arXiv:2303.04451v1 [cs.RO] 8 Mar 2023

# 通过多类型手势句向机器人伙伴传达人类意图

Petr Vanc[1] Jan Kristof Behrens[1] Karla Stepanova[1] Vaclav Hlavac[1]

**摘要—家庭和工业工作空间中的人机协作日益普及。然而，机器人与人类之间的通信仍是瓶颈。尽管人们会结合使用不同类型的手势来补充语音，但只有少数机器人系统利用手势进行通信。本文提出了一种手势伪语言，并展示了如何组合多种类型的手势，向机器人表达人类意图（即同时表达期望的动作及其参数——例如，指向某个物体，并示意应将该物体清空到碗中）。系统对所示手势和感知到的桌面场景（由 CosyPose 检测物体位姿）进行实时处理，以提取人类意图。我们利用行为树生成反应式机器人行为，以处理世界可能处于的各种状态（例如，在将物体放入抽屉之前必须先打开抽屉），并从错误中恢复（例如，场景发生变化时）。此外，我们的系统支持在末端执行器的直接遥操作与使用所提出手势句的高层操作之间进行切换。我们使用一台真正的 7 自由度 Franka Emika Panda 机械臂，在复杂程度逐步增加的任务上对该系统进行了评估。与直接遥操作相比，通过动作手势控制机器人最多可将执行时间缩短 60%。**

索引词—**人机协作、意图识别、场景感知、手势检测**

## I. 引言

通过手势控制机器人有许多方式。常用方式包括遥操作 [1] 和指向（指示手势） [2]。直接遥操作使用户能够实时控制机器人，适用于操作人类无法进入或存在危险的场所中的机器人（例如机器人手术，以及在恶劣环境中开展的安全与救援行动）。然而，这需要用户全神贯注，在通信延迟较高的情况下则难以实施。空间机器人学就是一个典型例子：通信延迟使遥操作（地面人员参与控制）过程中必须具备一定的自主能力 [3]。

本文提出一种系统，使用户能够通过手势向机器人传达高层次意图。了解用户的意图后，机器人便能执行许多自主执行所需的动作。机器人和人类都可以请求更改控制模式。这样，系统便能够消除指令中的歧义或获取更具体的要求。

不同的模态能够更高效地传达不同的概念，并相互补充。语言可能擅长建立顺序，而手势

[1]布拉格捷克理工大学，捷克信息学、机器人与控制论研究所，petr.vanc@cvut.cz，jan.kristof.behrens@cvut.cz，karla.stepanova@cvut.cz，vaclav.hlavac@cvut.cz

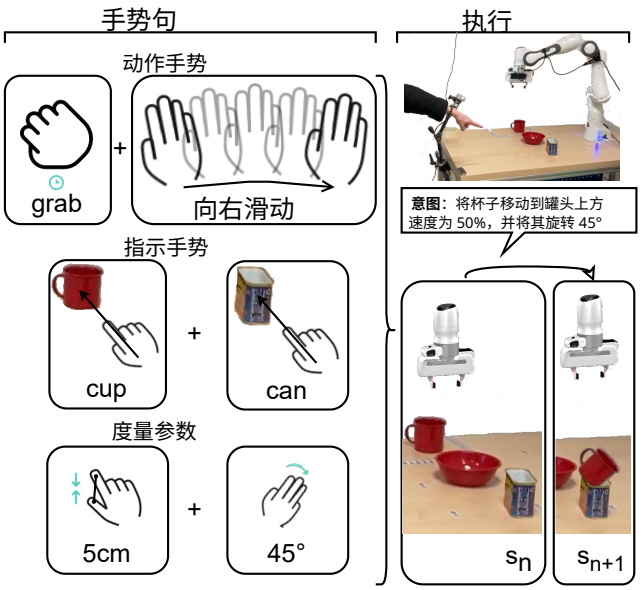



图 1：人类通过手势操控机器人。

在指代物体和位置、表达角度相关参数或轨迹分配方面更加便捷。总体而言，手势的词汇量比语音更有限。因此，一个手势可能不足以表达复杂的用户意图，必须将多个手势组合起来。本文提出了一种手势句，使用户能够通过手势伪语言向机器人表达复杂的用户意图。

在本研究中，我们提出了一种表现力丰富的手势伪语言，旨在提高机器人遥操作的效率与舒适度。我们表明，在桌面操作任务的应用中，与其他系统相比，操作人员花费在发出指令上的时间更少。本文的主要贡献如下：

- 定义了一种由不同类型的手势（即动作、指示手势和度量手势）组成的手势句。
- 扩展手势框架 [4]，使其能够处理各种类型的手势，消除不同手势含义之间的歧义，并根据观察到的手势句确定用户意图。
- 在模拟环境中日益复杂的任务以及在真实环境中的评估。

源代码、配套视频及其他材料可在项目网页 http://imitrob.ciirc.cvut.cz/gestureSentence.html 上获取。

## II. 相关工作

目前利用手势与人类交流的系统，主要侧重于机器人的遥操作（例如， [5]），或者使用手势向机器人发出指令，却无法将人类手势与场景当前状态关联起来（例如， [6], [7], [8]）。Zhang 等人 [5] 与我们类似，使用 Leap Motion传感器检测手部动作并对单个手势进行分类。他们使用一组基本手势遥操作机械臂，但无法将其与给定场景关联起来。文献 [6], 使用可穿戴传感器检测动态和静态手势，并向用户提供视觉反馈。文献 [7] 提出的 Meguru 系统使用户能够组合多个动作，并将手势与更高层次的动作关联起来。然而，这些系统都不支持用户灵活指定要操作的物体（例如，通过指向指定），也无法灵活设定所执行动作的参数（例如，旋转角度）。此外，这些研究都不支持在执行过程中灵活切换机器人直接遥操作与基于手势的操作。

在其他研究中，例如 [9], ，手势被用于指向应当被操作的物体。总体而言，当前研究中的手势通常被用作单词式命令，而不是真正包含动词、主语及其他参数的句子。这意味着它们无法将人类意图准确传达给机器人系统。

有一些研究专注于人类意图推断 [10]。在 [11], 中，研究者使用了能够融合多次观测的递归贝叶斯推理，构建了一个数学模型，并基于人类的操纵杆输入估计人类意图。随后，估计出的意图影响了人类与机器人之间共享自主性的程度。在 [12],中，研究者同样采用了共享自主性方法，机器人主动展示其选择的约定，以加快新手对系统的熟悉过程。他们的辅助方法基于操纵杆输入的观测结果，相较于无辅助情况提高了任务的成功率。在这些研究中，意图是基于操纵杆操作进行估计的。更自然的交流方式，例如人类手势以及当前场景的上下文，并未被纳入考虑。在我们的方法中，我们融合了更多输入模态（例如手势和场景信息）。

在我们之前的工作中 [4] ，我们展示了如何根据场景和手势向量的组合来估计人类意图。然而，这一概念并未在真实环境中进行测试，而且一次只使用一种类型的手势，并配合模拟的感兴趣区域。本文提出一种更通用的手势伪语言，它使用由多种类型（即静态、动态、指示手势、遥操作）手势构成的序列来定义人类意图。我们在真实机器人上使用一组复杂度逐渐增加的任务对该系统进行了评估。

## III. 问题表述

问题定义：给定一个观测$\mathbf{o}$，计算人类意图$\mathbf{i}$ ，并据此生成一系列机器人动作$\mathbf{a}$以实现该意图。

观测元组定义为$\mathbf{o} = [\mathbf{h}, \mathbf{p}, \mathbf{s}]$ ，由手势特征$\mathbf{h}$ 和指示手势特征$\mathbf{p}$的时间序列组成，两者均根据原始手部数据计算得出。$\mathbf{s}$ 表示情境的上下文以及系统的状态，其中包括物体和机器人的构型（机器人末端执行器、物体位置等）。

人类意图$\mathbf{i}$ 表示对世界状态的期望改变（例如，场景中的物体应如何移动、机器人夹爪是否应打开等）。为了能够根据观测 $\mathbf{o}$估计人类意图，我们还必须知道（或学习）映射 $\mathcal{M}$: $\mathbf{i} = \mathcal{M}(\mathbf{o})$。意图通过映射 $\mathcal{A}$: $\mathbf{a} = \mathcal{A}(\mathbf{i})$映射为机器人动作。

单个手势可以传达不同类型的信息（例如，期望执行的动作、动作的参数或目标物体）。我们的系统会将多个手势示范融合为一个手势句子，其中包含主语、动词、物体以及其他定量和定性参数。下面列出主要的手势类型。有关各个手势的更多信息，请参见第 IV 节：

1）动作手势（静态和动态）：单个手势会映射到各种预期动作，例如抓取、打开或推动。基于预训练手势集 $\mathbf{G}$、手部观测 $\mathbf{h}$ 和分类器 $\mathcal{G}$，我们的系统会输出各个手势$\mathbf{g}$的概率，

$$\mathbf{g} = \mathcal{G}(\mathbf{h}, \mathbf{G}).$$

2）指示手势。此手势使用户能够选择场景中的感兴趣区域（物体）。方向向量 $\mathbf{p}_{\text{line}}$ （指向手指的方向或手掌方向）是根据跟踪到的手部构型提取的。结合场景中已知的物体位姿 $\mathbf{o}^{\text{poses}}$，可以基于 $\mathbf{to}_{\text{dists}}$估算各个物体成为目标物体的概率。向量$\mathbf{to}_{\text{dists}}$ 表示每个物体到 $\mathcal{O}$ 至方向向量的距离：

$$\mathbf{to}_{\text{dists}} = \mathcal{D}(\mathbf{p}_{\text{line}}, \mathbf{o}_{\text{poses}})$$

单个距离通过计算物体 $s_i \in \mathcal{O}$ 到方向直线的最短距离得出：

$$d_i^2 = \frac{|(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_1 - \mathbf{s})|^2}{|\mathbf{p}_2 - \mathbf{p}_1|^2},$$

其中，$\mathbf{p}_1$ 和 $\mathbf{p}_2$ 是三维直线上的两个随机点，$\mathbf{s}_i$ 是物体 $i$的位置，所有向量均为三维笛卡尔向量。

3）度量信息。这些手势是可选的，可增强输入手势信息（否则将使用默认值）。它们是连续变量，例如，食指与拇指之间的距离会映射为相应的参数值，如推行动作的距离或抓取物体时的角度（参见图 2 中的示例）。

4）机器人的直接遥操作。某些动作手势可能会启动遥操作模式。激活后，跟踪到的手掌中心位置及其绕



图 2：度量手势示例，它们是辅助且可选的。捏合距离（左）、指向方向（中）、双手间距（右）

$z$ 轴分别直接映射到机器人的位置和第七自由度旋转。

最后，手势句 $S$ 可定义为一个元组，其中包含观测到的动作手势、由指示手势指定的物体，以及由度量手势指定的度量参数：

$$\mathbf{S} = (\textbf{action}, object(s), metric\ parameters). \quad (1)$$

虽然**动作** 始终是必需的，但仅特定动作需要指定物体，而度量参数则是辅助项（若未定义，则使用默认值）。

各个句子对应于用户意图**i** （如 [4]中所介绍），即目标动作*ta*、目标物体*to*、目标位置以及其他辅助度量参数**ap**：

$$\mathbf{S} \sim \mathbf{i} = (\underbrace{\underset{\mathcal{O}}{\operatorname{argmax}}(\mathbf{i}_{ta})}_{ta}, \underbrace{\underset{\mathcal{O}}{\operatorname{argmin}}(\mathbf{to}_{\text{dists}})}_{to}, \mathbf{ap}), \quad (2)$$

其中$\mathbf{i}_{ta}$ = $\mathcal{M}_a$**(g)**是由手势向量**g**生成的目标动作意图 向量概率，$\mathbf{to}_{\text{dists}}$ =$\mathcal{D}_a$**(p**$_{\text{line}}$,**o**$_{\text{poses}}$**)**是目标物体 向量概率。 $ta$表示目标动作意图， $to$ 目标物体，而**ap** 表示辅助参数集合。

手势句示例如下：

$$S = (\textbf{thumbsup}, [mug,\ bowl],\ [5cm]). \quad (3)$$

这对应于用户意图（”以*50%*的速度将马克杯移到碗边”）：

$$i = (\textbf{move}, mug,\ [bowl,\ 50\%]) \quad (4)$$

根据估计的人类意图**i**生成机器人动作序列 **a** 。在本研究中，我们通过行为树[13]来解决这一问题（参见图3中的示例）。整体结构见图5。

### A. 手势句复杂度

手势句的复杂度各不相同，取决于指定给定机器人动作所需的变量数量。该手势句只需要一个参数（$ta$）。我们将与具有 $N$ 个参数**p** （$a$**(p)**）的机器人动作$a$相对应的手势句 $S$ =的复杂度定义为$ta$，**至**，**ap**）：

$$S_c = N = len(\mathbf{to}) + len(\mathbf{ap})$$

。简单的机器人动作只需指定动作类型，无需其他参数，也不依赖物体（例如，根据网格世界步长向右移动）。在这种情况下，手势句的复杂度为 $S_c = 0$。对于动作例如 拿起杯子，我们必须指定动作和目标物体（$S_c = 1$）。复杂的机器人动作可能依赖两个或更多场景物体（例如，交换两个物体的位置）。

我们的系统包含以下机器人动作，其参数数量依次递增：

- 0-参数：笛卡尔移动、旋转、放置
- 1-参数：拿起、放置、倾倒、打开、关闭
- 2-参数：放置、倾倒、交换

请注意，例如，动作放置 可以有 1 个参数（如果我们正拿着目标物体，则只需指定动作和目标位置），也可以有 2 个参数（如果夹爪是空的，则必须指定动作、目标物体和目标位置）。

## IV. 材料与方法

第 3 节所述的问题可以分为几个部分：用户手部检测与骨骼重建、手势检测、用户意图识别，以及机器人动作生成及其执行。在我们的系统中，我们不考虑手部检测和骨骼重建。整体系统如图 5 所示。

我们的系统包括：1）手势识别 $\mathcal{G}$、2）意图分类 $\mathcal{M}$，以及 3）机器人动作生成 $\mathcal{A}$。系统的各个部分将在后续章节中作更详细的介绍，最后一个小节将把所有方法联系起来（第 IV-F 节）。

### A. 实时手势识别

我们将手势分为静态手势和动态手势。静态手势仅包含一个时间帧，但通常具有更多特征；动态手势包含更多时间帧，特征则较少。特征从手部骨骼结构中提取。我们使用 Leap Motion 传感器 [14]获取手部特征数据，并使用 Gesture Toolbox [15]管理给定会话中的手势。手势集合可根据当前需求进行调整。

静态手势 特征集由人工设计，用于处理手部骨骼结构，包含 57 个特征：1）指尖位置之间的距离，以及掌心中心位置；2）手部骨骼之间的关节角度 $\alpha_1, \ldots, \alpha_n$，这些角度由方向向量之间的夹角构成。方向向量由每根骨骼的起始位置和结束位置生成。实验证明，最适合对静态手势进行分类的方法是概率神经网络模型 [16]。我们还测试了其他方法，包括卷积神经网络和确定性方法（手动设定阈值），但其表现较差（见第 VI-B 节）。

动态手势 被评估为随时间变化的手部轨迹。所用的样本配置仅由每个时间步的掌心中心位置（笛卡尔坐标）组成：

$$\mathbf{h} = \left[ \begin{vmatrix} x_1 \\ y_1 \\ z_1 \end{vmatrix}, \begin{vmatrix} x_2 \\ y_2 \\ z_2 \end{vmatrix}, \cdots \begin{vmatrix} x_N \\ y_N \\ z_N \end{vmatrix} \right], \quad (5)$$

其中， $N$ 是以约 90Hz 的频率记录得到的轨迹点数量，我们将

接收到的帧的频率降至 $f = 20$Hz，而不会造成显著损失 [17]。我们将检测频率设为 10Hz，每次检测都会使用最近一个可调节长度的时间窗口（此处设为 1s）。

动态手势从一组示范中学习，并表示为概率运动基元 [18]。作为分类器，我们采用动态时间规整方法（DTW）[19], ，它可以比较两个运动，而不受速度和变形的影响。比较参数为距离 $\delta$。每个手势都有其保存的代表性运动（$\mathbf{h}_1, \cdots, \mathbf{h}_D$），其中 $D$ 是当前集合中的动态手势数量，而**h** 的定义见第5节。随后，通过 $\arg\min((\delta)_1^D) = \text{argmin}(DTW((\mathbf{h})_1^D, \mathbf{h}_s))$ 得到概率最大的手势，其中$\mathbf{h}_s$ 是我们要分类的样本运动。

### *B.* 指示手势

指示手势是一种特殊类型的静态手势，采用特殊的方法来检测预期的目标点或物体（参见第1节，了解如何找到最接近方向直线的物体）。如果我们想选择桌面上的某个位置，而不是目标物体（例如，将物体放置到某个位置时），指示手势方法会求出指向三维直线与工作空间桌面的交点。

每一帧都可以计算该解（最近的物体），但只有当用户指向某个物体或位置时，才会使用其信息。该手势与动作手势结合，使我们能够检测到所需的人类意图。我们在第 VI-A 节的另一项实验中评估了该手势的准确度和可用性。

### *C.* 机器人直接遥操作

启用直接遥操作后，系统会将用户的手上的选定点（例如手掌中心）实时映射到机器人末端执行器。为实现平滑控制，我们以至少 10 Hz 的频率向机器人更新数值。我们采用速度控制外环下的位置控制，以避免机器人运动时改变其运动轨迹而引起抖动。

映射特征向量由手掌笛卡尔位置（$p = [x, y, z]$）和绕 $z$ 轴旋转的手部旋转 $\theta$ 组成，以适配机器人的第 7 个自由度。当我们希望以不合适的构型抓取物体时，这一点十分重要。我们测试过映射人类手部的完整旋转，但由于用户需要专注于过多角度，遥操作体验和准确度都较差。

直接遥操作可通过对未知低层运动基元的示范，教会机器人一项新技能。
之后可以将其映射为更高层的命令（手势）。此外，当物体检测结果不够准确时，它还能帮助调整机器人的位置。

### *D.* 意图动作估计

预期的目标动作$\mathbf{i}_{ta}$ 由手势**g** 和场景**s** 等上下文信息估计，与 [4]中相同。此映射任务（$\mathcal{M}$）被视为分类问题；由于输入的场景特征较少，因此采用浅层概率神经网络即可满足需求。该网络包含二个全连接隐藏层。输出为分类分布 [20]，通常用于生成离散输出。自动微分变分推断（ADVI）[21]使用 Kullback-Leibler散度损失函数拟合网络权重。我们使用 PyMC [16] 框架。

### *E.* 动作序列生成与执行

意图**i** 元组（目标动作、目标物体以及（可选的）辅助参数）被映射为一组机器人动作**a**。我们使用行为树来编码这一高层策略，使我们能够处理更复杂的任务，例如抓取一个无法直接看见的物体。行为树的高层策略会针对每个已执行的动作重新计算所有任务，因此能够处理前一个动作执行失败的情况。行为树通过生成前置条件动作来实现所需意图，无需手动完成这些动作（有关我们实验中使用的行为树示例，请参见图 3）。

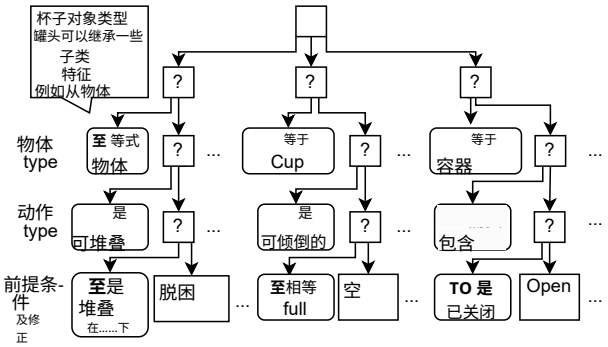



图 3：行为树结构示例可通过满足物体状态前置条件来解决。例如，要拿起一个与其他物体堆叠在一起的物体，必须先移开其上方的所有物体。

### *F.* 人机交互

人类与机器人之间的交互使用简单的手势句，即连续的手势组成一个手势句（见公式 1），并定义用户意图（见公式 2）。用户通过手势表达意图，同时从图形用户界面和机器人本身获取视觉反馈（例如，检测到的手势、选中的物体等）（参见随附视频）。人机交互示意图如图 5 所示。用户在一次简短的交互中表达每一项信息（即动作意图、物体意图、辅助参数），我们将这类交互称为片段。

我们将片段 定义为人类的手出现在手部检测区域内的时间窗口；当手不再可见或时间超过上限（设为$t = 3s$）时，该片段结束。在一个片段中，我们会获取已触发手势的列表。静态手势和动态手势被设置为两个独立线程，我们从每个线程获取识别数据。要将手势标记为已检测到，其持续时间必须超过给定阈值，也就是说，手势超过

阈值上限的时长必须大于 0.3$s$。图 4 展示了一个作为可选反馈提供给用户的手势检测示例。

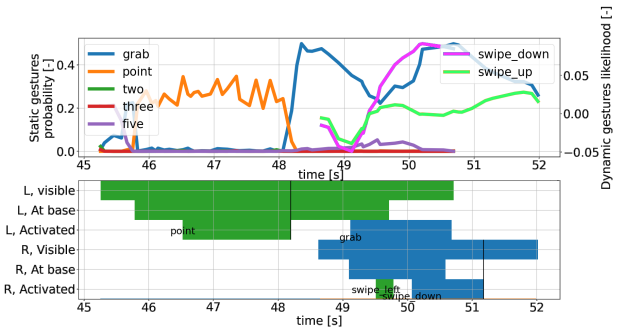



图 4：单个片段的手势检测输出。上方图表显示静态手势（左侧图例）和动态手势（右侧图例）的可能性。下方图表显示手部的可见性以及手势的激活情况。首先检测到静态指向手势，随后检测到抓取手势和动态向下滑动手势。

交互包含以下步骤（见图 5）：1）用户通过做出动作手势来表示预期动作，该手势由手势分类器 $\mathcal{G}$ 识别。只要手部可见，识别器就会一直运行。2）估计动作意图。虽然可能检测到多个动作，但在这些实验中，使用最后检测到的手势来确定目标动作。3）根据预期动作，系统会要求用户提供有关所执行动作的补充信息，并添加目标物体和辅助参数（例如目标位置、距离等）。4）句子信息填写完整后（或用户不愿提供更多信息时），意图元组即填写完毕。5）随后可以生成动作行为树，并执行一系列动作，引导机器人完成预期任务。

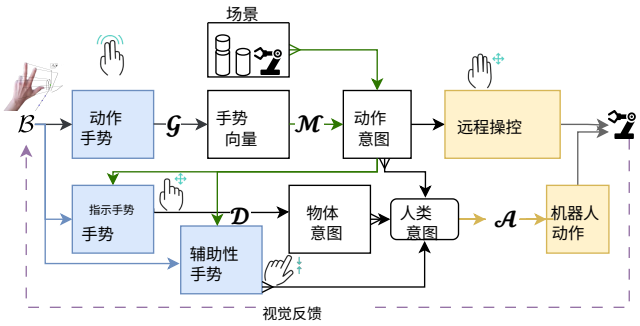
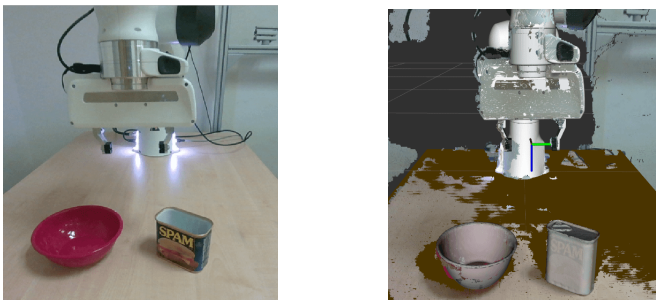



图 5：系统示意图。 $\mathcal{G}$ 表示手势分类器， $\mathcal{M}$是从手势到人类意图的映射， $\mathcal{A}$ 是使用行为树生成机器人动作，$\mathcal{D}$ 是指示手势执行。蓝色方块表示用户输入，黄色方块表示机器人输出模式。



（a）真实场景　　（b）置信度可视化。

图 6：Cosy Pose [24] 使用 RViz 可视化的物体检测视图，其中显示了检测到的两个物体。在该可视化中，检测到的物体网格（灰色）叠加显示在彩色点云上。

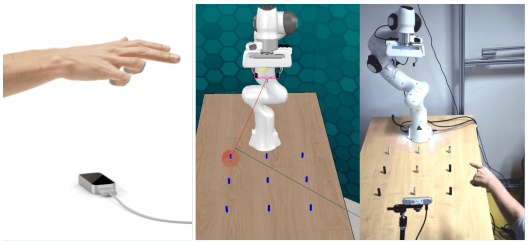


图 7：实验设置。Leap Motion 控制器（左），Coppelia Sim 中的仿真场景（中），真实场景（右）。

## V. 实验设置

### A. 机器人环境

我们的系统在模拟和真实环境中进行了测试（见图 6 和图 7）。对于模拟器，我们使用 Coppelia Sim [22] 以及 PyRep 扩展工具 [23]。场景中包含一台具有 7 个自由度的 Franka Emika Panda 机器人和若干操作物体。一台位于机器人对面的 Intel Realsense D455 RGBD 相机用于物体检测。LeapMotion 传感器安装在桌面上、相机旁边。

我们的真实环境设置（见图 1）采用了物体位姿检测器 CosyPose [24]（见图 6），使我们能够实时获取物体位姿。我们使用包含多种常见家居物品的 YCB 数据集 [25]，从而能够创建各种场景构型。

### B. 手势分类

我们的手部追踪设备包括 LeapMotion Controller[14]，，它能够实时提供手部骨骼结构。我们利用手势分类框架 [15]管理手势集、处理手势识别和各手势概率的可视化，并为用户提供图形界面和反馈。

实验中使用的静态和动态手势包括：抓取、捏合、指向、二、三、四、五、竖起大拇指、向上滑动、向下滑动、向左滑动和向右滑动手势。无手势 用于表示动态手势中的无移动状态，因此不在列表中。

### C. 实验场景描述

我们进行了一系列典型的操作实验，手势句复杂度逐步增加（更多详情见第 III-A 节，$x$ 表示夹爪中的物体）：

- $\mathbf{S_c = 0}$, $\mathbf{S}$ = (**ta** –, –) : $(rotate,\ x,\ –)$, $(place,\ x,\ –)$, $(move,\ –,\ –)$
- $\mathbf{S_c = 1}$, $\mathbf{S}$ = (**ta**, **to** –) : $(pick,\ can)$, $(open,\ drawer)$, $(close,\ drawer)$
- $\mathbf{S_c = 1}$, $\mathbf{S}$ = (**ta** –, **loc**) : $(pour,\ x,\ bowl)$, $(put,\ x,\ bowl)$
- $\mathbf{S_c = 2}$, $\mathbf{S}$ = (**ta**, **to**, **loc**) : $(pour,\ spam,\ bowl)$, $(swap,\ can,\ bowl)$, $(put,\ spam,\ drawer)$
- $\mathbf{S_c = 3}$, $\mathbf{S}$ = (**ta**, **to**, **loc**, **par**) : $(rotate,\ spam,\ ang = 180°)$, $(pour,\ can,\ bowl,\ ang = 60°)$
- **多步任务**：堆叠 3 个物体、整理 3 个物体、将 2 个物体倾倒入碗中
- **不可行任务**：将物体移动到被占用的碗中、将物体放入已关闭的抽屉

我们关注任务完成的总时间、每个场景的成功率以及用户交互时间。所有这些任务的执行过程都可以在随附的视频中看到。在用户实验中，我们选取了以下任务的一个子集：1）将物体放入碗中；2）交换物体；3）将物体放入被其他物体占用的碗中，以展示我们的系统能够先自动移出其中的物体。我们评估了以下三种辅助模式：基于三种辅助模式对任务进行测试：1）直接遥操作；2）通过动作手势进行低级控制；3）通过动作手势进行高级控制。

## VI. 实验结果

我们展示以下实验的结果：
1）指示手势——1）完成时间；2）准确度；2）静态手势和动态手势的准确度；3）场景完成情况

所有场景和任务均经过设置，使机器人能够利用场景中的物体完成任务。

### A. 指示性手势评估

我们在一张工作台上创建了一个由 9 个小物体组成的粗略网格（见图 7）。两个物体之间的网格距离设置为 $d = 0.2m$。检测预期（指向的）目标物体的总体准确度取决于 Leap Motion 的手部检测准确度以及各项变换的准确度（例如，从传感器到基座，或从手部到传感器的变换）。

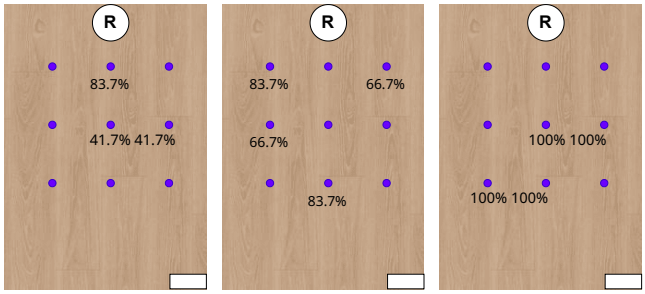
首先，我们跟踪了手指指向时的指骨位置，但观察到不同测量结果之间存在较大差异，因为该方法要求指向手指与手掌方向平行，且需要经验丰富的用户。因此，我们采用了手掌的数值。

两名用户在网格中检测目标物体的准确度列于表 I 中。模拟器的反馈（见图 7）向用户显示当前指向的是哪个物体。

比观察显示屏上的数值更好的方式，或许是操作机器人在当前作为目标的物体上方移动。这提高了整体系统人体工程学和沉浸感，系统。图8展示了各个位置的准确率等更详细的结果。

表I：在5名用户的9物体网格上评估指示手势的总体准确度。首先，用户接受了盲测，期间没有反馈。其次，他们观察了屏幕上的反馈。最后，经过训练后，再次在没有反馈的情况下评估准确度。

| 准确度 [%] | 组员 | 其他人 |
|---|---|---|
| 未经训练的，无反馈 | 59.3 | 40.7 |
| 已训练，无反馈 | 68.5 | 66.7 |
| 有反馈 | 100 | 100 |




(a) 未训练　(b) 已训练　(c) 有反馈

图 8：指示性手势评估结果。R 表示机器人底座，Leap 表示手部传感器设备，蓝色圆圈表示目标物体。百分比表示两位受测用户针对每个物体的准确度。

### B. 手势分类评估

手势集需要正确选择，任意两个手势之间不能相互重叠，这样分类器才能区分每个手势并正确分类。对于我们的示例手势集，我们使用了 8 种静态手势（抓取、捏合、指向、二、三、四、五、竖起大拇指），结果见表 II。

动态手势集由 5 种手势组成，其中包括无手势。这些手势表示不同方向的滑动，并针对右手使用进行了调整；训练数据集包含 ~300 个示范。基准测试（见表 II）基于训练数据集进行，因为此处没有发生学习。

表 II：手势分类基准测试。准确度表示平衡准确率 $BA = \frac{1}{2} \cdot (\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$，其中 T/F 分别表示真/假，P/N 分别表示正/负（更多信息见 [链接]），并在测试数据集上进行评估。静态类别约有 12000 个样本，其中 4000 个为测试样本。

| Type | 方法 | 准确度 |
|---|---|---|
| 静态 | 确定性 | 84.3% |
| | 概率神经网络 | 99.1% |
| 动态 | 欧拉 | 72.0% |
| | DTW | 87.3% |

### C. 场景评估

向用户介绍了可用于指挥机器人的手势。熟悉执行过程后，用户便能顺利完成低层任务（例如抓取、倾倒）。大多数错误都源于物体误检：例如，CosyPose 未能识别场景中的物体、在场景中随机添加幽灵物体，或发生崩溃。第 V-C 节所列场景的执行示例可在随附视频和项目网页 http://imitrob.ciirc.cvut.cz/gestureSentence.html 上查看。

对三个选定场景（见第 V-C 节）进行了更深入的评估。记录了完成时间、成功率和人工干预比例等参数。图 9 展示了使用遥操作、低级动作和高级动作时各场景的完成时间。可以看出，直接遥操作机器人所需的执行时间最长，而且用户还必须接受适当的培训。在单独的试验过程中，完成任务所需的时间逐渐缩短。

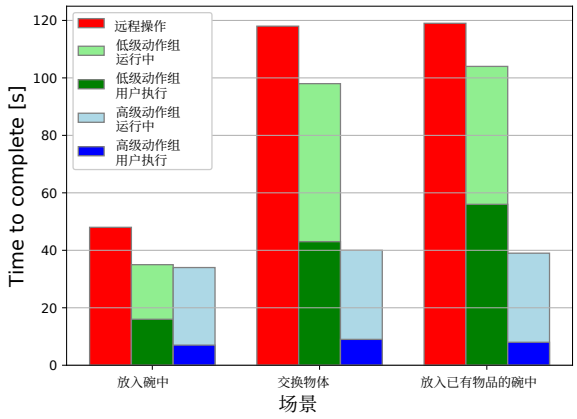



图 9：各场景的任务完成时间。仅统计成功的尝试。与总运行时间相比，该系统将用户所需时间减少了约 50%。不可行的放置任务最难通过指令控制机器人。因此，该任务的失败率最高；可以看出，控制机器人所需的时间也更长。

## VII. 结论与讨论

我们提出了一种系统，它能够处理多种手势表达，处理不同类型的手势（动作手势、指示手势等），并结合这些信息来表达和执行用户意图。

对用户进行的指示手势测试让我们了解到，视觉反馈是此类系统的关键组成部分。

从场景测试中可以看出，预先编程的通用高层任务在手势执行正确时，完成任务所需的时间最短。遥操作是最慢的方法，但可以处理更多细粒度任务。未来，它或许可以用来教会我们机器人新的技能。记住该技能后，低层动作手势便可映射到这项技能。这样就无需手动编程动作意图。与直接遥操作方法相比，通过动作手势控制机器人可将任务执行时间缩短多达 60%。

## VIII. 致谢

本研究获得欧洲区域发展基金资助，项目为“面向工业4.0的机器人技术”（注册号：$CZ.02.1.01/0.0/0.0/15\ 003/0000470$）；并获得 MPO TRIO 项目资助，项目编号 FV40319，以及捷克科学基金会资助（项目编号 GA21-31000S）。P.V. 获捷克理工大学学生资助机构资助（注册号：SGS23/138/OHK3-027/23）。

## 参考文献

[1] A. D. Dragan 和 S. S. Srinivasa，《辅助遥操作形式化》。MIT出版社，2012年7月，第376卷。[2] A. Kobsa、J. Allgayer、C. Reddig、N. Reithinger、D. Schmauks、K. Harbusch 和 W. Wahlster，“结合指示手势与自然语言进行指称对象识别”，载于《第11届计算语言学会议论文集》，COLING ’86系列。美国：计算语言学协会，1986年，第356–361页。[ ] 10 3115 991365 991471 在线。可访问：[3] https://doi.org/ 。 / 。 I. A. Nesnas、L. M. Fesq 和 R. A. Volpe，“空间机器人的自主性：过去、现在与未来”，《当代机器人报告》，第2卷，第3期，第251–263页，2021年9月。[4] P. Vanc、J. K. Behrens 和 K. Stepanova，“利用手势片段实现上下文感知机器人控制”，载于《2018年IEEE/RSJ国际机器人与自动化会议》，2023年。[5] W. Zhang、H. Cheng、L. Zhao、L. Hao、M. Tao 和 C. Xiang，“一种用于柔顺机器人运动的基于手势的遥操作系统”，《应用科学》，第9卷，第24期，第5290页，2019年。[6] P. Neto、M. Sim¯ao、N. Mendes 和 M. Safeea，“制造业中用于人类辅助的基于手势的人机交互”，《国际先进制造技术杂志》，第101卷，第119–135页，2019年。[7] C. Nuzzi、S. Pasinetti、R. Pagani、S. Ghidini、M. Beschi、G. Coffetti 和 G. Sansoni，“Meguru：面向元协作工作站的基于手势的机器人程序构建器”，《机器人与计算机集成制造》，第68卷，第102085页，2021年。[8] O. Mazhar、B. Navarro、S. Ramdani、R. Passama 和 A. Cherubini，“一种具备鲁棒背景不变手势检测功能的实时人机交互框架”，《机器人与计算机集成制造》，第60卷，第34–48页，2019年。[9] J. K. Behrens、K. Stepanova、R. Lange 和 R. Skoviera，“通过自然语言和示范指定双臂机器人规划问题”，《IEEE机器人与自动化快报》，第4卷，第3期，第2622–2629页，2019年。[10] D. P. Losey、C. G. McDonald、E. Battaglia 和 M. K. O’Malley，“实体人机交互共享控制中的意图检测、仲裁与通信方面综述”，《应用力学评论》，第70卷，第1期，2018年。[11] S. Jain 和 B. Argall，“共享控制机器人技术中的递归贝叶斯人类意图识别”，载于《2018年IEEE/RSJ智能机器人与系统国际会议（IROS）》，2018年，第3905–3912页。[12] A. Jonnavittula 和 D. P. Losey，“通过共享自主性传达机器人约定”，载于《2022年国际机器人与自动化会议（ICRA）》。IEEE，2022年，第7423–7429页。[13] M. Colledanchise 和 P. ¨Ogren，《机器人与人工智能中的行为树》。CRC出版社，2018年7月。 [在线]。可访问：https: // doi.org/ 10 1201 2 9780429489105 。% F F. Weichert、[14] D. Bachmann、B. Rudak 和 D. Fisseler，“Leap Motion 控制器的准确度与鲁棒性分析”，《传感器》，第13卷，第55期，第6380–6393页，2013年5月。[15] P. Vanc，“手势遥操作工具箱 v.0.1”，https://github.com/ imitrob/teleopgesturetoolbox/， 2022，[在线；访问于2-March- 2023]。[16] D. Emaasit，“Pymc-learn：Python中的实用概率机器学习”，编号 arXiv:1811.00542，2018年10月，arXiv:1811.00542 [cs, stat]. [在线]。可访问：http://arxiv.org/abs/1811.00542

[17]K. Forbes 和 E. Fiume，“一种使用加权主成分分析处理运动数据的高效搜索算法”，载于《2005年ACM SIGGRAPH/Eurographics计算机动画研讨会论文集 - SCA '05》。加利福尼亚州洛杉矶：ACM出版社，2005年，第67页。 [在线]。可访问：http://portal.acm.org/citation.cfm?doid
=1073368 1073377 。A.
[18]
Paraschos、C. Daniel、J. R. Peters 和 G. Neumann，“概率运动基元”，载于《神经信息处理系统进展》，第26卷。Curran Associates, Inc.，2013年。 [在线]。可访问：https://proceedings.neurips.cc/paper/2013/hash/e53a0a2978c28872a4505bdb51db06dc-Abstract.html[19] 《动态时间规整》。柏林，海德堡：Springer，2007年，第69–84页。
[ ] 10 1007 978 3 540 74048 3 4 在线。可访问：
[20]
https://doi.org/ 。 / - - - - K. P. Murphy，《机器学习：概率视角》。马萨诸塞州剑桥[等]：MIT出版社，2013年。[在线]。可访问：https://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/
0262018020 = 1 2 = 8 =1336857747 =8 2 /ref sr ?ie UTF &
[21]
qid &sr - A. Kucukelbir、D. Tran、R. Ranganath、A. Gelman 和 D. M. Blei，“自动微分变分推断”，《机器学习研究杂志》，2017年。[22] E. Rohmer、S. P. N. Singh 和 M. Freese，“V-rep：一种通用且可扩展的机器人仿真框架”，载于《2013年IEEE/RSJ智能机器人与系统国际会议》，2013年11月，第1321–1326页。[23] S. James、M. Freese 和 A. J. Davison，“Pyrep：将 V-REP 引入深度机器人学习”，编号 arXiv: 1906 11176, 2019, 1906 11176 。
[ ] [ ] 1906
6月 arXiv: 。 cs 。在线。可访问：http://arxiv.org/abs/ 11176
[24] ´ Y. Labbe、J.
Carpentier、M. Aubry 和 J. Sivic，“CosyPose：一致的多视角多物体6D姿态估计”，载于《计算机视觉——ECCV 2020》第28卷，2020年，16 23 16：第XX届欧洲会议，英国格拉斯哥，8月——论文集，第XVII部分。Springer，第–页。[25] B. Calli、A. Walsman、A. Singh、S. Srinivasa、P. Abbeel 和 A. M. Dollar，“操作研究中的基准测试：使用耶鲁大学-卡内基梅隆大学-加州大学伯克利分校物体与模型集”，《IEEE机器人与自动化杂志》，第22卷，第3期，第36–52页，2015年。