基于 VAE 优化的 YOLO-ResNeXt 二阶段草莓熟度分析方法

田宏伟[1]　徐云龙[1]　杨艳红[1]　刘雪兰[2]　任　艳[1]

[1](苏州大学应用技术学院　江苏 苏州 215325)

[2](江苏农牧科技职业学院农业信息学院　江苏 泰州 225300)

摘　要　草莓作为高价值经济作物,其自动化采摘需要进行目标发现及熟度判断,传统草莓采摘分析方法主要使用色度和大小分析等简单图像处理方法,误报率高。提出二阶段检测网络 YOLO-ResNeXt,并根据互联网图片及产地实拍创建 Strawberry3000 数据集,在此基础上,创新性采用变分自编码器(Variational Auto-Encoder,VAE)进行网络部分结构的快速搜索,该方案效率高且对简单结构搜索起到了较好的效果。经测试,该算法能够有效检测草莓目标并分析草莓熟度,在准确率及召回率等指标上对比通用计算机视觉算法有着很大提高,将有效促进高价值经济作物采摘工作的发展。

关键词　计算机视觉　深度学习　目标检测

中图分类号　TP3　　**文献标志码**　A　　DOI:10.3969/j.issn.1000-386x.2024.10.023

TWO STAGE MATURITY ANALYSIS OF STRAWBERRY BASED ON YOLO-RESNEXT OPTIMIZED BY VAE

Tian Hongwei[1]　Xu Yunlong[1]　Yang Yanhong[1]　Liu Xuelan[2]　Ren Yan[1]

[1](*Applied Technology College of Soochow University*, *Suzhou 215325*, *Jiangsu*, *China*)

[2](*School of Agricultural Information*, *Jiangsu Agri-animal Husbandry Vocational College*, *Taizhou 225300*, *Jiangsu*, *China*)

Abstract　As a high-value economic crop, strawberry's automatic picking requires target detection and maturity judgment. Traditional strawberry picking analysis methods mainly use simple image processing methods such as color and size analysis, which has high false alarm rate. In this paper, a two-stage detection network YOLO-ResNeXt is proposed. The Strawberry3000 dataset was created according to the Internet images and the actual farmland photos. On this basis, this paper innovatively used the variational auto-encoder (VAE) to search the network structure quickly, which had high efficiency and good effect on the simple structure search. According to the test results, the algorithm can effectively detect strawberry target and analyze strawberry maturity. Compared with the traditional computer vision algorithm, the accuracy and recall rate are greatly improved, which will effectively promote the development of high-value economic crop picking.

Keywords　Computer vision　Deep learning　Object detection

0　引　言

农业领域以往面临着信息化发展较差、作业环境露天恶劣、高新自动化设备应用困难的问题。随着物联网技术及第五代移动通信技术的发展,目前农村农业信息化技术已成为当前研究热点,在高价值经济作物领域对于农业信息化提出了更高的要求。在目前人工智能领域应用已日渐成熟的今天,有必要针对目前已有的高价值经济作物自动化采摘领域应用深度学习技术。当前的草莓自动化采摘领域仍然属研究发展初期,存在采摘正确率低、速度慢的问题。因草莓作物较

收稿日期:2021 - 04 - 13。国家自然科学基金项目(61472262)。**田宏伟**,高工,主研领域:计算机视觉,信号处理,嵌入式系统。**徐云龙**,副教授。**杨艳红**,讲师。**刘雪兰**,讲师。**任艳**,副教授。

为脆弱,轻微磕碰会使得草莓作物价格大幅贬值,因此视觉识别的定位准确率是草莓自动化采摘所面临的核心问题。视觉识别方案在农业复杂、随机的生产环境中可以有较高的鲁棒性,在当前国内外的自动化采摘领域研究中有着广泛的研究。日本 Kawamura 等采用五自由度机械臂应用于番茄采摘,采用近红外光检测番茄果实及果实熟度分析,成功率达到 80%,平均采摘用时 5 分钟[1]。国内相关研究起步较晚,且较高程度依赖机械结构实现自动化,如李牧等[2]研制的林木球果采摘机器人,通过采摘爪齿梳聚拢枝条,并梳下果实的方式进行采摘,该方法依赖机械与果树物理特性实现自动化采摘,但对果树本身有较大损伤。近年来,李扬等[3]进行了柑橘采摘的检测定位研究,采用 VGG16 与 SVM 算法实现果实识别和分割定位,机械臂获取定位信息后通过咬合结构末端切断果实果梗,实现柑橘采摘,经实验采摘成功率达 80%,障碍物的成功避障率达到 60%。Jan Bontsema 团队采用遮光棚稳定照明下的彩色相机及 TOF 相机结合的视觉方案进行了多品种水果的检测采摘研究。该方法有效避免了逆光强光照等问题,但该方法显著增大了末端执行机构体积,使得部分果树内藏果实采摘困难[4]。

综上目前的自动化采摘视觉分析方法仍存在有计算资源消耗大、准确率召回率低等问题。本文针对该问题提出 YOLO-ResNeXt 二阶段网络进行草莓目标检测及熟度分析,该网络采用轻量单阶段 YOLO 模型与 ResNeXt34 轻量模型进行异步二阶段检测,并在数据集方面采用互联网图片与产地拍摄相结合进行合理分配,创建 Strawberry3000 数据集,并在数据集基础上应用当前先进的图像增强方法强化网络目标检测性能。经过实际产地拍摄测试图片测试,结果表明,本文算法在指标上优于其他研究算法,在自动化采摘领域有着广阔的应用前景。

1 Strawberry3000 数据集与数据增强

草莓的自动化采摘工作中,进行草莓目标识别方向的研究当前面临着数据杂乱,网络相关图像数据不能与采摘机器人实际工作环境相匹配,拍摄质量参差不齐且过度美化图片等问题。因此本文采用了自建数据集方式,针对目前现有的网络图片使用自动化爬虫进行草莓采摘实景图片爬取,并结合草莓产地实拍模拟采摘近景图片构建了 Strawberry3000 数据集,该数据集有着数据量较大、拍摄真实、贴近农业生产一线的特点。在标注过程中,采用 VOC 格式进行目标检测标注。数据集标注可视化如图 1 所示,数据集包含丰富的草莓目标与不同成熟度,从而有助于训练高水平的监测模型。

图 1 Strawberry3000 部分标注图片

为进一步分析草莓熟度,在目标检测标注的基础上对自然生长草莓进行熟度标注。草莓的成熟过程通常分为四个阶段,即绿熟期、白熟期、转色期和红熟期,其最重要的指标即果面着色程度,依次可将着色面积 0%、25%、50%、75%、全着色作为熟度判断依据[5],本文以此作为标准进行基于深度学习图像分析的草莓熟度分类预测。

研究当前的农业一线生产过程,本文发现因草莓经常于大棚或露天种植,采摘过程中不可避免可能遇到阴雨昏暗天气、强日光照射、逆光等光照变化问题。同时由于草莓本身大小差别与机械平台在复杂作业环境移动等原因,使得草莓目标在相机画面内的实际画面大小差别悬殊。因此有必要通过数据增强手段进行模拟,从而加强模型鲁棒性。本文采用随机亮度增强及马赛克(Mosaic)增强两种增强手段进行数据集增强,以增强模型训练的鲁棒性。马赛克增强方法是缩放并拼接多幅图片及其标签,使得目标大小不确定性放大或缩小,从而人为地变化目标尺寸、扩充图像内大尺寸及小尺寸草莓目标数量[6]。马赛克增强示例如图 2 所示。随机亮度增强则是对数据集的图片进行随机亮度调节,使得同一幅图片呈现不同明暗程度,模拟强光或阴暗天气,增强示例如图 3 所示。通过两种增强方法使得数据集数量得到一定扩充的同时,更多地通过图片干扰形式对实际生产环境中的干扰进行了适应性的调整。

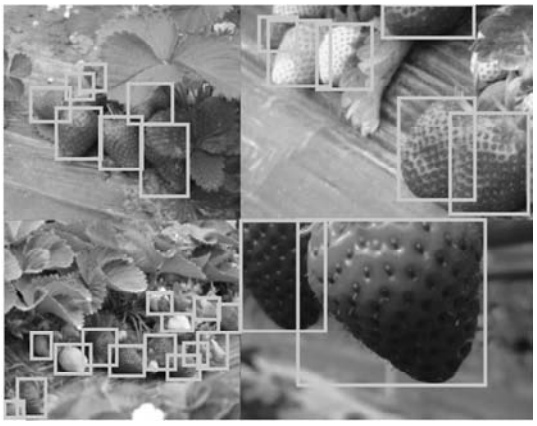

图 2　马赛克增强方法示例

图 3　随机亮度增强对比示例

2　YOLO-ResNeXt 二阶段网络

单一采用 YOLOv4 进行草莓目标检测并进行熟度分析方案存在有小目标区域在特征层特征少、准确率不高的问题。因此本文采用 YOLOv4、ResNeXt34 两个轻量级网络进行二阶段分析,算法通过 YOLOv4 进行草莓目标检测后,将目标区域原图上采样后送入 ResNeXt34 网络进行草莓熟度回归,从而在目标检测基础上更好地分析熟度。图 4 所示为 YOLO-ResNeXt 网络检测过程的流程。

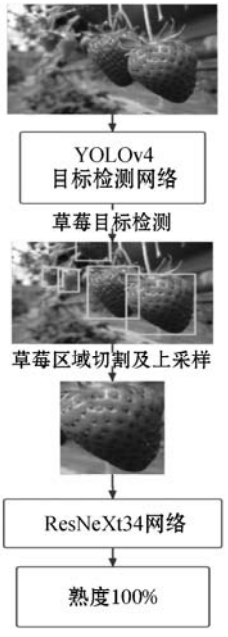

图 4　YOLO-ResNeXt 网络检测流程

图 5 为 YOLOv4 模型的简化结构,网络通过 CSP-Darknet53 网络三阶段特征进行上下采样后输出多层级的目标检测标签与回归框位置。图像采用 608 × 608 大小 RGB 图像作为输入,并在特征图大小为 76 × 76、38 × 38、19 × 19 大小时分别输出特征图进行上下采样并最终得出不同大小目标的输出位置与类别标签[7]。

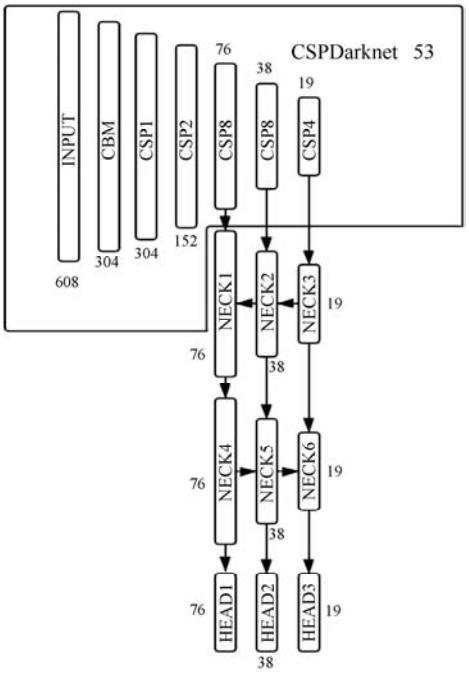

图 5　YOLOv4 网络结构简图

在使用 YOLOv4 进行草莓目标检测时,为解决图像中大量草莓重叠覆盖问题,边框回归采用 CIoU 作为定位损失函数。如式(1)所示,R_{CIoU}为 CIoU 回归损失;b 表示预测框;b_{gt}表示目标框;$\rho(b,b^{\mathrm{gt}})$表示预测框与目标框中心点距离;c 表示目标框对角线长度;w、h 表示分别预测框宽高;w^{gt}、h^{gt}则表示目标框宽高;υ 度量预测框与目标框之间的一致性,并使用 α 作为平衡参数防止 υ 过度干涉 IoU 计算。同时,式(1)解释了 IoU 计算即为 b 与 b_{gt}的交并比[7-9]。在完成草莓目标检测后,将预测草莓目标框输出,并通过图像区域分割及立方插值上采样方式进行图像缩放,以适应 ResNeXt 网络输入[10-12]。

$$\begin{cases} R_{\mathrm{CIoU}} = I_{\mathrm{oU}} - \dfrac{\rho^2(b,b^{\mathrm{gt}})}{c^2} - \alpha\upsilon \\ \upsilon = \dfrac{4}{\pi^2}\left(\arctan\dfrac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}} - \arctan\dfrac{w}{h}\right)^2 \\ \alpha = \dfrac{\upsilon}{(1-I_{\mathrm{oU}})+\upsilon} \\ I_{\mathrm{oU}} = \dfrac{b\cap b^{\mathrm{gt}}}{b\cup b^{\mathrm{gt}}} \end{cases} \tag{1}$$

随后使用 ResNeXt34 进行熟度预测可视为分类

问题,所需数据集由目标检测图像的目标区域提取后重新标注得到。从目标数量统计可得如表 1 所示的 Strawberry3000 标注目标熟度分布。由表 1 可知该数据集中全着色熟度草莓数量比例较大,整体分布较为不均衡,当前主要采用过抽样对低数量类别进行过抽样训练,但该方法仅对低数量类别样本多次抽样,仍易造成模型过拟合,改进的过抽样方法通过添加各类噪声,仍然难以改善模型过拟合现象。因此为在训练层面消除数据不平衡影响,采用多分类 Focal Loss 进行不同类别的损失函数再平衡。

表 1 Strawberry3000 草莓目标熟度分布表

熟度/%	草莓目标数量	总占比/%
0	1 734	10.09
25	2 954	17.20
75	3 370	19.62
100	9 120	53.09

多分类 Focal Loss 公式如式(2)所示。

$$\text{Focal Loss}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{2}$$

式中:α_t为不同分类下的逆类别频率,相当于不同分类下的分类权重;r 表示因类别均衡调节正负类别的超参数;p_t表示模型自身对样本分类预测的置信程度,在多分类 Focal Loss 中将模型输出为该类的类别视为 1,非该类的类别视为 0,此基础上 p_t有如式(3)表述。

$$p_t = \begin{cases} p & y=1 \\ 1-p & \text{其他} \end{cases} \tag{3}$$

式中:y 表示标注标签;p 表示多分类输出下对应该类的输出值。本文所采用的 ResNeXt34 网络的模型参数表如表 2 所示,表 2 中 Conv 模块即为 ResNeXt 的分组残差卷积块。模型最后通过全局平均池化操作提取图片 1 000 维特征后,使用全连接神经网络输出最终特征,并连接至输出层,维度为 4,即四种熟度预测。

表 2 ResNeXt34 模型参数表

模型结构名称	输出张量尺寸	网络层超参
Input	224×224	—
Conv1	112×112	7×7,64,stride 2
		3×3,max pooling,stride 2
Conv2_x	56×56	$\begin{bmatrix}1\times1,128\\3\times3,128,C=32\\1\times1,256\end{bmatrix}\times3$
Conv3_x	28×28	$\begin{bmatrix}1\times1,256\\3\times3,256,C=32\\1\times1,512\end{bmatrix}\times4$

续表 2

模型结构名称	输出张量尺寸	网络层超参
Conv4_x	14×14	$\begin{bmatrix}1\times1,512\\3\times3,512,C=32\\1\times1,1\ 024\end{bmatrix}\times6$
Conv5_x	7×7	$\begin{bmatrix}1\times1,1\ 024\\3\times3,1\ 024,C=32\\1\times1,2\ 048\end{bmatrix}\times3$
Average pooling	1×1×1 000	—
Feature	550	1 000×550+550 Relu Activation
Output	4	SoftMax,Activation Function

同时,本文发现,在普通 CNN 分类器的最后一层全连接网络中,使用 1 000 维输出的全连接神经网络并非最优解,且该层网络参数占全部参数 10% 以上,有着充分的精细调节空间。因此本文创新性提出采用 VAE 内部编码器空间维度的调参方法。

首先,图 6 所示为 ResNeXt34 网络经最后一层池化所输出的一维向量特征,并经过 N 维的全连接网络输出为网络的特征向量。

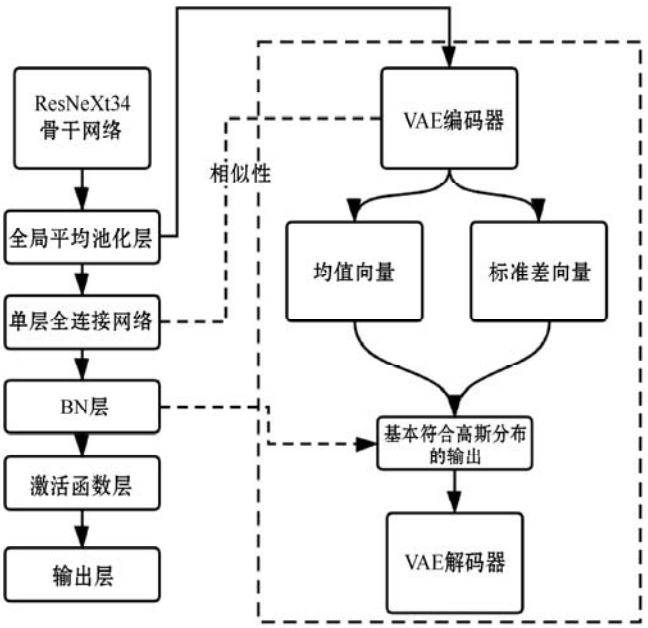

图 6 VAE 中编码器对比含有 BN 层的 ResNeXt 网络输出特征

而全连接网络中的 BN 层则会在训练中逐批次训练 μ 与 σ 值,使得输出尽可能符合正态分布输出,即全连接网络的输出基本能够符合正态分布的输出(激活函数前)。而此实际输出与 VAE 中的编码器将原样本映射至正态分布空间,并进行采样输出中间变量的思路较为接近。因此将 ResNeXt34 中的 N 维全连接网络视为 VAE 编码器存在有较大相似性。因此冻结 ResNeXt34 池化层输出特征,并以此作为 VAE 训练样

本,调节 VAE 编码器内部的维度设置(VAE 解码器为固定 1 000 维),使得 VAE 编码器生成的逐样本的均值与标注差通过固定解码器可获取的特征与池化层输出特征 KL 散度损失最小时,则认为该编码器是将 ResNeXt 池化层输出特征向正态分布空间映射的最佳维度,同理也应当是特征提取的最佳维度。

本文实践中首先使用 1 000D 全连接网络训练 ResNeXt34,随后冻结网络权重,采用 VAE 网络训练方式重建 1 000D 特征,并评估 VAE 不同维度编码器下的损失,本文尝试了 50 ~ 1 100 编码器维度下的 VAE 网络,并统计其重建损失如图 7 所示。

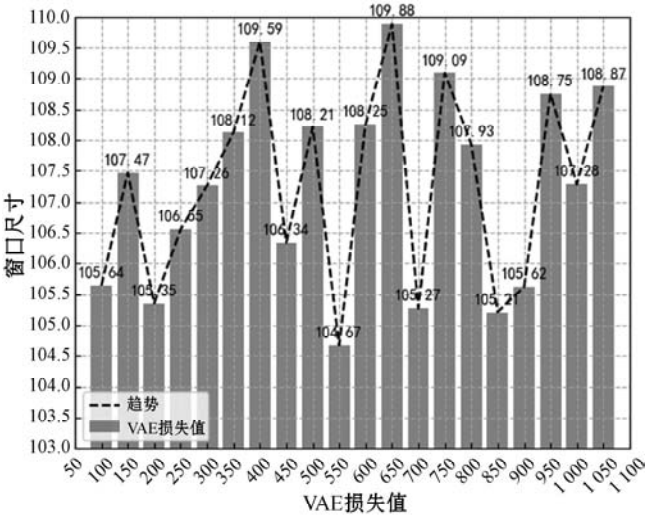

图 7　VAE 编码器训练损失与编码器维度

可以看出,在 VAE 的编码器维度为 550 时,VAE 重建损失最低,同时相比 1 000 维度,该维度可节省约 6.5×10^6 模型参数量及 1.5×10^7 FLOPs 模型计算量,在提升速度的同时进一步降低模型运算量。

本文测试了多个 VAE 编码器维度与 ResNeXt34 全连接特征维度相同情况下 ResNeXt34 网络的表现情况,发现 VAE 编码器维度与全连接特征维度确实存在有强关联性,多轮消融实验结果如图 8 所示,图 8 中分别测试了 400、500、550、650 四个 VAE 编码器表现较为悬殊的维度,以及原始的 1 000 维度作为全连接层神经元维度。可以发现 VAE 编码器效果较好的维度应用于网络全连接层同样取得更佳效果。与图 7 中 VAE 编码器维度表现对照,图 8 中 450 维、550 维 ResNeXt34 特征维度调参同样取得了更好的损失函数表现,其中使用 550 作为全连接层神经元维度的优化结构网络最终取得了测试集 0.153 的 Loss 与 98.26% 的测试集正确率,远超过未优化结构的 ResNeXt34 网络(全连接层维度 1 000D),这一表现和 VAE 编码器各个维度下测试结果基本相匹配。

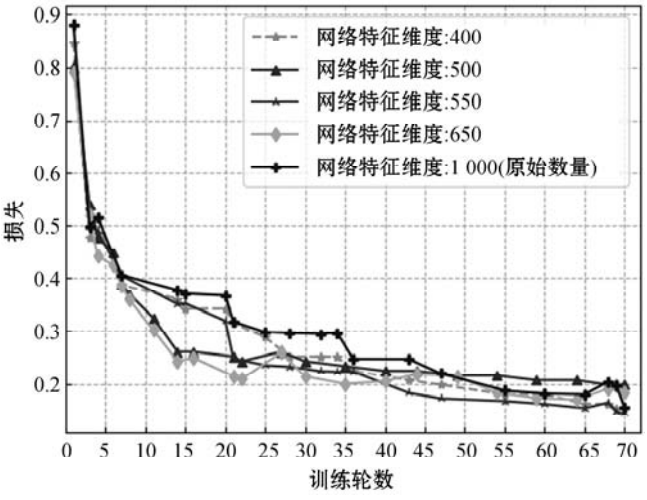

图 8　原始 ResNeXt34 网络与优化网络的测试集损失图

实际模型训练过程中,本文使用 AdamW 优化器,并采用线性 WarmUP 训练策略,训练学习率从 2×10^{-3} 上升至 10^{-2},随后线性下降直至 5×10^{-5} 学习率并保持。

训练集与测试集划分中,YOLOv4 模型与 ResNeXt 网络统一将 80% 图像划分为训练集,20% 图像划分为测试集,YOLO 使用划分图像的 mAP 作为评价指标,ResNeXt 则将训练集测试集中的草莓目标数量下的准确率作为评估标准。

表 3 所示为 YOLOv4-ResNeXt 二阶段网络对比其他算法在 Strawberry3000 数据集上的综合评价指标平均多类平均正确率(mean Average Precision,mAP)表现及单次推理平均耗时。其他算法采用一阶段多目标检测方式进行单次推理获取草莓熟度。通过比较可得出结论,采用 YOLOv4-ResNeXt 在本文自建 Strawberry3000 数据集上在算法精确度指标即耗时上均取得了优异的结果表现,相比 YOLOv4 亦能大幅提高检测正确率。

表 3　多目标检测算法对比横评表

算法类型	mAP	时间/ms
YOLOv3	37.2	78
SSD-ResNet50	29.7	89
YOLOv4	50.2	**51**
YOLOv4-ResNeXt	**56.9**	67

3　应用与部署

基于本文的 YOLO-ResNeXt 算法开发的软件根据自动化采摘实际需求,应用于原生支持深度学习的英伟达 Jetson TX2 嵌入式开发板,从而实现节能且便于采摘机器人轻负载部署的目标[13]。YOLO-ResNeXt 二

阶段算法采用 PyTorch 深度学习框架进行开发以及部署。同时为方便远程监控或即时功能切换,采用 QT 技术进行界面开发,从而支持农业生产中的便捷使用与远程调度。QT 界面如图 9 所示,软件通过 OpenCV 视频流协议软件可支持 RGB 及 YUV 格式远程视频推流或本地摄像头实时识别。

图 9 草莓目标检测与熟度分析系统软件界面

4 结 语

针对当前高价值的草莓作物自动化采摘面临的计算机视觉识别问题,本文对草莓目标检测与熟度分析设计 YOLO-ResNeXt 二阶段网络。同时收集、标注 Strawberry3000 数据集,且面向生产实际光学环境进行相应图像增强。本文通过采用 VAE 特征重建方法进行特征输出层的神经元选择分析,并应用 Focal Loss 解决目标类别不均衡分布问题。通过在 Strawberry3000 上对比实验其他多目标检测算法,本文算法取得了远超其他通用单阶段目标检测算法性能及推理耗时的效果。本文最后构建应用该算法的可视化分析系统软件,可以预见该技术未来将在自动化采摘领域发挥重要作用。

参 考 文 献

[1] 何蓓,刘刚. 果树采摘机器人研究综述[C]//2007 年中国农业工程学会学术年会,2007.

[2] 李牧,陆怀民,方红根,等. 我国农林机器人的研究现状及发展趋势[J]. 森林工程,2003,19(5):39 - 41.

[3] 李扬,杨长辉,胡友呈,等. 基于凸壳及距离变换的重叠柑橘目标识别与定位方法[J]. 现代制造工程,2018(9):82 - 87.

[4] Nguyen T T, Kayacan E, Baedemaeker J D, et al. Task and motion planning for apple harvesting robot[J]. IFAC Proceedings Volumes,2013,46(18):247 - 252.

[5] 刘怡. 不同光质对草莓生理特性及果实品质的影响[D]. 雅安:四川农业大学,2019:1 - 15.

[6] Yun S, Han D, Chun S, et al. CutMix: Regularization strategy to train strong classifiers with localizable features [C]//International Conference on Computer Vision,2019.

[7] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[EB]. arXiv: 2004.10934,2020.

[8] 孙剑云,袁兴,闻平. 基于蓝牙技术的蘑菇智能采摘多运动调度控制系统的研究[J]. 科技创新导报,2019,16(1):142 - 143.

[9] 郑刚,刘佳,李旭. 现代温室采摘机器人发展概况[J]. 农业工程技术,2019,39(31):35 - 40.

[10] 赵德安,吴任迪,刘晓洋,等. 基于 YOLO 深度卷积神经网络的复杂背景下机器人采摘苹果定位[J]. 农业工程学报,2019,35(3):164 - 173.

[11] Liu Y, Hao Y. Research and development in agricultural robotics: A perspective of digital farming[J]. Science of The Total Environment,2018(7):1 - 11.

[12] 杨长辉,王卓,熊龙烨,等. 基于 MaskR-CNN 的复杂背景下柑橘树枝干识别与重建[J]. 农业机械学报,2019,50(8):22 - 30,69.

[13] Joshi K. A review on apple detection methods for harvesting robot[J]. International Journal of Multimedia and Ubiquitous Engineering,2017,12(2):95 - 106.

(上接第 115 页)

[8] 邓带雨,李坚,张真源,等. 基于 EEMD-GRU-MLR 的短期电力负荷预测[J]. 电网技术,2020,44(2):593 - 602.

[9] 李鹏,何帅,韩鹏飞,等. 基于长短期记忆的实时电价条件下智能电网短期负荷预测[J]. 电网技术,2018,42(12):4045 - 4052.

[10] 柯铭,刘凯,赵宏. 基于 LSTM 的滚动预测风机发电量研究[J]. 计算机应用与软件,2020,37(5):67 - 71.

[11] 仰继连. 基于 MMAE 指数的高光谱影像序列微弱变化信息提取[J]. 计算机工程,2016,42(7):261 - 266.

[12] Burnham K P, Anderson D R. Multimodel inference: understanding AIC and BIC in model selection [J]. Sociological Methods & Research,2004,33(2):261 - 304.

[13] Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts [J]. International Journal of Forecasting,2016,32(3):669 - 679.

[14] Zhou M, Wang H, Huo Z. A new prediction model for grain yield in northeast China based on spring north Atlantic oscillation and late-winter Bering sea ice cover[J]. Journal of Meteorological Research,2017,31(2):409 - 419.