注;Chetty 等人[9]通过潜在语义分析和 CCA 方法对人脸和语音生物特征进行跨模态关联,从而使得身份验证系统达到了较好的预防反欺骗性(anti-spoofing)攻击的目的;Chakravarty 等人[10]利用跨模态监督学习方法(cross-modal supervision)对语音信息进行当前说话人检测,取得了鲁棒性的结果.研究发现,这些浅层特征相关性学习方法缺乏从非线性异构特征中提取有意义跨模态关联的本质特征能力,从而导致其相应跨模态关联匹配效果有所欠缺.

近年来,多模深度学习可以有效对多模态数据逐级提取从低层到高层的语义特征,展现出了强大本质特征学习的能力.据文献研究,现有挖掘人脸信息和语音信息之间关联的深度学习方法大致可以分为 2 类:1)基于分类损失的学习方法;2)基于距离度量损失学习的方法.基于分类损失的学习方法通常把跨人脸-语音跨模态匹配问题定义为分类问题.典型代表为 Nagrani 等人[3]提出的多分支卷积神经网络(convolutional neural network, CNN)结构方法,该方法首先采用多分支 CNN 分别提取人脸图片和语音数据的特征,接着将提取好的人脸特征和语音特征拼接起来,输入到 softmax 层以获得分类概率.该模型然在 1∶2 匹配任务上的表现可以媲美人类,但由于是针对特定匹配任务设计的,需要调整子网络数目才能应用于其他任务,模型灵活性欠佳;近期,Wen 等人[5]采用不相交映射网络(disjoint mapping network, DIMNet)来监督跨模态表示的学习,在 1∶2的跨模态匹配任务上获得了较好的准确率,超过了人类主观水平,但在一些挑战性较高的匹配任务上,如 1∶N 匹配及跨模态检索,该模型的表现有待提高.

基于距离度量损失的学习方法通常利用多模态神经网络将人脸样本和语音样本映射到欧氏空间,并通过优化网络距离度量损失,使得同一个体的人脸样本和语音样本对应的特征表达在欧氏空间中的距离足够近,不同人的人脸样本和语音样本对应的特征表达在欧氏空间中的距离足够远[11].Nagrani 等人[12]通过刻画个人身份节点方式(person identity nodes, PINs)来描述身份,并采用对比损失来约束正负样本对之间的距离,该方法提出了一种基于 Curriculum 的策略构造样本对,但这种策略构造的负样本对中可能存在噪声数据;Xiong 等人[13]采用了三元组损失来引导人脸-语音跨模态表示的学习,然而该方法只构造了跨模态三元组样本,忽略了许多其他类型有益于训练的三元组样本,其跨模态关联效果还有所欠缺.

2 跨人脸-语音特征学习框架

如图 1 所示,本文提出的结合双流深度网络和双向五元组损失的跨人脸-语音特征学习框架采用

了常见的双流网络架构,包含人脸和语音 2 个分支网络.其中,人脸子网络和语音子网络的权重是各自独立的,用来提取模态特有的特征;双流网络顶端的多模态加权残差网络的权重由 2 个模态共享,用来挖掘模态间语义关联,生成模态不变的跨模态表示;双向五元组损失用于进一步挖掘模态间关联,提高数据利用率和模型泛化性能;ID 损失用于保证跨模态表示的可分性,促进模型收敛.

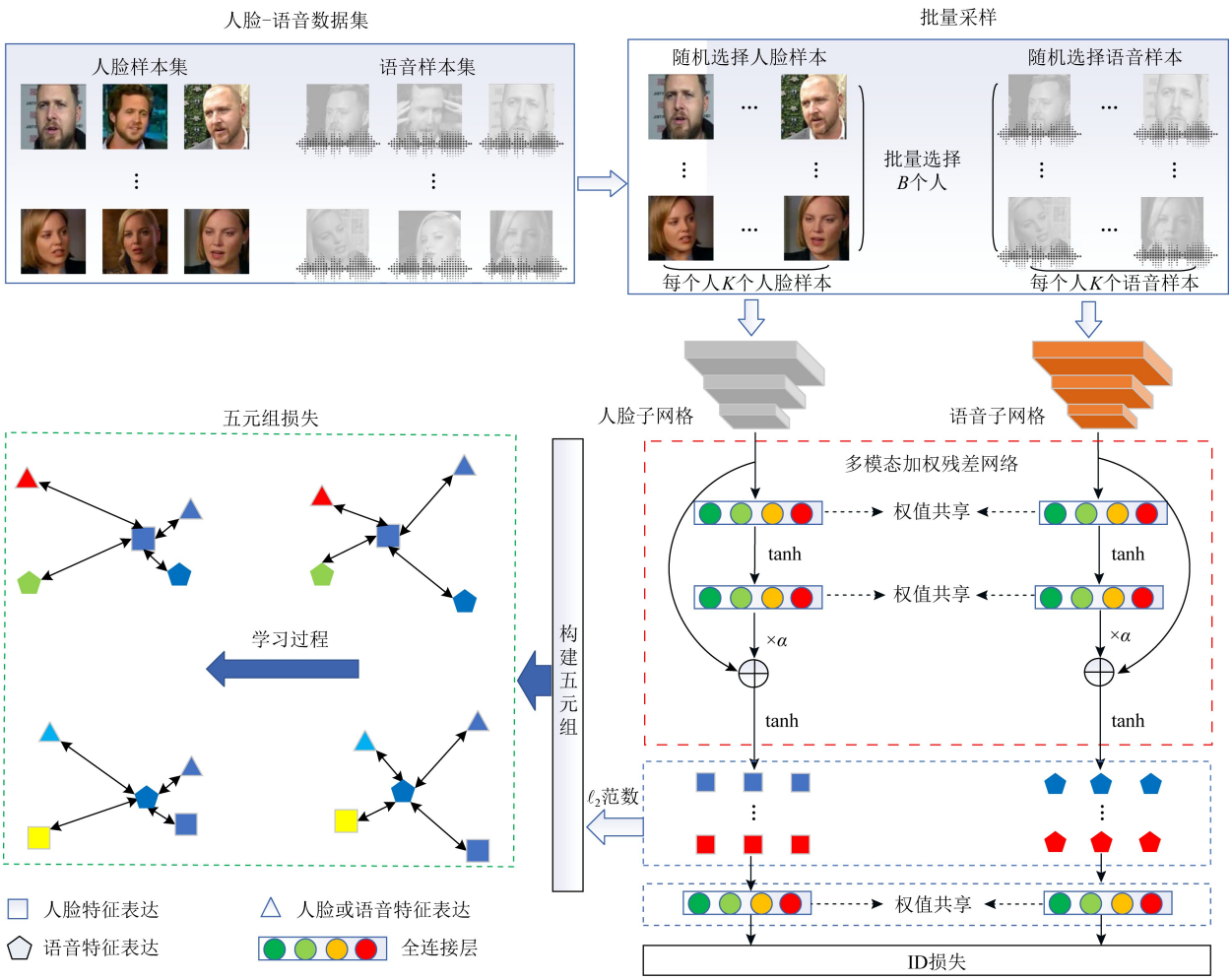


Fig. 1　The overall framework of the proposed cross face-voice matching method

图 1　本文提出跨人脸-语音匹配方法的整体学习框架

2.1　形式化定义

本文用 $X^f=\{x_i^f\}_{i=1}^N$ 表示人脸数据,$X^v=\{x_i^v\}_{i=1}^N$ 表示语音数据集,其中,x_i^f 表示第 i 张人脸图片,x_i^v 表示第 i 条语音数据,N 表示样本总数.X^f 和 X^v 有着共同的标签集 $Y=\{y_i\}_{i=1}^N$,$y_i\in\{1,2,\cdots,c\}$,c 表示类别总数.进一步,人脸子网络和语音子网络分别用 $\phi^f(\cdot)$ 和 $\phi^v(\cdot)$ 表示,$f_i=\phi^f(x_i^f)$ 表示人脸图片 x_i^f 对应的高层次人脸特征表示,$v_i=\phi^v(x_i^v)$ 表示语音数据 x_i^v 对应的高层次语音特征表示.

2.2　多模态加权残差网络

针对 2 种模态的关联性挖掘,双流深度网络能够有效地进行异构特征的兼容性学习.在跨模态特征表示学习中,文献[14]为了挖掘文本模态和图像模态之间的关联并生成模态不变的跨模态表示,选择在双流网络的顶端引入了权重共享的单层全连接.然而,一方面,单层全连接拟合能力有限,同时没办法解决非线性映射问题,因而挖掘到的模态间关联可能极为有限.另一方面,单纯地增加全连接层数有时会使得网络训练和优化起来越来越复杂和困难.

为解决上述问题,受残差网络[15]思想的启发,本文在双层全连接网络的输入层与输出层之间引入了一种新的加权残差连接,并采用权重共享策略保证生成的跨模态表示形式在同一个特征表示子空间中.本文提出的这种网络结构简称为多模态加权残

差网络(multi-modal weighted residual network, MWRN).实验结果表明,该网络可以有效地加强模态共享信息、挖掘模态间关联,促使模型生成更好的跨模态表示.

该网络的结构如图 2 所示,其中第 1 层全连接的权重用 W_1 表示,第 2 层全连接的权重用 W_2 表示,令 $\mathcal{F}(x)=W_2\sigma(W_1x)$,$x$ 表示输入,σ 表示激活函数 tanh,则对于输入人脸高层表示 f_i 和语音高层表示 v_i,其对应的最终输出可分别表示为

$$f_i^c=\sigma(f_i+\eta\mathcal{F}(f_i)), \tag{1}$$

$$v_i^c=\sigma(v_i+\eta\mathcal{F}(v_i)), \tag{2}$$

其中,η 为缩放因子,在网络中是一个可学习的参数.根据文献[16]中的设计,在网络训练中缩放因子初始值设为 0,用来避免训练初始阶段出现过分的梯度波动造成的不稳定,从而使得模型在训练初期更加的稳定,促进整个网络的训练平稳性和鲁棒性.

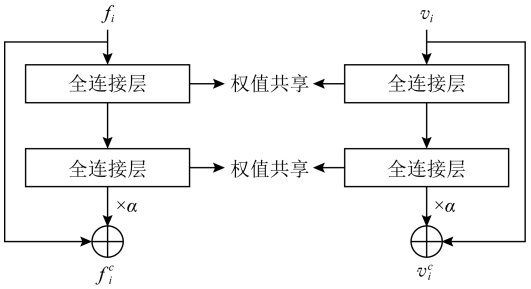


Fig. 2 Multi-modal weighted residual network

图 2 多模态加权残差网络

为方便后续网络描述的便捷性,若把人脸子网络及多模态加权残差网络看作一个整体并记为 $\mathcal{F}^f(\cdot)$,则第 i 张人脸图片对应的语义特征可以表达为 $f_i^c=\mathcal{F}^f(x_i^f)$;若把语音子网络及多模态加权残差网络看作一个整体并记为 $\mathcal{F}^v(\cdot)$,则第 i 条语音数据对应的语义特征 $v_i^c=\mathcal{F}^v(x_i^v)$.

2.3 双向五元组损失

本文提出了一种双向五元组损失函数进行人脸-语音的语义关联性学习约束,该损失由多个改进的三元组损失构成.具体地,三元组损失是一种常见的距离度量损失,其形式化定义为

$$\mathcal{L}_{\text{tri}}=\sum_{\substack{a,p,n\\ y_a=y_p\neq y_n}}(d_{a,p}-d_{a,n}+m)_+, \tag{3}$$

其中,a 表示固定(anchor)样本,p 表示与 a 属于同一类别的正(positive)样本,n 表示与 a 属于不同类别的负(negative)样本,$d_{a,p}$ 表示固定样本 a 与正样本 p 对应的特征表达之间的距离,$d_{a,n}$ 表示固定样本 a 与负样本 n 对应的特征表达之间的距离,m 表示最小间隔(margin).针对每个三元组$\langle a,p,n\rangle$,三元组损失的优化目标是让 $d_{a,p}$ 尽可能小,$d_{a,n}$ 尽可能大,并且要让 $d_{a,p}$ 与 $d_{a,n}$ 之间有一个最小的间隔 m,T 表示三元组样本集合,形式化描述为

$$d_{a,p}+m<d_{a,n},\ \forall\langle a,p,n\rangle\in T. \tag{4}$$

然而,若要在整个训练集上构造三元组,随着训练集样本数目增多,可能的三元组数量将呈立方级增长.因此,当训练集非常大时,训练将会非常耗时.同时,随着训练的深入,大量的简单的三元组不具有判别性,并对模型表现的提升毫无贡献.为解决此问题,文献[17]提出了 TriHard 损失,它是一种采用基于批量(batch)的在线难样本采样策略的三元组损失.对于每个训练 batch,随机挑选 B 个 ID,每个 ID 随机挑选 K 个样本,则每个 batch 中有 $B\times K$ 个样本,其样本集合记为 X_{batch}.对于每个样本 a,挑选一个最难正样本 p 和最难负样本 n,最难正样本 p 是 X_{batch} 中与样本 a 属于同一 ID 的样本中距离样本 a 最远(语义最不相关)的样本,最难的负样本 n 是 X_{batch} 中与样本 a 属于不同 ID 的样本中距离样本 a 最近(语义最相关)的样本.TriHard 损失的形式化定义为

$$\mathcal{L}_{\text{TH}}=\frac{1}{B\times K}\sum_{a\in X_{\text{batch}}}(\max_{p\in P}d_{a,p}-\min_{n\in Q}d_{a,n}+m)_+, \tag{5}$$

其中,P 表示 X_{batch} 中与 a 的 ID 相同的样本集合,Q 表示 X_{batch} 中与 a 的 ID 不同的样本集合,m 表示最小间隔.可以看到,由于 TriHard 损失是在每个 batch 上构造三元组,而不是在整个训练集上,因而大大提高了模型的采样效率与训练效率,同时,借助难样本的重点性采样策略,简单的样本将被过滤掉,因而提高了模型的鲁棒性.为了提高模型训练效率,同时促使足够多的有益于训练的三元组样本参与训练,本文提出了多种样本对构造策略,进而提出了跨模态 TriHard 损失以及混合模态 TriHard 损失.采用双向训练策略的跨模态 TriHard 损失和混合模态 TriHard 损失一起构成一种新的双向五元组损失.

本文采用基于身份 ID 的采样策略构建 batch 中的样本数据,进而利用 batch 中的样本构建双向五元组.首先,每一个 batch 会随机挑选 B 个 ID,对于每个 ID,从人脸数据集 X^f 中随机挑选 K 张人脸图片构成小批量人脸数据集 X_{batch}^f,$X_{\text{batch}}^f=\{x_j^f\}_{j=1}^{B\times K}\subseteq X^f$;从语音数据集 X^v 中随机挑选 K 条语音数据构

成小批量语音数据集 X^v_{batch},$X^v_{\text{batch}}=\{x^v_j\}_{j=1}^{B\times K}\subseteq X^v$.同时,为了保证模型的稳定收敛,本文对 batch 中所有样本特征表达都进行了 ℓ_2 归一化,其表示形式为

$$\ell_2(x)=\frac{x}{\|x\|_2}. \tag{6}$$

2.3.1　跨模态 TriHard 损失

传统的 TriHard 损失仅应用于单模态问题,为解决跨人脸-语音匹配问题,本文提出了一种跨模态 TriHard 损失.接下来介绍跨模态 TriHard 三元组的构建过程以及跨模态 TriHard 损失的定义.具体地,以 batch 中的任意 1 张人脸图片 x^f_{anc} 为固定样本,构建跨模态 TriHard 三元组 $(x^f_{\text{anc}},x^v_{\text{pos}},x^v_{\text{neg}})$,其中 x^v_{pos},x^v_{neg} 和距离的定义形式为

$$x^v_{\text{pos}}=\underset{x^v_j\in X^v_{\text{batch}},y_{\text{anc}}=y_j}{\arg\max}\ d(x^f_{\text{anc}},x^v_j), \tag{7}$$

$$x^v_{\text{neg}}=\underset{x^v_j\in X^v_{\text{batch}},y_a\neq y_j}{\arg\min}\ d(x^f_{\text{anc}},x^v_j), \tag{8}$$

$$d(x^f_{\text{anc}},x^v_j)=\|\ell_2(\mathcal{F}^f(x^f_{\text{anc}}))-\ell_2(\mathcal{F}^v(x^v_j))\|_2. \tag{9}$$

跨模态 TriHard 损失定义为

$$\ell_{\text{CTH}}(x^f_{\text{anc}},x^v_{\text{pos}},x^v_{\text{neg}})=$$
$$[d(x^f_{\text{anc}},x^v_{\text{pos}})-d(x^f_{\text{anc}},x^v_{\text{neg}})+m]_+. \tag{10}$$

跨模态 TriHard 损失可以有效缩减人脸模态和语音模态数据之间的“异构鸿沟”,从而使得同一个 ID 的人脸样本和语音样本对应的特征表达之间的距离足够近,不同 ID 的人脸样本和语音样本对应的特征表达之间的距离足够远.

2.3.2　混合模态 TriHard 损失

跨模态 TriHard 损失中,正负样本是从与固定样本模态异构的小批量数据集中采样得到,因而只能构造跨模态的 TriHard 三元组.本文接着采用正负样本模态扩展策略,提出了一种新的混合模态 TriHard 损失.混合模态 TriHard 损失是跨模态 TriHard 损失的扩展版本,它将跨模态 TriHard 损失中的正样本和负样本的采样范围扩展至 2 个模态,因而可以构造更多表现形式更丰富的 TriHard 三元组.同样,以 X^f_b 中的任意一张人脸图片 x^f_{anc} 为固定样本构建混合模态三元组 $(x^f_{\text{anc}},x_{\text{pos}},x_{\text{neg}})$ 的样本,其中:

$$x_{\text{pos}}=\underset{\substack{x_j\in X^f_{\text{batch}}\cup X^v_b,\\ y_{\text{anc}}=y_j}}{\arg\max}\ d(x^f_{\text{anc}},x_j), \tag{11}$$

$$x_{\text{neg}}=\underset{\substack{x_j\in X^f_{\text{batch}}\cup X^v_b,\\ y_{\text{anc}}=y_j}}{\arg\min}\ d(x^f_{\text{anc}},x_j). \tag{12}$$

值得注意的是,由于正负样本是在 2 个模态数据集上采样得到的,因而 x_{pos} 和 x_{neg} 可能来自人脸模态,也可能来自语音模态,因而称之为混合模态三元组.以这种方式构建的三元组称之为有效的混合模态 TriHard 三元组,满足条件:

$$d(x^f_{\text{anc}},x_{\text{pos}})>d(x^f_{\text{anc}},x_{\text{neg}}). \tag{13}$$

有效的混合模态 TriHard 三元组有 4 种表现形式,如图 3 所示.其中,方形表示人脸样本,圆形表示语音样本,不同的颜色深浅代表不同的 ID.从图 3 中可以看出:第 1 种三元组是 2.3.1 节提到的跨模态 TriHard 三元组,固定样本来自人脸模态,正负样本均来自语音模态;第 2 种三元组中固定样本与负样本来自同一模态,与正样本来自不同模态;第 3 种三元组中固定样本与正样本模态相同,与负样本模态不同;第 4 种三元组中固定样本及正负样本均来自同一模态,是一种模态内 TriHard 三元组.综合这些不同形式的三元组,混合模态 TriHard 损失的形式化定义为

$$\ell_{\text{MTH}}(x^f_{\text{anc}},x_{\text{pos}},x_{\text{neg}})=$$
$$[d(x^f_{\text{anc}},x_{\text{pos}})-d(x^f_{\text{anc}},x_{\text{neg}})+m]_+. \tag{14}$$

混合模态 TriHard 损失综合考虑了模态间和模态内的多种距离约束,极大地提高了模型的泛化能力.

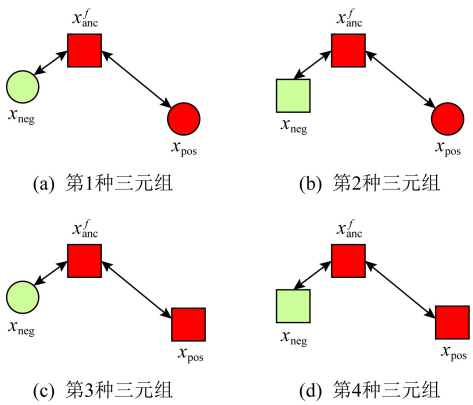


Fig. 3　Mixed-modal triplets with different forms

图 3　不同表现形式的混合模态三元组

2.3.3　双向训练策略

前面 2.3.1 节和 2.3.2 节在构建 TriHard 三元组的过程中,默认采用人脸样本作为固定样本,但实际上人脸样本和语音样本并没有角色上的区别,同时,很多跨模态任务都是双向的,为了学习更适合于最终任务的跨模态表示,同时为了更多的有效的三元组样本得以参与训练,本文基于双向训练策略提出了双向五元组损失.

考虑到模态间存在巨大的鸿沟,同时混合模态 TriHard 三元组中可能只存在少量的跨模态 TriHard 三元组,因而,构建跨模态 TriHard 三元组

是非常必要的,即混合模态 TriHard 三元组并不能替代跨模态 TriHard 三元组,而是跨模态 TriHard 三元组的重要补充.特别地,混合模态 TriHard 三元组(x^f_{anc},x_{pos},x_{neg})和跨模态 TriHard 三元组(x^f_{anc},x^v_{pos},x^v_{neg})可以合并成一个以人脸样本 x^f_{anc} 为固定样本的五元组组合形式,标记为(x^f_{anc},x^v_{pos},x^v_{neg},x_{pos},x_{neg}),其损失定义为

$$\ell_Q(x^f_{anc},x^v_{pos},x^v_{neg},x_{pos},x_{neg})=\ell_{CTH}(x^f_{anc},x^v_{pos},x^v_{neg})+\ell_{MTH}(x^f_{anc},x_{pos},x_{neg}). \quad (15)$$

采用双向训练策略可以得到以任一语音样本 x^v_{anc} 为固定样本的跨模态 TriHard 三元组(x^v_{anc},x^f_{pos},x^f_{neg})及混合模态 TriHard 三元组(x^v_{anc},$\bar{x}_{pos}$,$\bar{x}_{neg}$),进而得到以 x^v_{anc} 为固定样本的五元组损失:

$$\ell_Q(x^v_{anc},x^f_{pos},x^f_{neg},\bar{x}_{pos},\bar{x}_{neg})=\ell_{CTH}(x^v_{anc},x^f_{pos},x^f_{neg})+\ell_{MTH}(x^v_{anc},\bar{x}_{pos},\bar{x}_{neg}). \quad (16)$$

整个 batch 的双向五元组损失定义为

$$\mathcal{L}_{Bi\text{-}Q}=\frac{1}{B\times K}\sum_{i=1}^{B\times K}\ell_Q(x^{f_i}_{anc},x^v_{pos},x^v_{neg},x_{pos},x_{neg})+\frac{1}{B\times K}\sum_{j=1}^{B\times K}\ell_Q(x^{v_j}_{anc},x^f_{pos},x^f_{neg},\bar{x}_{pos},\bar{x}_{neg}). \quad (17)$$

双向五元组损失可以同时优化 2 个方向的五元组距离度量损失,从而可以极大地提高人脸-语音跨模态表示的鲁棒性和模型的泛化能力.

2.4 ID 损失

为确保人脸样本和语音样本在嵌入到公共表示空间之后的模态内判别性得以保留,即保持良好的可分性,本文提出了基于 ID 损失的约束.每一个个体的身份 ID 可以看作一个类别,通过在特征层后面附上一个全连接层 $\phi_C(\cdot)$ 来实现对语义特征所属的身份类别的预测.具体地,ID 损失的定义为

$$\mathcal{L}_{ID}=\frac{1}{B\times K}\sum_{i=1}^{B\times K}l(\phi_C(f^c_i),y_i)+\frac{1}{B\times K}\sum_{i=1}^{B\times K}l(\phi_C(v^c_i),y_i), \quad (18)$$

其中,$l(\cdot,\cdot)$表示交叉熵损失函数,f^c_i 和 v^c_i 分别为第 i 个人脸样本和语音样本,y_i 表示相应的身份类别标签.同时,在后面的消融分析实验中发现,单独采用双向五元组损失训练时模型比较难以收敛,结合了 ID 损失后模型训练时很快就开始收敛并趋于稳定,因而结合 ID 损失和双向五元组损失可以加速模型收敛,泛化能力较强.

2.5 模型训练

本文提出方法的整体损失函数形式为

$$\mathcal{L}=\mathcal{L}_{Bi\text{-}Q}+\mathcal{L}_{ID}. \quad (19)$$

本文采用 mini-batch 的训练方式,mini-batch 可以在训练过程中引入随机性,同时可以提升模型训练速度,每个 batch 会随机挑选 16 个 ID,接着每个 ID 随机挑选 4 张人脸图片和 4 条语音数据.同时,本文采用结合权重衰减和动量技术的随机梯度下降(stochastic gradient descent, SGD)方法来优化模型,其中,一方面,权重衰减($weight\_decay=0.000\,5$)用来调节模型复杂度对损失函数的影响,以防止过拟合;另一方面,动量($momentum=0.9$)用来加速模型收敛过程.本文采用了一种动态的学习率调整策略,学习率会随着训练轮数的增加而衰减,训练共需 50 轮,训练过程中学习率将从初始学习率 10^{-3} 衰减到 10^{-8}.

3 实验与结果

为了充分评估本文所提出方法的有效性和鲁棒性,本文在公开的 Voxceleb1 音视频数据集上进行实验测试,下面具体介绍实验详情.

3.1 数据集介绍

Voxceleb1[18]是公开的大规模音视频数据集,由上传到 YouTube 的采访视频中提取的 1 251 个名人的音视频短片组成.该数据集总计包含 10 万多条音频、2 万多条视频.文献[12]采用 SyncNet[19]方法从该数据集中提取出超过 10 万条说话人人脸轨迹片段.在本文实验中,实验所选取的数据集是由该文献作者处理好并在其官网公开发布的 Voxceleb1 数据集,其数据集划分方式也与作者在文献[12]中的描述相同.

3.2 实现细节

本文方法依据 Pytorch 深度学习框架进行配置和实现,其中双向五元组损失中的间隔设置为 $m=0.6$,人脸子网络采用 Inception-ResNet-v1 模型,并用标准的 VGGFace2[20]数据集上的预训练权重进行初始化,输入的人脸图片采用了与 PINs[12]方法和 SSNet[21]相同的预处理技术;语音子网络采用与 DIMNet-Voice[5]方法相同的结构,并使用在 Voxceleb1 上的预训练权重进行初始化.人脸特征和语音特征输出维度为 256.

3.3 实验场景及评价指标

为了全面验证本文方法的有效性,本文设计了 4 种不同的跨人脸-语音匹配任务,分别为跨模态验证、1∶2 匹配、1∶N 匹配以及跨模态检索任务.

1) 跨模态验证

跨模态验证是指给出1张人脸图片和1条语音数据,判断该数据对是否属于同一个人,其评价标准采用AUC值作为量化指标.

2) 1:2匹配

跨模态1:2匹配是指给出1张人脸图片和2条语音数据,2条语音数据中只有1条与给定的人脸图片属于同一个人,模型的任务是预测与给定的人脸图片匹配的那条语音数据的位置编号,本文称之为人脸到语音(face to voice, F-V)下1:2匹配;类似地,可以定义语音到人脸(voice to face, V-F)下1:2匹配.F-V下1:2跨模态匹配和V-F下1:2跨模态匹配均采用百分制匹配准确率作为评价指标,匹配准确率计算方式与文献[2]相同.

3) 1:N匹配

跨模态1:N匹配是1:2匹配任务的扩展版本,它将不匹配的样本数量扩增至N个.随着N的增大,任务的难度也将不断增大.同样地,1:N匹配也有2个实验场景,均采用匹配准确率作为评价指标.

4) 跨模态检索

跨模态检索任务中可以有一个或多个样本与给定的查询样本匹配,因而匹配任务难度更大.本文采用随机结果(Chance)作为参照依据,并利用标准的百分制平均准确度(mAP)作为该任务的评价指标.

3.4　实验对比结果

为了验证本文方法生成的跨模态表示的有效性,本文将其应用于3.3节中提到的4种任务.

1) 跨模态验证

在跨模态验证任务上,本文方法与现有方法的实验结果对比如表1所示.其中,"U"分组中是没有分层的测试数据,"G"分组中每个测试对中的人脸图片和语音数据来自性别相同的2个人,"N"分组中每个测试样本对中的人脸图片和语音数据来自国籍相同的2个人,"A"分组中每个测试对中的人脸图片和语音数据来自年龄相同的2个人,"GNA"分组中每个测试对中的人脸图片和语音数据来自性别、国籍、年龄均相同的2个人.

Table 1　Comparison with Other Methods on Verification Task

表1　与现有方法在跨模态验证任务上的实验结果对比 %

方法	U	G	N	A	GNA
PINs[12]	78.5	61.1	77.2	74.9	58.8
SSNet[21]	78.8	62.4	53.1	73.5	51.4
DIMNet-I[5]	82.5	71.0	81.1	77.7	62.8
DIMNet-IG[5]	83.2	**71.2**	81.9	78.0	62.8
本文方法	**85.1**	69.5	**84.9**	**80.3**	**63.5**

注:黑体数据表示最好的实验结果.

从实验结果可以得到,本文方法在各个分组上的各项指标几乎全面超越了现有方法,取得了较好的跨模态验证结果.例如,本文提出方法在"U""N""A"上取得了优于现有方法的跨模态验证结果.同时也注意到,相比其他分组,本文方法和其他方法在"G"分组中的表现都稍弱,这说明性别信息对模型执行跨模态验证任务有较大影响.

2) 1:2匹配任务

在1:2匹配任务上,本文方法与现有方法的实验结果对比如表2所示.实验结果中1:2匹配任务包含"F-V"和"V-F"2个跨模态匹配场景,并且"U""G""N"和"GN"代表的含义与本节跨模态验证部分描述一致.本文方法在这2个场景中的表现均优于现有方法,表明本文方法具有较好的鲁棒性.

Table 2　Comparisons on 1:2 Cross-Modal Matching Task

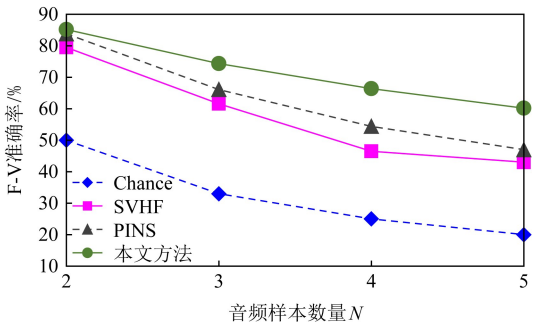
表2　跨模态1:2匹配任务上的实验对比结果 %

方法	F-V				V-F			
	U	G	N	GN	U	G	N	GN
SVHF[3]	79.50	63.40			81.00	63.90		
Horiguchi's[22]	77.80	60.80			78.10	61.70		
Kim's[23]	78.60	61.60			78.20	62.90		
PINs[12]	83.80							
DIMNet-I[5]	83.52	71.78	82.41	70.90	83.45	70.91	81.97	69.89
DIMNet-IG[5]	84.03	71.65	82.96	70.78	84.12	71.32	82.65	70.39
本文方法	**85.18**	**72.17**	**84.77**	**70.82**	**86.85**	**76.05**	**85.70**	**73.10**

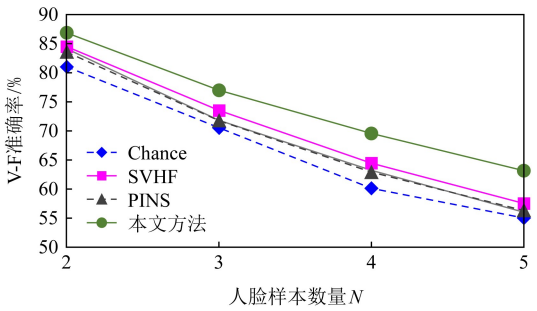
注:黑体数据表示最好的实验结果.

3) 1∶N 匹配

图4展示了本文方法与现有方法在1∶N匹配任务上的实验结果对比.从实验结果可以看到,本文方法无论是在"F-V"匹配任务上还是在"V-F"匹配任务上均轻松超越现有方法.当N取较大值时,本文方法表现仍然比其他方法好,表明本文方法相比其他方法可以更好地解决一些比较困难的任务.



(a) F-V的1:N匹配结果



(b) V-F的1:N匹配结果

Fig. 4 Comparisons with other methods on 1∶N matching task

图4 不同方法在1∶N匹配任务上的实验结果对比

4) 跨模态检索

表3展示了本文方法与现有方法在跨模态检索任务上的实验结果对比.从实验结果中可以看出本文方法在模态检索任务上的表现远远超过随机水平,并优于现有的对比方法.因此,实验结果充分表明本文提出模型能够有效学习人脸-语音间的语义关联.同时,本文方法在"F-V"和"V-F"检索场景中的表现均优于现有方法,表明了本文方法的优越性.

Table 3 Comparison with Other Methods on Cross-modal Retrieval Task

表3 与现有方法在跨模态检索任务上的实验对比 %

方法	mAP		
	Chance	F-V	V-F
Horiguchi's[22]	0.46	2.18	1.96
DIMNet-I[5]	1.07	4.17	4.25
DIMNet-IG[5]	1.07	4.23	4.42
VFMR3[13]	2.15		5.00
本文方法	0.73	**6.14**	**6.96**

注:黑体数据表示最好的实验结果.

跨模态检索任务是1∶N匹配任务的拓展,旨在将候选样本规模从N个(本文实验中$N\leqslant 5$)扩展到整个测试集,同时候选样本中匹配样本数量N_m也从一个增加到若干个($1<N_m\leqslant N$),因而任务难度急剧增加.同时,人脸和语音信息的复杂多样性和非静态性也给跨模态检索带来了极大挑战,故现有方法在该任务上的表现普遍不佳,仍是一项挑战性任务.

在测试集上执行的4个基于人脸-语音的跨模态匹配任务中,本文方法几乎全面超越现有方法,表明本文方法拥有很好的泛化性能.为了进一步验证采用本文方法得到的跨模态表示的有效性,首先,本文从测试集中随机挑选了8个人,每人挑选40条语音数据;接着,使用训练好的模型提取它们的特征;最后,采用t-SNE[24]技术对提取的特征进行可视化,可视化结果如图5所示.可以看到,同一个人的语音样本对应的语音特征聚到了一起,不同人的语音样本对应的语音特征相距较远,表明采用本文方法提取到的跨模态表示具有较好的判别性和可分性.

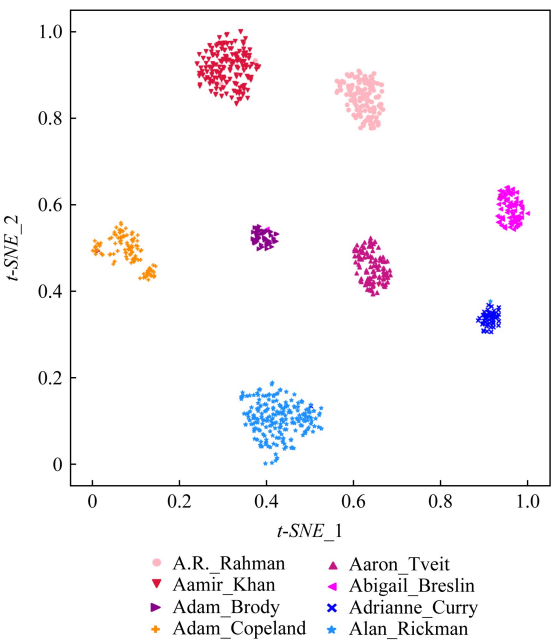


Fig. 5 Visualization of the deep voice embeddings

图5 语音深度特征可视化结果图

4　消融分析

为了探究本文提出的多模态加权残差网络、双向五元组损失以及其中的超参数对模型最终表现的影响,本文针对跨模态验证、1:2匹配、1:3匹配以及跨模态检索任务设计了一系列消融分析实验.

4.1　多模态加权残差网络的影响

为探究多模态加权残差网络(MWRN)对模型表现的影响,本文分别用单层全连接(SFC)、双层全连接(DFC)以及引入残差连接的双层全连接网络(DFC-R)替换 MWRN 进行实验,在 4 个不同任务上的实验结果如表 4 所示.可以发现,当把全连接层数由 1 层增加到 2 层时,模型表现有所下降,表明更深的网络可能更难训练和优化;引入残差连接后模型的表现有大幅提升,表明残差连接可以很好地解决上述问题;接着在残差连接中引入可学习的缩放因子后,模型表现又有一定幅度的提升,表明可学习的缩放因子可以进一步地减轻网络训练的难度,建立更有效的跨模态关联,进而促使网络收敛到更优的值.

表 4　Cross-modal Matching Performance Performance Under Different Network Settings

Table 4　不同网络配置下模型的跨模态匹配准确率　%

网络配置	验证	1:2匹配	1:3匹配	检索
SFC	79.89	84.95	74.24	5.80
DFC	77.17	82.99	68.89	4.06
DFC-R	**85.22**	86.32	76.06	6.43
MWRN	85.11	**86.85**	**76.97**	**6.96**

注:黑体数据表示最好的实验结果.

4.2　各项损失函数的影响

本节将探讨本文提出的双向五元组损失(Bi-Q损失)和 ID 损失对模型表现的影响.图 6 展示了本文模型采用不同损失函数训练时在验证集上的 1:2 匹配任务准确率在前 15 轮的变化曲线.

可以看到,单独采用 ID 损失时,随着训练轮数的增加,模型表现虽然总体呈稳定上升趋势,但最终表现并不是特别好;单独采用双向五元组损失时模型并不收敛;当把 ID 损失与双向五元组损失结合起来使用时,模型很快就收敛了,并取得了不错表现,表明在模型训练过程中嵌入身份 ID 信息可以保证模型训练过程的稳定性,促进模型收敛.

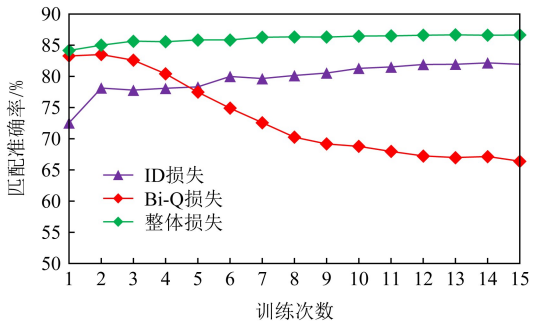


Fig. 6　Performance of our method with different loss

图 6　采用不同损失函数时本文模型的表现

4.3　间隔 m 取值的影响

本节探讨五元组损失中间隔 m 的取值对模型表现的影响.图 7 展示了当 m 取不同值时模型在 4 个不同跨模态匹配任务上的表现.

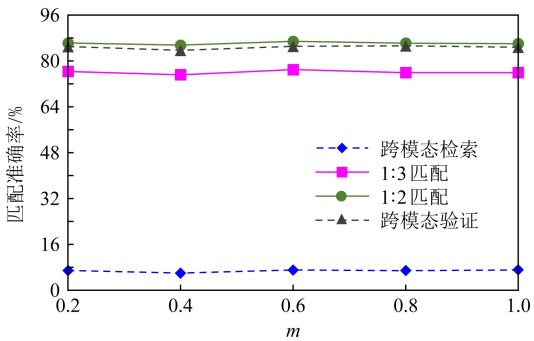


Fig. 7　Performance of our method with different m

图 7　间隔 m 取不同值时模型表现

从实验中可以看到,当 $m=0.6$ 时本文提出的模型的表现最佳,同时,对于每个跨模态匹配任务,当 m 取不同的值时模型表现的波动范围很小,表明本文方法受 m 取值的影响并不大,具有较好的稳定性.

5　结　　论

针对跨人脸-语音匹配挑战性问题,本文提出了一种结合双流网络和双向五元组损失的跨人脸-语音特征学习框架,使用该框架学到的跨模态特征可直接应用于多种人脸-语音的跨模态匹配任务.在公开名人多模态数据集上的实验结果表明:本文提出的网络模型能够对不同场景下名人影像数据进行跨模态标注,效果显著,取得了对面部姿态变化和样本多样性的鲁棒性,并在这些任务上的表现几乎全面超越了现有方法,实验验证了本文提出方法的有效

性.另外,除了人脸和语音2种模态外,本文方法预期也同样适用于其他类型的视听媒体样本进行跨模态匹配.

作者贡献声明:柳欣负责算法设计与实验;王锐负责模型优化和编码;钟必能负责模型可行性分析;王楠楠负责实验的多样性分析.

参 考 文 献

[1] McGurk H, MacDonald J. Hearing lips and seeing voices [J]. Nature, 1976, 264(5588): 746-748

[2] Ellis A W. Neuro-cognitive processing of faces and voices [M] //Handbook of Research on Face Processing. Amsterdam: Elsevier, 1989: 207-215

[3] Nagrani A, Albanie S, Zisserman A. Seeing voices and hearing faces: Cross-modal biometric matching [C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8427-8436

[4] Liu Xin, Geng Jiajia, Ling Haibin, et al. Attention guided deep audio-face fusion for efficient speaker naming [J]. Pattern Recognition, 2019, 88: 557-568

[5] Wen Yandong, Ismail M A, Liu Weiyang, et al. Disjoint mapping network for cross-modal matching of voices and faces [J]. arXiv preprint, arXiv: 1807.04836, 2018

[6] Zhang Lu, Wang Huabin, Tao Liang, et al. Cross-media clustering by share and private information maximization [J]. Journal of Computer Research and Development, 2018, 55(1): 151-162 (in Chinese)
(张露, 王华彬, 陶亮, 等. 基于分类距离分数的自适应多模态生物特征融合[J]. 计算机研究与发展, 2018, 55(1): 151-162)

[7] Hasan B A S, Valdes-Sosa M, Gross J, et al. Hearing faces and seeing voices: Amodal coding of person identity in the human brain [J]. Scientific Reports, 2016, 6: No.37494

[8] Li Dongge, Dimitrova N, Li Mingkun, et al. D. Multimedia content processing through cross-modal association [C] // Proc of ACM Int Conf on Multimedia. New York: ACM, 2003: 604-611

[9] Chetty G, Wagner M. Liveness detection using cross-modal correlations in face-voice person authentication [C] //Proc of European Conf on Speech Communication and Technology. Berlin: Springer, 2005: 2181-2184

[10] Chakravarty P, Tuytelaars T. Cross-modal supervision for learning active speaker detection in video [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2016: 285-301

[11] Yan Xiaoqiang, Ye Yangdong. Cross-media clustering by share and private information maximization [J]. Journal of Computer Research and Development, 2019, 56(7): 1370-1382 (in Chinese)
(闫小强, 叶阳东. 共享和私有信息最大化的跨媒体聚类[J]. 计算机研究与发展, 2019, 56(7): 1370-1382)

[12] Nagrani A, Albanie S, Zisserman A. Learnable PINs: Cross-modal embeddings for person identity [C] //Proc of European Conf on Computer Vision. Berlin: Springer, 2018: 71-88

[13] Xiong Chuyuan, Zhang Deyuan, Liu Tao, et al. Voice-face cross-modal matching and retrieval: A benchmark [J]. arXiv preprint, arXiv: arXiv:1911.09338, 2019

[14] Zhen Liangli, Hu Peng, Wang Xu, et al. Deep supervised cross-modal retrieval [C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 10394-10403

[15] He Kaiming, Zhang Xiangyu, Ren Shaoqi, et al. Deep residual learning for image recognition [C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[16] Bachlechner T, Majumder B P, Mao H H, et al. Rezero is all you need: Fast convergence at large depth [J]. arXiv preprint, arXiv: 2003.04887, 2020

[17] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification [J]. arXiv preprint, arXiv: 1703.07737, 2017

[18] Nagrani A, Chung J S, Zisserman A. Voxceleb: A large-scale speaker identification dataset [J]. arXiv preprint, arXiv: 1706.08612, 2017

[19] Chung J S, Zisserman A. Out of time: Automated lip sync in the wild [C] //Proc of Asian Conf on Computer Vision. Berlin: Springer, 2016: 251-263

[20] Cao Qiong, Shen Li, Xie Weidi, et al. VGGFace2: A dataset for recognizing faces across pose and age [C] //Proc of the 13th IEEE Int Conf on Automatic Face & Gesture Recognition . Piscataway, NJ: IEEE, 2018: 67-74

[21] Nawaz S, Janjua M K, Gallo I, et al. Deep latent space learning for cross-modal mapping of audio and visual signals [C/OL] //Proc of Int Conf on Digital Image Computing: Techniques and Applications. Piscataway, NJ: IEEE, 2019 [2020-05-12]. https://ieeexplore.ieee.org/document/8945863

[22] Horiguchi S, Kanda N, Nagamatsu K. Face-voice matching using cross-modal embeddings [C] //Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 1011-1019

[23] Kim C, Shin H V, Oh T H, et al. On learning associations of faces and voices [C] //Proc of Asian Conf on Computer Vision. Berlin: Springer, 2018: 276-292

[24] Maaten L, Hinton G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 86(9): 2579-2605

Liu Xin, born in 1982. PhD, professor. Senior member of CCF. His main research interests include information retrieval, pattern recognition, and deep learning.

柳　欣,1982 年生.博士,教授,CCF 高级会员.主要研究方向为信息检索、模式识别和深度学习.

Zhong Bineng, born in 1981. PhD, professor. Senior member of CCF. His main research interests include computer vision, pattern recognition, and machine learning.

钟必能,1981 年生.博士,教授,CCF 高级会员.主要研究方向为计算机视觉、模式识别和机器学习.

Wang Rui, born in 1995. Master candidate. Student member of CCF. His main research interests include multimedia analysis, data mining, and deep learning.

王　锐,1995 年生.硕士研究生,CCF 学生会员.主要研究方向为多媒体分析、数据挖掘和深度学习.

Wang Nannan, born in 1986. PhD, professor. Senior member of CCF. His main research interests include image processing, pattern recognition, and machine learning.

王楠楠,1986 年生.博士,教授,CCF 高级会员.主要研究方向为图像处理、模式识别和机器学习.