Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2023, 12(3), 181-188
Published Online August 2023 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2023.123021

基于目标识别与定位技术的机械手研究

于佳弘，张骥祥，张　军

天津职业技术师范大学，电子工程学院，天津

收稿日期：2023年6月1日；录用日期：2023年7月28日；发布日期：2023年8月7日

摘　要

针对生活场景中使用机械手进行物体抓取的操作，本文提出了基于Yolov5算法的多目标检测系统，此算法相较于低版本的Yolo算法及R-CNN算法具有计算量小、准确率高的特点。市面中多采用单目相机赋予机械手二维视觉模块，本文在单目相机的基础上，相较于双目相机在获取三维数据时成本高、受环境影响时特征点无法进行匹配易产生误差的问题，结合结构光模块，生成单目结构光系统，更好地赋予机械手获取物体的深度信息的能力，最终得到物体的世界坐标值。通过实验现象表明，获得的平均精度均值为96.5%，定位精度相较于双目系统更为精准，最终能够较好地满足机械手完成对于物体在三维空间中定位与抓取的需要。

关键词

Yolov5，目标识别，单目结构光，定位

Research on Robotic Arms Based on Target Recognition and Positioning Technology

Jiahong Yu, Jixiang Zhang, Jun Zhang

School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin

Received: Jun. 1st, 2023; accepted: Jul. 28th, 2023; published: Aug. 7th, 2023

Abstract

This article proposes a multi-objective detection system based on the Yolov5 algorithm for object grasping using robotic arms in real-life scenarios. Compared to earlier versions of the Yolo algorithm and R-CNN algorithm, this algorithm has the characteristics of low computational complexity and high accuracy. Monocular cameras are often used in the market to endow the manipulator with a two-dimensional vision module. This paper, based on the monocular camera, compared with the binocular camera, which has a high cost in acquiring three-dimensional data, and the

文章引用: 于佳弘, 张骥祥, 张军. 基于目标识别与定位技术的机械手研究[J]. 人工智能与机器人研究, 2023, 12(3): 181-188. DOI: 10.12677/airr.2023.123021

problem that the feature points cannot be matched when affected by the environment, is prone to errors. Combined with the structured light module, a monocular structured light system is generated to better endow the manipulator with the ability to obtain the depth information of the object, and finally obtain the world coordinate value of the object. Through experimental phenomena, it has been shown that the average accuracy obtained is 96.5%, and the positioning accuracy is more accurate than that of the binocular system. Ultimately, it can better meet the needs of the robotic arm for positioning and grasping objects in three-dimensional space.

Keywords

Yolov5, Target Recognition, Monocular Structured Light, Positioning

1. 引言

随着社会的进步与发展，人们对于科技的要求水平逐步增高。在人们日常生活中，物品识别抓取工作已然成为了新型工业、物流分类、无人车间、快速作业等领域必不可少的重要环节[1]。科技不成熟时，工厂普遍使用较为落后的分拣作业方法——利用大量人工进行物品分类堆积。此方法虽然满足了就业需要，但也暴露了效率低、错误率高的弊端[2]。赋予机械手更加智能化的操作成为了当今世界国内外研究的热潮，而对物体的目标识别及定位技术更是智能化的重中之重。依赖于计算机算力的提升，深度学习已经逐步取代了传统的机器学习，识别准确率极大地提升，机器视觉也随之得到了更好地发展。

为了提高机械手对于物体的检测能力并能够准确将物体抓取到特定区域，本文选用 Yolov5 算法进行识别，使用型号为 Astro Pro 的相机搭建单目结构光系统获取深度信息，实现物体的三维空间定位。工作流程为：对单目结构光相机采集到的数据集进行标注后，搭建 PyTorch 环境，使用 Yolov5 模型进行训练，将散斑图与参考平面散斑图使用 SGBM 算法进行图像匹配获取视差图，将视差图转深度图，深度图转点云图，得到物体三维坐标，最终实现抓取。

2. Yolov5 算法原理概述

当今流行的目标识别算法大多分为 one-stage 与 two-stage 两类，two-stage 算法代表是 R-CNN [3]系列，one-stage 算法代表是 Yolo [4]系列。two-stage 算法将两步分别进行，原始图像先经过候选框构成网络，例如 Faster R-CNN 中的 RPN 网络[5]，再对候选框的内容进行分类；one-stage 算法将两步同步进行，输入图像只经过一个网络，生成的结果中同时包含位置与分类信息。two-stage 与 one-stage 相比，虽精确度提高，但运算量增大，速度慢。

Yolov5 是一种端到端的深度学习模型，可以直接从原始图像中检测和定位目标。它使用卷积神经网络(CNN)来学习图像中物体的特征，并使用多尺度预测和网格分割来检测和定位目标。Yolov5 在 Yolov4 [6]的基础上进行了改进，是其工程化的版本。Yolov5 给予使用者了 10 个不同版本的模型，差异点在于网络的深度和宽度，其模型各项参数如图 1 所示。

我们可以将 Yolov5 算法的结构划分为四个部分：输入端、骨干网络(Backbone)、颈部网络(Neck)、预测输出(Prediction)。Yolov5 算法整体结构如图 2 所示。

Model	size (pixels)	mAP^{val} 0.5:0.95	mAP^{val} 0.5	Speed CPU b1 (ms)	Speed v100 b1 (ms)	Speed v100 b32 (ms)	params (M)	FLOPs @640 (B)
YOLOv5n	640	28.0	45.7	45	6.3	0.6	1.9	4.5
YOLOv5s	640	37.4	56.8	98	6.4	0.9	7.2	16.5
YOLOv5m	640	45.4	64.1	224	8.2	1.7	21.2	49.0
YOLOv5l	640	49.0	67.3	430	10.1	2.7	46.5	109.1
YOLOv5x	640	50.7	68.9	766	12.1	4.8	86.7	205.7
YOLOv5n6	1280	36.0	54.4	153	8.1	2.1	3.2	4.6
YOLOv5s6	1280	44.8	63.7	385	8.2	3.6	12.6	16.8
YOLOv5m6	1280	51.3	69.3	887	11.1	6.8	35.7	50.0
YOLOv5l6	1280	53.7	71.3	1784	15.8	10.5	76.8	111.4
YOLOv5x6	1280	55.0	72.7	3136	26.2	19.4	140.7	209.8
+ TTA	1536	55.8	72.7	-	-	-	-	-

Figure 1. Yolov5 model
图 1. Yolov5 模型

Figure 2. Overall structure of Yolov5 algorithm
图 2. Yolov5 算法整体结构

输入端：Yolov5 算法相较于基本的数据增强方法，改进为将 1~4 张图片进行随机裁剪、缩放，随后随机拼凑生成一张图片，增加目标数量，提升训练速度，降低内存需求。新增自适应锚框计算、自适应图片缩放的方法，训练时根据所选择数据集的不同，自动调整所需锚框的长宽值，并将其进行更新，填充灰色，预测时缩减黑边，使得特征提取更加快速。

骨干网络(Backbone)：骨干网络用于提取特征，并不断缩小特征图，设计了 Focus 结构[7]和 CSP 结构。对于 Focus 结构，图片进入 Backbone 前，对图片进行切片操作，在图片中每隔一个像素取一个值，类似于邻近下采样，通过此操作拿到 4 张互补的图片，将 W、H 信息输入通道，即原先图片的 RGB 三通道模式变成了进行拼接后的 12 个通道模式的照片，最后将得到的新图片再经过卷积操作，最终获取到

没有信息缺失情况下的二倍下采样特征图。相较于 Yolov4 中只有在骨干网络中使用了 CSP 结构，Yolov5 则设计了 CSP1_X 结构使用在骨干网络，其外的 CSP2_X 结构使用于颈部网络中，将原输入分成两个分支，分别进行卷积操作使得通道数减半，通过残差结构再次卷积，将两个分支通过融合后进行正态分布，激活后进行 CBS，使得模型学习到更多的特征，增大感受野[8]。

颈部网络(Neck)：颈部网络通常位于骨干网络与 Head 模块之间。图形特征来源于卷积神经网络浅层的特征，如颜色、轮廓、纹理、形状等；语义特征来源于卷积神经网络深层的特征，语义性虽强但却丢失了简单图形。Neck 结构就实现了浅层图形特征和深层语义特征的融合，使特征图的尺度变大，以便获取更加完整的特征。Yolov5 算法中将 SPP 更换成了 SPPF，并在 Pan 结构中加入了 CSP。进一步提升了算法模型对于不同大小图片的检测。

预测输出(Prediction)：head 层为 Detect 模块，由三个 1 * 1 卷积构成。通过升维或降维，增大感受野，增加通道数，获取更多浓缩的特征信息。搜索局部极大值，消除重叠部分，获取最佳边界框。消除 gird 敏感度，通过 CIOU_LOSS 损失函数计算损失值，输出预测结果。

3. 单目结构光

随着机器视觉、自动驾驶、无人车间等颠覆性技术的逐步发展，采用 3D 视觉技术进行物体识别、场景建模等方面的应用越来越多。相机模组部分直接等同于机械手的眼睛，普通的 RGB 相机所拍摄的图片只能简单获取二维平面的信息，仅仅只能从图像语义中得知距离远近，若想要还原真实场景，增加机械手的智能性及准确性，我们还需要获取到物体距离相机距离的真实数据，从而通过结合图像中每个点的像素信息，获取三维空间坐标。在 3D 视觉技术中主流的三大类分别是：结构光(Structured-light)、双目视觉(Stereo)和飞行时间法(TOF)。基于结构光的构成如图 3 所示。

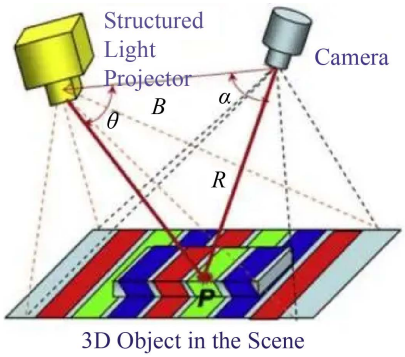

Figure 3. Composition of Structured light
图 3. 结构光的构成

不论是双目算法还是结构光算法，从历史发展角度出发，其基本物理原理都是基于双目三角法。单目结构光法由简单的单目图像处理深入，其发展基于三角测量技术[9] [10]。其处理过程为，利用红外线激光机，将不同特征结构的不可见红外光线透过掩膜投影到空间物体上，再根据物体所处位置的深度不同，通过红外相机收集到不同图像的相位信息，最后通过运算单元讲这种结构的变化换算成深度信息。单目结构光优势在于条件成熟，所需硬件体积小，测量视野较大，并且范围内精确度较高。本文采用散斑结构光的方法对邻域内的散斑分布实现空间编码，如图 4 所示，展示了散斑结构光的数学模型。

其中 object 为物体表面，reference plane 为参考平面，P 点为红外激光发射机位置，C 为 IR 相机位置，CP 为红外激光发射机和相机所处的平面，C 点和 P 点之间的距离 b 为基线长度，设定红外激光发射机和

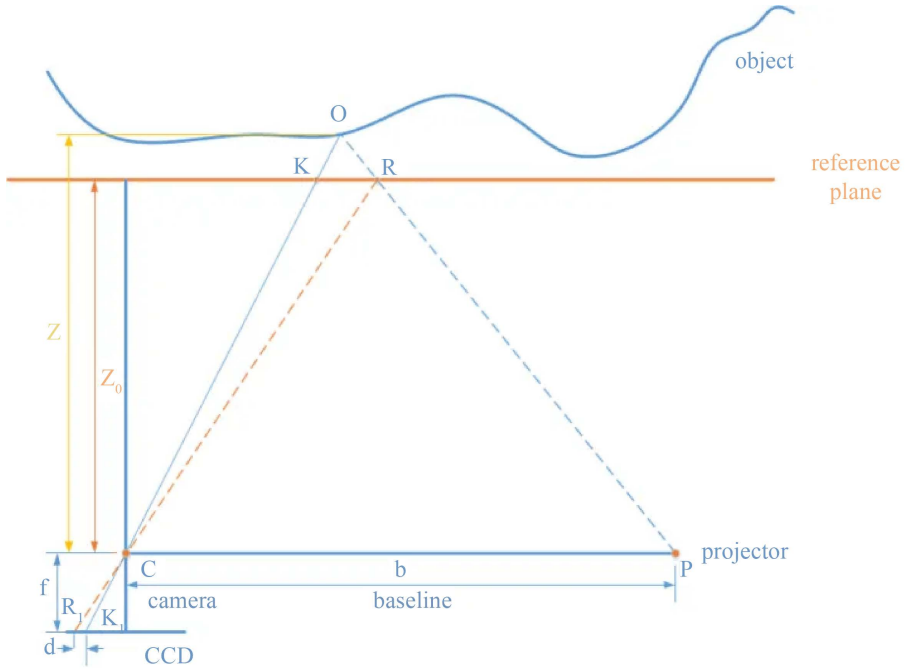

Figure 4. Mathematical model of speckle Structured light
图 4. 散斑结构光的数学模型

IR 相机的连线与参考平面平行。由红外激光发射机透过掩膜投影到被测空间物体的 O 点上，处于不同距离的发射的散斑被 IR 相机所捕获，此时在 IR 相机光心中拥有两个成像点 K_1 、R_1， K_1 为射线 PO 照射到被测物体反射的散斑成像点，与参考平面相交与 K 点，R_1 为发射的同一射线 PO 在不同距离下反射的散斑成像点，与参考平面相交与 R 点， K_1 、R_1 两个不同像素点之间的距离 d 为视差。根据固定可知参数：焦距 f、参考平面与基线平面之间距离 Z_0、基线长度，以及根据匹配算法求得的视差 d，利用三角测量技术可以得到物体上 O 点到相机平面的深度 Z。

运用相似三角形定理，根据△RKC∽△R_1K_1C，可以得到公式：

$$\frac{R_1K_1}{RK}=\frac{f}{Z_0} \tag{1}$$

由△OKR∽△OCP，可以得到公式：

$$\frac{RK}{b}=\frac{Z-Z_0}{Z} \tag{2}$$

联立公式(1)与公式(2)，可以得到深度公式，完成深度测量：

$$Z=\frac{Z_0}{1+\frac{Z_0}{f*b}*d} \tag{3}$$

结构光系统中，系统标定将直接影响测量的准确性。相机标定要完成由像素坐标系－图像坐标系－相机坐标系－世界坐标系的转换，为此，可以利用数学推导公式实现将空间中任意一点的坐标转换为像素坐标，在上述坐标系的基础上，坐标(u, v)就是像素所在的行和列，设点 O(u, v)为像素坐标点，代表像素的行数和列数，在世界坐标系中的任意一点 P(X_w, X_w, Z_w)，在相机坐标系中对应点 P(X_c, Y_c, Z_c)，P 点由像素坐标系转换至世界坐标系的完整推导公式如下：

$$
\begin{aligned}
Z_c\begin{bmatrix}u\\v\\1\end{bmatrix} &= \begin{bmatrix}\frac{1}{dx} & 0 & u_0\\ 0 & \frac{1}{dy} & v_0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}f & 0 & 0 & 0\\ 0 & f & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}R & T\\ 0 & 1\end{bmatrix}\begin{bmatrix}X_w\\Y_w\\Z_w\\1\end{bmatrix}\\
&= \begin{bmatrix}\frac{f}{dx} & 0 & u_0 & 0\\ 0 & \frac{f}{dy} & v_0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}R & T\\ 0 & 1\end{bmatrix}\begin{bmatrix}X_w\\Y_w\\Z_w\\1\end{bmatrix}\\
&= NM\begin{bmatrix}X_w\\Y_w\\Z_w\\1\end{bmatrix}
\end{aligned}
\tag{4}
$$

公式对应参数：f 为焦距，u_0、v_0 表示图像坐标系原点相对像素坐标系原点的偏移量，R 为旋转矩阵，T 为偏移矩阵，N 为相机内参矩阵，M 为外参矩阵，NM 为相机坐标系到世界坐标的转换矩阵[11]。

4. 实验结果及分析

4.1. 目标识别实验

基于 Yolov5 的目标识别需要创建一定数量的数据集进行训练，利用本文所提出的单目结构光深度相机，通过根据物体摆放位置的不同以及相机支架开合角度的不同，对被测物体进行拍摄，选取 1000 张图片制作成为所需数据集，利用标注软件进行信息标注，将标注后的数据集按照 8:2 的训测比分配，训练使用线程数为 10，循环次数设为 100，输入图片大小设置为 640 * 480，选用 Yolov5s 模型，所用 CPU 为 AMD Ryzen 7 5800H with Radeon Graphics，GPU 为 NVIDIA GeForce RTX 3050 Ti Laptop GPU，深度学习框架为 pytorch。

实验统计量由 Accuracy (准确率)、Precision (准确率)和 Recall (召回率)组成。最终实验结果由 mAP (平均精度均值)进行度量，其数学公式为：

$$
\mathrm{mAP} = \int_0^1 P(R)\mathrm{d}(R) \tag{5}
$$

简单来说，mAP 是准确率和召回率构成的 PR 关系曲线的下方面积，越接近于 1，代表模型的效果越好。图 5 为训练获得的 Precision 和 Recall 对比图，可以看出随着训练次数的增加，数值最终归于平稳，并且趋近于 1，得到的 mAP 值达到 96.5%，准确性较好。

4.2. 相机标定

利用张正友标定法对相机进行标定，OpenCV 对于该算法进行了优化，可直接利用 OpenCV 使用棋盘格板进行标定，得到本文所使用的相机参数如表 1 所示。

4.3. 定位技术实验

结构光算法的核心在于将散斑图与参考平面散斑图进行图像匹配，通过匹配代价计算－代价聚合－视差计算－视差优化/后处理四个步骤的技术寻找到像素的视差，使用 OpenCV 库中的 SGBM 半全局匹配

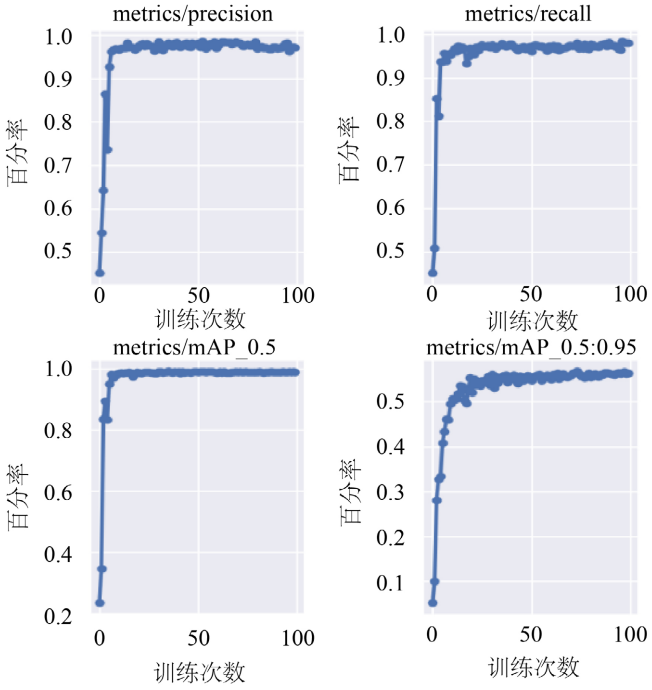

Figure 5. Precision and Recall comparison chart
图 5. Precision 和 Recall 对比图

Table 1. Camera parameters
表 1. 相机参数

	RGB 相机	IR 相机
内参	[[602.005 0. 333.335] [0. 600.856 233.002] [0. 0. 1.]]	[[584.336 0. 323.739] [0. 583.84 240.568] [0. 0. 1.]]
畸变参数	[[0.023 0.498 −0.004 −0. −1.483]]	[[−0.101 0.376 -0.002 0.002 −0.442]]
旋转矩阵 R	[0.9980544085026888, −0.0005053133907268852, −0.06234695122237745; −0.00192747042416588, 0.9992391545866735, −0.03895377772019822; 0.06231919869600642, 0.03899816148599567, 0.9972940694071134]	
偏移矩阵 T	[−0.003415024268825395; −0.01214055388563491; 0.02662079621125524]	

Table 2. Comparison of positioning accuracy between monocular Structured light system and binocular system
表 2. 单目结构光系统和双目系统定位精度对比结果

实际测量距离 (mm)	结构光定位距离 (mm)	结构光定位平均误差 (mm)	双目定位距离 (mm)	双目定位平均误差 (mm)
600	596		593	
700	695		697	
800	794	7.4	795	8
900	890		889	
1000	988		986	

算法最终可以有效地得到 Z 轴上的深度信息。实验时将物体放置于不同位置，使用单目结构光系统和双目系统定位精度对比结果如表 2 所示。结果表明单目结构光系统相较于双目系统的定位平均误差更小，定位效果更好，且结果符合抓取要求。

5. 总结

本文提出的基于目标识别与定位技术的机械手的系统，为机械手准确抓取物体提供了支持，利用 Yolov5 算法提高识别准确率，并在其基础上引入单目结构光系统完成物体的三维定位，最终实现空间物体的识别与抓取，通过实验表明该系统能够较好地满足抓取要求。

参考文献

[1] 杨旭海, 周文皓, 李育峰, 等. 采摘机械臂路径规划算法研究现状综述[J]. 中国农机化学报, 2023, 44(5): 161-169.

[2] 黄贤振, 彭淑萍, 等. 基于双目视觉的目标识别与定位及机械臂的抓取研究[J]. 自动化与仪表, 2022, 37(9): 32-35.

[3] Wang, X.B., Zhu, X.Y. and Yao, M.H. (2021) Target Detection Method Based on Improved Faster RCNN. *Chinese High Technology Letters*, **31**, 489-499.

[4] Joseph, R., Santosh, D., Ross, G., *et al*. (2016) You Only Look Once: Unified, Real-Time Object Detection. *Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, June 2016, 779-788.

[5] Ren, S.Q., He, K.M., Ross, G., *et al*. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[6] 林向会. 基于视频分析的铁路异物侵限检测系统的设计[D]: [硕士学位论文]. 贵州: 贵州大学, 2021.

[7] Sun, Q., Liang, L., Dang, X.H. and Chen, Y. (2022) Deep Learning-Based Dimensional Emotion Recognition Combining the Attention Mechanism and Global Second-Order Feature Representations. *Computers and Electrical Engineering*, **104**, Article 108469. https://doi.org/10.1016/j.compeleceng.2022.108469

[8] 肖志强, 周书民, 汪志成, 等. 结合目标检测与定位的物料抓取研究[J]. 电子技术与软件工程, 2022(5): 160-163.

[9] 李扬铭. 基于结构光的三维测量技术[D]: [硕士学位论文]. 吉林: 长春理工大学, 2022.

[10] 杨柳. 单目结构光三维测量精度优化技术研究[D]: [硕士学位论文]. 南京: 南京航空航天大学, 2017.

[11] 刘敬华. 基于双目视觉的运动工件识别与机器人抓取系统研究[D]: [硕士学位论文]. 山东: 山东科技大学, 2019.