先验信息驱动的跨模态通用特征空间构建与分析

孙 婧,苏剑波*
(上海交通大学自动化与感知学院,上海 200240)

摘 要: 面部识别和声纹识别是身份验证领域中两种核心的生物特征识别技术,广泛应用于多种场景. 尽管如此,关于这两种模态特征之间关联性的研究相对较少. 本研究旨在探索声音和面部特征之间的共通性. 不同于已有研究直接从实现特征对应方式出发寻找解决方案,本文从身份特征特性出发,从对身份信息的准确表示来主动获取通用特征空间,通过引入人脸识别任务中的身份特征间距离关系作为先验信息,在特征对应方法的基础上,保持身份相关关系不被破坏. 在声纹特征提取过程中,通过调整语音识别任务中的预训练参数,使模型更好地表示身份信息. 实验结果表明,在相同的特征对应方法下,使用语音 Transformer 模型作为声纹信号提取器,在验证任务上的表现相较于时延网络有显著提升. 此外,本文方法对数据要求较低,不需要额外训练分类器,在验证任务上能够取得与已有方法相近的表现. 未来的研究可进一步引入声纹特征的先验知识,以期进一步提升跨模态特征匹配的性能.

关键词: 声音-面容关联;跨模态特征;最优传输;对比学习
基金项目: 工业与信息化部专项项目(No.0747-2461SCCZA302)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)08-2614-10
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250022

Construction and Analysis of Cross-Modal General Feature Space Driven by Prior Information

SUN Jing, SU Jian-bo*
(*School of Automation and Intelligent Sensing*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: Facial recognition and voiceprint recognition are two core biometric technologies in the field of identity verification, widely applied in various scenarios. However, research on the correlation between these two modal features remains relatively limited. This study aims to explore the commonality between voice and facial features. Unlike the existing studies that directly look for solutions from the way of realising feature correspondences, this study starts from the identity feature characteristics and actively obtains the universal feature space from the accurate representation of identity information. The distance relationship between identity features in facial recognition tasks is introduced as prior information, ensuring that identity-related relationships are preserved while using feature correspondence methods. During the voiceprint feature extraction process, the pre-trained parameters from speech recognition tasks are adjusted to enable the model to better represent identity information. The experimental results demonstrate that the speech transformer model, when used as a voiceprint signal extractor with the same feature correspondence method, achieves significant improvement on verification task compared to the time-delay network. In addition, the method is able to achieve similar performance as the existing methods on the validation task with lower data requirements and no additional training of classifiers. Future studies could further incorporate prior knowledge of voiceprint features to enhance the performance of cross-modal feature matching.

Key words: voice-face association; cross-modal embeddings; optimal transport; contrastive learning
Foundation Item(s): Special Program of the Ministry of Industry and Information Technology (No.0747-2461SCCZA302)

1 引言

语音与面部之间的映射是人类智能的一个重要方面,因此研究语音与人脸之间的特征对应关系是计算智能的重要研究方向. 将声音和面部联系起来的能力

收稿日期:2025-01-06;录用日期:2025-05-02;责任编辑:覃怀银
*通讯作者:苏剑波

可以应用于犯罪调查和智能配音等需要对人类身份特征进行扩展的领域[1]. 在不同的识别任务中,人脸和声音线索是最常用的、非侵入性且易获取的线索之一[2],认知科学和神经科学的研究表明,人类在执行各种感知任务时倾向于整合视听信息[3]. 相关研究还证实,人类可以准确地将不熟悉的面部图像与相应的语音进行匹配,反之亦然[4]. 本文研究计算智能是否能匹配来自不同模态的数据特征,特别是人脸和声音模态的特征.

跨模态特征已经有大量研究,这些研究利用一种模态的特征来增强对另一种模态的检测[5,6],或借助某种模态的数据生成另一模态的信息[7]. 人脸-声音特征的相关关系就是一个跨模态问题,针对此问题也已有一定研究. 一些研究使用对比学习方法将属于同一个体的不同模态特征拉近,同时推远来自不同个体的特征[1,8]. 此外,还有一些研究者提出通过在特征提取网络中附加分类器可以隐式地限制两种分布之间的差异[9,10]. Wen等人[9]提出利用其他信息辅助可以消除不同模态特征之间的差异,而 Nagrani等人[10]则将音脸匹配任务作为一个多分类问题来处理. Cheng等人[11]提出使用对抗学习来实现跨模态特征对应,他们引入了一种对抗深度语义的匹配网络,该网络使用判别器来消除语音和人脸特征之间的差距,同时保持语义的一致性.

不同于已有的研究,本研究的核心观点在于,跨模态特征匹配不仅需要在样本层面建立对应关系,还需要与个体身份信息的精确表达相结合. 先前的研究表明,引入额外的身份识别器可以提高特征提取的效果,从而支持本研究的观点[9]. 为此,我们引入了身份识别数据集上的预训练模型作为先验信息,并在研究中进一步引入基于知识蒸馏的距离保持方法,以有效利用这些信息. 实验结果分析中,我们讨论了该方法对最终匹配效果的影响. 在特征对应方法上,本研究认为除了确保单个样本间的距离足够接近以满足匹配要求外,还必须在整体特征分布上实现域适应,以确保两种模态的身份特征均能有效表达身份信息,从而提升识别性能. 本文采用多层次的特征匹配方法来实现面部和声音模态之间的特征分布对齐,包含单样本对和整体分布的对应. 与现有方法中的隐式分布约束不同,本文采用最优传输(Optimal Transport,OT)来直接对整体特征分布的差异进行限制.

为了探究不同声音信号处理模型在此跨模态对应任务上表现的差异,我们将视线转向了在近期由于大语言模型的成功而备受关注的Transformer架构. Transformer在自然语言处理上的成功证明了其在处理时序信号时的强大能力[12,13]. 已有研究也发现,基于对比学习预训练的语音识别模型经过调整后,可以较好地适应身份识别、情绪识别等任务,提供了一种先验信息. 音频同样属于时序信号,在语音识别领域也有使用此结构的相关研究,这证明了Transformer结构在处理音频信号时存在显著优势. 由于其模型结构复杂,从头开始训练成本高,所以在应用过程中更多的是在充分的数据集上进行预训练,之后在特定任务上进行微调,获得最终的模型参数用于特征编码. 在本研究中,我们将对基于Transformer架构的音频信号编码网络在身份特征跨模态匹配任务上的表现进行研究与分析.

综上所述,本文主要关注如何将额外知识应用于跨模态身份特征匹配任务. 主要贡献在于揭示了身份信息的先验知识引入,以及语音识别模型预训练结果的调整有助于提升特征匹配效果. 先验知识包括两方面:一是引入人脸识别任务中的特征间相关关系,避免直接使用特征匹配手段破坏特征结构;二是通过对比学习预训练的语音识别模型.

2　相关工作介绍

本文的研究目标是建立面部与声音两个模态特征之间的对应关系,从而获得一个通用的特征空间,在该特征空间中,特征只在不同个体之间存在差异,而属于同一个体不同模态的特征则是相似的. 在过去已有的研究中,所使用的特征对应方式主要有以下三种:第一种使用对比学习方法[1,8],通过拉近同一个体的不同模态特征来实现匹配;第二种是通过给特征提取网络附加额外的分类任务来隐式地对特征分布进行限制[9,10];最后是使用对抗学习方法,通过添加额外的特征来源分类器消除不同模态特征间的差异[14]. 然而,这些方法存在一些局限. 首先,这些方法都是直接从实现特征匹配目标出发寻找解决方案的,通过直接的对比或者额外的分类器解决特征间差异性的问题,缺少对数据特性的分析,且可用成对音频-图像数据集有限. 其次,这些研究中所使用的模型结构比较单一,没有充分利用先进的特征提取模型. 与现有方法不同,本文分析了身份特征特性,从身份识别任务中引入先验信息,指导特征间差异的建模,从而获得有更好匹配效果的统一特征空间. 此外,本文在音频编码中使用了语音Transformer,以进一步提升特征提取的效果.

为实现跨模态特征对齐这一目标,首先需要完成对原始信号的编码,特征提取过程是其中一个关键环节. 由于不同模态数据有不同的特性,因此适合的特征提取网络也各不相同,需要使用不同的网络结构作为编码器[15~17];也有部分研究将不同模态的数据处理为相同尺寸,之后采用一个通用的单一流网络进行特征提取[18]. 本文为了提取更符合数据特征、更能够代表身份信息的特征,对不同模态的数据选择了不同的编码

网络结构. 为了验证本研究提出的声纹识别方法的有效性,并深入分析不同单模态特征提取网络在性能上的差异,本研究采用了多种声纹特征提取模型进行实验性评估. 这些模型包括基于时延神经网络的ECAPA-TDNN模型,以及基于Transformer架构的Wav2vec 2.0和Hubert模型. Transformer架构最初在论文"Attention Is All You Need"中被提出[19],它是一种依赖于自注意力机制的神经网络模型,摒弃了传统循环神经网络中对序列处理的循环依赖,转而采用自注意力机制来并行处理序列数据. 该架构由编码器和解码器两部分组成,编码器负责将输入序列映射到一个连续的向量空间,而解码器则基于编码器的输出以及之前的输出序列来构建最终的输出序列. 在先前的学术研究中,已有文献探讨了对Wav2vec 2.0和Hubert模型进行微调,以适应声纹识别等非语音识别任务[12,13]. 这些研究表明,Transformer架构在捕捉个体身份信息方面具有显著优势. 本研究进一步将这些模型应用于特征匹配任务中,以评估其在身份信息表征领域的应用潜力.

最优传输方法目前已有比较广泛的应用,如在图像处理、机器学习、统计学和计算流体力学等领域[20,21]. 也有部分研究将其用于域适应任务中,在这些研究中,来自不同数据集的数据被视为不同的域,通过域适应方法将使用一个领域的数据中学习得到的知识,转换到另一个领域中,从而在两个领域都取得良好的识别效果. 在说话人识别领域,最优传输被用于实现不同领域间的特征对齐[22],通过减小最优传输损失,减小两个领域数据特征空间的差异,弥补跨域说话人识别的损失. 这一任务和本文中研究都关注了身份特征的表示,并需要实现不同数据域上的特征对齐. 在本研究中,"不同域"的概念被扩展为"不同模态",最优传输方法被用于减小不同模态特征空间之间的距离差异,从而在这一跨模态问题中实现域对齐,并与基于对比学习方法的单个样本对上的特征对应相结合,共同实现跨模态特征的对应.

基于领域自适应理论的深度迁移学习算法通常依赖于预训练模型[23,24]. 由于深度网络的函数空间非常大,预训练过程可以有效缩小允许的函数空间,从而大大降低对目标域的泛化误差. 与现有的包含9 000多个个体的大规模人脸数据集相比[25],目前的语音-人脸对应数据集相对较小,仅有1 200多个个体[26],而预训练参数的使用结合了其他大量数据集的信息,使特征能更有效地代表数据中的身份信息. 本研究中,预训练的人脸识别模型不仅能够表示身份信息,而且在不同个体间具有区分度,因此被视为源域,旨在引导声纹识别模型的特征适应. 为了确保映射后的人脸特征与声纹特征在相同维度下仍能保留身份信息,本文引入了关系特征知识蒸馏方法. 该方法的核心理念在于,学习的本质不仅在于特征输出的结果,更在于层与层之间以及样本数据间的关系. 这种关系映射提供了一种恒等映射,使得学生模型能够更有效地学习教师模型的关系知识.

3 跨模态特征匹配任务

深度学习方法使用深度神经网络对原数据进行编码,以获得高维特征,之后将这些特征应用于下游任务中. 现有研究表明,个人的面部特征和声音特征之间存在相关性,这表明有可能将两种模式的特征空间统一起来. 通过使用对应的面部-语音样本对训练深度神经网络,并在训练过程中限制两种模态的特征差异,能够得到一个两种模态的高维特征的共享空间. 使用在这种条件下训练得到的模型对面部图像或语音进行编码,就能够通过特征间的距离确定给定的一对人脸和声音是否属于同一个人,或从一组人脸中确定哪张图像与特定声音对应. 图1是对此问题的示意图.

本研究中使用两个不同结构的深度神经网络分别提取面部和声音特征,随后进行特征对齐. 每个面部图像x_f和声音信号x_v分别由神经网络编码成高纬特征:$\boldsymbol{T}=\text{ImageEncoder}(X_f)$,$\boldsymbol{F}=\text{FeatureMapping}(\boldsymbol{T})$,$\boldsymbol{V}=\text{AudioEncoder}(X_v)$,ImageEncoder(·)与FeatureMapping(·)用于处理人脸图像,AudioEncoder(·)用于提取声纹特征. $\boldsymbol{F}_t$为人脸识别模型直接获得人脸特征,提供特征间需要学习的距离关系,$\boldsymbol{F}$与$\boldsymbol{V}$维度相同,用于特征对应. 要实现跨模态特征匹配,$\boldsymbol{F}$与$\boldsymbol{V}$应相互接近,同时两个模态的特征应有相似的分布.

4 多层次人脸-声音特征对齐

为了有效提取不同模态的特征,本文使用两个独立的网络对输入数据进行编码. 每个网络负责提取特定模态的特征,每个编码器的输出保证具有相同的特征维度,提取特征后采用特征对齐的方法来建立不同模态之间的对应关系. 整体网络结构如图2所示.

4.1 基于对比学习方法的单样本对特征对齐

要实现两个模态之间的特征对应,需要保证在不同模态间相同个体的特征相近,同时不同个体特征有差异,这一思路和对比学习方法是相符的. 对于两种模态上不同个体的特征$(\boldsymbol{f}_1,\boldsymbol{v}_1),(\boldsymbol{f}_2,\boldsymbol{v}_2),\cdots,(\boldsymbol{f}_n,\boldsymbol{v}_n)$,需要同时减小$(\boldsymbol{f}_i,\boldsymbol{v}_i)$之间的距离且增大$(\boldsymbol{f}_i,\boldsymbol{v}_j),\ i\neq j$之间的距离. 在一个大小为$N$的批次中,共包含$N$个图像样本及$N$个音频样本,其中对应的图像及音频互为正样本,需要减小正样本对之间的特征差异D_{p}^{i}:

$$D_{\text{p}}^{i}=\text{d}(\boldsymbol{f}_i,\boldsymbol{v}_i)=1-\cos(\boldsymbol{v}_i,\boldsymbol{f}_i) \tag{1}$$

同时对每个样本来说,存在N-1个对应模态的负

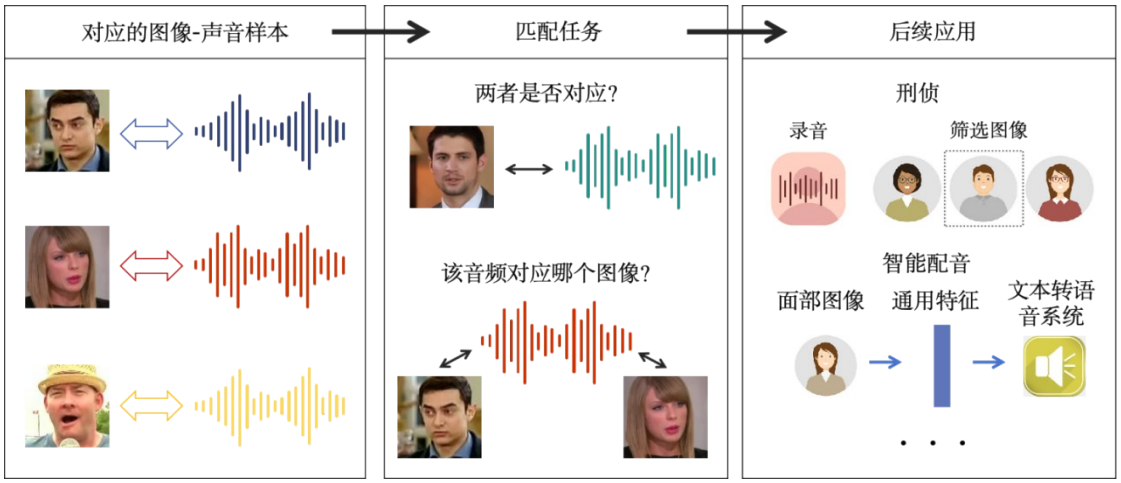

图1　扩模态特征匹配任务目标及应用

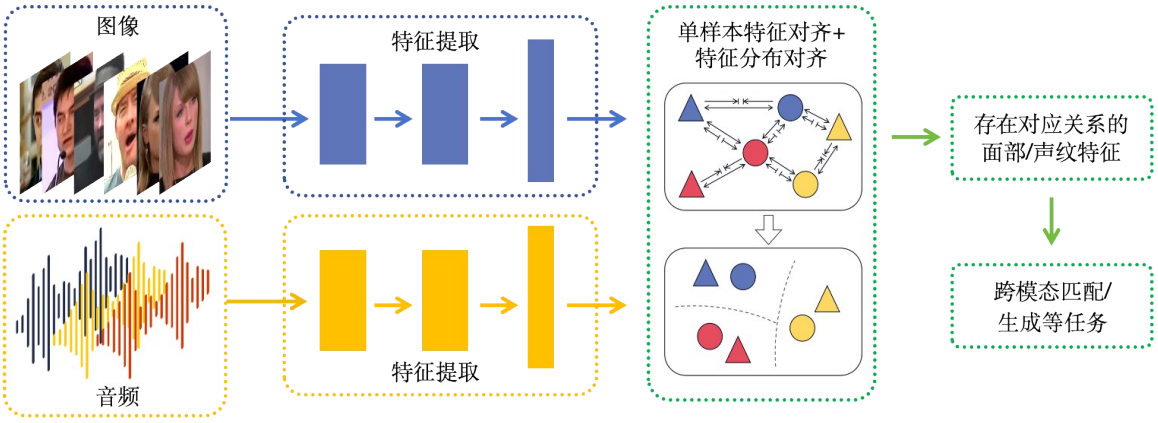

图2　多模态特征对齐框架

样本,需要增大这些负样本对之间的差异D_{n}^{i},这些负样本由两部分组成,分别对应语音模态和图像模态:

$$D_{\mathrm{n}}^{v_i}=\mathrm{d}(\boldsymbol{v}_i,\boldsymbol{f}_j)|_{f_j\in N_i^{\mathrm{F}}} \tag{2}$$

$$D_{\mathrm{n}}^{f_i}=\mathrm{d}(\boldsymbol{v}_j,\boldsymbol{f}_i)|_{v_j\in N_i^{\mathrm{V}}} \tag{3}$$

最终对比学习损失计算如下:

$$\begin{aligned}L_{\mathrm{C}}&=L_{\mathrm{C}}(\boldsymbol{v}_i,\boldsymbol{f}_i)+L_{\mathrm{C}}(\boldsymbol{f}_i,\boldsymbol{v}_i)\\&=-\sum_i\log\frac{\exp(D_{\mathrm{p}}^{i}/\tau)}{\exp(D_{\mathrm{p}}^{i}/\tau)+\exp(D_{\mathrm{n}}^{v_i}/\tau)}\\&\quad-\sum_i\log\frac{\exp(D_{\mathrm{p}}^{i}/\tau)}{\exp(D_{\mathrm{p}}^{i}/\tau)+\exp(D_{\mathrm{n}}^{f_i}/\tau)}\end{aligned} \tag{4}$$

这种方法只考虑了两种模态样本特征的直接对应关系,而没有考虑样本与总体分布的对齐情况. 这会导致模型收敛缓慢,即使在$i\neq j$时,$(\boldsymbol{f}_i,\boldsymbol{v}_j)$也可能相似. 如果出现这种情况,使用上述训练方法将导致较差的模型结果. 因此,有必要加入整体分布对齐方法. 本任务中需要实现两种不同的数据集上的特征存在一定的匹配关系,这和跨域学习的任务相似. 在跨域学习中,为了减少领域不匹配现象,已出现跨领域任务的领域适应算法,通过这种算法,源领域和目标领域的特征分布得以对齐. 为更好地进行两种模态间特征的对应,本文引入了最优传输方法实现特征在总体分布上的对齐.

4.2　基于最优传输的总体分布对齐

最优传输方法(OT)旨在找到一个全局最优传输计划,通过这个计划使得一种概率分布能够以最小的代价转换成另一种形状[27]. 该方法能够在两个分布之间高效传输数据,同时保持概率分布的质量水平. 同时,该方法定义了一种有效的几何感知的 Wasserstein 距离,并在匹配过程中保留了概率分布的形状.

在最优传输损失的计算过程中,我们注意到了潜在的错误传输问题[22]. 为规避此问题,本研究在计算最优传输损失之前引入了一种基于样本相似度的特征筛选机制. 具体而言,对于给定批次中任意样本对的特征向量$(\boldsymbol{f}_i,\boldsymbol{v}_i)$,我们通过比较其相似度函数$s(\boldsymbol{f}_i,\boldsymbol{v}_i)$与$s(\boldsymbol{f}_i,\boldsymbol{v}_j)$及$s(\boldsymbol{f}_j,\boldsymbol{v}_i)$(其中$i\neq j$)来确定样本对的区分性. 仅当$s(\boldsymbol{f}_i,\boldsymbol{v}_i)$大于$s(\boldsymbol{f}_i,\boldsymbol{v}_j)$和$s(\boldsymbol{f}_j,\boldsymbol{v}_i)$时,我们认为该样本对具有足够的区分度,从而在计算时避免了错误传输的问题,有助于提升模型训练的效率和准确性,这里

的相似度利用了对比学习中计算使用的余弦相似度.

为实现最优传输,首先需要解决一个数学问题,即“成本优化问题”. 这个问题的目标是确定一种最佳传输策略,使在两个给定分布之间传输数据的成本最小化. 在每个轮次中,不同特征$(\boldsymbol{f}_a, \boldsymbol{v}_b)$间的差异用欧氏距离表示:

$$D(\boldsymbol{f}_a, \boldsymbol{v}_b)=\|\boldsymbol{f}_a-\boldsymbol{v}_b\|^2 \tag{5}$$

需要求解使得下式最小的最优传输策略γ. 在这里γ是计算得出的,在每次计算过程中首先按式(6)对其进行求解,得到对这批样本确定的最优传输策略,之后使用该策略计算最优传输损失:

$$L_{\mathrm{OT}}^{*}=\sum_{a,b} D(\boldsymbol{f}_a, \boldsymbol{v}_b)\gamma(\boldsymbol{f}_a, \boldsymbol{v}_b) \tag{6}$$

最终两种模态之间分布的差异由下式表示:

$$L_{\mathrm{OT}}(\boldsymbol{f}, \boldsymbol{v})=\min_{\gamma\in\prod(\boldsymbol{f},\boldsymbol{v})}\sum_{a,b} D(\boldsymbol{f}_a,\boldsymbol{v}_b)\gamma(\boldsymbol{f}_a,\boldsymbol{v}_b) \tag{7}$$

其中,L是两个特征之间的距离,γ是两个模态领域之间的传输计划.

4.3 多层次特征对齐

在训练阶段需综合使用两种特征对齐方法,进行多层次的特征对齐. 只考虑L_{C}时,整体分布的对齐会被忽略. 同样如果只考虑L_{OT},就会缺少直接的样本特征对应信息. 因此,在网络训练过程中需要同时利用这两种损失,使用参数α来调整这两个参数的权重,最终的损失函数如下:

$$L=L_{\mathrm{C}}+\alpha L_{\mathrm{OT}} \tag{8}$$

通过采用此方法,我们能够实现样本对特征的直接匹配,并对两种模态的特征分布进行校准. 在模型训练阶段,参数α的动态调整对于平衡两种损失函数的贡献至关重要. 进一步的实验分析表明,这两种策略不仅各自有效,而且能够相互补充,从而协同增强模型训练的整体性能.

4.4 基于距离知识蒸馏的特征特性保持

为了保持特征向量在个体间区分关系的有效性,我们借鉴了知识蒸馏领域中的关系知识蒸馏(Relation Knowledge Distillation,RKD)[28]策略. 该策略侧重于特征间的距离关系,使得学生模型能够继承教师模型在特征空间中的距离结构. 传统的知识蒸馏方法主要关注单个样本的输出激活值,即学生模型通过模拟教师模型对单个样本的输出来学习. 与此不同,关系知识蒸馏侧重于数据样本之间的关系,通过传递样本间的结构关系(如距离和角度)来实现知识提炼. 这种方法能够更好地捕捉数据嵌入空间的结构信息,而不仅仅是单个样本的输出. 具体而言,RKD通过距离和角度等关系信息进行提炼,能够传递传统方法无法捕捉的高阶属性信息,例如样本之间的相对位置关系. 此外,由于RKD不依赖于单个样本的输出值,因此它对输出维度的差异不敏感,能够更好地适应不同架构的教师模型和学生模型. 实验表明,RKD在度量学习任务中表现优异,尤其在图像检索任务中,学生模型的性能甚至能够超越教师模型.

在本研究中采用了预训练的人脸识别模型作为教师模型,并通过一个多层感知器将其输出特征升维至与声纹特征相同维度,得到用于匹配的面部特征,随后对升维后的面部特征和声纹特征使用多层匹配方法. 通过保留升维后特征向量间的距离关系,确保了这些特征向量在个体间区分能力上的保留[29]. 最终,得到的统一特征不仅具有跨模态的对应关系,还保留了身份间的区分关系. 具体计算方法如下:

$$L_{\mathrm{RKD}}=\sum_{(x_i,x_j)} l_\delta\left(\Psi_{\mathrm{D}}(\boldsymbol{f}_i, \boldsymbol{f}_j), \Psi_{\mathrm{D}}(\boldsymbol{t}_i, \boldsymbol{t}_j)\right) \tag{9}$$

其中,l_δ为均方差损失,特征间的距离Ψ_{D}由余弦相似度衡量. 本节中所使用方法的示意图如图3所示,算法1为此方法的伪代码表示.

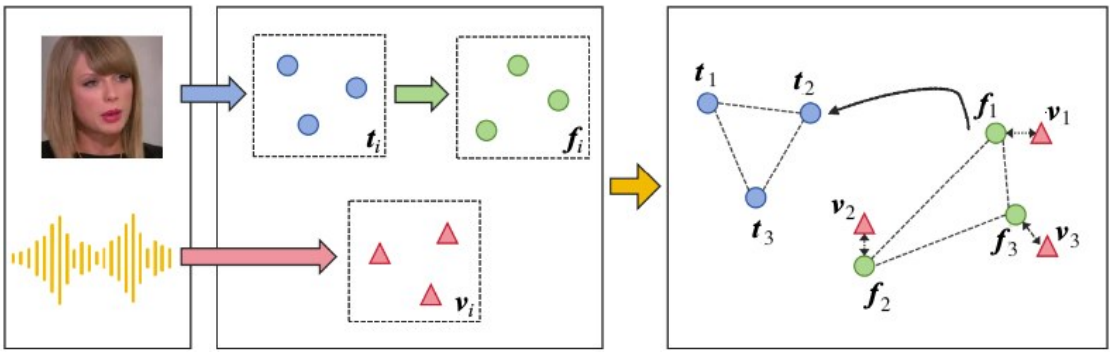

图3 本研究多层匹配方法示意图

5 特征提取和对齐的准备工作

本文中采用VoxCeleb1数据集进行实验与验证[26]. 该数据集包含从YouTube的上传视频中提取的1 251位不同说话人的21 063个视频片段,通过处理这些视频

算法1　基于先验信息的多层次跨模态特征匹配

输入：(X_f, X_v), ImageEncoder, FeatureMapping, AudioEncoder,Epoches

输出：ImageEncoder, FeatureMapping, AudioEncoder

```
for i in range (0, Epoches) do
    for j in range (0, M) do
        选择得到N对样本: (x_{f_1}, x_{v_1}),(x_{f_2}, x_{v_2}), ⋯, (x_{f_n}, x_{v_n})
        对图像和音频样本分别进行编码:
            T = ImageEncoder(X_f)
            F = FeatureMapping(T)
            V = AudioEncoder(X_v)
        对F, T计算RKD损失 L_RKD,如式(9)
        对F, V计算相似度,如式(1)~式(3)
        计算对比学习损失 L_C,如式(4)
        根据相似度筛选样本,使用筛选后样本计算最优传输损失 L_OT,如式(5)~式(7)
        计算损失函数 L = L_C + L_OT + L_RKD
        更新模型参数
    endfor
endfor
```

片段,就能够获得对应的面部图像和语音音频. 其中,面部图像是通过使用面部边界框裁剪视频帧生成的;语音片段是从视频原声带中提取的,采样率为16 kHz. 该数据集性别均衡,且包括不同种族、口音、职业和年龄的录制对象. 测试集中标注了各种人口统计属性,为进一步验证本文中方法,根据不同测试数据的属性创建了五个不同的测试组:无限制条件组(U)、同性别组(G)、同国籍组(N)、同年龄层组(A)以及同性别和国籍组(GN). 表1显示了每个分组的构成. 在验证实验中,每个字母对应一个不同限制条件的分组,在这些分组中,需要区分的个体均满足相同限制条件. 例如,在G组实验中,需要判断属于同一个人的图像.

在输入到编码神经网络前,每张原始面部图像被剪裁为122 × 122的统一尺寸,并被增加了随机水平翻转;声音信号首先被随机剪裁出2 s的片段,之后提取80维Mel频率倒谱系数,提取窗口大小为25 ms,帧移10 ms. 实验中对Log Mel频谱图使用了SpecAugment[30]作为增强步骤,该算法在时域随机屏蔽0~5个帧,在频域随机屏蔽0~10个通道.

表1　数据集的组成

分类	性别(G)		种族(E)						年龄段(A)							
	m	f	1	2	3	4	5	6	≤19	20~29	30~39	40~49	50~59	60~69	70~79	≥80
数量	150	100	1	10	19	13	189	18	2	27	77	58	43	21	14	8

注:括号中的字母与下表中的实验结果分组相对应.

5.1　特征提取网络结构

5.1.1　面部特征提取网络

ResNet-34是深度学习领域的一种重要网络结构,属于残差网络(ResNet)的一种. 残差网络结构通过引入残差连接,有效解决了深度神经网络中的梯度消失和表示瓶颈问题,从而提高了网络的深度和模型的性能. 使用ResNet-34能够在目标数据集上实现良好的人脸识别,证明该网络提取的特征包含有效身份信息且能够在不同个体间进行区分,可以用于本任务.

5.1.2　声纹特征提取网络

ECAPA-TDNN[16]是Desplanques等人提出的一种神经网络模型,主要用于说话人识别. ECAPA-TDNN利用扩展的TDNN-x-vector进行语音特征提取,它由多个帧级时延神经网络层、一个统计池层和两个句子级全连接层以及一个损失函数为交叉熵的softmax层组成. ECAPA-TDNN可以接受任意长度的输入,并将帧级特征合并为整句特征,已被证明能够有效实现说话人识别任务.

Wav2vec 2.0[31]通过多层卷积神经网络对语音音频进行编码,然后对得到的潜在语音表征的跨度进行屏蔽,类似于屏蔽语言建模. 潜表征被输送到Transformer网络,以建立上下文化的表征,并通过对比任务对模型进行训练,在对比任务中,真正的潜表征需要从干扰项中区分出来. 在对无标签语音进行预训练后,该模型将在有标签数据上进行微调,以用于下游语音识别任务. 研究表明Wav2vec 2.0适用于说话人识别任务,且语音识别任务上的预训练权重也可微调用于说话人识别[12]. 在本任务中,Wav2vec 2.0编码得到的特征首先被进行平均池化,之后再与相应的人脸特征进行比较,或计算训练过程中的损失并进行训练.

Hubert[32]关注了自监督语言表示学习,其使用类似于BERT(Bidirectional Encoder Representations from Transformers)中的预测损失,通过聚类步骤提供对齐的目标标签,在说话人识别中已有一定的应用. 已有研究使用微调的Hubert完成声纹识别、语音情绪识别及口语理解任务,探究了固定部分网络参数对实验结果的影响,证明Hubert能够完成这些任务. 在本任务中,对Hubert编码得到特征的处理方式与Wav2vec 2.0相同,首先进行平均池化得到一维特征,之后将该特征用于后续计算.

5.2 训练过程

所有模型都是使用Adam优化器,在使用ECAPA-TDNN作为声纹编码器时,采用介于10^{-8}和10^{-3}之间的阶跃学习率进行训练;在使用Wac2vec 2.0与HuBert的实验中,采用TriScheldluer学习率调整策略,学习率的初始值设定为10^{-7},随后逐步增加至最大值10^{-5},最终再降至10^{-7}. 在此过程中,学习率的增加、稳定和减少三个阶段的比例分别设定为15%、1%和7%. 为防止过度拟合,模型中的所有权重采用2×10^{-5}的权重衰减. 训练的批次大小为64.

6 实验结果与分析讨论

通过比较这些特征的相似性来完成验证或匹配任务. 在验证或匹配任务中,面部图像或包含语音的音频均被编码成相同大小的特征向量,通过比较这些特征间的相似度判断不同模态的数据间的对应关系. 相似度可以从多个不同的角度进行衡量,如欧氏距离、余弦相似度和相关系数等,通过这些方法可以量化两个特征之间的差异大小,从而确定它们是否属于同一个体. 除对两个数据间的相似度进行衡量外,需要设定一个阈值,当两个数据特征之间的距离小于这个阈值时,就认为这两个数据是属于同一个人. 在确定最佳阈值时,应综合考虑历史数据和实际应用场景. 为此,交叉验证是一个有效的工具,它可以帮助我们找到适合特定情境的阈值.

6.1 评估方法

在评估过程中,测试数据被分为不同的组别进行研究. 依照文献[8]中的标准,测试过程分别对已知样本(seen-heard)和未知样本(unseen-unheard)进行测试,在这两个类别下测试数据又被分为不同的组别. 为了研究特定的属性对本任务的影响,在选择负的测试样本时对这些属性进行限制,在对应分组中正负样本在以下属性中需保持一致:性别(G)、国籍(N)和年龄(A),其中性别和国籍标签来自维基百科,年龄分组是由将年龄分类器应用于视频中的人脸帧,并对每段视频进行平均得到的. 本文中从验证任务上对不同模态特征对应的效果进行了评估. 在跨模态样本验证任务中,输入数据是来自不同模态的两个样本,验证任务的目标是辨别这两个样本是否属于同一个个体. 这个辨别过程是通过设定相似度阈值实现的,使用不同的阈值会带来不同的错误接受率和错误拒绝率,最终的辨别效果由ROC曲线下面积(AUC)来体现. 分组限制条件越多,验证任务的难度越大,多个分组上的验证结果能反映特征在不同个体上的区分度. 对比实验中,所选方法的选择标准是确保这些方法在相同条件下进行实验,以保证公平对比. 所选方法涵盖了不同的特征匹配思想,PINs采用对比学习进行特征对应,DIMNet和SSNet通过相同任务和相同特征编码模型间接实现特征对应,ADSM则通过混淆模态分类器和对抗思想实现特征对应.

6.2 定量结果分析和讨论

6.2.1 对齐方法有效性分析

在本节中,我们通过一系列对照实验,系统性地验证前文提出的多层次特征对齐策略的有效性. 相较于单一对比学习框架,本研究引入了最优传输理论,旨在降低不同特征分布间的差异性,从而显著增强特征匹配的准确性. 如表2所示,对不同特征匹配方法对实验结果的影响进行了对比分析. 实验数据揭示,单纯依赖最优传输方法进行域适应,不足以有效完成特征匹配任务. 然而,将最优传输策略与对比学习框架相结合,能够在多组实验中显著提升特征匹配的效果,尤其是在“G”组和“GNA”组中,该多层次特征匹配策略的效果尤为显著. 具体而言,在挑战性较高的“G”组和“GNA”组的验证实验中,相较于单独使用对比学习,准确率分别提升了11.5%和8.8%. 这一发现不仅进一步证实了所提方法的有效性,而且为将来将该方法扩展至其他特征对应任务提供了理论基础. 通过实现跨领域的统一特征表示,该方法有望在更广泛的实际应用中发挥关键作用. 如表3所示,本研究采用Wav2vec 2.0作为特征编码器,并应用多层次特征匹配方法,结果显示本文方法在多个维度上均优于先前的方法. 值得注意的是,本文方法无需额外的数据标签信息,也无需在训练过程中进行聚类,因此,我们的方法在效率上更高,对计

表2 验证实验结果

方法	unseen-unheard					seen-heard				
	—	G	N	A	GNA	—	G	N	A	GNA
PINs[8]	78.5	61.1	77.2	74.9	58.8	87.0	74.2	85.9	86.6	74.0
DIMNet[9]	83.2	71.2	81.9	78.0	62.8	94.7	89.9	93.2	94.8	87.8
SSNet[18]	78.8	62.4	53.1	73.5	51.4	91.2	82.5	89.9	90.7	81.8
ADSM[10]	88.4	79.3	83.6	83.9	64.7	95.9	92.5	93.9	95.5	89.9
本文方法	84.4	72.2	84.5	81.1	66.6	95.5	89.1	95.7	95.7	88.2

注:表中字母表示不同的约束条件.

算资源的需求也更少.

为了进一步验证所提出的特征匹配方法的有效性,本研究采用了可视化技术对数据进行图形化展示. 具体而言,本研究选取了来自三个不同个体的声音样本,并对其进行了编码处理. 随后,利用主成分分析将编码后的数据降维至二维空间,并在二维坐标系中进行了图形绘制,相关图像详见图4. 实验结果表明,在未经微调的状态下,模型无法有效地区分不同个体的声纹特征. 然而,经过微调后,模型能够识别出不同个体间声纹特征的显著差异. 这一发现揭示了模型通过本文提出的跨模态特征匹配方法逐渐获得了区分不同个体声纹特征的能力.

表3　不同模型上不同方法表现比较

模型结构	方法		实验分组				
	对比	OT	—	G	N	A	GNA
Wav2vec 2.0	√	×	81.5	64.4	82.0	78.2	61.2
	×	√	51.8	52.0	51.7	51.4	50.8
	√	√	84.4	72.2	84.5	81.1	66.6

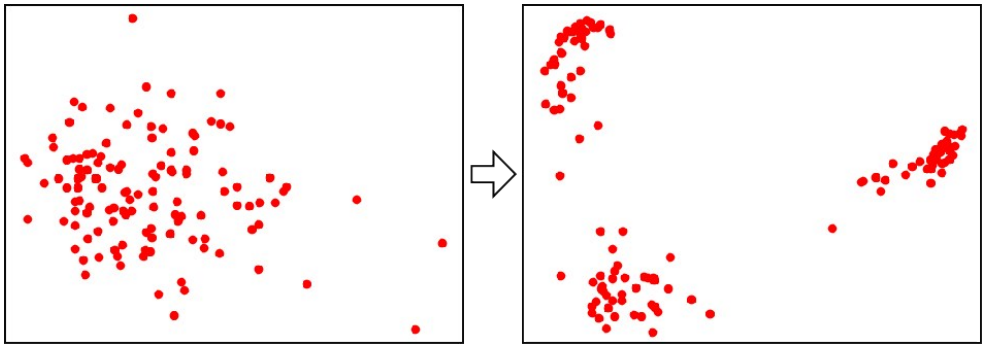

图4　Wav2vec 2.0微调前后特征分布变化示意图

6.2.2　关系知识蒸馏策略的影响

在本项研究中,我们分析模型的不对称性的影响,虽然经过训练后的声纹特征在不同个体间差异明显,但面部特征却易于混淆. 这一现象可能源于模型固有的差异以及微调阶段学习率配置的不一致性. 针对此问题,本研究提出了一种特征保持策略,该策略旨在通过调整特征空间来降低模型的不对称性. 具体来说,我们提出的策略通过优化特征选择机制,增强了模型对不同类别数据的辨识能力. 如图5所示,该策略在一定程度上有效地缓解了模型的不对称性,使得模型在处理不同类别数据时展现出更加均衡的性能. 然而,尽管模型的不对称性得到了一定程度的缓解,这种改进并未直接导致模型匹配效果的显著提升. 这可能是因为模型匹配效果受到多种因素的影响,如特征的表示能力、模型的泛化能力以及数据的复杂性.

尽管如此,我们的研究为如何有效地整合先验知识提供了新的视角. 通过特征保持策略,我们可以更有效地融合先验知识. 这种方法不仅有助于提升模型的预测精度,而且在理解模型的决策过程和增强模型的可解释性方面也具有重要意义. 因此,本研究不仅为解决模型不对称性问题提供了一种潜在的解决方案,而且为未来在更广泛的应用领域中利用先验知识奠定了理论基础,并提供了实践指导.

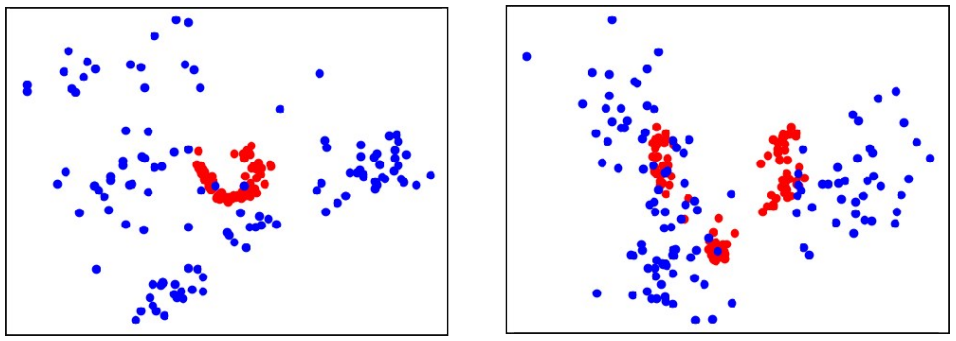

图5　距离知识蒸馏策略影响示意图

6.2.3　特征提取网络差异分析

在对表4所展示的数据进行分析后,发现Wav2vec 2.0和Hubert模型在验证任务中相较于ECAPA-TDNN模型表现出显著的性能提升. 此外,表3中Wav2vec 2.0模型的结果显示,与历史研究相比,该模型同样展现出了优越的性能. 这些发现体现了Transformer架构在提

取个体身份信息任务中具有应用潜力. 尽管我们也探索了将人脸特征提取模型替换为视觉Transformer的可能性,但经过多次实验和方法调整后,发现其在本任务中的应用效果并不理想. 这一现象可能归因于视觉Transformer主要在图像分类任务上进行预训练,因此难以通过简单的微调来适应身份信息的表示任务. 未来研究中,有必要通过进一步的实验来深入分析和解决这一问题.

表4 unseen-unheard验证任务中不同声音编码器表现

模型结构	方法	—	G	N	A	GNA
ECAPA-TDNN	Contrastive	60.7	55.2	60.7	59.4	54.7
	Overall Align	69.6	61.4	69.1	67.9	59.4
Wav2vec 2.0	Contrastive	81.5	64.4	82.0	78.2	61.2
	Overall Align	84.4	72.2	84.5	81.1	66.6
Hubert	Contrastive	82.3	71.9	81.6	79.3	66.2
	Overall Align	84.2	64.8	84.6	80.9	60.5

7 结论

本研究针对面部与语音特征的跨模态匹配问题,探索了将大规模数据集上学习到的先验知识迁移到该任务的方法. 在声纹特征的处理中,将语音识别任务中的Transformer模型迁移至跨模态身份特征匹配领域;在人脸特征领域,则利用能够表示身份信息的模型作为教师网络,结合关系蒸馏技术保持身份特征间关系. 与使用传统时延网络相比,本研究在验证任务中的效果显著提升,证明在跨模态特征匹配任务中,语音识别任务的预训练同样存在优势. 通过可视化方法的深入分析,本研究直观地展示了训练后声纹特征提取模型的效果,验证了其能够有效地区分不同个体. 这一结果支持了身份信息有效表示对于跨模态匹配任务的重要性. 本研究为未来的研究方向提供了新的视角和潜在的解决方案,不仅加深了对模型性能的理解,也为后续研究提供了参考.

参考文献

[1] ZHU B Q, XU K L, WANG C J, et al. Unsupervised voice-face representation learning by cross-modal prototype contrast[C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. Shen Zhen: International Joint Conferences on Artificial Intelligence Organization, 2022: 3787-3794.

[2] WELLS T, BAGULEY T, SERGEANT M, et al. Perceptions of human attractiveness comprising face and voice cues[J]. Archives of Sexual Behavior, 2013, 42(5): 805-811.

[3] AWWAD SHIEKH HASAN B, 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: 37494.

[4] JOASSIN F, PESENTI M, MAURAGE P, et al. Cross-modal interactions between human faces and voices involved in person recognition[J]. Cortex, 2011, 47(3): 367-376.

[5] MORGADO P, MISRA I, VASCONCELOS N. Robust audio-visual instance discrimination[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 12934-12945.

[6] ARANDJELOVIC R, ZISSERMAN A. Look, listen and learn[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 609-617.

[7] XIE Z W, LI L, ZHONG X, et al. Image-to-video person re-identification with cross-modal embeddings[J]. Pattern Recognition Letters, 2020, 133: 70-76.

[8] NAGRANI A, ALBANIE S, ZISSERMAN A. Learnable PINs: Cross-modal embeddings for person identity[M]// Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 73-89.

[9] WEN Y D, ISMAIL M AL, LIU W Y, et al. Disjoint mapping network for cross-modal matching of voices and faces[EB/OL]. (2018-07-16) [2025-05-06]. https://arxiv. org/abs/1807.04836v2.

[10] NAGRANI A, ALBANIE S, ZISSERMAN A. Seeing voices and hearing faces: Cross-modal biometric matching[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8427-8436.

[11] CHENG K, LIU X, CHEUNG Y M, et al. Hearing like seeing: Improving voice-face interactions and associations via adversarial deep semantic matching network[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 448-455.

[12] VAESSEN N, VAN LEEUWEN D A. Fine-tuning Wav2Vec2 for speaker recognition[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 7967-7971.

[13] ZHANG L, WANG Q, LEE K A, et al. Multi-level transfer learning from near-field to far-field speaker verification[C]//Interspeech 2021. Florida: ISCA, 2021: 1963-1967.

[14] WANG B K, YANG Y, XU X, et al. Adversarial cross-modal retrieval[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM, 2017: 154-162.

[15] DENG J K, GUO J, XUE N N, et al. ArcFace: Additive angular margin loss for deep face recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR). Piscataway: IEEE, 2019: 4685-4694.
[16] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//Interspeech 2020. Florida: ISCA, 2020: 3830-3834.
[17] XIAO Y, ZHOU A C, ZHOU L, et al. Automatic insect identification system based on SE-ResNeXt[J]. International Journal of Systems, Control and Communications, 2023, 14(1): 81.
[18] NAWAZ S, JANJUA M K, GALLO I, et al. Deep latent space learning for cross-modal mapping of audio and visual signals[C]//2019 Digital Image Computing: Techniques and Applications (DICTA). Piscataway: IEEE, 2019: 1-7.
[19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc, 2017: 5998-6008.
[20] COURTY N, FLAMARY R, TUIA D, et al. Optimal transport for domain adaptation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(9): 1853-1865.
[21] DAMODARAN B B, KELLENBERGER B, FLAMARY R, et al. DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 467-483.
[22] ZHANG R T, WEI J G, LU X G, et al. Optimal transport with a diversified memory bank for cross-domain speaker verification[C]//ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2023: 1-5.
[23] GE C J, HUANG R, XIE M X, et al. Domain adaptation via prompt learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 1160-1170.
[24] AGHAJANYAN A, GUPTA S, ZETTLEMOYER L. Intrinsic dimensionality explains the effectiveness of language model fine-tuning[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: USAACL, 2021: 7319-7328.
[25] CAO Q, SHEN L, XIE W D, et al. VGGFace2: A dataset for recognising faces across pose and age[C]//2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Piscataway: IEEE, 2018: 67-74.
[26] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A large-scale speaker identification dataset[C]//Interspeech 2017. Florida: ISCA, 2017: 2616-2620.
[27] PEYRÉ G, CUTURI M. Computational optimal transport: With applications to data science[J]. Foundations and Trends in Machine Learning, 2019, 11(5/6): 355-607.
[28] PARK W, KIM D, LU Y, et al. Relational knowledge distillation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 3962-3971.
[29] XUE Z, GAO Z, REN S, et al. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation[C]//The Eleventh International Conference on Learning Representations. Oxford: ICLR, 2023: 1.
[30] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: A simple data augmentation method for automatic speech recognition[C]//Interspeech 2019. Florida: ISCA, 2019: 2613-2617.
[31] BAEVSKI A, ZHOU Y, MOHAMED A, et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in Neural Information Processing Systems, 2020, 33: 12449-12460.
[32] HSU W N, BOLTE B, TSAI Y H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2021, 29: 3451-3460.

作者简介

孙　婧　女,2000年9月出生于河北省. 上海交通大学自动化系硕士研究生. 主要研究方向为模式识别.
E-mail: sunjing4231@sjtu.edu.cn

苏剑波　男,1969年11月出生于江苏省. 上海交通大学自动化系教授. 主要研究方向为机器视觉、机器学习与人机交互、多传感器信息融合与智能机器人等.
E-mail: jbsu@sjtu.edu.cn