智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

机器人视觉听觉融合的感知操作系统

王业飞, 葛泉波, 刘华平, 陆振宇

引用本文:
王业飞,葛泉波,刘华平,陆振宇. 机器人视觉听觉融合的感知操作系统[J]. 智能系统学报, 2023, 18(2): 381–389.
WANG Yefei,GE Quanbo,LIU Huaping,LU Zhenyu. A perceptual manipulation system for audio–visual fusion of robots[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(2): 381–389.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202111036

DOI: 10.11992/tis.202111036
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.tp.20221025.1544.004.html

机器人视觉听觉融合的感知操作系统

王业飞[1], 葛泉波[2], 刘华平[3], 陆振宇[4]

（1. 南京信息工程大学 电子与信息工程学院, 江苏 南京 210044; 2. 南京信息工程大学 自动化学院, 江苏 南京 210044; 3. 清华大学 计算机科学与技术系, 北京 100084; 4. 南京信息工程大学 人工智能学院, 江苏 南京 210044）

摘　要: 智能机器人面对复杂环境的操作能力一直是机器人应用领域研究的前沿问题, 指称表达是人类对指定对象定位通用的表述方式, 因此这种方式常被利用到机器人的交互当中, 但是单一视觉模态并不足以满足现实世界中的所有任务。因此本文构建了一种基于视觉和听觉融合的机器人感知操作系统, 该系统利用深度学习算法的模型实现了机器人的视觉感知和听觉感知, 捕获自然语言操作指令和场景信息用于机器人的视觉定位, 并为此收集了 12 类的声音信号数据用于音频识别。实验结果表明: 该系统集成在 UR 机器人上有良好的视觉定位和音频预测能力, 并最终实现了基于指令的视听操作任务, 且验证了视听数据优于单一模态数据的表达能力。

关键词: 视觉定位; 音频识别; 深度学习; 视觉感知; 听觉感知; 视听融合; 多模态数据; 主动操作
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)02−0381−09

中文引用格式: 王业飞, 葛泉波, 刘华平, 等. 机器人视觉听觉融合的感知操作系统 [J]. 智能系统学报, 2023, 18(2): 381–389.
英文引用格式: WANG Yefei, GE Quanbo, LIU Huaping, et al. A perceptual manipulation system for audio-visual fusion of robots[J]. CAAI transactions on intelligent systems, 2023, 18(2): 381–389.

A perceptual manipulation system for audio-visual fusion of robots

WANG Yefei[1], GE Quanbo[2], LIU Huaping[3], LU Zhenyu[4]

(1. School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China; 3. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; 4. School of AI, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: The ability of intelligent robots to function in complex environments has been a longstanding challenge in the field of robotic applications. Referential expressions are frequently utilized for object positioning, making this method a common approach in robot interactions. However, relying on a single visual modality alone is not adequate for all tasks in real-world scenarios. This study proposes a robot perception system based on the fusion of visual and auditory modalities. The system employs a deep learning algorithm model to realize the visual and auditory perceptions of the robot, and it processes natural language and scene information for visual positioning and collects data from 12 types of sound signals for audio recognition. The experimental results indicate that the system integrated into the UR robot has a strong visual positioning ability and audio prediction, and it has successfully carried out an instruction-based audio-visual operation task. The results confirm that audio-visual data has a higher expressive capability than single-modal data.

Keywords: visual positioning; audio recognition; deep learning; visual perception; auditory perception; audio-visual fusion; multi-modal data; active operation

机器人正在逐渐进入人类的生活当中, 为了有效地帮助人类, 机器人必须尽可能地学习人类的各项能力, 包括用视觉感知去观察世界、理解人类的自然语言指令, 甚至借助听觉、触觉等获取多模态的信息感受物理世界以进行更多复杂的任务。随着人工智能技术的不断发展, 在视觉识别[1-2]、自然语言系统[3-4]、三维场景建模[5-6]、操作

收稿日期: 2021−11−18.　**网络出版日期:** 2022−10−26.
基金项目: 国家自然科学基金项目 (U1613212).
通信作者: 刘华平. E-mail: hpliu@tsinghua.edu.cn.

抓取以及运动规划[7-8]方面都取得了极大的进展，使得各种先进的计算模型能够部署在机器人上帮助其更加智能化，从而高效稳定地辅助人类完成更加复杂困难的任务。如最先进的具身指示表达的机器人导航任务 (remote embodied visual referring expression in real indoor environments, REVERIE)[9]，该系统将视觉、语言和机器人的行为共同进行学习以帮助机器人探索环境来找到目标对象。这是一种十分具有挑战性的任务，因为它不仅需要对具体目标进行定位，还需要对目标和其位置关系进行高层次的语义理解，用以帮助区分正确的物体和不相关的指示物体。在此基础上，文献 [10] 开发了一种混合控制的机器人系统，它赋予了机器人更加复杂的操作能力，该系统能够根据自然语言的操作指令对目标物体进行拾取和放置。对于有歧义的操作指令或者任务场景，文献 [11] 设计了一种部分可观测的马尔可夫模型 (partially observable Markov decision process, POMDP) 用于观察历史操作记录以帮助机器人排除有歧义的目标。为了方便人与机器人更加有效直观的交互，文献 [12] 设计了一种不受限制的自然语言交互架构，能够在没有辅助信息的支持的情况下实现自然语言的消歧和查询。

然而，单纯依靠视觉信息并不足以支持机器人完成所有类型的任务。对于现实的物理世界，机器人需要配备不同类型的传感器获取更多的模态信息，如听觉信息[13-15]、触觉信息[16-17]、雷达信息[18-19]、多传感器融合信息[20-21]。为了提升机器人的自主导航探索能力，文献 [22] 在捕获视觉信息的基础上，结合音频感官信息嵌入到机器人的路径规划器当中，提高了机器人的导航精度。文献 [23] 通过给实际机器人配备听觉传感器，操作目标物体收集听觉数据，实现了对视觉上难以区分的目标的判别。在此基础上，文献 [24-25] 增加了触觉传感器，采集了不同材质的电压值信息作为触觉感知，构建了一个触觉和听觉融合的机器人分类系统，大大提高了机器人的工作能力。

上述研究虽然取得了很大的进步，但是缺少了部分与人的交互能力，如何让机器人接收人的操作指令，利用多模态信息共同决策操作行为还是一个很大的挑战。为使配备多传感器的机器人系统能够适应更加复杂的操作环境，本文借助于视觉传感器和听觉传感器，构建了一个视听融合的指称表达的机器人自主操作系统。该系统能够接收人类的自然语言操作指令，理解指令中的高级语义信息，结合视觉目标进行定位，并且根据听觉信息进一步判别目标类别。在真实的物理环境中，该系统能够在设计的实验下稳定地发挥性能。主要贡献如下：

1) 本文提出了一个新的视听操作任务，利用视觉信息和音频信息用于解释指示表达的操作指令。

2) 本文在构建的数据集下，实现了机器人的视觉定位和音频识别，用于完成目标操作任务。

3) 本文将实验系统应用在实际机器人中并进行实验验证，实验结果表明本多模态数据对于机器人操作效率有着显著的提升。

1　机器人视听系统

本文利用 UR 机械臂作为机器人平台构建了视听融合的具身操作系统，整个系统架构如图 1 所示。

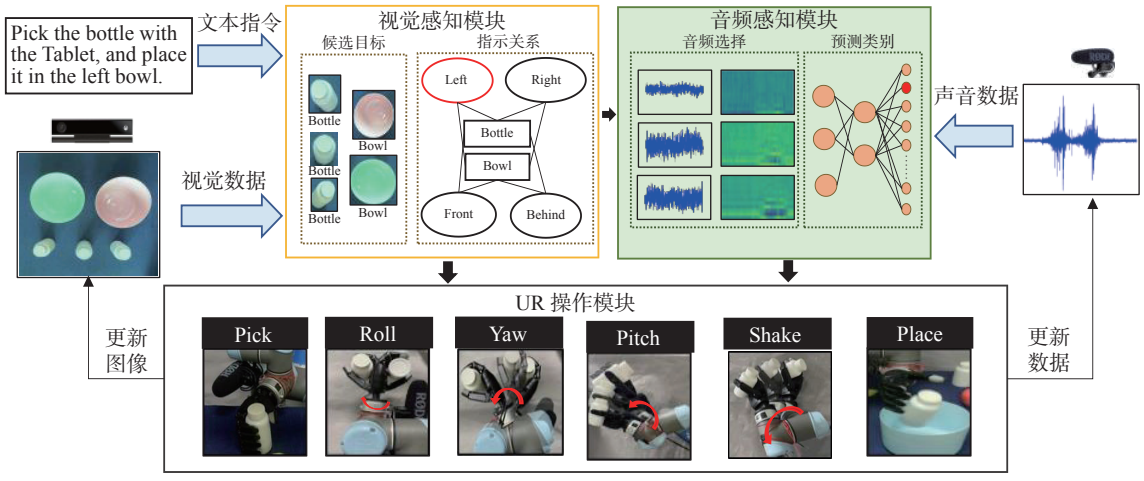


图 1　本文实验系统架构

Fig. 1　Experimental system architecture of this article

其中机器人的任务目标是根据复杂的自然语言指令完成操作任务，并且结合视觉信息定位任

务目标, 利用音频信息判别目标物体。当机器人接收到给定的自然语言指令, 如“拿起带有胶囊的瓶子, 并且放置在左边的盒子”。这就需要机器人通过捕捉视觉信息定位场景中的瓶子和盒子, 并且理解带有位置关系的语句找到左边的盒子。对于视觉上相同的瓶子, 机器人通过选择不同的操作行为, 采集瓶子晃动的声音信息, 进行判别, 最终找到带有胶囊的目标瓶子。整个系统要求机器人能够正确地理解给定的指令, 并且结合指示表达定位目标从而实现相应的操作。

本系统的架构主要分为 3 个模块, 分别是视觉语言感知模块、音频感知模块以及机器人操作模块。首先, 将文本指令和视觉信息输入到视觉语言模块当中, 对可能的目标对象进行定位。当视觉信息不足以判断目标物体的类别时, 机器人的操作模块会产生不同的动作摇晃目标, 声音传感器记录下声音信息, 音频感知模块进行分析, 识别指令中涉及的目标对象, 完成相应的操作任务。

2　机器人视听模型

对于不同的感知模块, 利用深度学习算法设计相应的网络构建整个系统。本文的模型分为指示表达模型、音频分类模型以及机器人的操作模型。

2.1　指示表达模型

不同于基础的目标检测, 本文利用操作指令中涉及到的物体指称关系与视觉信息进行匹配, 利用高级语义关系定位目标物体。

对于给定的图像 $\boldsymbol{I}$, 任务目标是定位图像 $\boldsymbol{I}$ 中的一个子区域, 该子区域对应操作指令中的语义信息。对于操作指令, 首先对其每个单词进行编码转成独热向量, 然后利用循环神经网络提取其编码后的文本特征。

对于图像部分, 利用在 ImageNET 上预训练好的卷积神经网络提取其图像特征和 YOLO 提取图像内的候选目标。对于完整的操作指令, 分为 3 个组成部分, 分别是主体描述、位置描述和关系描述, 对于不同的句子部分, 利用语言注意力机制网络提取其相应的权重与图像特征进行匹配。

本文指示表达模型如图 2 所示, 图像编码部分利用 Darknet53 和特征金字塔网络提取原始图像 $\boldsymbol{I}$ 不同尺度的特征 $f=(f_{v1},f_{v2},f_{v3})$。

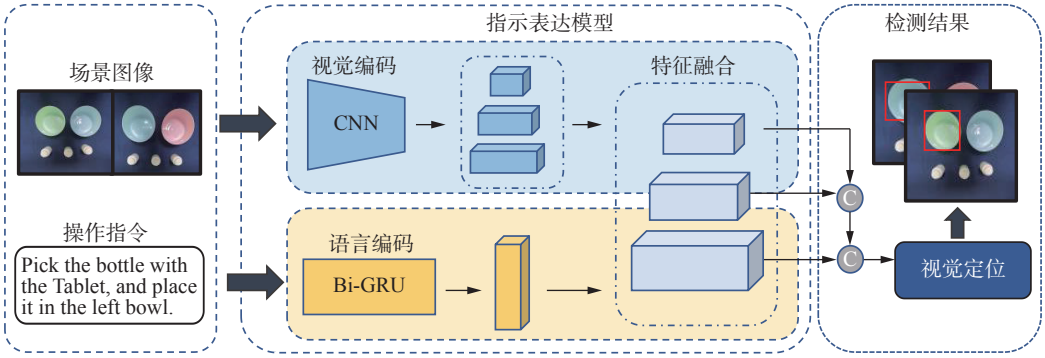


图 2　指示表达模型

Fig. 2　Referring expression model

在指令编码部分, 利用独热编码的方式将操作指令 $E=(w_1,w_2,\cdots,w_T)$ 转化成独热码的形式:

$$\boldsymbol{e}_t=\text{Onehot}(w_t)$$

将编码后的词向量 $\boldsymbol{e}_t$ 以正序和逆序的方式送入双向 GRU 网络中获取相应文本的特征:

$$\vec{\boldsymbol{h}}_t=\text{GRU}\left(e_t,\vec{\boldsymbol{h}}_{t-1}\right)$$

$$\overleftarrow{\boldsymbol{h}}_t=\text{GRU}\left(e_t,\overleftarrow{\boldsymbol{h}}_{t-1}\right)$$

将提取的文本特征相连以获取上下文语义特征向量 $\boldsymbol{h}_t=\left[\vec{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t\right]$。利用上采样过程将文本特征映射到与图像特征相同的维度进行融合:

$$\boldsymbol{f}_{m_i}=\sigma(\boldsymbol{h}_t\boldsymbol{W}_t)\odot\sigma(\boldsymbol{f}_{v_i}\boldsymbol{W}_{v_i})$$

式中: σ 为激活函数, $\boldsymbol{W}_t$ 和 $\boldsymbol{W}_{v_i}$ 为对应的参数矩阵, $\odot$ 为点乘。利用多模态融合特征 $\boldsymbol{f}_{m_i}$ 与目标检测器生成的候选区域 r_{ioc} 进行匹配得到候选区域:

$$u_{ioc}=\text{Softmax}\left((\boldsymbol{W}_m\boldsymbol{f}_{mi}+\boldsymbol{b}_m)\otimes(\boldsymbol{W}_r\boldsymbol{r}_{ioc}+\boldsymbol{b}_r)\right)$$

其中: $\boldsymbol{W}_m$ 和 $\boldsymbol{W}_r$ 是相应的学习参数, $\boldsymbol{b}_m$ 和 $\boldsymbol{b}_r$ 是对应的偏置系数, $\otimes$ 是矩阵相乘。最终目标区域选取两者得分最高的区域 $\mathcal{D}(u_{loc},r_{loc})$ 作为最终预测位置, 该区域用一个组合向量 $\{\boldsymbol{t}_x,\boldsymbol{t}_y,\boldsymbol{t}_w,\boldsymbol{t}_h\}$ 表示, 分别代表了预测框的坐标及尺寸。

2.2　音频分类模型

对于机器人的听觉感知部分, 本文设计了一个音频分类模型, 用于对收集的声音信号进行预

测分类。为了将结构化的声音输入进模型中，需要提取声音信号中特有的梅尔倒频谱系数（Mel-frequency cepstral coefficients, MFCC）特征，首先将时域上的信号$x(t)$进行预加重处理，通过滤波系数α=0.97 过滤掉其中的低频噪声，保留高频分量的信息：

$$x(t) = x_{t+1} - \alpha x_t$$

接着将处理后的特征进行 N 帧分割，利用汉明窗$x(t)\times w(n)$提取局部稳定的信号：

$$w(n) = \begin{cases} 0.54 - 0.46\cos[2\pi n/(N-1)], & 0 \leqslant n \leqslant N \\ 0, \quad \text{其他} \end{cases}$$

对分割后的信号进行短时傅里叶变换和梅尔滤波获取对数频率上的尺度的特征$L(m)$。为了减少特征之间的线性相关性，取低频系数进行离散余弦变换：

$$M(n) = \sum_{m=0}^{N-1} L(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1,2,\cdots,L$$

式中：M 为梅尔滤波器个数；L 为阶数，最后将变换后的特征进行归一化以增加音频信号的信噪比，得到最终可以输入模型的 MFCC 特征。

音频分类模型的网络主要结构如图 3 所示，将不同机械臂动作产生的音频信号提取 MFCC 特征进行拼接，为了保证声音信号的连续性，采用了双向 GRU 作为主要的特征处理网络，同样，在双向 GRU 网路中添加了残差边结构，缓解梯度爆炸的问题，保证整个音频分类模型的准确率。作为分类模型，添加了全连接层和 softmax 函数作为最终分类结果的预测。

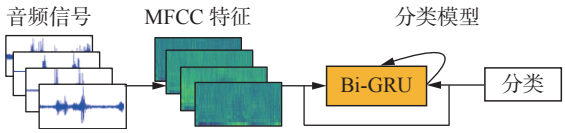


图 3　音频分类模型

Fig. 3　Audio classification model

2.3　机器人操作模型

机器人的操作模型主要是控制机械臂的各个轴的旋转从而产生机器人的各个行为动作，各个行为如图 4 所示，具体命名方式为：拿（Pick）、放（Place）、旋转（Roll）、摇晃 1（Yaw）、摇晃 2（Pitch）、摇晃 3（Shake）。分别包含了机器人对单个物体操作时的拿起与放置行为，以及操控对应机械手末端（x, y, z）轴不同的旋转角获取对应的摇晃动作。因此在设计的操控任务中，规划了机器人的操作动作空间为{Pick, Roll, Yaw, Pitch, Shake , Place}。对于本文的抓取目标统一设定为单一类别，因此设定固定的旋转角获取最佳的抓握姿势。根据不同的任务需求，机器人选择相应的动作完成操作命令。

(a) 目标坐标　(b) Pick　(c) Place　(d) Roll

(e) 机械手坐标　(f) Yaw　(g) Pitch　(h) Shake

图 4　机器臂行为动作

Fig. 4　Robotic arm manipulation behavior

当机器人在接收具体的控制指令后，经过视觉分析获取可以操控的目标点位，执行相应的控制行为。机械臂的具体流程图如图 5 所示，以当前场景的状态为初始状态，通过顺序决策依次执行动作空间中的各个动作，并判断任务是否完成，当执行为最后一个放置动作时，结束当前操作模型的行为。其中，任务操作坐标以及目标任务的坐标由视觉感知模块提供，即通过指示表达模型生成机械臂可以操作的目标位置；对于操作任务完成状态，需要得到正确的容器内的目标物品，音频感知模块可以将搜集的音频信号进行分类，一旦将操作指令中涉及到的目标物品进行正确分类时，则设定下一个状态为放置状态，否则放回原处，重新操作下一个目标物体。

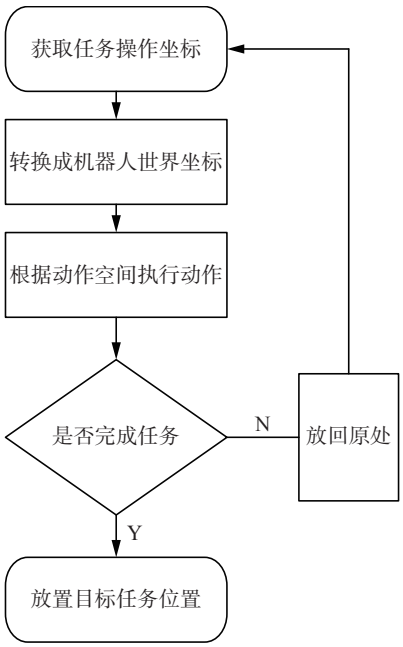


图 5　机器人操作流程
Fig. 5　Robot manipulation

3　实验

3.1　实验平台

本实验采用 UR5 机械臂和五指灵巧手作为整个实验平台的抓取设备, 其中灵巧手采用 5 指设计和连杆传动的方式, 并且具备 6 自由度, 可以保障瓶子的固定抓取。视觉上利用 Kinect 相机捕获 RGB 图像和深度图像, 听觉上利用 RODE 麦克风固定在机械臂的末端下收集接收晃动瓶子的声音信号, 这样可以缩短声源和采集设备的距离, 更方便捕获声音特征。整个实验数据的分析在带有 NVIDIA 2070 的 PC 机上进行处理。整个实验平台如图 6 所示。

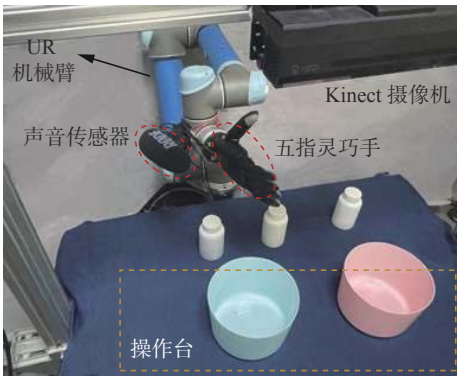


图 6　数据差异性分析
Fig. 6　Analysis of data differences

3.2　数据集采集

3.2.1　操作指令设计

本文设计的操作指令在机器人的行为动作上主要分为两类, 分别是{抓, 放}; 对于物体之间的位置关系, 由{左边, 右边, 中间, 前面, 后面}组成; 对于物体的自身属性, 主要选择了颜色属性, 包含{绿色、蓝色、红色}。操作指令根据物体的类别、属性、位置关系模板交叉组合生成, 符合实际的操作需求, 例句如“拿起中间的瓶子, 放进绿色的碗中”、“拿起带有山楂的瓶子, 放进左边的碗中”。机器人通过这些操作指令完成人类布置的操作任务。

3.2.2　交互听觉数据采集

根据各类中药材的不同特质, 本实验选取了常见的 12 类药用物品, 如图 7 所示, 分别是胶囊、酒精、红枣、药片、生山楂、药丸、决明子、生牡蛎、蜡丸、蝉蜕、颗粒以及空瓶。为了获取数据的多样性, 选取了 1/4、1/2、2/3 瓶子含量的数据。通过结合机械臂的运动特性, 选择{Roll, Yaw, Pitch, Shake} 4 种不同的动作分别对每类物品进行 20 次采样。每个声音信号以 44.1 kHz 进行采样, 根据机械臂的运动时长, 设定采集单个目标种类的音频时长为 6000 ms, 一共采集了 960 组数据作为声音数据集。

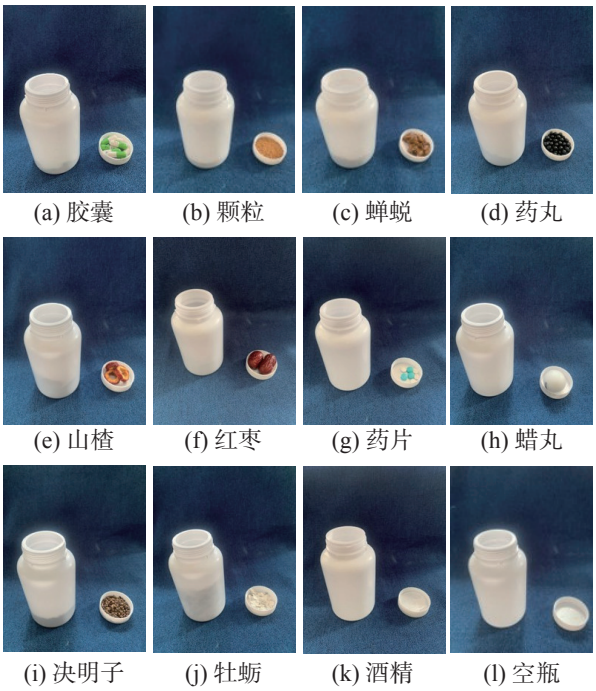


图 7　硬件架构
Fig. 7　Hardware architecture

对于不同类型的声音数据, 选取了具有代表性的物品的声音信号绘制了时域图和频谱图进行了对比。如图 8 所示, 在 4 种机器人的动作下, 山楂和药片的声音具有相当大的差异性。对于相似的声音信号, 在转化成频谱图后, 也在不同的区

域表现出明暗不一的差异，这为在后处理时送入循环神经网络进行分类提供了有效的保障。

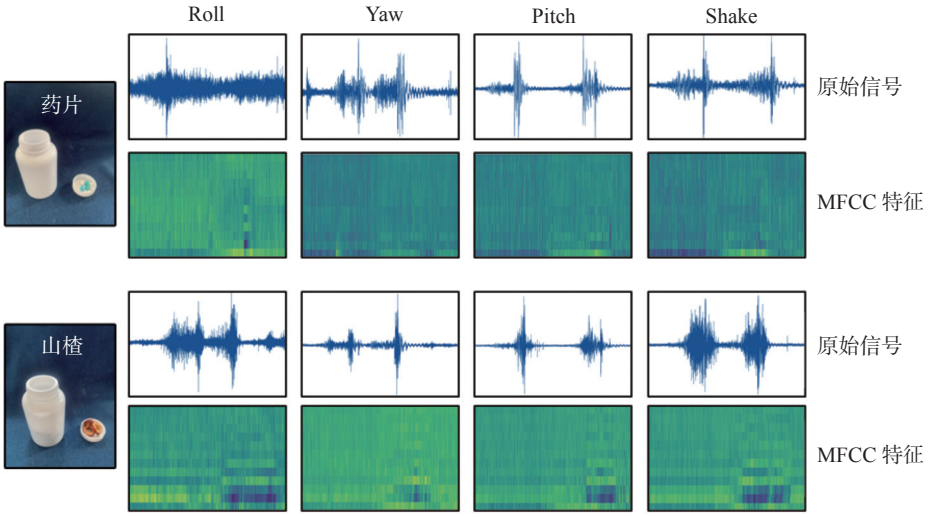
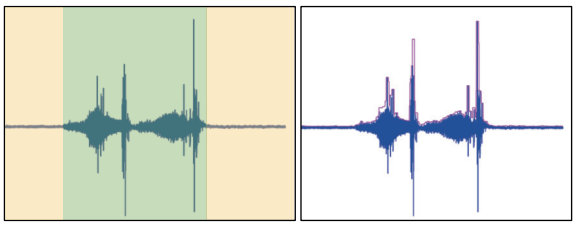


图 8　数据集种类

Fig. 8　Object dataset

3.2.3　对交互听觉数据进行预处理

在实际采集声音数据的过程中，机械臂在执行 4 种动作时有着不同的运动时长以及自身的运动噪声，这些噪声与运动的幅度、速度、加速度有关，会对采集的声音数据和实验结果产生一定的影响，为了减少来自非目标对象的声音的影响。本文采取了噪声抑制过程，如图 9(a)所示，对于音频信号没有超过阈值的区域标记为黄色区域，这部分区域将被剔除，而绿色的区域用于训练，通过利用信号包络线设定阈值，如图 9(b)所示，这样能够有效提供目标分类精度。



(a) 信号选择　　(b) 信号包络

图 9　信号区域选择

Fig. 9　Signal area selection

3.3　实验设置及评估指标

根据设计的操作指令和目标物体关系布置实验场景，场景操作任务结合视觉和听觉识别难易程度总共分为 3 类：第 1 类场景探索物体不同的位置关系，指令如“找到胶囊，放在左边的碗中”；第 2 类场景探索物体的属性关系，指令如“找到所有放胶囊的瓶子，放在绿色的碗中”；第 3 类探索物体的类别关系，指令如“找到放胶囊的瓶子，放在苹果旁边的碗中”。实验目的是根据给定的操作指令，整个机器人系统能够理解指令，并且定位出带有指示关系的物体，利用听觉系统进行分类预测目标，并且放置成功，完成最终的实验。本文定义了一种离线实验机制，设定场景状态{bottle1, bottle2, bottle3, bowl1, bowl2}，机械手的状态{Pick, Roll, Yaw, Pick, Shake, Place}，通过设定执行机械臂的动作序列，如{move bottle1>pick bottle1>check bottle1>place bowl1}，结合视听感知模型依次更新目标状态池{bottle1:Tablet>bowl1:left}完成操作任务。根据实验任务，定义 3 种实验指标：

目标识别率 (target recognition accuracy, TRA)：是否检测到正确操作的目标物体

音频识别率 (audio recognition accuracy, ARA)：是否正确识别了指令中的目标物体

整体任务准确率 (overall task success rate, OTSR)：是否完成了指令中的位置关系的检测。

通过对设计以上 3 种准确率指标来验证本文的视听系统的可行性。

4　实验结果

4.1　视觉检测结果

本文选取了部分操作指令在指定场景中进行检测，对于颜色形状大小相同的 bottle，本文选择用音频感知判别种类，对于 bowl 类别使用视觉指示表达来定位目标物体。操作指令主要包含目标的位置关系、颜色属性关系以及和其他目标的方位关系。准确率保持在 70% 以上，并且绘制了相

应的可视化结果, 如图 10, 在 3 类测试场景下红色框为对应操作指令的正确结果, 蓝色框为本文的视听模型在实际机械人系统下的检验结果。

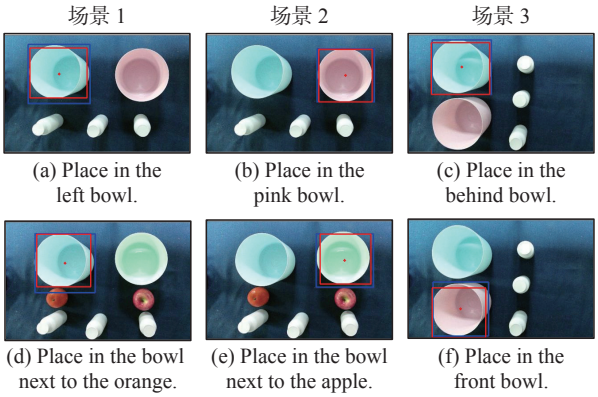


图 10　视觉检测结果

Fig. 10　Results of visual location

4.2　听觉分类结果

为了使整个机器人听觉系统能够有效地工作, 验证每种音频类别的精度是很有必要的。根据本文采集的 12 类音频数据, 在设计的循环神经网络模型上, 对分类结果进行计算, 得到最终的模型分类的混淆矩阵, 如图 11 所示。

从图 11 结果可知, 不同的药用物品准确率有所区别, 对于声音信号较为明显的药片和蜡丸类物品, 在我们设计的分类模型下, 保持了 90% 以上的准确率; 而对于牡蛎和空瓶而言, 由于这两类瓶内物体与瓶臂碰撞的声音较小, 更多的是环境噪声, 因此预测的实验结果准确性相对较低, 但也保持 70% 以上的准确率。因此, 本文设计的音频识别模型能够对不同类别的材料进行识别预测。

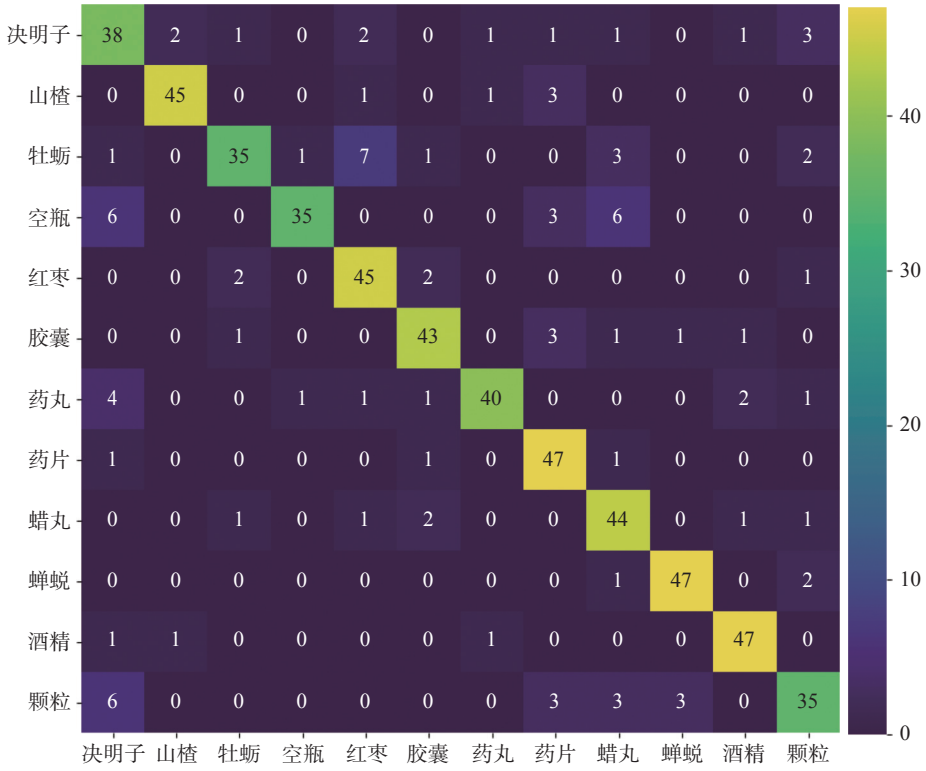


图 11　对所有类别计算混淆矩阵

Fig. 11　Calculate confusion matrix for all categories

4.3　总体任务结果

对于整个系统的操作任务, 本文根据 4.2 节设计的评估指标进行了验证。对于每个实验场景, 根据物体的类别、属性关系、方位关系进行测试, 并且单独测试目标准确率和音频识别准确率, 当两者都识别成功的情况下, 计算总体任务成功率结果如表 1 所示。

根据实验结果进行分析, 由于 3 种类型的测试场景整体比较干净清晰, 因此对于视觉上的指示表达关系来说, 目标准确率的结果相对较高。音频识别准确率根据不同的指令需求, 识别率也相对不一, 因此整个任务的总体成功率出现了不同的情况。在第 1 种操作指令下, 只需定位到单一类别目标, 不需要检索所有候选目标, 因此相对于第 2 种操作指令, 音频识别的准确率较高。而对于第 3 种探索性的指令, 由于物体种类增加, 场景的复杂度提高, 整个视觉的识别率相对下降, 而操作任务简单, 因此整个音频识别率相对提高。

表 1　总体实验结果

Table 1　Overall experiment results　%

场景	TRA	ARA	OTSA
场景1	79.2	72.9	65.3
场景2	75.0	60.4	51.2
场景3	50.0	64.5	46.1

为了进一步验证本文的机器人视听操作系统的实用性，在相同的场景和指令下，设计了无听觉检测的模块如图 12 所示，选取物品的方式依照均匀采样的规则完成操作任务。因为实验的目的是验证整个系统，而抓取任务不是研究的重点，当机械臂的五指灵巧手抓取失败时，选择把目标物体放在手掌以保证实验的顺利进行。在对比实验中，本文选取了相同的场景进行了测试，并且计算了 OTSR 指数，视听框架的准确率可以达到 45.4%，而无音频模块的只有 24.7%。可以发现，在结合多模态数据的情况下，机器人感知目标更加准确，能够有效提升任务成功率。

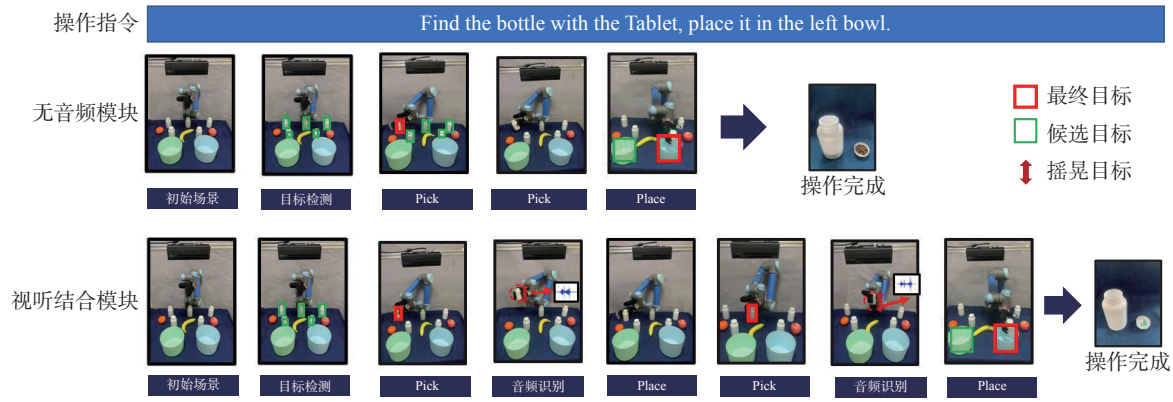


图 12　多模态系统实验对比流程

Fig. 12　Multi-modal system comparison process

5　结束语

为了给机器人提供更复杂的操作能力，本文构建了一个能够接收自然语言指令并具有视觉、听觉的多模态融合的机器人操作系统。其中视觉感知模块能够分析指令中的指示关系，并且定位到目标物体，听觉感知模块能够预测目标物体类别。对于每个感知模块，在构建的多模态数据集中进行了实验验证，结果表明本文的实验系统在接收多模态数据的情况下比单一模态的表现能力更强。然而目前本文设计的操作指令和场景单一，在未来的工作中，将继续增加目标物品的种类，以及设计更复杂且带有歧义的场景和操作指令，构建一个端对端的机器人行为框架。

参考文献：

[1] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2980−2988.

[2] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018−04−08)[2021−01−01].https://arxiv.org/abs/1804.02767.

[3] 周俊佐, 朱宗奎, 何正球,等. 面向人机对话意图分类的混合神经网络模型 [J]. 软件学报, 2019, 30(411): 3313–3325.
ZHOU Junzuo, ZHU Zongkui, HE Zhengqiu, et al. Hybridneural network models for human-machine dialogue intention classification[J]. Journal of software, 2019, 30(411): 3313–3325.

[4] CHOWDHARY K R. Natural language processing[J]. Fundamentals of artificial intelligence, 2020, 17(6): 603–649.

[5] MUREZ Z, VAN AS T, BARTOLOZZI J, et al. Atlas: end-to-end 3D scene reconstruction from posed images[M]// Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 414−431.

[6] HODAŇ T, BARÁTH D, MATAS J. EPOS: estimating 6D pose of objects with symmetries[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11700−11709.

[7] CORONA E, PUMAROLA A, ALENYÀ G, et al. GanHand: predicting human grasp affordances in multi-object scenes[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 5030−5040.

[8] QURESHI A H, SIMEONOV A, BENCY M J, et al. Motion planning networks[C]//2019 International Conference on Robotics and Automation. Montreal: IEEE, 2019: 2118−2124.

[9] QI Yuankai, WU Qi, ANDERSON P, et al. REVERIE: remote embodied visual referring expression in real in-

door environments[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9979−9988.

[10] GAO Chen, CHEN Jinyu, LIU Si, et al. Room-and-object aware knowledge reasoning for remote embodied referring expression[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3063−3072.

[11] ZHANG HANBO, LU YUNFAN, YU CUNJUN, et al. INVIGORATE: interactive visual grounding and grasping in clutter[EB/OL]. (2021−08−25)[2021−10−10].https://arxiv.org/abs/2108.11092.

[12] MI Jinpeng, LYU Jianzhi, TANG Song, et al. Interactive natural language grounding via referring expression comprehension and scene graph parsing[J]. Frontiers in neurorobotics, 2020, 14: 43.

[13] ONDRAS J, CELIKTUTAN O, BREMNER P, et al. Audio-driven robot upper-body motion synthesis[J]. IEEE transactions on cybernetics, 2021, 51(11): 5445–5454.

[14] LATHUILIÈRE S, MASSÉ B, MESEJO P, et al. Neural network based reinforcement learning for audio-visual gaze control in human-robot interaction[J]. Pattern recognition letters, 2019, 118: 61–71.

[15] HÖNEMANN A, BENNETT C, WAGNER P, et al. Audio-visual synthesized attitudes presented by the German speaking robot SMiRAE[C]//The 15th International Conference on Auditory-Visual Speech Processing. Melbourne: ISCA, 2019: 10−11.

[16] YAMAGUCHI A, ATKESON C G. Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision?[J]. Advanced robotics, 2019, 33(14): 661–673.

[17] 朱文霖, 刘华平, 王博文, 等. 开放环境下未知材质的识别技术 [J]. 智能系统学报, 2020, 15(1): 33–40.
ZHU Wenlin, LIU Huaping, WANG Bowen, et al. An intelligent blind guidance system based on visual-touch cross-modal perception[J]. CAAI transactions on intelligent systems, 2020, 15(1): 33–40.

[18] LALONDE J F, VANDAPEL N, HUBER D F, et al. Natural terrain classification using three-dimensional ladar data for ground robot mobility[J]. Journal of field robotics, 2006, 23(10): 839–861.

[19] 张新钰, 邹镇洪, 李志伟, 等. 面向自动驾驶目标检测的深度多模态融合技术 [J]. 智能系统学报, 2020, 15(4): 758–771.
ZHANG Xinyu, ZOU Zhenhong, LI Zhiwei, et al. Deep multi-modal fusion in object detection for autonomous driving[J]. CAAI transactions on intelligent systems, 2020, 15(4): 758–771.

[20] LIU Hongyi, FANG Tongtong, ZHOU Tianyu, et al. Deep learning-based multimodal control interface for human-robot collaboration[J]. Procedia cirp, 2018, 72: 3–8.

[21] YOO Y, LEE C Y, ZHANG B T. Multimodal anomaly detection based on deep auto-encoder for object slip perception of mobile manipulation robots[C]//2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021: 11443−11449.

[22] GAN Chuang, ZHANG Yiwei, WU Jiajun, et al. Look, listen, and act: towards audio-visual embodied navigation[C]//2020 IEEE International Conference on Robotics and Automation. Paris: IEEE, 2020: 9701-9707.

[23] JIN Shaowei, LIU Huaping, WANG Bowen, et al. Open-environment robotic acoustic perception for object recognition[J]. Frontiers in neurorobotics, 2019, 13: 96.

[24] JONETZKO Y, FIEDLER N, EPPE M, et al. Multimodal object analysis with auditory and tactile sensing using recurrent neural networks[M]//Communications in Computer and Information Science. Singapore: Springer Singapore, 2021: 253−265.

[25] 靳少卫, 刘华平, 王博文, 等. 开放环境下未知材质的识别技术 [J]. 智能系统学报, 2020, 15(5): 1020–1027.
JIN Shaowei, LIU Huaping, WANG Bowen, et al. Recognition of unknown materials in an open environment[J]. CAAI transactions on intelligent systems, 2020, 15(5): 1020–1027.

作者简介:

王业飞, 硕士研究生, 主要研究方向为计算机视觉、人机交互。

葛泉波, 教授, 博士生导师, 主要研究方向为工程信息融合方法及应用、人机混合系统智能评估。主持国家自然科学基金青年基金项目 1 项。

刘华平, 副教授, 博士生导师, 中国人工智能学会理事、中国人工智能学会认知系统与信息处理专业委员会秘书长, 吴文俊人工智能科学技术奖获得者, 主要研究方向为机器人感知、学习与控制、多模态信息融合。主持国家自然科学基金重点项目 2 项。发表学术论文 100 余篇。