DOI: 10.19650/j.cnki.cjsi.J2210833

面向人机交互的手势指向估计方法*

陈仁钧[1],费敏锐[1,2],杨傲雷[1,2]
(1. 上海大学机电工程与自动化学院 上海 200444; 2. 上海大学上海市电站自动化技术重点实验室 上海 200444)

摘 要:针对人机共融环境中机器人与人之间的交互问题,提出了一种面向人机交互场景的手势指向估计方法,通过人体指向手势,以实现机器人对工作平面上指向目标点的信息交互。首先,基于 RGB-D 相机与 VICON 人体动作捕捉系统,构建时间同步的视觉指向手势位姿数据集,其中的每个样本包含人体指向手势的 RGB-D 图像和指向手势的位姿真值;其次,提出融合语义与几何信息的指向手势位姿估计多层次神经网络模型;然后,设计融合位置点误差 ΔP 和方向角度误差 $\Delta\theta$ 的射线近似损失函数,并基于构建的数据集,对指向手势位姿估计模型进行训练;最后,在实验室环境中进行了人机交互实验与模型验证。实验结果表明,在距离相机 5 m 的范围内,指向手势检测的平均精度为 98.4%,指向手势位姿的平均位置误差为 34 mm,平均角度误差为 9.94°,进而实现工作平面上的手势指向目标点的平均误差为 0.211 m。

关键词: 人机交互;指向手势;目标检测;位姿估计
中图分类号: TP391 TH86 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Estimation of gesture pointing for human-robot interaction

Chen Renjun[1], Fei Minrui[1,2], Yang Aolei[1,2]
(*1. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China;*
2. Shanghai Key Laboratory of Power Station Automation Technology, Shanghai University, Shanghai 200444, China)

Abstract: To solve the problem of interaction between robot and human in the human-robot integration environment, this article proposes an estimation method of gesture pointing for human-robot interaction scenes by pointing gesture to achieve the information interaction between the robot and the target point on the workplane. First, based on the RGB-D camera and the human motion capture system VICON, a time-synchronized visual pointing gesture position dataset is established. Each sample contains the RGB-D image of the pointing gesture and the true value of the pointing gesture pose. Secondly, a multi-level neural network model is formulated for estimating pointing gesture pose by combining semantic and geometric information. Thirdly, a ray approximation loss function is designed, which combines the position errorΔP and direction angle error $\Delta\theta$. The pointing gesture pose estimation model is trained based on the constructed dataset. Finally, human-robot interaction experiments and model validation are implemented in the laboratory environment. In the range of 5 m from the camera, results show that the average precision of pointing gesture detection is 98.4%, the average position error of pointing gesture pose is 34 mm, and the average angle error is 9.94°. The average error of gesture pointing to the target point on the workplane is 0.211 m.

Keywords: human-robot interaction; pointing gesture; object detection; pose estimation

0 引 言

基于视觉的手势交互是人机交互的重要研究内容[1],具有广泛的应用前景[2]。人体手势根据形态的不同可以分成很多类别,其中指向手势是一种表示方向的通用交互语言。人类与机器人进行交互时,采用手势指向的方式与机器人进行信息传递,显得非常自然和方便。所以,手势指向的研究工作具有重要的理论和现实意义。

收稿日期:2022-12-05 Received Date: 2022-12-05
*基金项目:上海市自然科学基金(22ZR1424200)、国家自然科学基金(62203290)、111 引智基地项目(D18003)资助

基于视觉的手势指向研究可以分为指向方向分类[3]和指向方向估计[4]的研究。在指向方向的分类方面,研究人员已经提出了许多方法。Lai 等[5]使用 Kinect 相机获取骨架关节数据,依据设定的约束来判定出指向手势,然后通过肩关节和手关节的 3D 坐标连线来表示指向向量,最后量化到平面的 8 个方向来控制机器人方向的移动。Barbed 等[6]将手势的指向分为了 8 个平面方向类别和 26 个空间方向类别。比较了 3 种不同的分类方法,在先通过 Mask R-CNN 分割出人体,再通过改进的 MobileNetV2 进行方向分类的方法上取得了较好的分类效果。

在指向方向估计的研究方面,又主要分为结合人体结构特征的方向估计和针对指向手势手部的方向估计。结合人体特征的方向估计是根据人体其他部位,例如面部、肘部、肩部的特征结合手部的特征进行估计。文献[7]研究表明,面部到手部之间的视线为指向方向提供了非常可靠的估计。Raheja 等[8]则借助肩关节特征,采用 Kinect 相机通过骨架识别的方法获得肩关节和手关节的位置,通过两个关节点的位置关系来确定指向方向,实现桌面上指向物品识别的应用。Tölgyessy 等[9]选择手肘的特征,通过获得手肘与手腕关节点方向向量来表示指向方向,实现机器人对工作平面上指向目标点的确认。这种结合人体特征的估计方法,总是对用户的位置和指向方式有所限制,要求用户的全身或部分身体在相机的视野中。因此,在大多数情况下,更为典型的方法是从指向手势手部进行方向估计。

目前针对指向手势手部的研究工作主要探索了手是场景中的主要对象的近距离交互[10]。有一些方法试图从 2D 图推断 3D 方向[11],但由于没有深度数据,其精度受到了限制,只能在近距离得到一定的效果。而在许多人机交互的应用中,例如与移动机器人的交互,指向手势手部可能与机器人的距离较远,因此利用深度数据已成为这项任务中不可缺少的部分。Das 等[12]利用深度数据,通过一定的距离约束条件分割出手部区域,然后识别出指跟和指尖,根据两点连线得到方向,与相机的距离可以达到 1 m 以上。

面向通过人体指向手势实现机器人与工作平面上指向目标点信息交互的场景,提出了一种手势指向估计方法,该方法适用于任何安装了 RGB-D 相机的移动机器人。本文构建了一个带有指向手势位姿标签和指向手势检测框标签的视觉指向手势位姿数据集。该数据中的指向手势距离相机深度在 5 m 内。基于卷积神经网络构建了一个指向手势位姿估计模型,并设计了射线近似损失函数。通过融合 RGB-D 图中的语义和几何特征直接回归得到指向手势位姿。

1　问题描述与方法架构

1.1　问题描述及指向手势位姿的定义

为了描述机器人与人体手势指向目标点的交互场景,相关坐标系如图 1 所示。世界坐标系 $\{W\}$ 为任务空间的基坐标系。相机坐标系 $\{C\}$ 原点 O_C 为相机光心,Z_C 为光轴方向。相机坐标系 $\{C\}$ 相对于机器人坐标系 $\{R\}$ 的关系为 ${}^R_C\boldsymbol{T}$,与相机在机器人上的安装位置有关。机器人坐标系 $\{R\}$ 相对于世界坐标系 $\{W\}$ 的关系 ${}^W_R\boldsymbol{T}$,表示了机器人在世界基坐标系中的位置。因此,相机坐标系 $\{C\}$ 与世界坐标系 $\{W\}$ 的关系 ${}^W_C\boldsymbol{T}$ 已知,可描述为 ${}^W_R\boldsymbol{T}\times{}^R_C\boldsymbol{T}$。世界坐标系 $\{W\}$ 下指向目标点为 $M({}^Wm_x,{}^Wm_y)$。

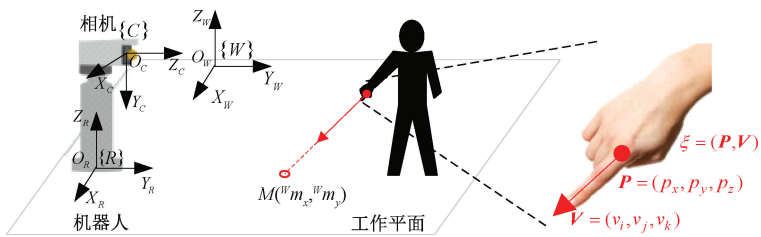


图 1　坐标系与指向手势位姿的定义

Fig. 1　Coordinate system and definition of pointing gesture pose

机器人需要根据携带的 RGB-D 相机检测到指向手势,并估计指向手势位姿,计算得到指向目标点,其中估计指向手势位姿是实现整个过程的关键。对于该指向手势位姿,主要关注的是其在空间中的位置和指向方向,所以在 3D 空间中用一条射线来定义指向手势位姿,该射线由一个 3D 位置点 $\boldsymbol{P}=(p_x,p_y,p_z)$,一个 3D 方向向量 $\boldsymbol{V}=(v_i,v_j,v_k)$ 组成。如图 1 所示,定义为:

$$\xi=(\boldsymbol{P},\boldsymbol{V}) \tag{1}$$

其中,$\boldsymbol{P}=(p_x,p_y,p_z)$,$\boldsymbol{V}=(v_i,v_j,v_k)$。

假设通过指向手势位姿估计模型已估计出指向手势在相机坐标系下 $\{C\}$ 的位姿为 ${}^C\xi=({}^C\boldsymbol{P},{}^C\boldsymbol{V})$。根据相机坐标系 $\{C\}$ 相对于世界坐标系 $\{W\}$ 的变换关系 ${}^W_C\boldsymbol{T}$,可计

算出世界坐标系{W}下指向手势位姿${}^{W}\xi=({}^{W}\boldsymbol{P},{}^{W}\boldsymbol{V})$，计算公式如下：

$$\begin{cases}{}^{W}\boldsymbol{P}={}^{W}_{C}\boldsymbol{T}^{C}\boldsymbol{P}\\{}^{W}\boldsymbol{V}={}^{W}_{C}\boldsymbol{R}^{C}\boldsymbol{V}\end{cases}\tag{2}$$

式中：${}^{W}_{C}\boldsymbol{R}$ 为${}^{W}_{C}\boldsymbol{T}$ 的旋转矩阵部分。

工作平面方程为世界坐标系{W}中的平面，由于机器人只在 $Z=0$ 地面上移动，本文定义的平面为 $Z=0$，联立射线方程和工作平面方程，即可解得指向目标点 $M({}^{W}m_x,{}^{W}m_y)$：

$$\begin{cases}\dfrac{{}^{W}m_x-{}^{W}p_x}{{}^{W}v_i}=\dfrac{{}^{W}m_y-{}^{W}p_y}{{}^{W}v_j}=\dfrac{{}^{W}m_z-{}^{W}p_z}{{}^{W}v_k}=t>0\\{}^{W}m_z=0\\{}^{W}v_k<0\end{cases}\tag{3}$$

1.2 指向目标点估计方法与架构

本文提出的手势指向估计方法是通过机器人上相机，采集场景中的 RGB-D 图，通过指向手势检测模型检测出指向手势，再通过设计指向手势位姿估计模型，构建相机坐标系{C}下 RGB-D 图中指向手势区域与指向手势位姿之间的映射关系，估计出相机坐标系下指向手势位姿，最后变换指向手势位姿到世界坐标系下，通过与工作平面方程的关系计算得到工作平面上的指向目标点。

综上所述，机器人与人体手势指向目标点的交互问题包括了指向手势检测，指向手势位姿估计，指向目标点计算问题。本文提出的方法架构如图 2 所示，主要由 3 部分构成：1）指向手势位姿数据集构建；2）指向手势检测模型与指向手势位姿估计模型的构建与评估；3）人体手势指向目标点的交互验证。部分 1）的作用是将相机采集的含有指向手势的场景 RGB-D 图序列与 VICON 系统捕捉的指向手势刚体位姿序列进行时间同步与匹配，根据刚体位姿计算定义的指向手势位姿 ${}^{C}\xi$，并手动标注 2D 图中指向手势手部区域，构建包含指向手势的 RGB-D 图，指向手势位姿标签与指向手势检测框标签的指向手势位姿数据集。部分 2）是根据获取的指向手势位姿数据集，构建并训练指向手势检测模型和指向手势位姿估计模型。部分 3）通过加载指向手势检测模型和位姿估计模型，对工作平面上人体手势指向目标点交互进行验证，计算并分析指向目标点的误差。

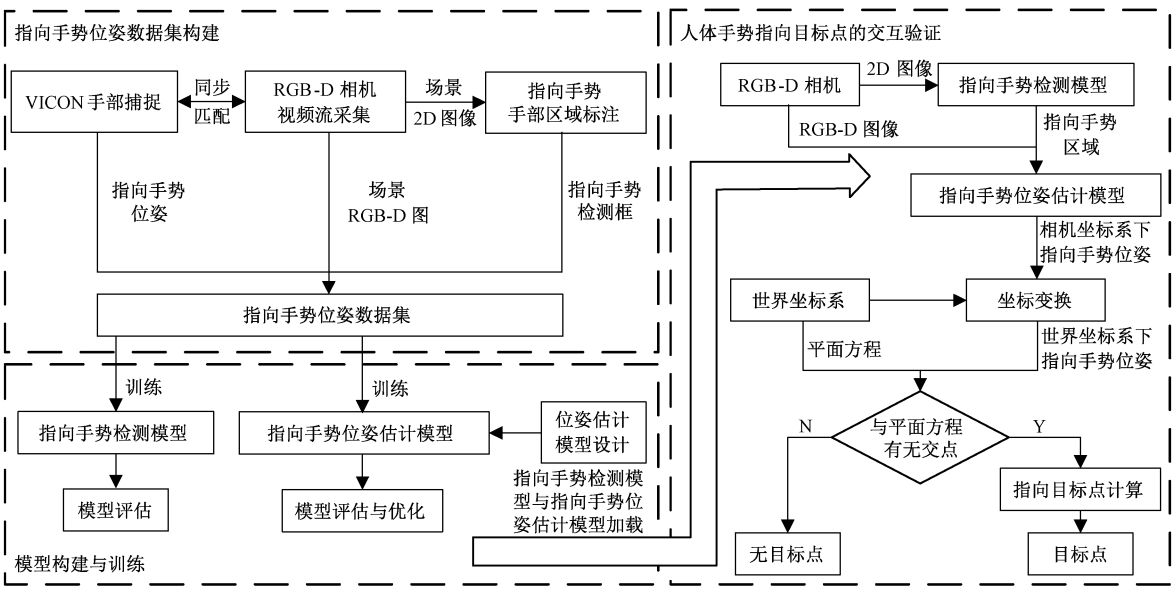


图 2 整体方法架构

Fig. 2 Architecture of the proposed method

2 指向手势位姿数据集构建

本文搭建的指向手势位姿数据采集场景如图 3 所示。数据采集场景上方布置了一套 VICON 运动捕捉系统，该系统是一种基于 marker 点识别的光学动作捕捉系统，能够计算得到由 marker 球所建刚体的位姿，具有高速度，高分辨率和高精度的特点。场景下方地面包含一个视觉相机和做出指向手势的人体对象。由于视觉相机可视范围以及场景大小的限制，人体运动范围在距离相机前方 5 m 的正方形区域内。人体在场景中不同位置，做出不同指向的指向手势，视觉相机采集含有指向手势不同位置和不同方向的 RGB-D 场景图，VICON 系统采集同步的指向手势位姿真值。指向手势位姿数据集构建主要包括 3 部分：1）相机坐标系{C}下指向手势位姿 ${}^{C}\xi$ 真值标签的获取；2）包含指向手势的 RGB-D 图对齐，并

与指向手势位姿真值标签进行同步与匹配;3）图像中指向手势区域的标注。

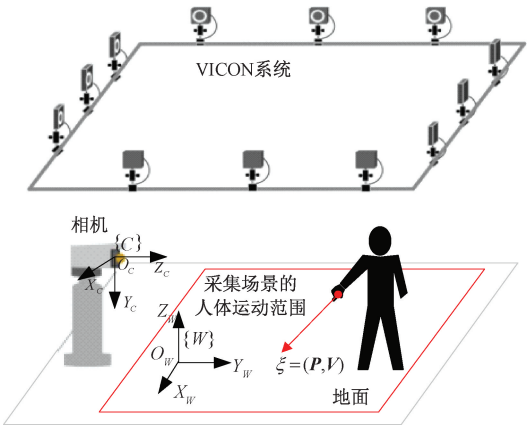


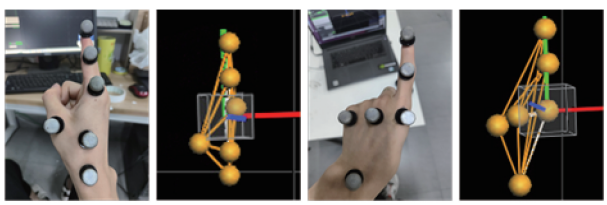
图 3　指向手势位姿数据的采集场景

Fig. 3　Acquisition scene of pointing gesture pose data

1）指向手势位姿真值标签的获取

初始化 VICON 系统基坐标系与世界坐标系 $\{W\}$ 重合。VICON 系统通过实时跟踪 marker 球建立的指向手势刚体位姿变换 ${}^{W}\boldsymbol{T}$ 计算指向手势位姿 ${}^{C}\xi$ 真值。

在获取正对、背对、侧对相机的各方向各姿态指向手势的同时,为了保证 VICON 系统的 marker 贴点不影响获取到的 2D 图和深度图质量,在指向手势的不同面进行刚体的建立。一共构建了 8 种指向手势刚体,左手与右手各 4 种刚体,分别在手背、手心、手左侧和手右侧进行建立。为了统一不同刚体得到所定义的指向手势位姿 ${}^{C}\xi$,在构建不同刚体时,忽略手指的粗细,将食指指跟作为刚体原点,食指指向作为刚体 Y 轴方向。图 4 所示为手左侧和手背面的现实 marker 贴点和 VICON 系统中所建指向手势刚体及其坐标系。



(a) 左侧贴点 (a) Left marker　(b) 左侧刚体 (b) Left body　(c) 背面贴点 (c) Back marker　(d) 背面刚体 (d) Back body

图 4　指向手势现实贴点及相应刚体

Fig. 4　Reality marker and rigid body of pointing gesture

获取到世界坐标系 $\{W\}$ 下的指向手势刚体位姿变换 ${}^{W}\boldsymbol{T}$ 后,需要根据刚体位姿先计算出世界坐标系 $\{W\}$ 下所定义的指向手势位姿 ${}^{W}\xi$。由于已将不同刚体统一到刚体原点与刚体 Y 轴方向一致,所以将刚体原点作为指向手势位姿 3D 位置点 $\boldsymbol{P}=(p_x,p_y,p_z)$,刚体 Y 轴方向作为指向手势位姿 3D 方向向量 $\boldsymbol{V}=(v_i,v_j,v_k)$,即得到世界坐标系 $\{W\}$ 下指向手势位姿 ${}^{W}\xi=({}^{W}\boldsymbol{P},{}^{W}\boldsymbol{V})$,如下:

$$\begin{cases} {}^{W}\boldsymbol{P}={}^{W}\boldsymbol{t} \\ {}^{W}\boldsymbol{V}={}^{W}\boldsymbol{R}\boldsymbol{V}_{Y}={}^{W}\boldsymbol{R}[0\quad 1\quad 0]^{\mathrm{T}} \end{cases} \tag{4}$$

式中:${}^{W}\boldsymbol{R}$ 与 ${}^{W}\boldsymbol{t}$ 分别为 ${}^{W}\boldsymbol{T}$ 的旋转矩阵和平移向量部分;$\boldsymbol{V}_Y$ 为刚体 Y 轴的方向向量。

相机的位置与 VICON 系统变换矩阵 ${}^{C}_{W}\boldsymbol{T}$ 可以通过视觉标定获得[13],因此通过坐标变换,最后得到相机坐标系 $\{C\}$ 下所定义的指向手势位姿 ${}^{C}\xi=({}^{C}\boldsymbol{P},{}^{C}\boldsymbol{V})$ 标签真值,如下:

$$\begin{cases} {}^{C}\boldsymbol{P}={}^{C}_{W}\boldsymbol{T}{}^{W}\boldsymbol{P} \\ {}^{C}\boldsymbol{V}={}^{C}_{W}\boldsymbol{R}{}^{W}\boldsymbol{V} \end{cases} \tag{5}$$

式中:${}^{C}_{W}\boldsymbol{R}$ 为 ${}^{C}_{W}\boldsymbol{T}$ 的旋转矩阵部分。

2）RGB-D 图与指向手势位姿真值的同步匹配

VICON 系统与图像采集系统在同一局域网下,通过网络套接字进行通信和时钟同步,实现采集同步和匹配。图像采集系统开始按照一定频率进行采集和对齐 RGB-D 图,并同步获取 VICON 系统下的指向手势位姿真值。

3）图像中指向手势区域标注

由于采集的是场景的 RGB-D 图,所以指向手势区域只是场景图中的一部分。为了训练指向手势检测网络和指向手势位姿估计网络,需要对图像中的指向手势区域进行标注。本文借助标注工具,通过手动标注的方法进行指向手势检测框标注。

3　指向手势位姿估计模型的建立

关于指向手势目标检测模型,直接通过训练成熟的目标检测网络模型得到。由于 YOLOv5 目标检测网络是单步回归网络,在保证高检测率的同时能达到较快的检测速度,在目标检测领域的应用已经十分广泛[14],因此训练该网络得到指向手势目标检测模型。而对于位姿估计模型,需要自主构建一个能从指向手势 RGB-D 区域映射到指向手势位姿的多层次神经网络模型。其模型构建主要包括数据预处理、网络设计和损失函数 3 个部分。

3.1　数据预处理与数据增强

在截取由指向手势检测模型得到的指向手势所在区域时,为了避免模型检测到的指向手势区域不完整,截取的范围基于指向手势检测的范围扩大 1.2 倍,并在训练过程中采用随机区域获取数据增强方法。

该方法描述如下:已有指向手势检测框坐标标签 (x_1,y_1,x_2,y_2),其中 (x_1,y_1) 与 (x_2,y_2) 分别为检测框左上角和右下角坐标。计算将检测框放大 1.2 倍和缩小 0.8 倍时的坐标标签 $(x_1^+,y_1^+,x_2^+,y_2^+)$ 和 $(x_1^-,y_1^-,x_2^-,y_2^-)$,这样就获得了左上角和右下角 x 和 y 各 3 个可能取值,可以随机组成 9 个左上角坐标和右下角坐标,进而能随机

获取 81 个区域。当随机组成的坐标标签为(x_1,y_1^-,x_2^-,y_2^+)时,截取的指向手势区域如图 5 所示(实线框)。

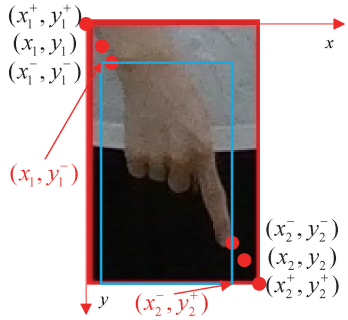


图 5 随机区域获取数据增强

Fig. 5 Random area acquisition for data enhancement

当直接截取 RGB-D 图中指向手势区域时,会失去指向手势在场景图中的位置信息,使得在场景图不同位置而指向方向相同的指向手势具有相同的位姿 ${}^C\xi$,而实际在空间中不同位置的指向手势应是不同的位姿 ${}^C\xi$。为了保留指向手势在场景图中的位置信息,将深度图通过相机内参矩阵转换为点云图,截取指向手势区域的 2D 图和点云图再输入到指向手势位姿估计网络模型中。

3.2 网络设计

2D 图和点云图作为两种异构数据,分别蕴含了不用形式的信息,2D 图主要包含了语义信息,点云图则主要包含了几何信息,所以对于这两张图的输入若进行简单的逐像素串联,是不能充分利用这两种信息的,应分别进行高维语义和几何特征提取后,再融合进行全局特征提取,最后通过全局特征来回归指向手势位姿 ${}^C\xi$。网络设计由 4 部分构成:1)编解码模块进行高维特征提取;2)坐标注意力模块加强特征图中重要位置点间联系;3)特征融合模块进行高维语义和几何特征的全局特征提取;4)回归模块进行指向手势位姿 ${}^C\xi$ 回归。网络结构如图 6 所示。

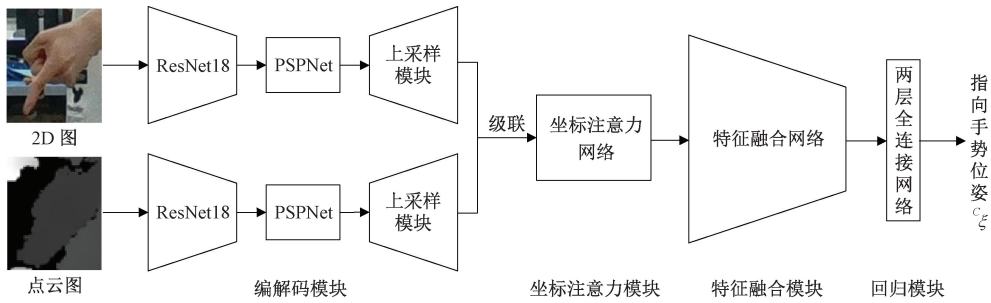


图 6 指向手势位姿估计网络结构

Fig. 6 The architecture of pointing gesture pose estimation network

1)编解码模块

本文采用由 ResNet18、PSPNet 和卷积神经网络(convolutional neural network,CNN)多级上采样构成的编解码结构网络模块分别对 2D 图和点云图进行像素级的高维特征提取,有利于提升网络对 2D 图中语义信息和点云图中几何信息的充分利用。

2)坐标注意力模块

坐标注意力[15]将通道注意力分解为两个一维特征编码过程,分别沿宽和高两个方向进行特征聚合。通过这种方式,可以沿一个方向捕获远程依赖关系,同时可以沿另一个空间方向保留精确的位置信息。这样便能有效地将空间坐标信息整合到生成的注意力图中,加强特征图中各重要位置点间的联系和敏感性。

3)特征融合模块

为了适应对不同尺寸的输入要求,全局特征提取的特征融合模块是由 ResNet50 进行一定的改进得到的。输入的高维语义和几何特征图尺寸是不固定的,本文在 ResNet50 的最后一层通过全局平均池化来提取全局特征,对于不同尺寸的输入统一了输出。并将整个网络中的批归一化(batch normalization, BN)层都改为组归一化(group normalization, GN)[16]层。这是因为 GN 层能在单个数据内部进行归一化的同时,在视觉任务模型上表现更好。

4)回归模块

提取了高维语义和几何特征融合的全局特征后,需要输出 6 个值来回归指向手势位姿 ${}^C\xi$,通过两个全连接层实现该回归模块,两个全连接层的输出分别为 256 维和 6 维。

3.3 射线近似损失函数

指向手势位姿 ξ 由空间位置点 $\boldsymbol{P}=(p_x,p_y,p_z)$ 和方向向量 $\boldsymbol{V}=(v_i,v_j,v_k)$ 表示的射线所定义,为了让估计射线与目标射线在距离上和方向上靠近,设计射线近似损失函数,该损失函数包含两部分,一部分为方向余弦相似度损失 $Loss_{Sim}$;另一部分为位置点间均方差损

失 $Loss_{Mse}$。

$$Loss_{Sim} = 1 - \cos\theta = 1 - \frac{{}^{C}\boldsymbol{V} \cdot {}^{C}\tilde{\boldsymbol{V}}}{\| {}^{C}\boldsymbol{V} \| \ \| {}^{C}\tilde{\boldsymbol{V}} \|} = 1 - \frac{\sum_{i=i,j,k} {}^{C}v_i \times {}^{C}\tilde{v}_i}{\sqrt{\sum_{i=i,j,k} ({}^{C}v_i)^2} \times \sqrt{\sum_{i=i,j,k} ({}^{C}\tilde{v}_i)^2}} \tag{6}$$

$$Loss_{Mse} = MSE({}^{C}\boldsymbol{P}, {}^{C}\tilde{\boldsymbol{P}}) = \frac{\sum_{i=x,y,z} ({}^{C}p_i - {}^{C}\tilde{p}_i)^2}{3} \tag{7}$$

则射线相似损失函数 $Loss_{Raysim}$ 为:

$$Loss_{Raysim} = Loss_{Sim} + Loss_{Mse} \tag{8}$$

式中:${}^{C}\boldsymbol{P} = ({}^{C}p_x, {}^{C}p_y, {}^{C}p_z)$ 和 ${}^{C}\boldsymbol{V} = ({}^{C}v_i, {}^{C}v_j, {}^{C}v_k)$ 为指向手势位姿 ${}^{C}\xi$ 标签对应的坐标点和方向向量;${}^{C}\tilde{\boldsymbol{P}} = ({}^{C}\tilde{p}_x, {}^{C}\tilde{p}_y, {}^{C}\tilde{p}_z)$ 和 ${}^{C}\tilde{\boldsymbol{V}} = ({}^{C}\tilde{v}_i, {}^{C}\tilde{v}_j, {}^{C}\tilde{v}_k)$ 为指向手势位姿估计网络估计的位姿 ${}^{C}\tilde{\xi}$ 对应的坐标点和方向向量。

4　手势指向目标点估计方法的验证及分析

4.1　实验平台

实验的算法训练环境为 Ubuntu16.04 LTS 系统的计算机,搭载 Intel® Core™ i7-7700 CPU @ 3.60 GHz×8,16 G RAM 和 8 GB 显存的 NVIDIA GeForce RTX2070 GPU。

4.2　数据集采集与构建

样本数据采集如图 3 所示。本文采用 RealSense D435i 相机,该相机是一款 RGB-D 相机,集成了一个基于结构光原理的深度模块传感器和一个彩色相机,能够输出的深度图分辨率最高可达 1 280×720,彩色图分辨率最高达 1 920×1 080,深度探测范围最高可达 10 m。通过该相机和 VICON 系统,在实验室环境下收集了一个带有指向手势位姿 ${}^{C}\xi$ 标签的多深度多位姿样本数据集。共有 18 967 个样本数据,每个样本数据包含分辨率为 1 280×720 的 RGB-D 图,指向手势检测框标签,指向手势位姿标签。将收集的指向手势数据集,按照 8:2的比例随机划分训练集和测试集。不同深度下的样本数量如表 1 所示。图 7 所示为数据集中的部分样本,为彩色图像、伪彩色化的深度图以及指向手势位姿 ${}^{C}\xi$ 标签的具体数值,并将检测框与位姿标签可视化到图像中。

表 1　指向手势位姿数据集

Table 1　Pointing gesture pose data set

指向手势距相机深度/m	训练集数量	测试集数量
0~2	3 630	895
2~3	8 567	2 105
3~5	2 974	796
0~5 (总计)	15 171	3 796

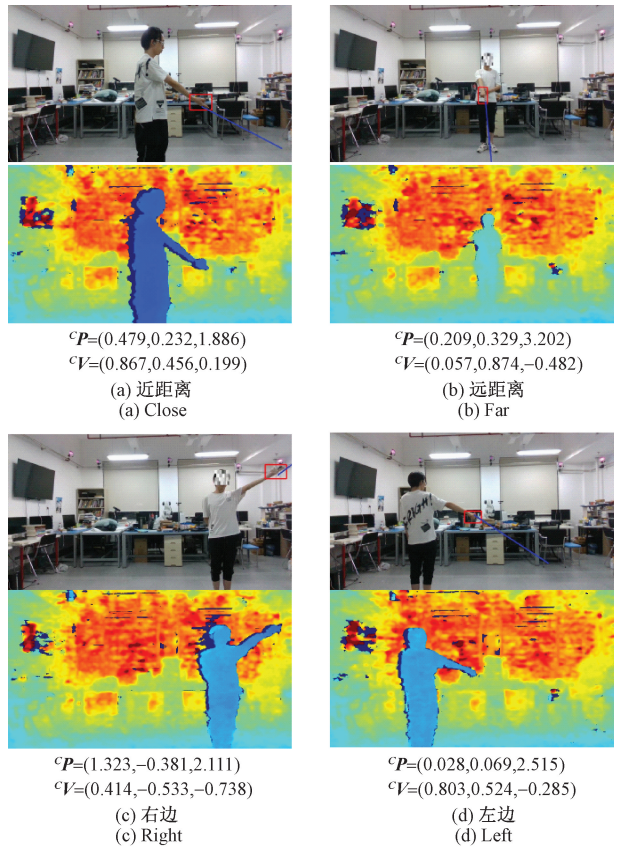


图 7　部分数据样本

Fig. 7　Partial data samples

4.3　指向手势检测模型的评估与分析

为了评估指向手势检测模型的效果,通过计算模型在测试集上的准确率(precision, P)、召回率(recall, R)和平均精度(average precision, AP)进行评估,如表 2 所示。

表 2　指向手势检测模型评估结果

Table 2　Pointing gesture detection results　%

评估数据集	*P*	*R*	*AP*
测试集	96.9	95.4	98.4

结果表明,该指向手势检测模型能够较好地检测到图片中的指向手势。识别精度达到了 98.4%。

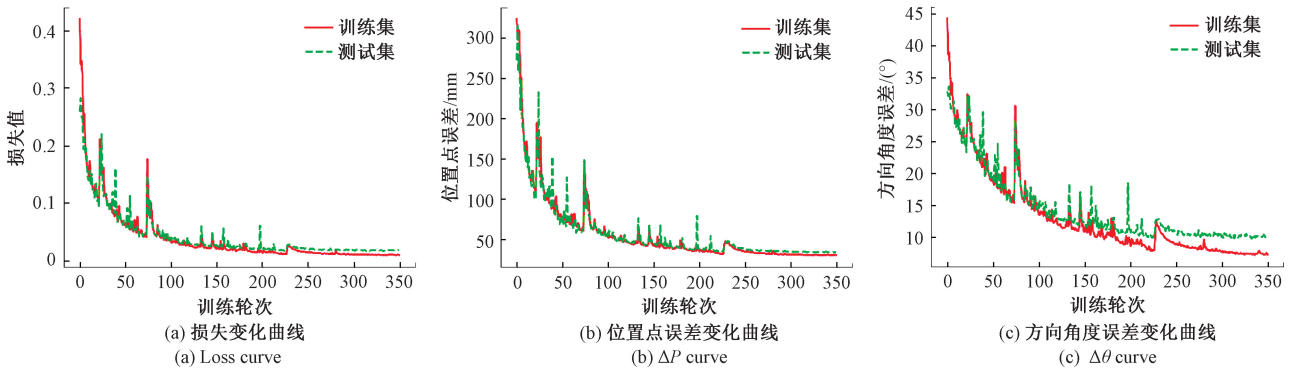
4.4　指向手势位姿估计模型的评估与分析

对于指向手势位姿估计模型的评估,本文通过计算测试集上平均位置预测误差 ΔP,平均方向角度预测误差 $\Delta\theta$ 指标来进行评估。

$$\Delta P = \sqrt{\sum_{i=x,y,z} (p_i - \tilde{p}_i)^2} \tag{9}$$

$$\Delta\theta = \frac{360}{\pi} \cos^{-1}\left(\frac{\boldsymbol{V} \cdot \tilde{\boldsymbol{V}}}{\| \boldsymbol{V} \| \ \| \tilde{\boldsymbol{V}} \|}\right) \tag{10}$$

训练网络时,在训练集和测试集上的损失变化曲线,位置点距离误差变化曲线,方向角度误差变化曲线如图 8 所示。在计算所有测试数据估计指标的同时,进行不同深度下的指向手势位姿估计结果比较,位姿估计结果如表 3 所示。



(a) 损失变化曲线 (a) Loss curve
(b) 位置点误差变化曲线 (b) ΔP curve
(c) 方向角度误差变化曲线 (c) $\Delta\theta$ curve

图 8 训练过程变化曲线

Fig. 8 Training process curves

表 3 不同深度下的指向手势位姿估计结果

Table 3 Pose estimation results at different depths

指向手势距离相机深度/m	ΔP/mm	$\Delta\theta$/(°)
0~2	26	7.69
2~3	33	9.94
3~5	46	11.76
0~5 (总计)	34	9.94

由表 3 可知,指向手势位姿估计模型的平均位置点估计误差 $\Delta P = 34$ mm,平均方向角度估计误差 $\Delta\theta = 9.94°$,模型对于指向手势位姿具有较好的估计结果。另外,通过比较距离相机不同深度的估计结果,距离相机越近,平均位置点估计误差和平均方向角度估计误差都更小,可知指向手势位姿估计模型对于近距离的指向手势位姿估计效果更好,这与距离相机越近获得的指向手势区域信息越丰富相符。

通过一定的消融和对比实验来评价模型的效果,进行了直接级联 2D 图与点云图的网络模型与去除注意力模块的网络模型训练,并与 PointNet++ 网络[17]训练的效果进行对比。PointNet++ 网络是一种经典的处理不规则点云数据的分层神经网络,可以提取点云的局部和全局特征,对物体姿态进行较好的估计。比较结果如表 4 所示。

由消融实验的结果可知,采用编解码器进行高维特征提取和引入坐标注意力模块都可以提高模型对指向手势位姿估计的效果,特别是对指向方向的估计效果。通过 GN 层的引入,能够明显对指向手势位姿估计起到更好的效果。通过与 PointNet++网络训练结果对比,说明了保存点云的图结构,通过所构建的估计方法

表 4 消融与对比实验结果

Table 4 Ablation and comparison experiment results

模型	ΔP/mm	$\Delta\theta$/(°)	ΔT/ms
直接级联+特征融合(GN)	58	13.95	16.23
编解码器+特征融合(GN)	36	11.28	35.89
编解码器+坐标注意力+特征融合(GN)(本文)	34	9.94	37.64
编解码器+坐标注意力+特征融合(无 GN)	110	16.26	24.75
PointNet++	41	12.56	145.15

优于不关注点云的图结构,通过 PointNet++的估计方法。在平均估计时间 ΔT 方面,所提出的模型为 37.64 ms,PointNet++网络模型为 145.15 ms,所提出的模型明显优于 PointNet++网络模型,更好地满足实时性的要求。

4.5 人体手势指向目标点的交互验证与分析

经过上述的指向手势检测模型和指向手势位姿估计模型,已经得到了相机坐标系 $\{C\}$ 下指向手势位姿 ${}^{C}\xi$。将相机坐标系 $\{C\}$ 下指向手势位姿 ${}^{C}\xi$ 通过坐标变换转换到世界坐标系 $\{W\}$ 下,并计算与工作平面方程 $Z=0$ 的交点来验证手势指向目标点估计方法,分析指向目标点的准确性。

由于世界坐标系 $\{W\}$ 下,只有方向指向下方的指向手势位姿 ${}^{W}\xi$,才能和工作平面相交形成指向目标点,本文只分析指向手势与世界坐标系 $\{W\}$ Z 轴负方向 θ_Z^- 呈 60°以下的指向手势。通过计算指向目标点的位置误差 ΔD 作为评估指标(式(11)),结果如表 5 所示。

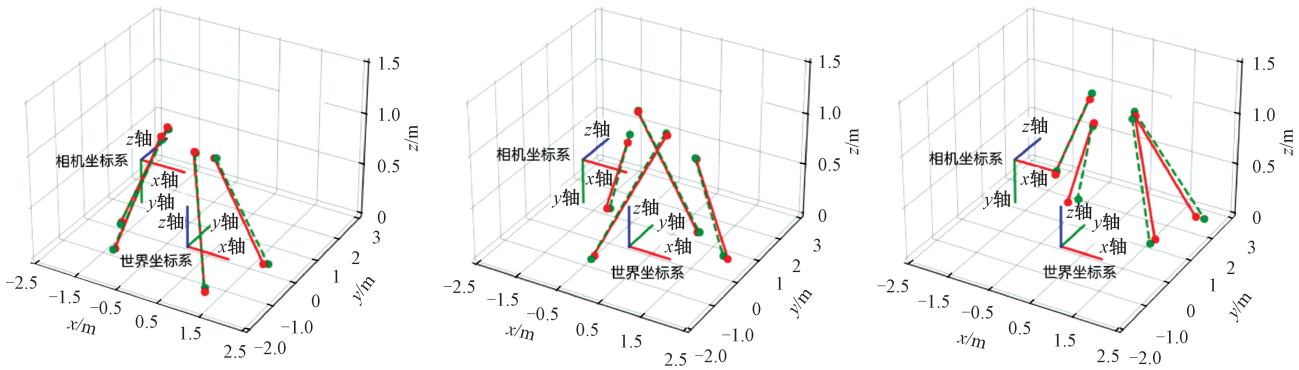
$$\Delta D = \sqrt{\sum_{i=x,y}(m_i - \tilde{m}_i)^2} \tag{11}$$

表 5　指向目标点估计结果

Table 5　Point to target estimation results

指向手势距离 相机深度/m	ΔD/m		
	$\theta_Z^- < 30°$	$\theta_Z^- < 45°$	$\theta_Z^- < 60°$
0~2	0.090	0.125	0.168
2~3	0.132	0.161	0.208
3~5	0.171	0.208	0.272
全部	0.130	0.162	0.211

表 5 数据表明,在距离相机 5 m 的范围内,指向手势指向目标点的总体平均估计误差为 0.211 m。这对于配备 RGB-D 相机的移动机器人来说,能在指向手势的交互下,得到前方至少 5 m×5 m 的区域内较为准确的目标点。并且指向手势距离相机越近,与 Z 轴负方向夹角越小,指目标点估计效果越好。随机可视化了部分不同深度下指向手势位姿和指向目标点的估计结果如图 9 所示。其中,实线为真值,虚线为估计值。



(a) 0~2 m　(b) 2~3 m　(c) 3~5 m

图 9　不同距离下估计结果可视化

Fig. 9　Visualization of estimation results at different distances

5　结　　论

本文提出了一种面向人机交互的手势指向估计方法,只通过指向手势手部区域就能估计出指向,不需要人体其他特征的辅助,具有更少的约束和更广的适用性。经过实验表明,该方法对手势指向具有较好的估计效果,能够用于机器人对工作平面上指向目标点的信息交互。本文方法是直接基于单张 RGB-D 图估计得到手势指向,在进一步研究中,可以结合滤波算法从多张 RGB-D 图的估计结果中得到更为可靠的指向。

参考文献

[1] 鹿智, 秦世引, 李连伟, 等. 智能人机交互中第一视角手势表达的一次性学习分类识别[J]. 自动化学报, 2021, 47(6): 1284-1301.
LU ZH, QIN SH Y, LI L W, et al. One-time learning classification recognition of first view gesture expression in intelligent human-computer interaction [J]. Acta Automatica Sinica, 2021, 47(6): 1284-1301.

[2] RASTGOO R, KIANI K, ESCALERA S. Sign language recognition: A deep survey [J]. Expert Systems with Applications, 2021, 164: 113794.

[3] MAZHAR O, NAVARRO B, RAMDANI S, et al. A real-time human-robot interaction framework with robust background invariant hand gesture detection [J]. Robotics and Computer-Integrated Manufacturing, 2019, 60: 34-48.

[4] YOO M, NA Y, SONG H, et al. Motion estimation and hand gesture recognition-based human-UAV interaction approach in real time [J]. Sensors, 2022, 22(7): 2513.

[5] LAI Y, WANG C, LI Y, et al. 3D pointing gesture recognition for human-robot interaction [C]. 2016 Chinese Control and Decision Conference (CCDC), IEEE, 2016: 4959-4964.

[6] BARBED O L, AZAGRA P, TEIXEIRA L, et al. Fine-grained pointing recognition for natural drone guidance[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020: 1040-1041.

[7] MASHITA T, SHINTANI K, KIYOKAWA K. Improving pointing direction estimation by considering hand-and ocular-dominance[J]. IEICE Transactions on Information and Systems, 2020, 103(10): 2168-2177.

[8] RAHEJA J L, CHANDRA M, CHAUDHARY A. 3D gesture based real-time object selection and recognition[J]. Pattern Recognition Letters, 2018, 115: 14-19.

[9] TÖLGYESSY M, DEKAN M, DUCHOŇ F, et al. Foundations of visual linear human-robot interaction via pointing gesture navigation[J]. International Journal of Social Robotics, 2017, 9(4): 509-523.

[10] 舒子超, 曹松晓, 谢代梁, 等. 基于三维视觉特征的数字手势语义识别新方法研究[J]. 电子测量与仪器学报, 2021, 35(6): 124-130.
SHU Z CH, CAO S X, XIE D L, et al. Research on a new method of digital gesture semantic recognition based on 3D visual features [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(6): 124-130.

[11] SHUKLA D, ERKENT O, PIATER J. Probabilistic detection of pointing directions for human-robot interaction[C]. 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2015: 1-8.

[12] DAS S S. Precise pointing direction estimation using depth data[C]. 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, 2018: 202-207.

[13] GARRIDO-JURADO S, MUÑOZ-SALINAS R, MADRID-CUEVAS F J, et al. Automatic generation and detection of highly reliable fiducial markers under occlusion[J]. Pattern Recognition, 2014, 47(6): 2280-2292.

[14] 闫钧华, 张琨, 施天俊, 等. 融合多层级特征的遥感图像地面弱小目标检测[J]. 仪器仪表学报, 2022, 43(3): 221-229.
YAN J H, ZHANG K, SHI T J, et al. Ground small object detection from remote sensing images fusing multilevel features [J]. Chinese Journal of Scientific Instrument, 2022, 43(3): 221-229.

[15] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 13713-13722.

[16] WU Y, HE K. Group normalization[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[17] QI C R, YI L, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in Neural Information Processing Systems, 2017, DOI:10.48550/arXiv.1706.02413.

作者简介

陈仁钧,2020 年于宁波大学获得学士学位,现为上海大学硕士研究生,主要研究方向为机器人视觉。
E-mail: crjunchen@163.com

Chen Renjun received his B. Sc. degree from Ningbo University in 2020. He is currently a M. Sc. candidate at Shanghai University. His main research interest is robotic vision.

费敏锐(通信作者),分别在 1984 年和 1992 年于上海工业大学获得学士学位和硕士学位,1997 年于上海大学获得博士学位,现为上海大学教授、博士生导师,主要研究方向为智能化网络控制理论、系统、仿真和安全,及其关键技术在智能机器人与机器视觉系统等中应用。
E-mail: mrfei@staff.shu.edu.cn

Fei Minrui (Corresponding author) received his B. Sc. degree and M. Sc. degree both from Shanghai University of Technology in 1984 and 1992, and Ph. D. degree from Shanghai University in 1997, respectively. Now he is a professor and a Ph. D. supervisor at Shanghai University. His main research interests include intelligent network control theory, systems, simulation, and security, and its key technologies in the application of intelligent robots and machine vision systems.