面向跨模态通信的信息恢复技术

徐建博[1,2],魏 昕[1,2],周 亮[1,2]
(1. 南京邮电大学通信与信息工程学院,江苏南京 210003; 2. 南京邮电大学宽带无线通信与传感网技术教育部重点实验室,江苏南京 210003)

摘 要: 针对多模态数据在传输过程中丢失、受到无线信道噪声污染而严重影响跨模态通信质量的问题,提出了一种面向跨模态通信的信息恢复技术,通过充分利用接收端已有数据,采用同模态一对一检索、跨模态一对一检索、跨模态一对多检索等方式,在接收端进行信息恢复. 所提方法在公共数据集以及实际跨模态通信平台上进行验证,实验表明,该方法可以实现精准的信息恢复,有效提升了跨模态通信质量.

关键词: 跨模态通信; 信息恢复; 语义融合; 多模态业务
中图分类号: TP302 **文献标识码:** A **文章编号:** 0372-2112(2022)07-1631-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210945

Information Recovery Technology for Cross-Modal Communications

XU Jian-bo[1,2], WEI Xin[1,2], ZHOU Liang[1,2]
(1. *School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*;
2. *Key Laboratory of Broadband Wireless Communication and Sensor Network Technology (Ministry of Education), Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210003, *China*)

Abstract: Aiming at the issues of multi-modal data loss and data pollution by noise of wireless channel during the transmission, which seriously affect the cross-modal communication quality, an information recovery technology for cross-modal communications is proposed. In this scheme, by making full use of the existing data at the receiving end, the information is recovered at the receiving end by means of one-to-one intra-modal retrieval, one-to-one cross-modal retrieval, one-to-many cross-modal retrieval, etc. Moreover, the proposed scheme is validated on an open data set and the practical cross-modal communication platform. Experimental results show that the scheme can achieve accurate multi-modal information recovery and effectively improve the quality of cross-modal communications.

Key words: cross-modal communications; information recovery; semantic fusion; multi-modal services

1 引言

随着以社交媒体、AR/VR、5G等为代表的无线通信与多媒体技术的快速发展,人们在视听需求得到极大满足的同时,开始追求更多维度、更高层次的感官体验[1,2]. 当前,触觉信息正逐步融入到以音频、视频为代表的传统多媒体业务中,形成了包含音频、视频、触觉信息等多模态业务. 例如,日产汽车公司将HaptX触觉手套与VR头显结合,实现方向盘、各种开关、后视镜调整等更为逼真细腻的架控操作[3]. 为了支撑多模态业务的发展,跨模态通信思想应运而生[4]. 与传统的多媒体通信、触觉互联网[5]相比,跨模态通信是以音频、视频、触觉信息协同传输与处理为典型特征,即充分利用不同模态码流间的相关性,实现多模态码流的高效传输以及处理.

表1给出了音频、视频、触觉码流的通信传输参数需求[4]. 在传输过程中,视频码流因其体积容量大而需要较大的传输带宽,而触觉码流则对低时延、高可靠性要求很高. 从表1中可以发现,触觉码流对抖动、丢包等特别敏感,这对通信环境提出了非常苛刻的要求. 此外,由于音频、视频和触觉码流尺寸差别大,传输时延与速率各不相同,也容易造成终端接收到的多模态码

收稿日期:2021-07-19;修回日期:2021-09-28;责任编辑:王天慧
基金项目:国家自然科学基金(No.62071254);江苏高校优势学科建设工程项目;南京邮电大学宽带无线通信与传感网技术教育部重点实验室开放课题(No.JZNY202111)

流之间存在严重的不同步问题. 并且,无线信道往往带有各种信道噪声,进一步影响了通信质量.

表1 音频、视频、触觉的通信传输参数[4]

特性	音频	视频	触觉
时延	≤150 ms	≤350 ms	[1 ~ 60] ms
抖动	≤30 ms	≤30 ms	[1 ~ 10] ms
数据丢失率	≈30%	≈20%	$\approx 10^{-5}$%
数据传输速率	[22~200] kb/s	[2.5~40] Mb/s	≈128 kb/s

为了解决上述问题,一方面,Yuan等[6]提出一种冗余设备到设备传输方案,可以实现网络的超低时延和超可靠性连接. Zhou等[7]针对海量多媒体业务调度难题,提出了基于数据驱动的高效调度算法. 然而,这些方法仅从功率控制、码流调度等传输角度解决问题,其局限性在于:传输中无线信道环境复杂并且是动态变化的,即使采用了一定机制保障了传输时延和可靠性,数据包受到的干扰、噪声等产生的丢失、时延增加等仍然难以避免.

因此,仅从传输角度考虑,无法有效地解决跨模态通信中面临的问题. 不同于现有复用、调度等传输策略,本文的出发点在于:可否运用信号处理的手段,对实际跨模态通信系统接收端存有已接收到的音频、视频、触觉信号等多模态数据加以合理利用,通过检索的方式,直接弥补接收终端处存在的某些模态数据包的丢失、数据延迟到达、数据不同步等问题,实现信息的有效恢复. 在信息检索方面,Mikolaj等[8,9]研究了关于行人和车辆的边缘无线图像检索问题,提出了一种基于自动编码器的融合信源和信道编码的检索模型,并将其应用于无线信道环境[8],该方法也是无线通信场景中关于信息检索的首个研究工作. 然而,该方法仅仅针对单模态图像的检索问题. 因此,目前尚未出现从信息检索的角度去恢复无线通信环境下的传输丢失或受噪声干扰的多模态码流,并且现有的无线通信环境下的单模态检索方法无法直接扩展到多模态信息检索. 究其原因,主要存在三大关键性挑战:第一,如何构建面向信息恢复的跨模态通信系统架构以充分利用接收端已有数据;第二,如何解决视频、音频、触觉三种不同模态信号之间的相互检索问题;第三,如何解决在无线信道环境下受到干扰或丢失的多模态码流的恢复问题.

为了应对上述挑战,本文提出了面向跨模态通信的信息恢复技术,具体贡献如下:

(1)提出了面向信息恢复的跨模态通信系统架构. 在发送端的边缘节点处将视频、音频、触觉信号三种模态进行预处理并传输,同时考虑数据在无线信道传输过程中受到的信道噪声污染等问题,利用接收端边缘节点处已有的多模态数据实现信息恢复.

(2)设计了一种视频、音频和触觉信号相互检索的信息恢复方法. 运用多模态语义融合技术,挖掘并关联蕴含在三种模态信息中的语义关联性,并将同模态一对一检索、跨模态一对一检索、跨模态一对多检索等方式统一于该方法下进行信息恢复.

(3)通过在公开的多模态数据集以及实际跨模态通信平台采集的数据上进行实验,分析多模态数据在无线信道传输条件下,信噪比和带宽限制对信息恢复效果的影响.

2 相关工作

2.1 跨模态通信

Zhou等[4]提出了跨模态通信思想,构建了一种跨模态流传输架构以及基于信号语义融合和共享的跨模态信号恢复、重建和渲染策略. 在此基础上,针对跨模态传输中出现的低延迟、高可靠、吞吐量高和复杂度低等需求,Zhou等[10]提出一种通用的跨模态流调度方案. 更进一步地,Gao等[11]提出一种面向边缘智能的跨模态流传输架构,将人工智能引入到通信、缓存、计算和控制能力中,并利用基于注意力机制的深度强化学习来解决跨模态流传输优化模型. 另外,与只关注于触觉这一种模态单独作用的触觉互联网不同,跨模态通信旨在音频、视频、触觉信息三者协同作用,使得码流高效传输并利用.

2.2 触觉表示

将机器人应用于触觉感知方面的研究工作也在不断地开展,不同设备采集到的触觉信号拥有不同的表现形式. Liu等[12]构建的数据集包含水果、瓶子等实物,由装有触觉传感器的机械手抓取实物并处理获得三维触觉序列. Luo等[13]构建的数据集包含扳手、剪刀等实物,由机械手握住触觉阵列传感器对实物进行按压并处理获得基于尺度不变特征变换的触觉描述子. Chu等[14]构建的数据集包含约60种实物相关的触觉信息,既有机械手在物体表面移动获得的触觉时间序列信号,也有由人类作为测试者进行收集构建的触觉形容词语料库(“硬的”“有弹性的”等). Ward等[15]开发一种同时采集视觉和触觉数据的装置,其中收集到的触觉信号主要表现为触摸点阵列. 但是,目前针对触觉信号还没有一种普适的预处理和特征提取方法.

2.3 跨模态检索

目前跨模态检索大部分研究工作主要集中在涉及两种模态的检索. 一方面,子空间法[16]通过最大化两种不同模态数据的投影特征来学习同一个公共子空间,其中较为典型的方法包括典型关联分析[17]和核典型关联分析[18]. 核典型关联分析在典型关联分析的基础上引入核函数,将低维数据映射至高维空间中使其具有非线性表达能力. 但是该方法仅仅局限于两种模态.

另一方面,深度学习[19]具有强大的非线性表达和底层特征提取能力,可以有效地提取不同模态的底层特征并在高层进行语义关联.Shang等[20]提出一种基于多模态自编码器的深度学习模型,通过共享层生成图像和文本两种模态的高层通用特征,用于多模态检索.Wang等[21]针对图像和文本这两种模态的高度非线性语义相关性,提出一种正则化深度神经网络模型来学习图像和文本的高层语义映射.然而,现有的跨模态检索方法却无法直接应用于伴有信道噪声污染的无线通信场景中.

为此,本文综合利用深度学习的特征提取能力以及通过公共子空间关联不同模态的语义信息,通过解决三种模态的检索难题,实现接收端的信息恢复,从而提升跨模态通信质量.

3　系统架构

图1展示了面向信息恢复的跨模态通信系统架构.发送端采集设备感知并采集视频、音频、触觉信号三路信息,完成同步后上传至边缘节点进行预处理.对于视频,直接采用视频中的每帧图像;对于音频和触觉信号,进行预加重、分帧、加窗、计算功率谱、滤波器组,将音频和触觉信号都转变成最终的图像[22].将三种模态预处理后得到的图像再进行编码,经过无线信道传输,接收端边缘节点在收到多模态码流后进行解码.考虑到信息可能受到无线信道噪声污染或者丢失问题,对此使用接收端已有的多模态数据替换该信息,实现信息恢复.最后,将该信息传输至接收端设备.与此同时,接收端的触觉控制设备可以向发送端发送位置坐标等指令,控制机械装置移动.

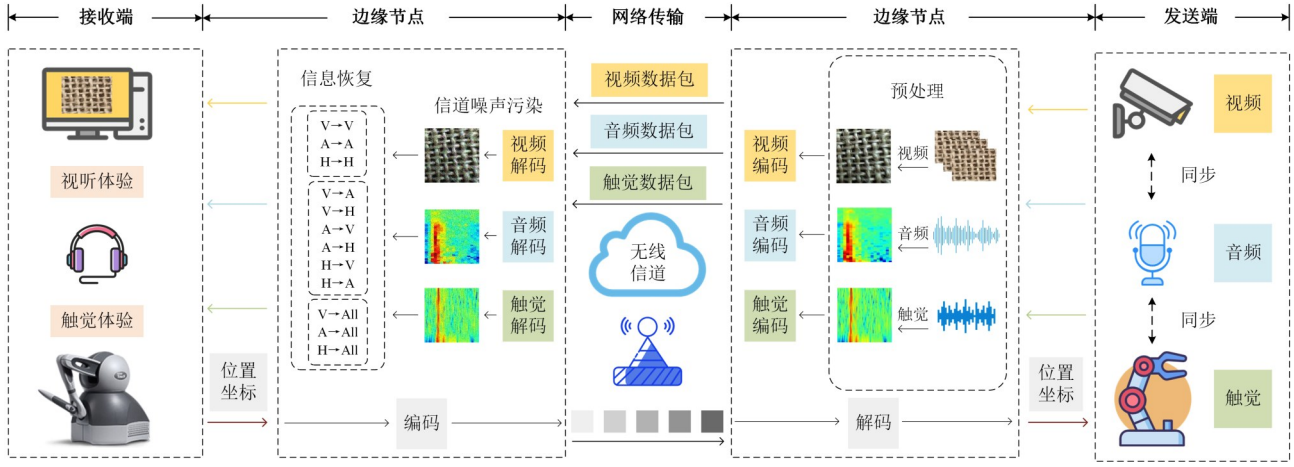


图1　面向信息恢复的跨模态通信系统架构

对于接收端边缘节点处的信息恢复模块而言,根据传输过程中的不同情况,分别采用不同的检索方式,但是我们假设传输过程的情况和对应的检索方式是已知的.具体而言,将视频、音频和触觉信号分别记为V、A和H,所有模态的集合记为All={V,A,H},分为三种典型的情形:同模态一对一检索,包括V→V、A→A、H→H;跨模态一对一检索,包括V→A、V→H、A→V、A→H、H→V、H→A;跨模态一对多检索,包括V→All、A→All、H→All.

情形1:在一段时间内,视频、音频、触觉三种模态信息都受到信道噪声污染.以视频模态为例,这时将受到噪声污染的视频模态信息作为查询样本,从终端的视频数据库中检索出同类视频,用这个检索出的同类视频替换掉传输过程中受到噪声污染的视频,达到信息恢复的目的.即采用同模态一对一检索.

情形2:在一段时间内,某一模态信息发生丢失.假设视频丢失,这时将音频作为查询样本,从终端的视频数据库中检索出一个与音频类别最相似的视频样本,将这个最相似的视频样本替补传输过程中丢失的视频,达到信息恢复的目的.即采用跨模态一对一检索.

情形3:在一段时间内,可能会出现两种模态的丢失.假设视频和音频丢失,触觉未丢失,这时将触觉作为查询样本,从终端数据库中检索出与触觉种类类似的视频或音频模态,替补传输过程中丢失的视频和音频模态.即采用跨模态一对多检索.

上述三种情形涵盖了跨模态通信中所产生的信号丢失情况.针对这三种情形,在接下来的章节中,设计了一种针对视频、音频、触觉信号的信息恢复方法.

4　信息恢复方法

视频、音频、触觉信号三种模态信息恢复方法的总体流程如图2所示.首先将三种模态预处理后得到的图像都输入至使用ImageNet权重的去除末端全连接层的VGG16网络中并扁平化(flatten),得到的一维向量作为各个模态的特征;然后将三种模态经过扁平

层输出的特征分别进行训练,学习更好的特征;再将所有模态共同输入至语义融合模块中训练,实现不同模态深层语义的相互关联;最后进行检索,从而实现信息恢复.

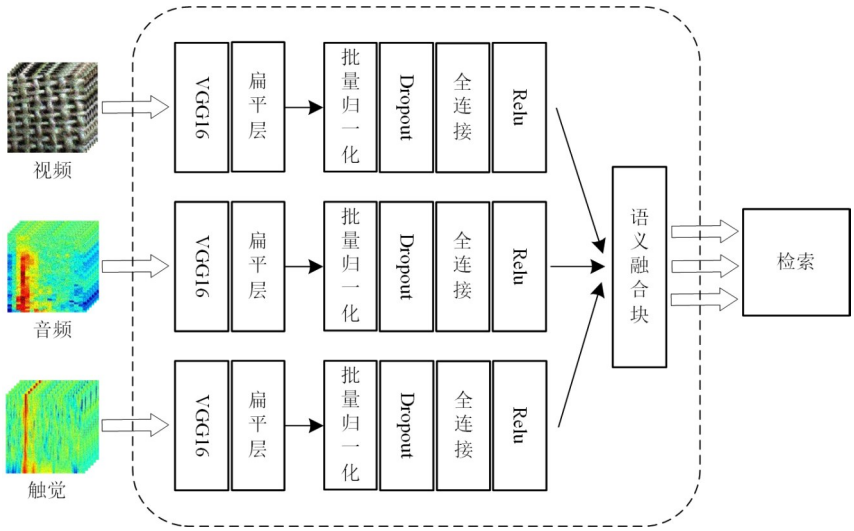


图2 信息恢复方法的总体流程

4.1 特征提取

在对不同模态进行语义融合前,需要将不同模态分别进行训练,进一步提取特征[23,24]. 图3展示了特征提取的网络结构,包括:批量归一化层(Batch Normalization, BN)、Dropout层1、全连接层1(Fully Connected layer, FC)、激活函数Relu、Dropout层2、全连接层2、Softmax函数. 批量归一化层可以对数据进行归一化,加速模型训练,并且具有正则化的效果;Dropout层可以防止模型过拟合;全连接层用来学习特征;激活函数Relu可以增加网络的非线性映射能力;Softmax函数用来对数据进行分类. 网络训练完毕后,移除Dropout2、全连接2、Softmax,并接入语义融合模块中.

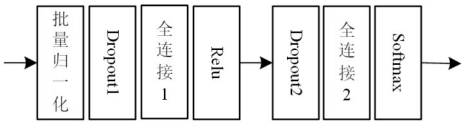


图3 特征提取结构

视频、音频、触觉三种模态的特征提取都采用多分类交叉熵损失函数进行各自的网络训练. 具体而言,视频模态的损失函数为f_{V},音频模态的损失函数为f_{A},触觉模态的损失函数为f_{H},如式(1)~(3)所示.

$$f_{\mathrm{V}}=\frac{1}{N_{\mathrm{V}}}\sum_{k=1}^{N_{\mathrm{V}}}\sum_{j=1}^{C_{\mathrm{V}}}1\left\{y_{\mathrm{V}}^{k}=j\right\}\log\left[p\left(x_{\mathrm{V}}^{k}\right)\right] \tag{1}$$

其中,N_{V}表示视频模态训练数据的样本总数,x_{V}^{k}表示视频模态的第k个样本经过特征提取结构中的Softmax函数输出得到的特征,y_{V}^{k}表示视频模态的第k个样本的标签,标签为$1,2,\cdots,C_{\mathrm{V}}$,其中$C_{\mathrm{V}}$的数值含义也表示视频模态的类别总数.$1\left\{y_{\mathrm{V}}^{k}=j\right\}$为符号函数,当$y_{\mathrm{V}}^{k}$属于标签$j$时,$1\left\{y_{\mathrm{V}}^{k}=j\right\}=1$,否则$1\left\{y_{\mathrm{V}}^{k}=j\right\}=0$,$1\left\{y_{\mathrm{V}}^{k}=j\right\}\log\left[p\left(x_{\mathrm{V}}^{k}\right)\right]$中的$p\left(x_{\mathrm{V}}^{k}\right)$表示视频模态的第$k$个样本$x_{\mathrm{V}}^{k}$的标签为$j$的概率,其中$j$的取值范围是$1,2,\cdots,C_{\mathrm{V}}$.

$$f_{\mathrm{A}}=\frac{1}{N_{\mathrm{A}}}\sum_{k=1}^{N_{\mathrm{A}}}\sum_{j=1}^{C_{\mathrm{A}}}1\left\{y_{\mathrm{A}}^{k}=j\right\}\log\left[p\left(x_{\mathrm{A}}^{k}\right)\right] \tag{2}$$

其中,N_{A}表示音频模态训练数据的样本总数,x_{A}^{k}表示音频模态的第k个样本经过特征提取结构中的Softmax函数输出得到的特征,y_{A}^{k}表示音频模态的第k个样本的标签,标签为$1,2,\cdots,C_{\mathrm{A}}$,其中$C_{\mathrm{A}}$的数值含义也表示音频模态的类别总数.$1\left\{y_{\mathrm{A}}^{k}=j\right\}$为符号函数,当$y_{\mathrm{A}}^{k}$属于标签$j$时,$1\left\{y_{\mathrm{A}}^{k}=j\right\}=1$,否则$1\left\{y_{\mathrm{A}}^{k}=j\right\}=0$,$1\left\{y_{\mathrm{A}}^{k}=j\right\}\log\left[p\left(x_{\mathrm{A}}^{k}\right)\right]$中的$p(x_{\mathrm{A}}^{k})$表示音频模态的第$k$个样本$x_{\mathrm{A}}^{k}$的标签为$j$的概率,其中$j$的取值范围是$1,2,\cdots,C_{\mathrm{A}}$.

$$f_{\mathrm{H}}=\frac{1}{N_{\mathrm{H}}}\sum_{k=1}^{N_{\mathrm{H}}}\sum_{j=1}^{C_{\mathrm{H}}}1\left\{y_{\mathrm{H}}^{k}=j\right\}\log\left[p\left(x_{\mathrm{H}}^{k}\right)\right] \tag{3}$$

其中,N_{H}表示触觉模态训练数据的样本总数,x_{H}^{k}表示触觉模态的第k个样本经过特征提取结构中的Softmax函数输出得到的特征,y_{H}^{k}表示触觉模态的第k个样本的标签,标签为$1,2,\cdots,C_{\mathrm{H}}$,其中$C_{\mathrm{H}}$的数值含义也表示触觉模态的类别总数.$1\left\{y_{\mathrm{H}}^{k}=j\right\}$为符号函数,当$y_{\mathrm{H}}^{k}$属于标签$j$时,$1\left\{y_{\mathrm{H}}^{k}=j\right\}=1$,否则$1\left\{y_{\mathrm{H}}^{k}=j\right\}=0$,$1\left\{y_{\mathrm{H}}^{k}=j\right\}\log\left[p\left(x_{\mathrm{H}}^{k}\right)\right]$中的$p(x_{\mathrm{H}}^{k})$表示触觉模态的第$k$个样本$x_{\mathrm{H}}^{k}$的标签为$j$的概率,其中$j$的取值范围是$1,2,\cdots,C_{\mathrm{H}}$.

4.2 语义融合块

图4展示了语义融合块的整体结构. 三种模态经

过特征提取后,共同输入至该模块中进行语义融合. 相对于三种模态使用不同的网络进行语义融合,使用同一个网络有利于降低模型的复杂性. 更重要的是,针对该模块设计了一种新的损失函数,以此进行不同模态的语义关联. 该损失函数由两部分组成,具体如式(4)所示.

图4　语义融合块结构

$$L=L_{\mathrm{cls}}+\lambda\cdot L_{\mathrm{cen}} \tag{4}$$

其中λ为损失函数的超参数.

部分1　模态内损失:分类约束

$$L_{\mathrm{cls}}=\frac{1}{N_{\mathrm{V}}}\sum_{k=1}^{N_{\mathrm{V}}}g\left(x_{\mathrm{V}}^{k},y_{\mathrm{V}}^{k}\right)+\frac{1}{N_{\mathrm{A}}}\sum_{k=1}^{N_{\mathrm{A}}}g\left(x_{\mathrm{A}}^{k},y_{\mathrm{A}}^{k}\right)+\frac{1}{N_{\mathrm{H}}}\sum_{k=1}^{N_{\mathrm{H}}}g\left(x_{\mathrm{H}}^{k},y_{\mathrm{H}}^{k}\right) \tag{5}$$

为了区分模态的类别,采用式(5)的分类约束函数. 将视频、音频和触觉信号三种模态分别记为V、A和H,N_i表示模态i训练数据的样本总数,x_i^k表示模态i的第k个样本通过语义融合块输出的特征,y_i^k表示模态i的第k个样本对应的标签,标签为$1,2,\cdots,C$,C的数值含义也表示模态i的类别总数,其中$i\in\{\mathrm{V},\mathrm{A},\mathrm{H}\}$. 式(5)中$g(\cdot)$为多分类交叉熵损失函数,如式(6)所示.

$$g\left(x_i^k,y_i^k\right)=\sum_{j=1}^{C}1\left\{y_i^k=j\right\}\log\left[p\left(x_i^k\right)\right] \tag{6}$$

其中,$g\left(x_i^k,y_i^k\right)$表示模态$i$的第$k$个样本的多分类交叉熵损失函数. $1\left\{y_i^k=j\right\}$为符号函数,当y_i^k属于标签j时,$1\left\{y_i^k=j\right\}=1$,否则$1\left\{y_i^k=j\right\}=0$,$1\left\{y_i^k=j\right\}\log\left[p\left(x_i^k\right)\right]$中的$p\left(x_i^k\right)$表示模态$i$的第$k$个样本$x_i^k$的标签为$j$的概率,其中$j$的取值范围是$1,2,\cdots,C$.

部分2　模态间损失:中心约束

$$L_{\mathrm{cen}}=\frac{1}{2}\sum_{m=1}^{N}\left\|x^m-c_m\right\|_2^2 \tag{7}$$

为了进一步使不同模态的相同类别更加紧凑,采用式(7)的中心约束函数. 其中,N表示视频、音频和触觉信号三种模态所有训练数据样本总数,即$N=N_{\mathrm{V}}+N_{\mathrm{A}}+N_{\mathrm{H}}$,$x_m$表示第$m$个训练数据的特征,$c_m$表示第$m$个训练数据对应类别的中心,该中心在模型训练过程中是不断变化的.

4.3　语义融合模块的网络优化

将视频、音频和触觉信号三种模态经过特征提取后,各自输出的n个样本分别记为$v=\left[v_1,v_2,\cdots,v_n\right]$,$a=\left[a_1,a_2,\cdots,a_n\right]$和$h=\left[h_1,h_2,\cdots,h_n\right]$,不同模态的相同类别的样本具有共同的类别标签$y=\left[y_1,y_2,\cdots,y_n\right]$. 本文中不同模态各自的样本数相同. 语义融合模块的优化目标函数如式(8)所示,学习网络映射函数$f(v,a,h;\theta)$,使得$y=f(v,a,h;\theta)$,其中v、a、h分别为视频、音频、触觉模态的n个样本,将样本$v_1,v_2,\cdots,v_n,a_1,a_2,\cdots,a_n,h_1,h_2,\cdots,h_n$依次输入语义融合模块中,并采用新设计的损失函数$L$进行网络优化,$\theta$为语义融合网络参数. 具体网络优化流程如算法1所示.

$$\min_{\theta}L=L_{\mathrm{cls}}+\lambda\cdot L_{\mathrm{cen}} \tag{8}$$

算法1　网络优化算法

输入:依次输入视频样本$v=\left[v_1,v_2,\cdots,v_n\right]$、音频样本$a=\left[a_1,a_2,\cdots,a_n\right]$、触觉样本$h=\left[h_1,h_2,\cdots,h_n\right]$;各自对应的类别标签$y=\left[y_1,y_2,\cdots,y_n\right]$;初始化网络参数$\theta$,一阶矩变量$s$=0,二阶矩变量$r$=0,矩估计的指数衰减速率$\rho_1$和$\rho_2$,用于数值稳定的常数$\delta$;设置批大小,步长$\mu$,超参数$\lambda$和迭代步数$K$

输出:映射函数$f(v,a,h;\theta)$

FOR k=1:K

　根据式(8)更新参数θ

　更新梯度:$g\leftarrow\frac{1}{n}\nabla_{\theta}\left(L_{\mathrm{cls}}+\lambda\cdot L_{\mathrm{cen}}\right)$

　更新有偏一阶矩估计:$s\leftarrow\rho_1 s+\left(1-\rho_1\right)g$

　更新有偏二阶矩估计:$r\leftarrow\rho_2 r+\left(1-\rho_2\right)g^2$

　修正一阶矩的偏差:$\hat{s}\leftarrow\frac{s}{1-\rho_1^{\,k}}$

　修正二阶矩的偏差:$\hat{r}\leftarrow\frac{r}{1-\rho_2^{\,k}}$

　更新网络参数:$\theta\leftarrow\theta-\mu\frac{\hat{s}}{\sqrt{\hat{r}}+\delta}$

END

经过算法1优化的语义融合模块后所输出的视频、音频、触觉特征,如其具有同一或相似语义,则在语义空间中也更为接近. 举个例子,材质为木板的视频、音频、触觉数据经过语义融合模块后,其输出的特征矢量比输入的特征矢量在语义空间中更加相似(因为它们都有共同的语义“木板”).

4.4　检索

将视频、音频、触觉这三种模态所有的样本经过整个模型后,各自模态输出的集合分别记为{V}、{A}和{H},并分别提供一个查询样本q_i和检索样本r_j,具体而言:

情形1:同模态一对一检索,包括$q_i\in\{\mathrm{V}\}\cup r_j\in\{\mathrm{V}\}$、$q_i\in\{\mathrm{A}\}\cup r_j\in\{\mathrm{A}\}$、$q_i\in\{\mathrm{H}\}\cup r_j\in\{\mathrm{H}\}$;

情形2:跨模态一对一检索,包括$q_i\in\{\mathrm{V}\}\cup r_j\in$

$\{A\}$、$q_i \in \{V\} \cup r_j \in \{H\}$、$q_i \in \{A\} \cup r_j \in \{V\}$、$q_i \in \{A\} \cup r_j \in \{H\}$、$q_i \in \{H\} \cup r_j \in \{V\}$、$q_i \in \{H\} \cup r_j \in \{A\}$；

情形3：跨模态一对多检索，包括 $q_i \in \{V\} \cup r_j \in \{V,A,H\}$、$q_i \in \{A\} \cup r_j \in \{V,A,H\}$、$q_i \in \{H\} \cup r_j \in \{V,A,H\}$.

$$\mathrm{Sim}\left\langle q_i, r_j \right\rangle = \frac{q_i \cdot r_j}{\left\| q_i \right\| \cdot \left\| r_j \right\|} \tag{9}$$

利用余弦相似函数度量两个向量之间的距离，如式(9)所示. 其中，i和j分别表示查询样本集合和检索样本集合中样本的序号. 固定i，将j遍历检索样本集合，并由式(9)计算两者之间的余弦值，按照距离从大到小的顺序进行排序，距离越大代表两个样本越相似，输出最相似的结果，替换受到噪声污染或丢失的信息，实现信息恢复.

5 公共数据集实验验证

5.1 多模态公共数据集

实验选取的多模态公共数据集为LMT108表面纹理材质数据集[25]，如图5所示. 该数据集收集了关于表面纹理材质的多模态数据，包括图像、声音信号、加速度信号、摩擦力信号和反射率扫描信号，其中声音和加速度信号又分别包含移动和击打两种采集方式获取得到的数据. 其中，采集加速度信号的装置是三轴加速度计(ADXL345)，其中配置范围是±78.5 m/s^2(±8 g)，采样频率为1 000 Hz. 将三轴加速度计和麦克风都集成在一支笔上，采集者手握这支笔，将笔尖在材料表面进行移动或者用笔尖击打材质得到加速度信号和声音信号. 该数据集共有108种不同的表面纹理材质，可以分为九大类，包括网格、石头、空白光滑表面、木材、橡胶、纤维、泡沫、箔纸类和纺织品面料类，每个大类含有5~17个小类，每种小类材质含有20个样本，即九种类别的材质各含有100~340个样本.

将数据集中的图像、击打得到的声音和加速度信号分别作为实验的视频、音频、触觉信号，并将每种材质的样本打乱，按3∶1∶1的比例重新划分出训练集、验证集和测试集，如表2所示，其中将测试集用于最后的检索，通过检索性能来评价信息恢复的效果.

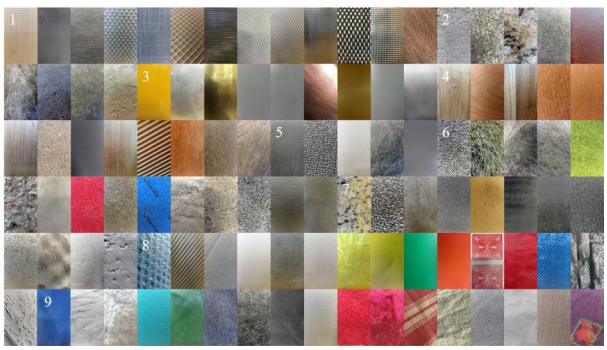


图5 公共数据集

表2 公共数据集划分

材质	训练集	验证集	测试集	总计
小类材质	12	4	4	20
所有材质	1 296	432	432	2 160

5.2 评价指标

实验采用常见评价指标平均精度均值[26](Mean Average Precision，MAP)来评估方法的优劣. 提供一个查询样本，根据检索集中的所有样本求出每个查询样本的平均准确率(Average Precision，AP)，然后对所有查询样本的平均准确率求均值，得到最终的MAP值.

$$\mathrm{MAP} = \frac{1}{Q}\sum_{j=1}^{Q}\mathrm{AP}_j \tag{10}$$

$$\mathrm{AP}_j = \frac{\sum_{i=1}^{R}\mathrm{Pre}(i)\mathrm{Rel}(i)}{\sum_{i=1}^{R}\mathrm{Rel}(i)} \tag{11}$$

计算式如式(10)和式(11)所示. 其中，AP_j表示第j个查询样本的AP值，Q表示查询样本总数，R表示检索集中样本总数，$\mathrm{Pre}(i)$是位置排序i处检索到的样本对应的精确率，$\mathrm{Rel}(i)$是位置排序i处查询样本与检索样本的相关度(如果两者属于同一类别，则$\mathrm{Rel}(i)=1$；否则，$\mathrm{Rel}(i)=0$).

5.3 实验条件

将图1中发送端边缘节点处三种模态预处理得到的图像调整成相同的分辨率，通过分辨率的不同间接地反映出无线信道带宽B的大小. 实验中将分辨率128×128×3、128×96×3、128×64×3分别看成是带宽为128、96、64的情形. 无线信道噪声采用加性高斯白噪声，讨论信噪比范围在-12~15 dB，以及理想信道(B=128、SNR=∞)情形，并且实验弱化了编码和解码过程.

5.4 实验结果与分析

首先，在理想信道情形下，通过MAP值的大小选择合适的全连接层长度. 在图3、图4的网络结构中，实验有九大类别，因此全连接2长度都为固定值9，损失函数分别为多分类交叉熵损失和新设计的损失函数L，优化器都为Adam优化器. 实验中分别选取全连接1长度为64、128、256、512和1 024. 从表3中可以看出，全连接1长度为256时，MAP值最大，其中MAP值表示三种一对一同模态检索和六种一对一跨模态检索MAP值的平均值. 因此，实验选取全连接1长度为256.

然后，对语义融合块中损失函数的超参数λ进行敏

表3　全连接层长度对MAP的影响

全连接1长度	MAP值
64	0.571
128	0.520
256	0.613
512	0.602
1 024	0.606

感性分析,讨论迭代步数(Epoch)和超参数λ对MAP值的影响. 同样,MAP值表示三种一对一同模态检索和六种一对一跨模态检索MAP值的平均值. 选取$\lambda\in\{0.001,0.01,0.1,1,10,100\}$,从图6可以看出:$\lambda\in\{10,100\}$时检索效果最差;$\lambda=1$时检索效果一般,并且在迭代步数为30时,MAP值的变化才趋于平稳;$\lambda\in\{0.001,0.01,0.1\}$时检索效果最好,而且MAP值随迭代步数变化比较平稳. 因此,实验选取$\lambda=0.001$.

接着,将所提方法与4种传统机器学习和3种深度学习方法在公共数据集上进行比较,包括典型关联分析[17](Canonical Correlation Analysis,CCA)、核典型关联分析[18](Kernel Canonical Correlation Analysis,KCCA)、

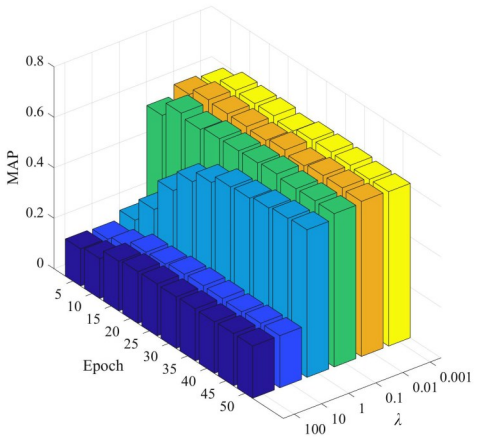


图6　参数λ敏感性分析

主成分分析[27](Principal Component Analysis,PCA)、独立成分分析[28](Independent Component Analysis,ICA)、自编码器[29](Auto-Encoder,AE)、变分自编码器[30](Variational Auto-Encoder,VAE)、注意力机制[31](Attention),其中KCCA的核函数选取高斯核函数. 从表4的实验结果表明,本文所提方法的MAP值远远优于其他方法.

表4　各种方法的MAP值比较

方法	V→V	V→A	V→H	A→V	A→A	A→H	H→V	H→A	H→H	V→All	A→All	H→All
CCA	0.179	0.140	0.138	0.143	0.168	0.139	0.139	0.138	0.185	0.141	0.140	0.146
KCCA	0.199	0.198	0.199	0.201	0.217	0.205	0.200	0.216	0.216	0.201	0.205	0.205
PCA	0.302	0.144	0.152	0.161	0.250	0.159	0.172	0.151	0.279	0.183	0.165	0.175
ICA	0.297	0.147	0.144	0.152	0.247	0.178	0.140	0.178	0.268	0.188	0.175	0.181
AE	0.303	0.154	0.165	0.154	0.242	0.151	0.157	0.145	0.290	0.188	0.162	0.178
VAE	0.296	0.149	0.155	0.151	0.248	0.161	0.153	0.154	0.295	0.187	0.170	0.186
Attention	0.673	0.453	0.367	0.485	0.425	0.338	0.420	0.353	0.338	0.506	0.403	0.355
本文	0.921	0.702	0.588	0.712	0.560	0.474	0.636	0.479	0.451	0.762	0.580	0.510

最后,考虑误码对实验性能的影响. 具体而言,考虑训练集和测试集的样本具有相同的信噪比,分析带宽B和信噪比SNR对信息恢复效果的影响. 图7展示了在加性高斯白噪声信道条件下,带宽、信噪比和MAP值的关系. 实验结果表明:一方面,带宽越大,MAP值越大,信息恢复效果越好;另一方面,当信噪比低于0 dB时,曲线增长比较陡峭,而当信噪比高于0 dB时,曲线增长相对较平缓,但是都低于理想信道情形. 总之,信噪比越大,带宽越大,信息恢复效果越好.

6　跨模态通信平台性能验证

6.1　跨模态通信平台

图8展示了实际搭建的跨模态通信系统平台. 在该平台的发送端,采用4 K高清HDMI摄像头采集视频,机械手指装有麦克风和TeckScan薄膜压力传感器分别采集音频和触觉信号,其中TeckScan薄膜压力传感器的采样频率约200 Hz,通过该机械手在材料表面移动,TeckScan薄膜压力传感器会收集到机械手指尖按压材料得到的压力信号. 在平台的接收端,用户使用Geomagic Touch力反馈设备控制UR3机械臂和机械手触摸材质,可以获得触觉体验,并在装有64位操作系统的笔记本电脑上进行实时显示;与此同时,根据在Unity 3D里搭建的虚拟环境,用户可以获得视听体验. Touch力反馈设备和UR3机械臂机械手之间进行双向通信,力反馈设备向机械装置传递位置坐标,机械装置向接收端笔记本和力反馈设备传递视频、音频和触觉信号三路信息.

6.2　所接收到的多模态数据

图9展示了跨模态通信平台接收到的部分表面纹理材质数据,从左往右依次为石板、木板、纸板、丝绸、泡沫、黄铜、亚麻布、气泡膜和孔状塑料片,以及对应的音频和机械手指压力信号,其中机械手指压力信号作为实验所需的触觉信号.

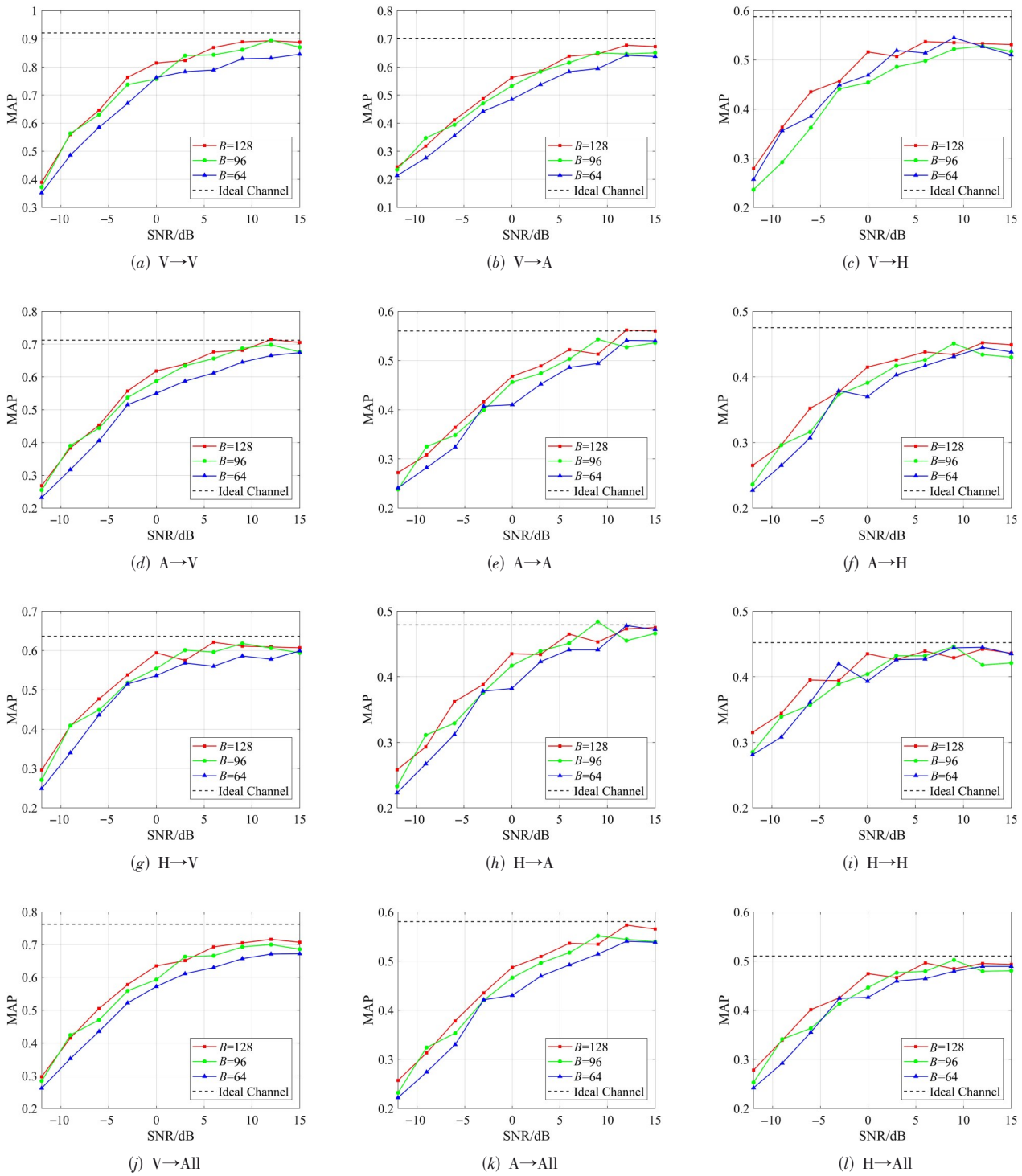


图7 加性高斯白噪声信道下MAP值比较

6.3 实验结果与分析

对于实际跨模态通信系统平台，考虑误码对实验性能的影响. 固定带宽，考虑训练集和测试集的样本具有不同的信噪比. 图10展示了在加性高斯白噪声信道条件下，不同信噪比的训练集、测试集和MAP值的关系. 通过提供低信噪比的查询样本，在接收端已有数据中检索出高质量的样本. 从图10的曲线可以看出，当用于检索的样本质量越来越高（即信噪比越来越大）时，曲线总体上呈上升趋势，说明检索效果越来越好. 当训练样本的信噪比过低时，比如-3 dB，检索性能可

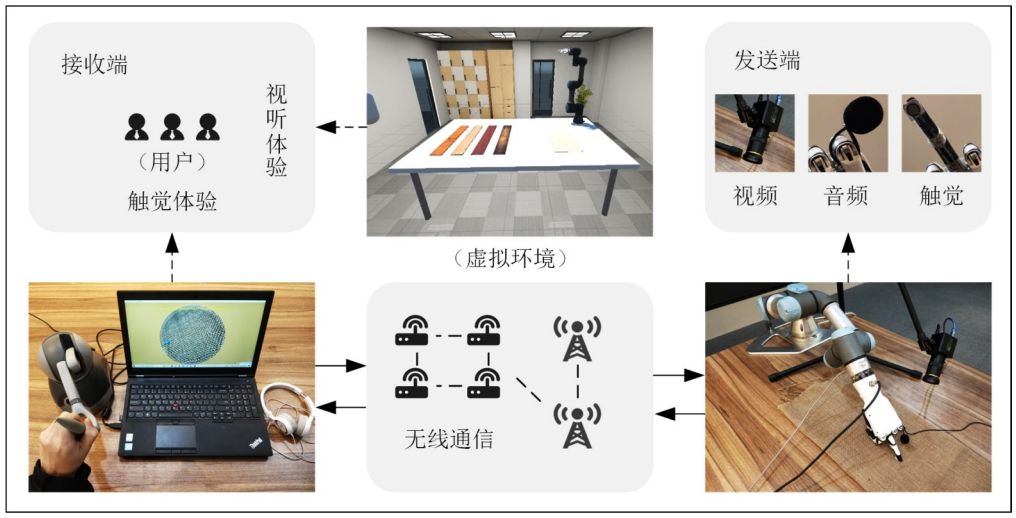
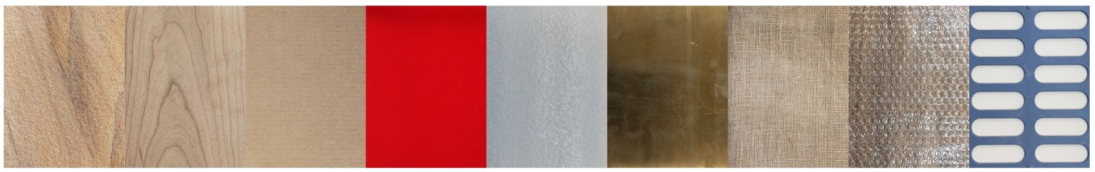
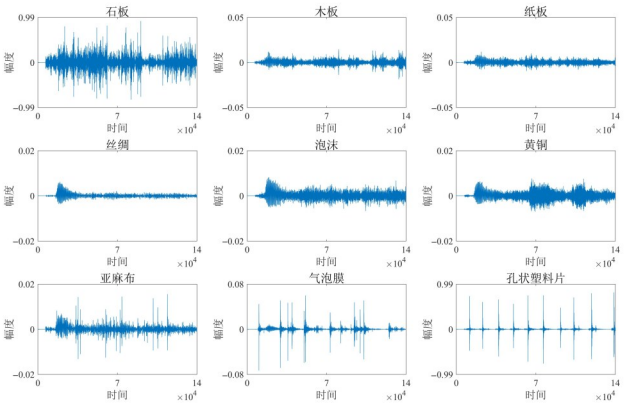


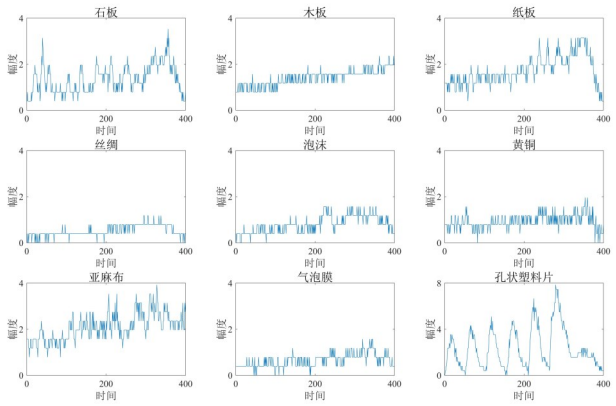
图8　实际跨模态通信系统平台



(*a*)　九种表面纹理材质



(*b*)　九种材质的音频



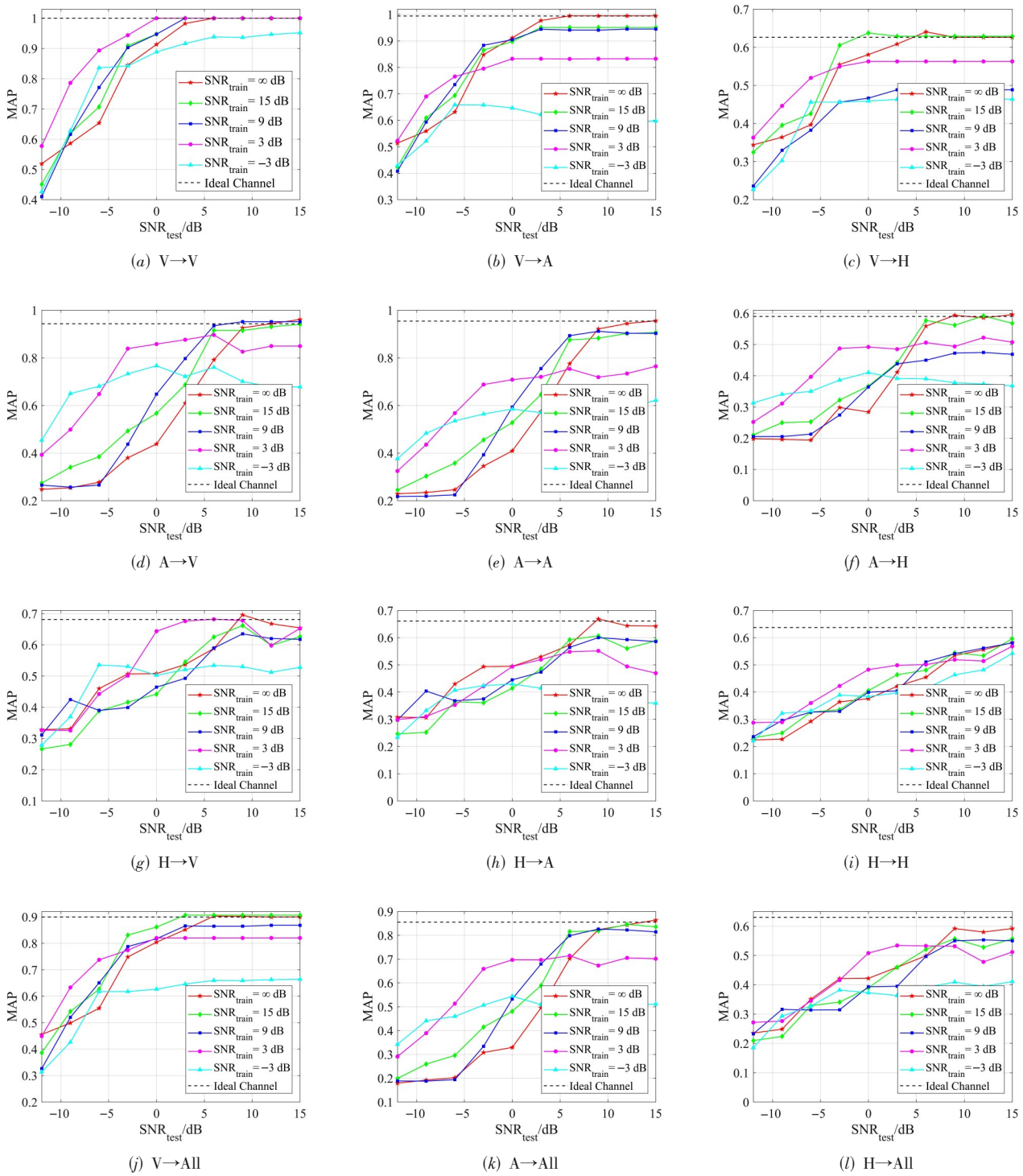
(*c*)　九种材质的触觉信号

图9　所接收到的多模态数据

能会发生急剧恶化. 当测试集和训练集样本的信噪比都比较低时,MAP值较小;当训练样本的信噪比低,即使测试样本质量高,也获得较低的MAP值;当训练集和测试集都有较高的信噪比时,MAP值较高. 因此,如果模型无法在拥有充足的高质量数据下进行训练时,可以适当弱化训练数据的质量,这也会获得较好的检索性能.

7　结束语

本文提出了面向跨模态通信的信息恢复技术,通过利用模态间的语义相关性实现跨模态信息恢复,以解决多模态码流传输过程中的丢失以及受到的无线信道噪声污染问题. 首先,提出了一种面向信息恢复的跨模态通信系统架构,并针对数据在无线信道传输过程中遇到的不同问题,讨论了同模态一对一检索、跨模态一对一检索、跨模态一对多检索等方式. 接着,设计了具体的信息恢复方法,挖掘视频、音频、触觉信号三种模态间的深层语义关联,实现上述三类检索方式,通过检索达到信息恢复的目的. 最后,在带宽受限、存在无线信道噪声的信道条件下,通过在公开的多模态数据集和实际跨模态通信平台采集的数据上进行实验,分析了不同的信噪比和带宽限制对信息恢复效果的影



(a) V→V (b) V→A (c) V→H

(d) A→V (e) A→A (f) A→H

(g) H→V (h) H→A (i) H→H

(j) V→All (k) A→All (l) H→All

图10 加性高斯白噪声信道下MAP值比较

响,仿真结果验证了所提方法的有效性.

未来工作将进一步探讨在实际跨模态通信平台上音频、视频、触觉信号三种模态的信息恢复问题. 具体包括三个方面:第一,在实际系统方面,目前跨模态通信平台采集触觉数据的方式仅为机械手的单根手指(单点),而在实际中只有通过多点触摸才能更加全面地了解物体的质感,因此需要对触觉信息采集装置做进一步的改进,这将间接性地影响到信息恢复性能;第二,在算法方面,目前的检索方法一般忽略了丰富的上下文信息,如何充分关联上下文信息进行细粒度跨模态信息恢复也是未来需要研究的问题;第三,在跨模态通信架构方面,目前实验忽略了编码和解码过程,未来

将讨论在融合编解码过程和信道条件下的信息恢复效果.

参考文献

[1] CHEN S Z, LIANG Y C, SUN S H, et al. Vision, requirements, and technology trend of 6G: How to tackle the challenges of system coverage, capacity, user data-rate and movement speed[J]. IEEE Wireless Communications, 2020, 27(2): 218-228.

[2] 张宏科, 冯博昊, 权伟. 智融标识网络基础研究[J]. 电子学报, 2019, 47(5): 977-982.
ZHANG H K, FENG B H, QUAN W. Fundamental research on smart integration identifier networking[J]. Acta Electronica Sinica, 2019, 47(5): 977-982. (in Chinese)

[3] CARLTON B. Nissan partners with HaptX to bring realistic touch to vehicle design[EB/OL]. (2019-03-08)[2022-04-26]. https://vrscout. com/news/nissan-haptx-vr-vehicle-design/.

[4] ZHOU L, WU D, CHEN J X, et al. Cross-modal collaborative communications[J]. IEEE Wireless Communications, 2020, 27(2): 112-117.

[5] MOSKVITCH K. Tactile Internet: 5G and the cloud on steroids[J]. Engineering & Technology, 2015, 10(4): 48-53.

[6] YUAN Z, WEI X, CHEN J X, et al. Ultra-reliability connectivity with redundant D2D transmission scheme for tactile Internet[C]//2019 IEEE Globecom Workshops. Waikoloa, HI: IEEE, 2019: 1-6.

[7] ZHOU L. On data-driven delay estimation for media cloud [J]. IEEE Transactions on Multimedia, 2016, 18(5): 905-915.

[8] JANKOWSKI M, GÜNDÜZ D, MIKOLAJCZYK K, et al. Wireless image retrieval at the edge[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(1): 89-100.

[9] JANKOWSKI M, GÜNDÜZ D, MIKOLAJCZYK K. Deep joint source-channel coding for wireless image retrieval[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 5070-5074.

[10] ZHOU L, WU D, WEI X, et al. Cross-modal stream scheduling for eHealth[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(2): 426-437.

[11] GAO Y, WEI X, KANG B, et al. Edge intelligence empowered cross-modal streaming transmission[J]. IEEE Network, 2021, 35(2): 236-243.

[12] LIU C F, HUANG W B, SUN F C, et al. LDS-FCM: A linear dynamical system based fuzzy C-means method for tactile recognition[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(1): 72-83.

[13] LUO S, MOU W X, ALTHOEFER K, et al. Novel tactile-SIFT descriptor for object shape recognition[J]. IEEE Sensors Journal, 2015, 15(9): 5001-5009.

[14] CHU V, MCMAHON I, RIANO L, et al. Robotic learning of haptic adjectives through physical interaction[J]. Robotics and Autonomous Systems, 2015, 63(3): 279-292.

[15] WARD-CHERRIER B, PESTELL N, LEPORA N F. NeuroTac: A neuromorphic optical tactile sensor applied to texture recognition[C]//2020 IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020: 2654-2660.

[16] 李志欣, 凌锋, 张灿龙, 等. 融合两级相似度的跨媒体图像文本检索[J]. 电子学报, 2021, 49(2): 268-274.
LI Z X, LING F, ZHANG C L, et al. Cross-media image-text retrieval with two level similarity[J]. Acta Electronica Sinica, 2021, 49(2): 268-274. (in Chinese)

[17] HARDOON D R, SZEDMAK S, SHAWE-TAYLOR J. Canonical correlation analysis: An overview with application to learning methods[J]. Neural Computation, 2004, 16(12): 2639-2664.

[18] AKAHO S. A kernel method for canonical correlation analysis[J]. Tsukuba, Japan, 2006: 263-269.

[19] 周沛, 陈后金, 于泽宽, 等. 跨模态医学图像预测综述[J]. 电子学报, 2019, 47(1):220-226.
ZHOU P, CHEN H J, YU Z K, et al. Review of cross-modality medical image prediction[J]. Acta Electronica Sinica, 2019, 47(1): 220-226. (in Chinese)

[20] SHANG X D, ZHANG H W, CHUA T S. Deep learning generic features for cross-media retrieval[C]//MMM 2016: Proceedings, Part I, of the 22nd International Conference on MultiMedia Modeling. Miami, FL: Springer, 2016: 264-275.

[21] WANG C, YANG H J, MEINEL C. Deep semantic mapping for cross-modal retrieval[C]//2015 IEEE 27th International Conference on Tools with Artificial Intelligence. Vietri sul Mare: IEEE, 2015: 234-241.

[22] FAYEK H. Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients(MFCCs) and what´s in-between[Z/OL]. (2016-04-21) [2022-04-26]. https://haythamfayek. com/2016/04/21/speech-processing-for-machine-learning.html.

[23] SALVADOR A, HYNES N, AYTAR Y, et al. Learning cross-modal embeddings for cooking recipes and food im-

ages[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 3068-3076.

[24] HORIGUCHI S, KANDA N, NAGAMATSU K. Face-voice matching using cross-modal embeddings[C]//Proceedings of the 26th ACM international conference on Multimedia. Seoul: ACM, 2018: 1011-1019.

[25] STRESE M, SCHUWERK C, IEPURE A, et al. Multimodal feature-based surface material classification[J]. IEEE Transactions on Haptics, 2017, 10(2): 226-239.

[26] 张峰, 钟宝江. 基于兴趣目标的图像检索[J]. 电子学报, 2018, 46(8):1915-1923.
ZHANG F, ZHONG B J. Image retrieval based on interested objects[J]. Acta Electronica Sinica, 2018, 46(8): 1915-1923. (in Chinese)

[27] LU C Y, FENG J S, CHEN Y D, et al. Tensor robust principal component analysis with a new tensor nuclear norm [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(4): 925-938.

[28] ECKERT M A, KAMDAR N V, CHANG C E, et al. A cross-modal system linking primary auditory and visual cortices: Evidence from intrinsic fMRI connectivity analysis[J]. Human Brain Mapping, 2008, 29(7): 848-857.

[29] VINCENT P, LAROCHELLE H, LAJOIE I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. Journal of Machine Learning Research, 2010, 11 (12): 3371-3408.

[30] WU Y L, WANG S H, HUANG Q M. Multi-modal semantic autoencoder for cross-modal retrieval[J]. Neurocomputing, 2019, 331: 165-175.

[31] 秦姣华, 黄家华, 向旭宇, 等. 基于卷积神经网络和注意力机制的图像检索[J]. 电讯技术, 2021, 61(3): 304-310.
QIN J H, HUANG J H, XIANG X Y, et al. Image retrieval based on convolutional neural network and attention mechanism[J]. Telecommunication Engineering, 2021, 61 (3): 304-310. (in Chinese)

作者简介

徐建博 男,1996 年 12 月出生,江苏高邮人. 南京邮电大学通信与信息工程学院硕士研究生. 主要研究方向为多媒体通信.
E-mail: xujianbo8881996@163.com

魏 昕 男,1983 年 1 月出生,江苏南京人. 博士,南京邮电大学教授,硕士生导师. 主要研究方向为多媒体通信.
E-mail: xwei@njupt.edu.cn

周 亮(通讯作者) 男,1981 年 11 月出生,安徽芜湖人. 博士,南京邮电大学教授,博士生导师. 主要研究方向为多媒体通信.
E-mail: liang.zhou@njupt.edu.cn