doi:10.6041/j.issn.1000-1298.2024.08.040

基于深度归一化的任意交互物体检测方法研究

黄玲涛　孔紫静　杨　帆　张红彦
（吉林大学机械与航空航天工程学院，长春 130022）

摘要：交互物体的检测识别是实现人机交互的一项关键技术，针对人机交互过程中交互物体检测范围受限的问题，本文利用深度归一化提高深度图像质量，提出了一种基于图像分割的任意交互物体检测方法。该方法针对操作人员侧向和正向姿态，分别采用基于显著性检测的图像处理和人体姿态引导的区域生长算法分割目标区域，锚定目标物体边框实现物体检测。最后，进行了交互物体检测实验及不同深度区间位置测距和跟随实验。实验结果表明，所提出的物体检测方法能够实现任意交互物体检测，在交互物体检测方面具有广泛适用性；较小深度区间的归一化能够使物体位置误差变小，提高了物体检测距离精度及机器人跟随效果。

关键词：目标物体检测；深度归一化；图像分割；人机交互
中图分类号：TP242　　**文献标识码**：A　　**文章编号**：1000-1298(2024)08-0428-09　　**OSID**：

Arbitrary Interactive Object Detection Method Based on Deep Normalization

HUANG Lingtao　KONG Zijing　YANG Fan　ZHANG Hongyan
(*School of Mechanical and Aerospace Engineering*, *Jilin University*, *Changchun* 130022, *China*)

Abstract: Detection and recognition of interactive objects is a key technology to realize human-computer interaction, and in order to solve the problem of limited types of interactive objects in the process of human-computer interaction, an arbitrary interactive object detection method was proposed based on image segmentation. Firstly, for the depth image, after filtering out the data outside the range of the original depth data, the min - max scale normalization method was used to improve the quality of the depth image. Secondly, the target area was segmented by using the image processing method based on saliency detection and the human pose-guided region growth algorithm for the operator's side-to-side camera and front-facing camera posture, respectively. Then, the pixel set of the target object obtained by the above segmentation was input into the image processing functions, and the minimum external rectangle of the area point set was obtained, and the rotating bounding box of the target object was anchored. Then, for the depth image, after filtering out the data outside the range of the original depth data, the min - max scale normalization method was used to improve the quality of the depth image. Finally, the detection experiments of arbitrary interactive objects and the ranging and following experiments of different depth intervals were carried out. Experimental results showed that the proposed object detection method had a lower detection cost and a higher degree of freedom in the detection category of interactive objects, which can realize the detection of arbitrary interactive objects, and had wide applicability in the detection of interactive objects. The normalization of the small depth interval can effectively improve the depth image quality, make the object position error smaller, and improve the accuracy of the object detection distance and the following effect of the robot in the human-computer interaction experiment.

Key words: target object detection; depth normalization; image segmentation; human-computer interaction

收稿日期：2024-04-17　修回日期：2024-05-15
基金项目：吉林省重点研发计划项目(20200401130GX)和国家自然科学基金项目(51575219)
作者简介：黄玲涛(1979—)，男，副教授，主要从事机器人控制和主从控制及力反馈技术研究，E-mail：hlt@jlu.edu.cn
通信作者：张红彦(1978—)，女，副教授，主要从事机器人技术和机器视觉处理研究，E-mail：Zhanghy@jlu.edu.cn

0　引言

随着“工业 4.0”[1]的到来，机器人技术得到迅速发展[2]，人机交互技术得到了广泛研究。基于视觉模态的人机交互侧重于两方面研究，首先从图像或视频中根据人和物体的关系对交互物体进行检测，其次根据人及物体动作进行意图识别并与机器人进行交互。目前对于交互物体检测仍存在交互物体种类受限的问题[3]，需要引入新的交互物体检测技术。

针对交互物体检测所用到的深度图像质量提升，国内外学者进行了大量研究。由于深度传感器硬件本身条件限制、采集环境光照限制[4]等原因，得到的深度图像分辨率低，在前景背景边缘遮挡处会产生较多空洞[5]。目前主流的修复方法是传统的图像滤波修复和基于深度卷积神经网络的图像修复[6]。文献[7]提出基于双边滤波实现深度图像修复，文献[8]选择对彩色图像提取边缘辅助滤波修复。随着深度学习的发展，文献[9]选择从彩色图像中提取边界信息，利用对抗生成网络进行深度图像修复，但是此方法无法做到高实时性。综上所述，现有深度图像质量提升都是在生成深度图像后采用各种技术手段修复。

目标检测[10]是计算机视觉的重要研究课题之一，应用领域广泛，常用于人机交互任务，文献[11－12]将目标检测算法应用于交互物体检测。对于目标检测算法发展，相关学者进行了大量研究。文献[13－15]提出了深度卷积神经网络概念，奠定了计算机视觉目标检测领域基础。文献[16]提出了单阶段的 YOLO v1 算法，使用单个神经网络直接从完整图像上预测边界框和类别概率。文献[17]提出了 SSD 算法，提出了区域概念，直接回归物体位置和类别。文献[18]提出的 Mask R－CNN 算法模型，添加了一个用于预测目标类别、边界框坐标和像素级掩膜的 FCN 多任务学习分支。文献[19]提出了 YOLO v7 算法模型，该模型进一步提升了目标检测性能和灵活性。然而，上述算法模型在处理方向角任意、背景复杂、长宽比较大的物体时泛化能力较差[20]。针对对比度低、边缘模糊图像检测问题，文献[21]提出了一种单像素边缘跟踪策略。针对旋转目标检测网络，文献[22]首次将 RPN 结构引入旋转候选框，但计算量仍较大且能检测旋转角度有限。文献[23]提出了一种基于语义注意力和特征金字塔的模型 Mask OBB，文献[24]提出了一个基于 PointRPN 和 PointReg 的航空图像无角度旋转目标检测框架 Point RCNN，实现了任意旋转目标检测，但仍受到训练数据集种类限制。从上述研究可以看出，基于深度学习的目标检测只能针对特定训练过的物体，在人机交互任务中，会存在交互物体检测范围受限的问题。

本文在深度图像生成前，选取合适深度区间对深度数据进行归一化，生成深度图像，结合现有技术进一步修复，以提高深度图像质量和对比度。提出一种基于图像分割的任意交互物体检测方法。该方法针对操作人员侧向和正向姿态，分别采用基于显著性检测的图像处理和人体姿态引导的区域生长算法，分割目标区域，锚定目标边框，以期实现任意交互物体检测。

1　彩色与深度图像配准

为提高系统实时性和物体检测及定位精度，需要将彩色图像与深度图像进行配准。本文使用 Kinect V2 红外单目深度相机，其彩色相机与深度相机位置、视角和拍摄范围存在差异，彩色和深度图像尺寸也不同，因此彩色像素和深度像素不能一一对应。针对该问题，需要建立彩色和深度像素的映射关系。利用 Kinect 官方提供的 MapColorFrameToDepthSpaceUsingIntPtr 函数，可以得到一个与彩色图像像素个数相同的 1 920 × 1 080 矩阵，实现彩色图像中的像素与深度的对应，具体效果如图 1 所示。

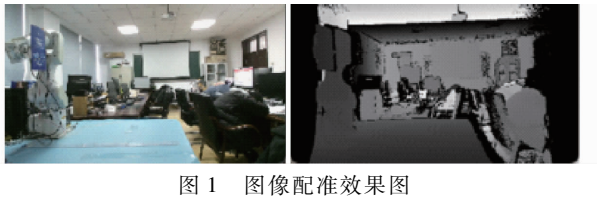


图 1　图像配准效果图

Fig. 1　Image registration

使用上述方法能够实现深度图像与彩色图像的配准，但仍存在大片空洞、无效点及噪点，并且实时性差，获得彩色、深度图像速度 6 ~ 7 f/s，影响系统图像处理实时性。本文通过相机彩色图像与深度图像映射关系，得到像素坐标转换矩阵，生成同深度图像尺寸相同的彩色图像。Kinect V2 彩色相机与深度相机安装在同一平面内，深度相机处于彩色相机 X 轴 5.1 cm 处。设彩色相机坐标系下点 P_c 坐标为 $(x_c, y_c, z_c, 1)$，对应深度相机坐标系下点 P_d，两相机间位姿转换矩阵为 ${}^{d}_{c}\boldsymbol{T}$，则深度相机坐标系坐标与彩色相机坐标系坐标转换关系为

$$\boldsymbol{P}_d = {}^{d}_{c}\boldsymbol{T}\boldsymbol{P}_c = \begin{bmatrix} 1 & 0 & 0 & 0.051 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{P}_c \tag{1}$$

设三维世界坐标系中某点 P_w 坐标为(X_w,Y_w,Z_w,1),该点在彩色像素坐标系中对应像素点 P_c 坐标为(u_c,v_c),在深度图像中对应像素点 P_d 坐标为(u_d,v_d),如图 2 所示。根据彩色像素坐标系—彩色相机坐标系—世界坐标系的转换方式,对应深度像素转换方式为:深度像素坐标系—深度相机坐标系—世界坐标系。彩色相机内参为 $\boldsymbol{K}_c$,外参为${}^c_0\boldsymbol{T}$;深度相机内参为 $\boldsymbol{K}_d$,外参为${}^d_0\boldsymbol{T}$,则有

$$d_c\boldsymbol{P}_c = \boldsymbol{K}_c {}^c_0\boldsymbol{T}\boldsymbol{P}_w \tag{2}$$

$$d_d\boldsymbol{P}_d = \boldsymbol{K}_d {}^d_0\boldsymbol{T}\boldsymbol{P}_w \tag{3}$$

式中 d_c——点 P_w 到彩色相机镜头平面的距离

d_d——点 P_w 到深度相机镜头平面的距离

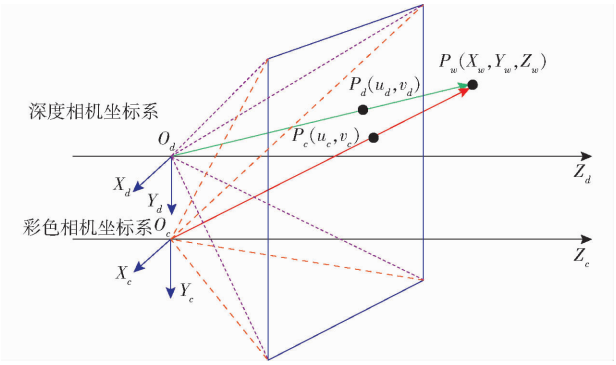


图 2 彩色和深度相机坐标系转换

Fig. 2 Coordinate transformation between color and depth camera

以点 P_w 为中间量,得到两幅图像像素转换矩阵为

$$\boldsymbol{P}_w = d_c {}^c_0\boldsymbol{T}^{-1}\boldsymbol{K}_c^{-1}\boldsymbol{P}_c = d_d {}^d_0\boldsymbol{T}^{-1}\boldsymbol{K}_d^{-1} \tag{4}$$

彩色相机和深度相机在同一平面,因此 $d_c = d_d$。式(4)中${}^d_0\boldsymbol{T}$可通过彩色相机外参及式(1)获得,$\boldsymbol{K}_d$可通过出厂参数获得。因此,可由式(4)得 P_w 的深度像素与 P_w 的彩色像素映射关系为

$$\boldsymbol{P}_c = \boldsymbol{K}_c {}^d_c\boldsymbol{T}^{-1}\boldsymbol{K}_d^{-1}\boldsymbol{P}_d \tag{5}$$

深度像素与对应彩色像素坐标转换矩阵只与$\boldsymbol{K}_c$、$\boldsymbol{K}_d$、${}^d_c\boldsymbol{T}$有关。因为深度图像尺寸较小,所以为完成有效映射,需以深度图像为基准。按上述方法对彩色图像和深度图像进行映射,得到叠加效果如图 3 所示,采用这种方法得到的叠加图具有较好彩色和深度配准效果。因此,在确定像素坐标转换矩阵后,对式(4)只需计算一次并记录,就能明确深度图像与原彩色图像(1 920 像素 ×1 080 像素)的像素位置对应,然后直接读取彩色图像中对应的映射位置,即可生成彩色图像,并且彩色图像的生成速度更快、实时性更好。

2 深度归一化

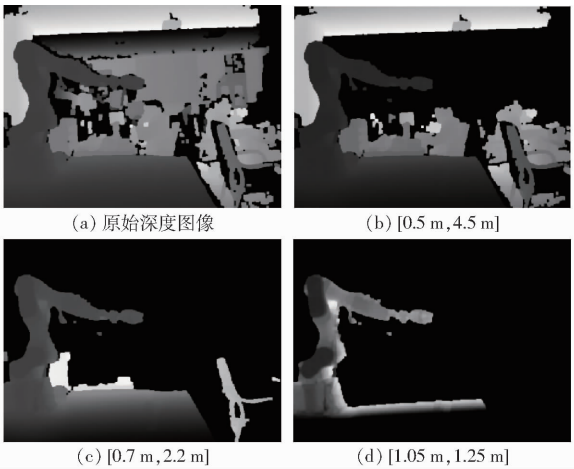
Kinect V2 相机利用发射红外光线并计算其飞行时间的方式进行深度数据采集,因为深度传感器硬件本身条件限制、采集环境、范围限制等原因,得到的深度图像分辨率低,噪声较多,在前景背景边缘遮挡处会产生较多空洞。本文通过归一化方程生成深度图像,并采用中值和联合双边滤波对深度图像进行修复。

图 3 本文的图像配准效果

Fig. 3 Image registration by proposed method

Kinect V2 相机深度传感器有效范围为 0.5 ~ 4.5 m,而生成的深度数据中存在小于 0.5 m 和大于 4.5 m 的数据,为保证可靠性,需滤除超出该范围的数据;结合机器人硬件尺寸和工作范围,保留实际中会涉及到的深度空间区域。对深度数据进行归一化处理时,应滤除超出范围的深度数据,只保留实际使用中所涉及到的深度信息。将剩余数据映射为深度图像后,深度图像上有效数据间的差异会变得更加明显,也有利于提高算法处理图像所得结果的精度。在获得深度图像深度后,需要进行深度像素坐标到世界坐标转换,这是人机交互的一个关键环节。通过归一化方法得到的交互物体深度会更加精确,进而得到更加准确的空间位置,实现对目标物体的高精度作业。

深度图像与深度数据是线性映射,在滤除原始深度数据中超出范围的数据后,采用最小-最大缩放归一化方法进行深度归一化。选取 3 种归一化范围,分别为[0.5 m,4.5 m]、[0.7 m,2.2 m]与[1.05 m,1.25 m],未归一化和归一化后的深度图像如图 4 所示,图 4a 为未归一化深度图像,图 4b 为[0.5 m,4.5 m]区间归一化后深度图像,图 4c 为[0.7 m,2.2 m]区间归一化后深度图像,图 4d 为[1.05 m,1.25 m]区间归一化后深度图像。从图 4 可以看出,不同深度区间的深度图像具有不同的深度信息和清晰度,并且深度图像的清晰度随映射范围的缩小而升高。归一化后不同区间的深度图像单位灰度像素所表示的距离也存在较大差异,如表 1 所示。随着归一化范围的缩小,单位灰度所表示距离也在减小,读取深度精度得到显著提高,物体定位就会越准确。



(a) 原始深度图像　(b) [0.5 m,4.5 m]　(c) [0.7 m,2.2 m]　(d) [1.05 m,1.25 m]

图 4　原始深度图像及归一化后深度图像

Fig. 4　Original depth image and normalized depth images

表 1　不同深度区间精度

Tab. 1　Depth accuracy in different depth intervals

归一化范围/m	单位灰度表示距离/mm
未归一化	17.647 0
[0.5,4.5]	15.686 3
[0.7,2.2]	5.882 4
[1.05,1.25]	0.784 3

3　任意交互物体检测

交互物体检测是实现人机交互的关键技术，基于深度学习的目标检测算法需提前采集数据集进行训练，并只能对已训练物体完成检测。为了实现任意交互物体检测，本文提出一种基于图像分割的任意交互物体检测方法，该方法由图像分割和物体旋转边框锚定两部分组成，图像分割又分为基于显著性检测的图像处理算法和人体姿态引导的区域生长算法，图像分割算法以人体骨骼关键点检测结果为基础。

3.1　人体骨骼关键点检测

基于图像进行人体关键点检测技术是计算机视觉领域中一种重要技术，旨在通过识别和定位图像中人体各关键部位如面部、肢体、手等以帮助计算机理解人体肢体语言。MediaPipe 是 Google 开发的具有轻量级、跨平台特点的机器学习框架，提供了姿态检测、手部检测等解决方案。本文使用 MediaPipe 模型基于彩色图像提取人体骨骼关键点，重点关注手腕、手掌中指关节坐标的检测，其检测结果如图 5 所示。为抑制人手抖动及噪声对检测结果的影响，对检测结果进行低通滤波平滑处理。

3.2　基于显著性检测的图像分割

基于显著性检测的图像分割适用于分割实验人员侧对相机时的任意交互物体。显著性检测是采用模型检测图像中最令人类感兴趣的区域，它在目标

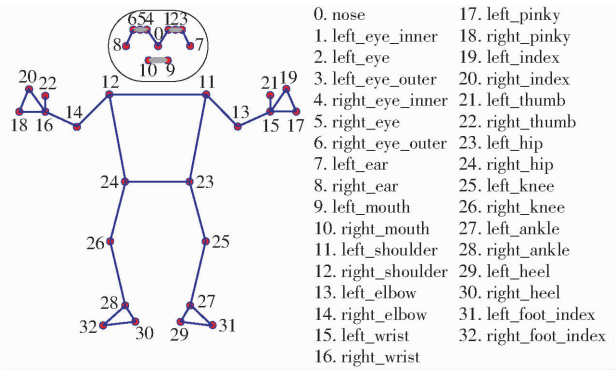


图 5　MediaPipe 检测的人体关键点

Fig. 5　Keypoints of human body by MediaPipe

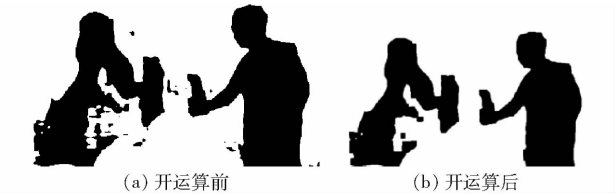
检测和识别、关键点定位、视觉跟踪、语义分割等计算机视觉应用方面发挥着重要的作用。本文选择基于层次深度感知的 RGB－D 显著目标检测网络 HiDAnet，该网络根据深度图像质量分配不同的权值，使显著性检测结果鲁棒性更好。将人机交互场景的彩色图像和对应深度图像输入网络，得到显著性检测结果如图 6 所示，该图中包含人体区域、交互物体区域以及其他背景区域。

(a) 彩色图像　(b) 深度图像　(c) 显著性检测结果

图 6　彩色图像、深度图像和显著性检测结果

Fig. 6　Color image, depth image, and saliency detection

为实现从显著性区域中分割交互物体区域，首先对显著性检测结果进行二值化处理，过滤显著性置信度较低的区域。选取二值化阈值为 0.5，将显著性区域置为黑色，过滤区域置为白色。通过二值化处理虽去除大部分背景噪声，但仍存在较多连续噪声区域，并且边界也不够平滑。为消除小干扰和模糊连接区域，对图像进行开运算。取核大小为 5，进行 2 次开运算，结果如图 7 所示，主要噪声得到消弭，轮廓边界更加平滑。



(a) 开运算前　(b) 开运算后

图 7　图像开运算前后对比

Fig. 7　Comparison image before and after opening operation

为分割交互物体区域，图 7 采用 cv2.findCounters( ) 获取图像所有边界点，根据人体骨骼关键点提取手腕坐标，过滤人体一侧的边界点，计算剩余边界点集中距离手腕坐标最近的点集，该点集即为交互物体所在区域的边界，边界点集筛选

结果如图 8 所示。

图 8 边界点集筛选结果

Fig. 8 Result of boundary point

3.3 人体姿态引导区域生长算法

人体姿态引导区域生长算法适用于分割实验人员正对相机时的任意交互物体。区域生长算法是一种基于区域寻找的传统图像分割算法,该算法简洁、直观,适合具有明显区域特征的图像。

获得人体骨骼关键点检测模型后,将手掌中指关节像素坐标作为算法起始种子点,使用待生长像素点与当前种子点灰度差值作为筛选条件,灰度差值小于阈值作为相似性准则,当区域内没有待生长的像素点时生长停止。结果如图 9 所示,图中红色区域为算法分割得到的物体区域。

图 9 区域生长算法物体分割效果

Fig. 9 Object segmentation by seeded region growing

图 9 仅根据当前种子点与相邻像素点对于生长规则的满足性来分割区域会造成分割区域扩大,需进一步修正。以手掌姿态和手臂范围对算法生长方式进行修正。设手腕坐标为(x_w,y_w),中指关节坐标为(x_m,y_m),则手掌当前角度 θ_h 可表示为

$$\theta_h = |\arctan((y_m - y_w)/(x_m - x_w))| \quad (6)$$

当 θ_h 较小时,根据手腕坐标和中指关节坐标的 x 坐标判断手掌方向,从而决定生长方向是向左还是向右。当 θ_h 较大时,比较手腕坐标和中指关节坐标 y 坐标,从而决定生长方向是向上还是向下。

实际分割时,仅以手掌姿态引导生长会舍弃手腕一侧的所有像素点,在实验人员手中物体朝手腕另一侧倾斜时会产生分割区域不完整的问题。因此,需要设定引导生长范围,手腕坐标一定范围视为手臂范围,在该范围内像素点按照引导方向生长,在此范围外像素点自由生长,从而避免了物体部分分割问题。进行范围限制后的区域生长算法效果与仅人体姿态引导的生长算法效果对比如图 10 所示。

3.4 物体旋转边框锚定

图像目标物体检测过程中,通常用矩形边界框对检测物体进行标记。锚框可在不同位置、尺度上生成多样化的边界框以帮助准确检测出不同目标。借助 OpenCV 库,将分割得到的像素点集输入到 cv2. minAreaRect( ) 函数,获得区域点集的最小外接矩形,函数返回值包含该区域的最小外接矩形数据信息。使用 cv2. drawCounters( ) 函数绘制目标框体,再标识矩形中心点,完成物体旋转边框锚定,锚定结果如图 11 所示。

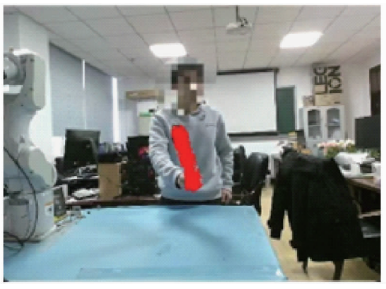


图 10 生长范围限制分割效果对比

Fig. 10 Seeded region growing with range limitation

图 11 交互物体锚定结果

Fig. 11 Object anchoring

在人机交互任务中,物体被操作人员握持在手中,因此必定会有部分物体区域被手掌遮盖,目标检测算法对物体的检测识别会受到手掌不同程度的影响,对于较小的物体,手掌遮盖产生的影响更大。同样,对本文检测方法而言,较小物体在区域分割时更容易受到手掌干扰,使检测结果准确性降低。因此本文方法也只适用于交互物体比操作人员手掌大的物体,对于体积较小物体检测效果不佳。本文检测方法对物体的形状、纹理和颜色关注度较低,更关注于物体深度、显著性特征,无需提前采集数据集进行训练。

4 实验及分析

4.1 实验平台

实验平台如图 12 所示,其包括 EPSON C4 六自由度机器人、Kinect V2 相机、图像处理计算机和机器人控制计算机。图像处理计算机负责处理 Kinect V2 相机的图像信息,进行交互物体目标检测;将处理好的物体坐标传送至机器人控制计算机;机器人控制计算机控制机器人跟随交互物体

运动。平台安置在室内，为避免光照对实验效果的影响，在中午阳光充足且屋内白炽灯全部打开的条件下开展实验。

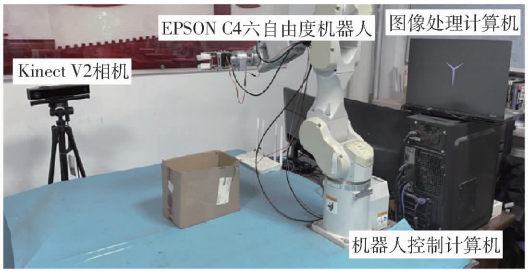


图 12　实验平台

Fig. 12　Experimental platform

4.2　物体检测实验

为了验证本文交互物体检测方法对任意交互物体的检测效果，选用 YOLO v7 目标检测网络与本文的检测方法进行对比实验。实验物体为：白桃苏打水瓶、饼干盒、白色水瓶、白色水杯、罐头瓶、啤酒罐、杯子盒、快递盒、洗发水、卫生纸、巧克力、椰汁瓶，如图 13 所示。因为本文检测方法不用采集数据集进行训练，可直接对物体进行检测识别。但是 YOLO v7 算法需要采集数据集进行训练，为了对比验证 YOLO v7 算法与本文检测方法的交互物体检测范围，YOLO v7 算法只针对白桃苏打水瓶、饼干盒、白色水瓶 3 种物体采集数据并进行训练，其余 9 种物体不进行训练。YOLO v7 算法训练过程：首先，采集白桃苏打水瓶、饼干盒、白色水瓶 3 个物体各 400 幅图像作为原始图像，然后，经过旋转、添加噪声、高斯模糊处理等方式扩增数据集，获得 3 个类别共 5 400 幅图像的数据集，以比例 8∶2把数据集划分为训练集与测试集对网络进行训练，网络设置训练轮数为 300。

图 13　实验物体种类

Fig. 13　Object type

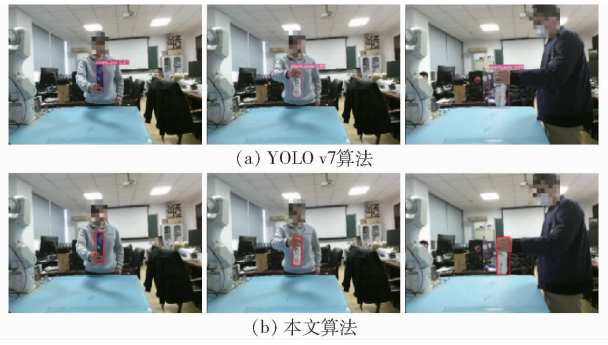
对 12 种实验物体采集 60 幅人机交互场景图像（每种物体的不同交互场景图像采集 5 次），分别使用 YOLO v7 算法和本文算法进行交互物体检测，结果如表 2 所示。

表 2 中可检测次数为算法检测到物体存在的总次数，低置信度次数为物体种类识别错误次数，成功率为可检测次数与总次数比值。由表 2 可知，相较于 YOLO v7 算法，对于未知物体本文所提出的检测方法成功率更高，具有更加广泛的检测能力。对于在训练集中的物体，两种算法都有足够的检测效果，如图 14 所示。由图 14 可知，YOLO v7 算法较本文检测方法还拥有确定物体具体类别功能，语义更清晰。

表 2　算法检测能力对比

Tab. 2　Algorithm detection ability comparison

参数	YOLO v7 算法	本文算法
可检测次数	20	56
低置信度次数	7	
成功率/%	33. 33	93. 33



(a) YOLO v7算法

(b) 本文算法

图 14　两种算法对训练物体检测

Fig. 14　Detection/recognition results for training objects

实验中 YOLO v7 算法出现部分误检测情况与大量无法检测情况，如图 15 所示。YOLO v7 算法在检测有不同图案表面的物体时，存在物体各个表面训练图像不充分导致无法识别的问题；此外 YOLO v7 算法只能识别训练过的物体，对于未经训练的物体无法识别；对于相似物体，YOLO v7 算法存在种类识别错误的问题。这些原因是造成如表 2 所示 YOLO v7 算法物体检测成功率低的原因。

图 15　YOLO v7 算法异常检测

Fig. 15　Anomaly recognition by YOLO v7

本文检测方法对不同物体检测结果如图 16 所示，本文方法无需训练即可完成任意交互物体检测。实验中本文所提的检测方法出现了一些未检测情

图 16 本文算法的物体检测效果

Fig. 16 Object recognition results by proposed method

况,原因是部分交互物体体积较小,算法出现了 4 次误判,因此,本文检测方法适合于比操作人员手掌更大的物体。

与 YOLO v7 等深度学习目标检测算法相比,本文所提出的方法对物体纹理关注度较低,更关注于物体深度、显著性特征,检测代价相对更低,在交互物体的检测范围上自由程度高,对不同任务具有广泛适用性。将该检测方法部署在便携式计算机上,计算机配置信息为:Windows 11 操作系统,Intel Core i7 - 12700H 处理器,系统运行内存为 16 GB,显卡为 NVIDIA GeForce RTX 3060 Laptop,显存大小为 6 GB。测得该检测方法在该便携式计算机上处理实时交互任务检测速度约为 18 f/s,可以满足实时性要求。

4.3 不同深度区间物体测距实验

为分析不同深度区间物体测距的影响,本文对 3 种不同深度区间下物体进行了静态物体测距实验。实验过程为:在实验台上放置 3 个方盒,方盒表面与世界坐标系 X 轴平行,分别选取 3 个盒子表面的一个像素点,在不同深度区间下测量该像素点深度,每个物体在每个深度区间采集 30 组数据,计算其平均值和方差,结果如表 3 所示。所获得的距离为相机表面到物体距离,该距离无法准确测量其数值,因此将深度转换为三维空间坐标。由于选取像素点三维世界坐标人工测量误差较大,因此只选取 X 坐标,转换后物体 X 坐标值与人工测量物体 X 坐标值差值为误差。

表 3 不同深度区间物体距离测量结果

Tab. 3 Object distances in different depth intervals

距离/mm	1 242.01	1 239.85	1 237.26	1 183.59	1 182.35	1 178.51	1 114.69	1 112.37	1 111.77
物体 1 方差/mm^2	0.37	0	0	0.45	0	0	0.39	0	0
物体 1 误差/mm	1.39	3.51	4.06	0.09	3.21	6.0	0.50	2.42	3.42

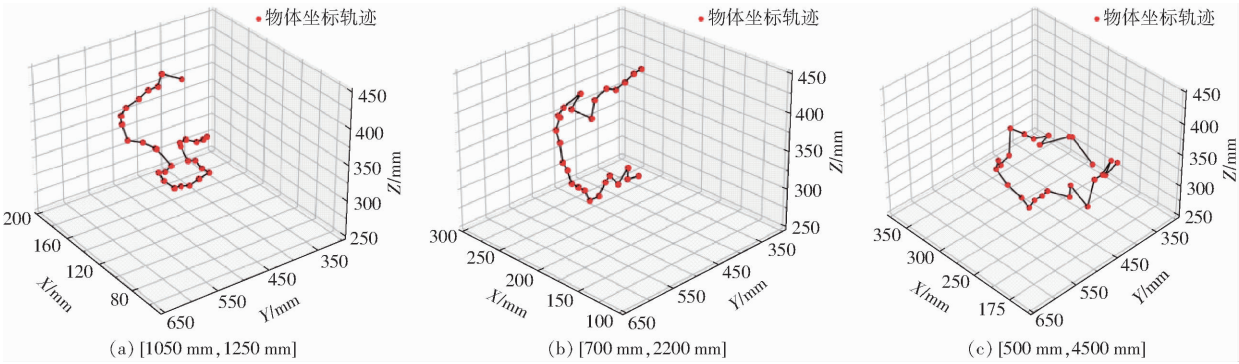
从表 3 可知,随着深度区间变大,物体位置测量精度有较大变化。当深度区间较小时,测得物体距离存在一定方差,而深度区间扩大后方差消失,这是因为深度图像是用深度数据合成,深度图像通过某点深度数据映射到图像像素范围获得某点的深度,该过程存在小数到整数的转换;随着深度区间的扩大,单位灰度所表示的距离也增大,当距离小于单位灰度表示的距离时,该距离无法表示,因此深度区间变小后其方差反而变大。但是,随着物体检测距离精度的提高,其转换后物体位置误差变小,系统物体定位精度得到提高。

4.4 不同深度区间位置跟随实验

为了验证不同深度区间对物体位置跟随能力的影响,本文开展了物体跟随实验。实验人员手持物体在实验区域进行缓慢自由移动,相机实时采集图像,利用本文的检测方法进行物体识别,然后进行坐标变换[25],控制机器人跟随实验人员手中物体的运动。为防止实验人员手抖动及数据处理的噪声影响(测量误差、光照等),当实时物体测算位置与上一保存物体位置差值大于阈值(8 mm)时,机器人末端才跟随物体运动。

选用深度区间[1 050 mm,1 250 mm]、[700 mm,2 200 mm]和[500 mm,4 500 mm]分别进行 5 次实验,选取各区间一次实验数据绘制移动轨迹,如图 17 所示。从图 17 观察到,随着深度区间缩小,检测物体位置连续性更好、更加密集,波动程度也更小。计算各区间每段轨迹的点与点最小、最大和平均间距,再求取各区间所有轨迹的综合平均最小、最大和平均间距,结果如表 4 所示。

深度区间[1 050 mm,1 250 mm]的轨迹最小间距与运动阈值差值为 0.05 mm,深度区间[500 mm,4 500 mm]与运动阈值差值为 5.74 mm,对比可知物体跟随精度随深度区间缩小得到了较大提升,系统可以更快对交互物体位置变化做出判断。而 3 个深度区间的轨迹平均间距随深度区间的增大而增大,交互物体跟随的连贯性随深度区间的增大而下降。可知不同深度区间对位置精度有较大影响,选取合适深度区间有利于提升交互系统跟随性能。



(a) [1050 mm, 1250 mm]　(b) [700 mm, 2200 mm]　(c) [500 mm, 4500 mm]

图 17　不同深度区间的跟随轨迹

Fig. 17　Tracking trajectory in different depth intervals

表 4　跟随轨迹统计

Tab. 4　Tracking trajectory statistics results　mm

区间	最小值	最大值	平均值
[1 050, 1 250]	8. 05	26. 68	16. 27
[700, 2 200]	8. 49	31. 42	20. 70
[500, 4 500]	13. 74	42. 65	25. 79

表 4 中最大值和阈值相差太大的原因有：① Kinect摄像头测量物体深度存在误差，如表 1 所示，通过不同深度区间，可以减小这种误差。②光照对 Kinect 相机的深度信息也有影响，为避免该影响，实验设定在中午阳光充足且屋内白炽灯全部打开的条件下开展。③实验中物体移动速度影响，所使用机器人二次开发程度不高，其 API 运动控制函数约有 0. 1 s 的耗时，这期间手持物体移动速度会对其产生影响。但随深度区间的缩小跟随轨迹点间距最大值也对应变小，因此小深度区间能够有效提高跟随效果。

5　结束语

采用不同深度区间对深度数据进行归一化，提出了一种基于图像分割的任意交互物体检测方法，该方法包括图像分割和物体旋转边框锚定两部分，根据人体与相机的位置分别采用基于显著性检测的图像处理和人体姿态引导的区域生长算法进行图像分割。在搭建的实验平台上，开展了物体检测实验及不同深度区间位置测距和跟随实验。实验结果表明，与 YOLO v7 相比，所提出的物体检测方法检测代价较低，在交互物体的检测类别方面自由程度更高，具有广泛适用性；利用较小深度区间归一化处理深度图像，能够提高系统物体定位精度及机器人跟随效果。

参　考　文　献

[1] 吕铁, 韩娜. 智能制造:全球趋势与中国战略[J]. 人民论坛(学术前沿), 2015(11): 4 - 17.
LÜ Tie, HAN Na. Intelligent manufacturing: global trends and China's strategy[J]. Frontiers, 2015(11): 4 - 17. (in Chinese)

[2] 王天然. 机器人技术的发展[J]. 机器人, 2017, 39(4): 385 - 386.
WANG Tianran. Development of the robot technology[J]. Robot, 2017, 39(4): 385 - 386. (in Chinese)

[3] SUN Z, KE Q, RAHMANI H, et al. Human action recognition from various data modalities: a review[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2023, 45(3): 3200 - 3225.

[4] 肖志刚, 周猛祥, 袁洪波, 等. 光照强度对 Kinect v2 深度数据测量精度的影响[J]. 农业机械学报, 2021, 52(增刊): 108 - 117.
XIAO Zhigang, ZHOU Mengxiang, YUAN Hongbo, et al. The influence of light intensity on the measurement accuracy of Kinect v2 depth data[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(Supp): 108 - 117. (in Chinese)

[5] 吕朝辉, 沈萦华, 李精华. 基于 Kinect 的深度图像修复方法[J]. 吉林大学学报 (工学版), 2016, 46(5): 1697 - 1703.
LÜ Chaohui, SHEN Yinghua, LI Jinghua. Deep image restoration method based on Kinect[J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 46(5): 1697 - 1703. (in Chinese)

[6] 强振平, 何丽波, 陈旭, 等. 深度学习图像修复方法综述[J]. 中国图象图形学报, 2019, 24(3): 447 - 463.
QIANG Zhenping, HE Libo, CHEN Xu, et al. Review of deep learning image restoration methods[J]. Journal of Image and Graphics, 2019, 24(3): 447 - 463. (in Chinese)

[7] ESFAHANIR M, POURREZA H. Kinect depth recovery based on local filters and plane primitives[C] // Integral Methods in Science and Engineering, 2017: 53 - 63.

[8] 苏东, 张艳, 曲承志, 等. 基于彩色图像轮廓的深度图像修复方法[J]. 液晶与显示, 2021, 36(3): 456 - 464.
SU Dong, ZHANG Yan, QU Chengzhi, et al. Deep image restoration method based on color image contour[J]. Chinese Journal of Liquid Crystals and Displays, 2021, 36(3): 456 - 464. (in Chinese)

[9] 刘坤华,王雪辉,谢玉婷,等. Edge-guided GAN:边界信息引导的深度图像修复[J]. 中国图象图形学报, 2021,26(1):186 - 197.
LIU Kunhua, WANG Xuehui, XIE Yuting, et al. Edge-guided GAN: boundary information-guided deep image restoration[J]. Journal of Image and Graphics, 2021, 26(1): 186 - 197. (in Chinese)

[10] 郭庆梅,刘宁波,王中训,等. 基于深度学习的目标检测算法综述[J]. 探测与控制学报,2023,45(6): 10 - 20,6.
GUO Qingmei, LIU Ningbo, WANG Zhongxun, et al. Survey of object detection algorithms based on deep learning[J]. Journal of Detection & Control, 2023, 45(6): 10 - 20,6. (in Chinese)

[11] 贺文涛,黄学宇,李瑶. 基于 Openpose 和 Yolo 的手持物体分析算法[J]. 空军工程大学学报(自然科学版), 2021, 22(6):82 - 89.
HE Wentao, HUANG Xueyu, LI Yao. Handheld object analysis algorithm based on Openpose and Yolo[J]. Journal of Air Force Engineering University (Natural Science Edition), 2021,22(6):82 - 89. (in Chinese)

[12] 吴超,吴绍斌,李子睿,等. 基于人机交互的免锚检测和跟踪系统设计[J]. 兵工学报,2022,43(10):2565 - 2575.
WU Chao, WU Shaobin, LI Zirui et al. Design of anchorless detection and tracking system based on human computer interaction[J]. Acta Armamentarii, 2022, 43(10): 2565 - 2575. (in Chinese)

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks [J]. Advances in Neural Information Processing Systems, 2017,60(6):84 - 90.

[14] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C]// 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009: 248 - 255.

[15] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPP), 2014: 580 - 587.

[16] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the Computer Vision & Pattern Recognition, 2016.

[17] BERG A C, FU C Y, SZEGEDY C, et al. SSD: single shot multibox detector[Z]. 2015

[18] HE K M, GKIOXARI G, DOLLáR P, et al. Mask R - CNN[C]// Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), 2017:22 - 29.

[19] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// Proceedings of the arXiv, 2022.

[20] 王旭,吴艳霞,张雪,等. 计算机视觉下的旋转目标检测研究综述[J]. 计算机科学,2023, 50(8): 79 - 92.
WANG Xu, WU Yanxia, ZHANG Xue, et al. A review of rotating target detection under computer vision[J]. Computer Science, 2023, 50(8): 79 - 92. (in Chinese)

[21] 刘浩,任宏,赵丁选,等. 基于亚像素定位的图像边缘检测策略研究[J]. 农业机械学报,2024,55(2):242 - 248,294.
LIU Hao, REN Hong, ZHAO Dingxuan, et al. Research on image edge detection strategy based on subpixel localization[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024,55(2): 242 - 248,294. (in Chinese)

[22] MA J, SHAO W, YE H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11):3111 - 3122.

[23] WANG J, DING J, GUO H, et al. Mask OBB: a semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images[J]. Remote Sensing, 2019, 11(24): 2930.

[24] ZHOU Q, YU C. Point RCNN: an angle-free framework for rotated object detection[J]. Remote Sensing, 2022, 14(11): 2605.

[25] 黄玲涛,王彬,倪涛,等. 基于 Kinect 的机器人抓取系统研究[J]. 农业机械学报,2019,50(1):390 - 399.
HUANG Lingtao, WANG Bin, NI Tao, et al. Research on Kinect based robot grasping system[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(1): 390 - 399. (in Chinese)