DOI:10.13196/j.cims.2024.0139

基于大语言模型的人机交互移动检测机器人导航方法

王 湉,范峻铭,郑 湃+
(香港理工大学 工业及系统工程学系,香港特别行政区 999077)

摘 要:在工业制造领域,移动机器人的广泛应用已成为提高作业安全和效率的关键。然而,现有的机器人系统只能完成预定义的导航任务,无法适应非结构化场景。为了突破这一瓶颈,提出一种基于大语言模型(LLM)的人机交互移动检测机器人导航方法,可代替操作人员进入工业环境中的危险区域进行检测,并且可以根据人类自然语言指令完成复杂的导航任务。首先,通过高分辨率网络(HRNet)模型进行场景语义分割,并在点云融合阶段将语义分割结果渲染到重建的三维场景网格模型中,得到三维语义地图;然后利用大语言模型让机器人可以理解人类的自然语言指令,并根据创建的三维语义地图生成 Python 代码控制机器人完成导航任务。最后,通过一系列非结构化场景下的实验验证了该系统的有效性。

关键词:人机交互;大语言模型;视觉语言导航;智能制造;工业 5.0
中图分类号:TP249; TH166 **文献标识码**:A

Large language model-based approach for human-mobile inspection robot interactive navigation

WANG Tian, FAN Junming, ZHENG Pai+
(Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China)

Abstract: In the manufacturing field, the wide application of mobile robots has become the key to improving operational safety and efficiency. However, most existing robotic systems can only complete predefined navigation tasks, and cannot be adapted to the unstructured environment. To overcome this bottleneck, an interactive navigation method for mobile inspection robots based on large language models was introduced, which replaced operators in conducting inspections within hazardous industrial areas, and to execute complex navigation tasks based on verbal instructions. The High-Resolution Net (HRNet) model was utilized for semantic scene segmentation, integrating the segmentation results into the reconstructed 3D scene mesh during the point cloud fusion phase to create a comprehensive 3D semantic map. A large language model was used to make the robot comprehend human natural language instructions and generate Python code based on the 3D semantic map to complete navigation tasks. A series of experiments had been conducted to validate the effectiveness of the proposed system.

Keywords: human-robot interaction; large language model; vision and language navigation; smart manufacturing; Industry 5.0

0 引言

在工业制造领域,特别是在存在潜在危险的环境中,移动机器人的应用已成为提高安全和效率的关键。尽管其技术已经取得了显著进展,但大多数现有系统仍然依赖于预定义的导航任务和路径规划,人机交互环节的缺乏限制了它们在非结构化场景中的应用[1],而目前已有的基于编码的人机交互方式增加了

收稿日期:2024-03-18;**修订日期**:2024-05-04。Received 18 Mar. 2024; accepted 04 May 2024.
基金项目:香港研究资助局资助项目(15210222,15206723)。**Foundation items**: Project supported by the General Research Fund (GRF) from the Research Grants Council of the Hong Kong Special Administrative Region, China (No. 15210222, 15206723).

机器人操作的复杂性和对专业知识的要求，这限制了这些方法在实际工业场景下的应用[2-3]。视觉语言导航是一个有前景的解决方案，它允许人类用自然语言与机器人交流，机器人根据对场景和语言命令的理解完成导航任务[4]。然而，目前大多数关于视觉语言导航的工作都面临模型缺乏泛化性的问题，无法适应人类语言的复杂性和多样性，从而限制了这些视觉语言导航系统在实际智能制造环境中的应用[5]。鉴于这些挑战，本文提出了一种基于大语言模型(Large Language Model，LLM)的人机交互移动检测机器人导航方法，该方法利用大语言模型强大的泛化、理解和推理能力，允许机器人理解和执行基于自然语言的指令，从而为用户提供了一种更加直观、灵活的交互方式。这种能力使得机器人可以更有效地适应非结构化的作业环境，并能够根据实时情况和用户需求进行快速调整。

首先，利用高分辨率网络(High-Resolution Net，HRNet)模型进行高效的场景语义分割，并将分割结果与点云数据融合，以构建详尽的三维语义地图。这样的地图不仅提供了环境的详细视图，还增强了机器人理解和导航的能力。其次，通过利用大语言模型，该导航系统能够解析复杂的自然语言指令，并将其转化为具体的导航任务，由此生成相应的 Python 代码以指导机器人行动。再者，为了验证所提系统的有效性和实用性，本研究对每个组成部分进行了详细的实验分析，旨在全面评估系统性。实验结果证明了该方法的灵活性、可靠性和用户友好性，为工业环境中的移动检测和导航任务提供了一种新的解决方案。最后，讨论了目前研究工作的不足，并指出了未来的发展方向。

1 相关工作

由于工业制造领域时常会伴随危险因素，人类进行设备检查的风险极高，在这种背景下，移动机器人成为了解决方案的关键，它们能够在这些恶劣的环境中取代人类执行检查任务[1]。针对这些机器人，自主导航技术尤其是路径规划和 3D 场景重建，已经获得重点关注并得到了显著发展[2-6]。此外，一些创新的研究通过集成视觉检测和监控，提高了机器人应对非结构化环境的智能化水平[3]。同时，研究人员也在不断探索改进机器人的结构设计和系统控制，以适应充满障碍物和狭窄空间的特殊环境[7-8]。然而，大多数研究未能充分考虑人类在紧急应对和复杂决策方面的独特优势，这些是目前的机器人技术尚未能完全实现的功能。因此，尽管机器人能独立执行基础的检查任务，但缺乏与人类的有效交互仍然是一个明显的短板。

随着工业 5.0 的发展，以人为本的智能制造开始受到重视，进而促进了对人机交互的关注[9]。人类在领导、创造、解决问题和决策方面具有独特优势，而机器人则在速度、耐力、准确度、可扩展性及知识储备方面表现突出。在人与机器人协作的系统中，人类不仅是机器人的指导者，还是创新和决策的中心[10]。同时，机器人则通过增强人类的认知和身体能力，在人机互动中促进双方能力的提升，共同完成更加复杂的任务。为了实现自然的人机交互，多模态融合的交互方式必不可少，多模态通信和控制策略可以保证安全、高效和智能的人机交互体验[4]。目前视觉和语言两种模态是最常见的交互方式，视觉为机器人配备了人、物体和环境识别能力，而人类则使用文本或口头交流与机器人系统进行交互。因此在计算机和机器人导航领域，视觉语言导航是一个新的快速发展的方向[11-12]。

视觉语言导航指机器人可以根据人的自然语言指令完成导航任务，其要求机器人具备对环境和自然语言融合理解的能力[4]。近年来，视觉语言导航受到越来越多的关注，PASHEVICH 等[12]提出了 Episode Transformer，增加了机器人与人类在导航方面通过视觉与文字交互的能力；YAN 等[11]提出一套记忆视觉语音室内导航系统(Memory Vision Voice Indoor Navigation system，MVV-IN)，允许人类可以用语音控制机器人完成室内导航。然而，视觉语言导航目前遇到的主要问题在于模型的泛化能力有限，特别是指令泛化能力和行为泛化能力[5]。在训练过程中，由于数据集采集和标注的限制，对于训练机器人的指令和对应的行为总是有限的，无法囊括真实智能制造任务中所有的场景。如何应对实际应用中复杂和多样的指令和导航动作是目前大多数视觉语言导航研究的主要挑战之一。大语言模型的出现为自然的人机交互带来了新的可能性，其强大的语义理解能力可以帮助机器人更加准确地理解人类的自然语言指令[13]。而场景理解技术则可以使机器人对环境进行感知和理解，两者的结合可以更好地适应非结构化的工业制造场景。因此，本文旨在将视觉语言导航引入工业制造场景，融合大语言模型与场景理解技术来实现基于视觉语言交互的

机器人导航任务。

2　方案设计

为了实现通过人与机器人的视觉语言交互来引导机器人完成导航任务,本文提出基于大语言模型的检测机器人导航策略,框架如图 1 所示。系统的输入是具有深度信息和色彩信息的视频流,还有人类发出的自然语言导航指令。系统主要部分为理解推理环节,主要包含两个模块:①根据输入的视频流进行场景理解从而创建出三维语义地图;②基于大语言模型的导航推理,根据三维语义地图和自然语言指令,大语言模型可以推理出导航目的地坐标,再生成可以控制机器人运动的 Python 代码。最后机器人根据生成的地图和代码执行导航任务。

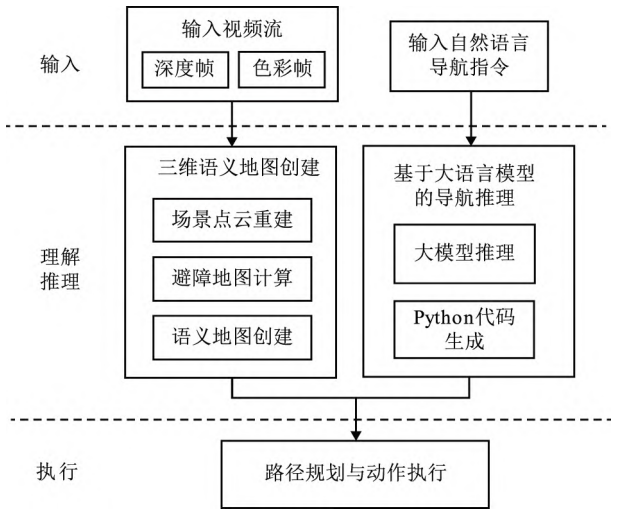


图 1　系统结构

2.1　三维语义地图创建

2.1.1　场景点云重建

重建三维场景的一个关键因素是准确估计摄像机的姿态,这直接影响到重建结果的精度。在制造环境中,周围环境往往杂乱无章、噪声很大,使用惯性测量单元(Inertial Measurement Unit,IMU)和加速度计等传感器对移动机器人的摄像头姿态进行精确估算极具挑战性,无法达到所需的精度。因此,利用基于 RGBD 图像信息的视觉里程计算来估算摄像头姿态。具体来说,本文参考了文献[14-15]的重建流程,整个重建过程分为两个阶段以减少里程漂移的影响。在第一阶段,输入视频帧被分为多段,每段包含 100 对 RGBD 图像。重建过程在每个片段上进行,得到对应数量的点云片段。第二阶段需要在这些重建的片段上应用类似的重建过程。最后,根据优化后的姿态图,应用基于截断地带符号距离函数(Truncated Signed Distance Function,TSDF)点云融合算法将所有视频帧融合为一个整体的三维场景模型。

2.1.2　避障地图计算

基于重建的三维场景网格模型,Navmesh 可以用于表示机器人导航的最基本的避障地图。Navmesh 是一种数据结构,用于计算虚拟环境中的可行走区域并生成导航路径。它将环境划分为一系列可行走的三角形,每个三角形代表一个适合导航的区域。该算法的基本方法是将环境建模为三维网格,然后处理网格以生成导航路径。在计算导航网格时,需要配置几个基本参数,包括网格细化程度、确定每个网格单元的大小和高度、机器人的尺寸、机器人的最大爬升高度和斜坡角度、导航网格生成区域的最小和最大尺寸,以及导航网格边缘允许的最大长度和误差。计算出的 Navmesh 可用于后续路径规划,及将机器人依次导航至三维人机协作(Human-Robot Collaboration,HRC)场景环境中的子目标,从而完成导航任务。

2.1.3　语义地图创建

语义地图创建分为两个部分:①语义分割,利用 HRNet[16] 对输入的视频流中每一帧的 RGB 图片进行二维语义分割;②将分割结果渲染到重建的三维模型中得到三维语义分割结果。基于创建出来的语义地图,可以对拥有相同语义信息的所有点的三维坐标计算平均值,从而得到每一个导航目标中心的三维坐标。

语义地图创建的第一步是利用 HRNet 进行二维语义分割。HRNet 是并行多分辨率结构,它从一个高分辨率子网开始,逐步加入较低分辨率子网,但同时保持对高分辨率特征的链接。这种并行结构可以提取不同分辨率的特征并且进行有效融合,从而实现空间中高精度的语义分割。HRNet 结构示意图如图 2 所示。

通过语义分割可以得到每一张二维的色彩图片每一个像素的语义分割信息,利用 c_{ij} 来表示像素 I_{ij} 处的语义信息,其中 c_{ij} 为 8-bit 的 RGB 数据。分割数据的语义信息渲染的重点是将每一帧二维图片的像素 I_{ij} 和三维模型中的体素 V_{xyz} 对齐,然后将语义信息在 TSDF 点云融合[17] 阶段渲染到三维网格模型表面。从摄像机帧到世界坐标系的变换为:

$$P_{wrd} = R_k P_{cam} + T_k \text{ 。}$$

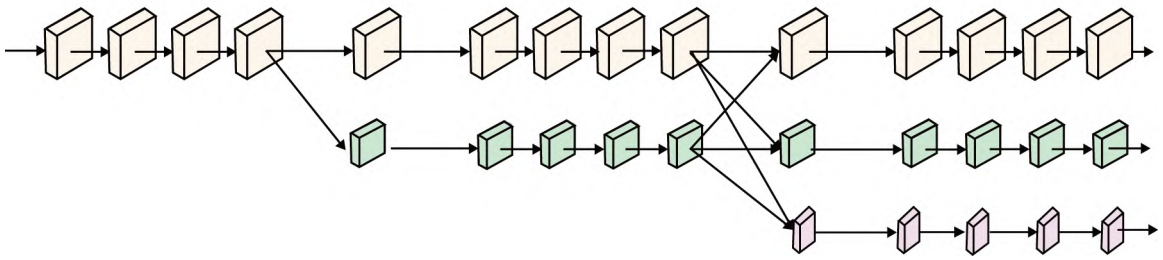


图 2 HRNet 结构示意图

其中：R_k 和 T_k 表示经过三维点云鲁棒性重建计算和优化的精确摄像机姿态，P_{cam} 表示摄像机坐标，P_{wrd} 表示世界坐标。通过这种变换，每个帧的语义信息都可以与三维几何信息整合在一起。然后，应用 TSDF 点云融合方法将所有帧融合到全局点云中。对于全局点云中位于 (x,y,z) 的体素 V_{xyz}，其世界坐标为 $P_{V_{xyz},wrd}$，可根据体素长度求得，根据上述公式可以计算出相应的摄像机坐标 $P_{V_{xyz},cam}$，因此根据相机成像原理可以得到体素 V_{xyz} 相对于摄像机 $Cam_z(V_{xyz})$ 的深度，从而将二维图片的像素 I_{ij} 和三维模型中的体素 V_{xyz} 对齐。对于每一个体素，sdf 值计算公式为：

$$sdf(V_{xyz}) = D(I_{ij}) - Cam_z(V_{xyz})。$$

其中 $D(I_{ij})$ 为像素 I_{ij} 的深度信息。体素的 $tsdf$ 值 $tsdf_{V_{xyz}}$ 计算公式为：

$$tsdf_{V_{xyz}} = \max\left[-1, \min\left(1, \frac{sdf(V_{xyz})}{t}\right)\right]。$$

其中 t 为 $tsdf$ 阶段距离，是一个固定参数。通过这种方式，几何信息 $tsdf_{V_{xyz}}$ 和语义信息 c_{ij} 都存储在一个体素 V_{xyz} 中，通过求平均值可以得到整个模型表面的几何信息和语义信息：

$$tsdf_{V_{xyz}} = \frac{1}{N}\sum_{k=1}^{N} tsdf_{Vxyz}(k)；$$

$$c_{ij} = \frac{1}{N}\cdot\sum_{k=1}^{N} c_{ij}(k)。$$

$tsdf$ 值为 0 的区域代表物体表面，这可以使用 Marching Cubes 算法来识别。根据语义信息值分配颜色，最终生成三维语义地图。在生成的三维场景语义网格中，属于同一物体的所有顶点都具有相同的 RGB 值。通过平均这些具有相同 RGB 值的顶点的坐标，就可以得到物体的中心坐标。这样就生成了三维场景语义，其中包括每个导航对象的语义描述及其对应的中心坐标。

2.2 基于大语言模型的导航推理

在空间目标导航任务中，GPT-3.5 模型扮演着将复杂的自然语言指令转换为可执行的 Python 代码的关键角色，流程图如图 3 所示。

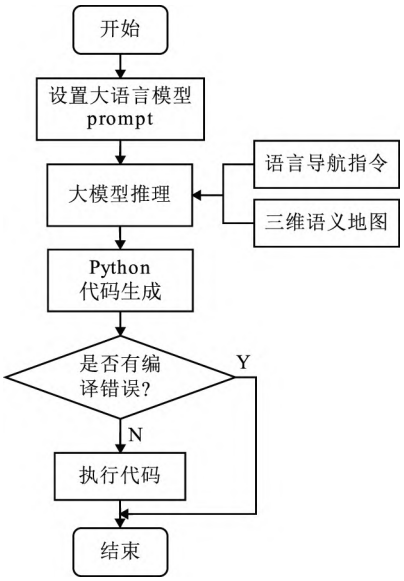


图 3 LLM 流程图

为了实现将复杂的自然语言指令转换为可执行的 Python 代码，首先构建了一组精心设计的初始提示，这些提示旨在概述结合视觉和语言信息进行导航的整体任务框架。这些初始提示不仅定义了任务的范围，还设定了预期的输入输出格式，确保了后续操作的准确性和高效性。随后，通过提供一系列的对话示例来进一步说明自然语言输入与 Python 代码输出之间的具体映射关系。这些示例不仅作为使用指南，帮助用户理解如何与系统交互，还提供了对系统理解能力和执行能力的直观展示。在用户开始输入指令之后，这些指令会被嵌入到预先设计好的提示中，并通过接口调用发送给 GPT 模型。GPT 模型的作用是解析这些指令并转化为具体的 Python 代码。关键的函数如'move_to_obj(object)'

被预先定义并嵌入到模型中,该函数负责处理与空间导航相关的各项任务,如计算对象的坐标索引、规划到目的地的最佳路径,以及生成控制机器人从当前位置移动到指定目标的具体命令。大语言模型在这一过程中扮演着至关重要的角色,它不仅能从自然语言指令中准确识别出各种导航对象,还能根据这些对象特点生成相应的 Python 调用函数代码。一旦生成的代码没有编译错误,它就会被立即执行,从而实现机器人的实时导航操作。这一过程极大地提高了空间目标导航的灵活性和可用性,使得机器人能够更加智能地响应人类指令,实现复杂的导航任务。

3　实验结果与分析

为了验证提出的基于大语言模型的检测机器人导航系统的有效性,本文针对系统的三维语义地图创建和基于大语言模型的导航推理两个部分,分别进行了一系列的实验以全面评估系统性能。

3.1　三维场景重建实验

本实验旨在验证三维重建的集合精度。数据采集是在真实环境中通过英特尔 RealSense L515 摄像头完成的,其中环境被细分为 3 个区域,每个区域都包含了数百帧的 RGBD 视频,这些视频被用于进行场景重建。在场景重建过程中,利用 Nvidia RTX 3080 图形卡加快计算速度。本文采用手机端的激光雷达对场景进行扫描,扫描结果将作为 ground truth,并用两个点云之间的倒角距离来评估基于视觉的三维重建精度,实验结果数据和点云如表 1 和图 4 所示,其中,图 4a 为激光雷达扫描结果,图 4b～图 4d 为基于视觉的重建结果。

表 1　三维重建实验结果

场景编号	倒角距离/m	视频帧数
区域 1	0.085	847
区域 2	0.069	492
区域 3	0.062	646
平均值	0.072	—

图 4a 中的重构点云是手机端激光雷达的扫描结果,图 4b～图 4d 为 3 个子场景的 RGBD 点云重建结果。表 1 列出了 3 个区域重建结果与激光雷达重建结果之间的倒角距离的计算结果,区域 1、区域 2 和区域 3 分别对应图 4b、图 4c 和图 4d。由表 1 可知,各区域的倒角距离均未超过 10 cm,平均距离为 7.2 cm。考虑到整个场景的面积达到 15 m^2,这样的重建误差相对来说是在可接受的范围内。

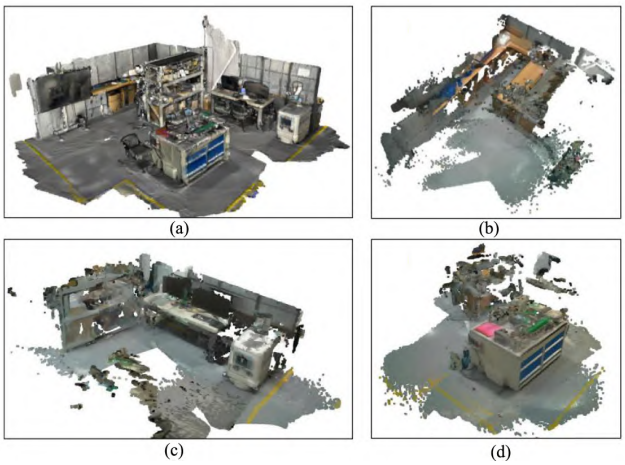


图 4　三维重建点云

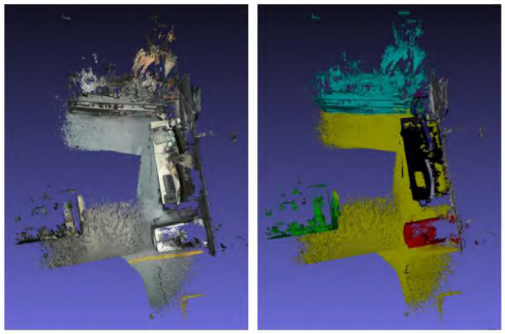
3.2　语义分割和渲染实验

针对语义分割模型,笔者在自己制作的数据集上进行了对比实验以验证其语义分割精度。数据集包含 115 张图片,训练集 75 张,测试集 20 张,验证集 20 张。本文将 HRNet 与另外两个著名语义分割模型 U-Net[18] 和 DeepLabV3+[19] 在此相同数据集上进行了微调并利用 Nvidia RTX 3090 GPU 进行加速,实验结果如表 2 所示。

表 2　语义分割实验结果

语义分割模型	主干网络	mIoU/%	*mPA*/%	Accuracy/%
U-Net[18]	VGG	47.32	61.63	75.76
DeepLabV3+[19]	MobileNet	50.17	62.11	80.68
HRNet[16]	HRNetV2	58.03	68.11	86.41

上述 3 种模型的训练参数为:epoch=340,batch size=4,选用 Adam(adaptive moment estimation)优化器,学习率分别为 1e−4(UNet),5e−4(DeepLabV3+)和 5e−4(HRNet)。从表 1 可以明显看出,在 3 种语义分割模型中,HRNet 在所有评价指标上都表现最优,具体而言,在测试集上的效果,它的平均交并比(mIoU)为 58.03%,平均像素准确率(*mPA*)为 68.11%,总体准确率(Accuracy)达到了 86.41%。这表明 HRNet 在分割任务上具有较高的效率和准确性,尤其是在处理复杂背景和细节方面表现出色。语义分割渲染到三维模型如图 5 所示。



a 场景三维重建结果　　b 语义信息渲染结果

图 5　三维语义地图

图 5a 表示三维重建几何信息，图 5b 为将语义信息渲染到三维模型中的结果。从图中可以看出，针对场景中的分割目标如地板，桌子等，其语义信息可以准确地被渲染。图 5b 三维模型中每一个属于相同分割目标的顶点都被赋予了同一个颜色，因此根据每一个顶点的颜色信息，计算拥有相同颜色信息的所有顶点的平均坐标，即可得到每一个导航目标的三维坐标。

3.3　基于大语言模型的导航实验

在模拟器里进行了基于大语言模型的导航实验，及输入自然语言指令，测试 LLM 是否可以正确解析指令并生成机器人导航代码，从而控制机器人完成导航任务。本文测试了 GPT-3.5 是否能根据不同的自然语言指令准确生成对应的代码，并成功执行它们。具体来说，根据指令中包含的导航子目标的数量对自然语言指令进行了分类，最终形成了 4 个类别，这些类别覆盖了从 1～4 的子目标数量。对于每个类别，使用 50 个不同的自然语言指令进行了测试，并统计了成功率，结果如表 3 所示。

表 3　基于大语言模型的导航实验结果　　%

子目标数量	导航成功率
1	100
2	84
3	78
4	72
平均值	83.5

表 3 中的数据代表了系统在整体导航任务中的成功率。随着自然语言指令中包含的子目标数量的增加，成功率逐渐降低，平均成功率为 83.5%。这表明，当任务变得更加复杂，即指令中涉及到更多的子目标时，系统完成任务的效率会有所下降。

如图 6 所示为真实工业场景下移动机器人根据自然语言指令完成导航任务的例子。

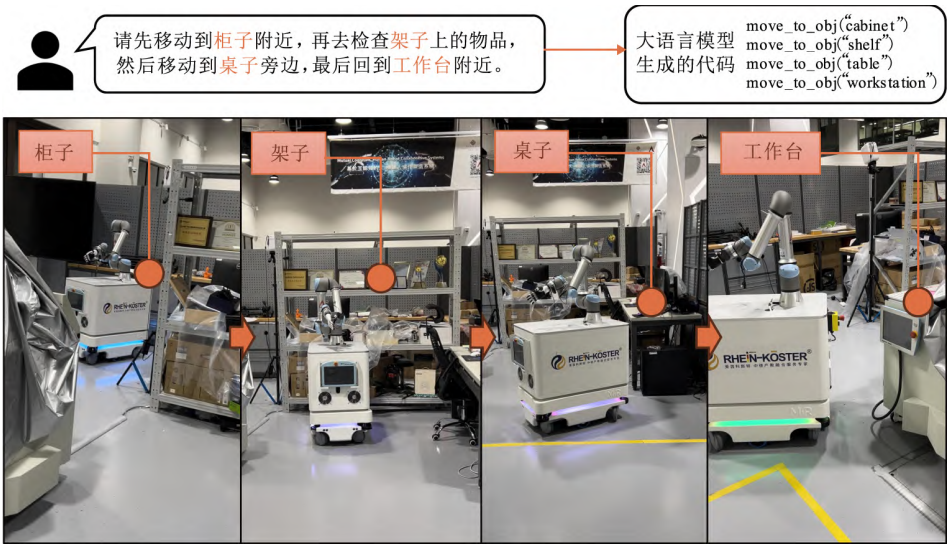


图 6　基于大语言模型的机器人导航示例

人类自然语言指令包含柜子、桌子、架子和工作台 4 个导航目标，而大语言模型可以准确地从自然语言指令中识别到这些导航目标，调用函数生成代码，控制机器人按照准确的顺序依次到达 4 个导航目标，完成导航任务。这一实验结果表明，大语言模型在视觉语言导航中具有很高的准确性和可靠性，能够有效地将人类的自然语言指令转化为机器人可以理解和执行的导航任务。这为将视觉语言导航技术应用于工业制造场景提供了有力的支持，为实现基于视觉语言交互的机器人导航任务打下了坚实的

基础。通过利用大语言模型的语义理解能力和泛化能力,可以实现更加智能化、灵活化的机器人导航任务,从而提高工业制造领域的自动化水平和生产效率。这一研究成果为工业制造领域的智能化发展提供了新的思路和方法,为实现基于视觉语言交互的机器人导航任务奠定了坚实的基础。

4　结束语

本文提出并验证了一种基于大语言模型的人机交互移动检测机器人导航方法,该方法能够理解人类的自然语言并在工业环境中执行复杂导航任务。通过结合先进的语义分割技术和大语言模型,不仅提高了机器人在非结构化环境中的表现,还拓宽了人机交互下移动机器人的自适应能力。未来研究将致力于将视觉语言大模型引入自主导航系统中,并探索人机交互下的移动机器人高效自主学习方法等。

参考文献:

[1] ZHAO Shanshan, MI Zengzhen, CHEN Ren. Moving target detection and 3D reconstruction based on binocular vision[J/OL]. Computer Integrated Manufacturing Systems, 2024: 1-17[2024-02-07]. http://kns.cnki.net/kcms/detail/11.5946.TP.20221103.1148.002.html(in Chinese). [赵珊珊,米曾真,陈　韧. 基于双目视觉的运动目标检测与三维重建[J/OL]. 计算机集成制造系统, 2024: 1-17[2024-02-07]. http://kns.cnki.net/kcms/detail/11.5946.TP.20221103.1148.002.html.]

[2] MATHEW R, HIREMATH S S. Control of velocity-constrained stepper motor-driven Hilare robot for waypoint navigation[J]. Engineering, 2018, 4(4): 491-499.

[3] YUAN C, XIONG B, LI X Q, et al. A novel intelligent inspection robot with deep stereo vision for three-dimensional concrete damage detection and quantification[J]. Structural Health Monitoring, 2022, 21(3): 788-802.

[4] WANG T, ZHENG P, LI S F, et al. Multimodal human-robot interaction for human-centric smart manufacturing: A survey[J]. Advanced Intelligent Systems, 2024, 6(3): 2300359. DOI:10.1002/aisy.202300359.

[5] NIU Kai, WANG Peng. Survey on the research progress and development trend of vision-and-language navigation[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(12): 1818-1827(in Chinese). [牛　凯,王　鹏. 视觉-语言导航的研究进展与发展趋势[J]. 计算机辅助设计与图形学学报, 2022, 34(12): 1818-1827.]

[6] JEONG S, KIM M G, PARK J Y, et al. Long-term monitoring method for tunnel structure transformation using a 3D light detection and ranging equipped in a mobile robot[J]. Structural Health Monitoring, 2023, 22(6): 3742-3760.

[7] CHEAH W, GROVES K, MARTIN H, et al. MIRRAX: A reconfigurable robot for limited access environments[J]. IEEE Transactions on Robotics, 2022, 39(2): 1341-1352.

[8] KAKOGAWA A, MA S. An in-pipe inspection module with an omnidirectional bent-pipe self-adaptation mechanism using a joint torque control[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Washington, D. C., USA: IEEE, 2019: 4347-4352.

[9] LENG J W, SHA W N, WANG B C, et al. Industry 5.0: Prospect and retrospect[J]. Journal of Manufacturing Systems, 2022, 65: 279-295.

[10] MIAO Xiu, HOU Wenjun. Human-computer interaction multi-task modeling based on implicit intent EEG decoding[J/OL]. Computer Integrated Manufacturing Systems, 1-27[2024-02-07]. https://doi.org/10.13196/j.cims.2023.HI12(in Chinese). [苗　秀,侯文军. 基于隐式意图脑电解码的人机交互多任务建模研究[J/OL]. 计算机集成制造系统, 1-27[2024-02-07]. https://doi.org/10.13196/j.cims.2023.HI12.]

[11] YAN L Q, LIU D F, SONG Y X, et al. Multimodal aggregation approach for memory vision-voice indoor navigation with meta-learning[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Washington, D. C., USA: IEEE, 2020: 5847-5854.

[12] PASHEVICH A, SCHMID C, SUN C. Episodic transformer for vision-and-language navigation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington, D. C., USA: IEEE, 2021: 15942-15952.

[13] ZHU Rui, XIAO Honghao, LI Wenxin, et al. Automatic business process generation based on abstract label sequence and large language model[J/OL]. Computer Integrated Manufacturing Systems, 2024: 1-16[2024-02-07]. https://doi.org/10.13196/j.cims.2024.BPM13(in Chinese). [朱　锐,肖鸿浩,李文鑫,等. 基于抽象标签序列与大语言模型的业务过程自动生成[J/OL]. 计算机集成制造系统, 2024: 1-16[2024-02-07]. https://doi.org/10.13196/j.cims.2024.BPM13.]

[14] WANG H, WANG J, LIANG W. Online reconstruction of indoor scenes from RGB-D streams[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, D. C., USA: IEEE, 2016: 3271-3279.

[15] PARK J, ZHOU Q Y, KOLTUN V. Colored point cloud registration revisited[C]//Proceedings of the IEEE International Conference on Computer Vision. Washington, D. C., USA: IEEE, 2017: 143-152.

[16] WANG J D, SUN K, CHENG T H, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3349-3364.

[17] NEWCOMBE R A, IZADI S, HILLIGES O, et al. Kinectfusion: Real-time dense surface mapping and tracking[C]//Proceedings of the 10th IEEE International Symposium on

Mixed and Augmented Reality. Washington, D. C., USA: IEEE, 2011: 127-136.

[18] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]// Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin, Germany: Springer-Verlag, 2015: 234-241.

[19] CHEN L C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Germany: Springer-Verlag, 2018: 801-818.

作者简介:

王 淞(2000—),**女,四川成都人,博士研究生,研究方向:视觉语言人机交互、人机协作**,E-mail:tianna.wang@connect.polyu.hk;

范峻铭(1992—),**男,重庆人,博士后,博士,研究方向:计算机视觉、6D 姿态估计、人机协作**;

+**郑 湃**(1988—),**男,江苏扬州人,副教授,博士,博士生导师,研究方向:人机协作制造系统、智能产品服务系统、工业人工智能等,通讯作者**,E-mail:pai.zheng@polyu.edu.hk。