安小松,宋竹平,梁千月,等.基于CNN-Transformer的视觉缺陷柑橘分选方法[J].华中农业大学学报,2022,41(4):158-169.
DOI:10.13300/j.cnki.hnlkxb.2022.04.020

基于CNN-Transformer的视觉缺陷柑橘分选方法

安小松[1],宋竹平[1],梁千月[1],杜璇[1],李善军[1,2,3]

1.华中农业大学工学院,武汉 430070; 2.国家柑橘保鲜技术研发专业中心,武汉 430070;
3.农业农村部长江中下游农业装备重点实验室,武汉 430070

摘要 针对产线分拣缺陷柑橘费时费力等问题,以柑橘加工生产线输送机上随机旋转的柑橘果实为研究对象,开发了一种基于卷积神经网络(CNN)的检测算法Mobile-citrus,用于检测和暂时分类缺陷果实,并采用Tracker-citrus跟踪算法来记录其路径上的分类信息,通过跟踪的历史信息识别柑橘的真实类别。结果显示,跟踪精度达到98.4%,分类精度达到92.8%。同时还应用基于Transformer的轨迹预测算法对果实的未来路径进行了预测,平均轨迹预测误差达到最低2.98个像素,可用于指导机器人手臂分选缺陷柑橘。试验结果表明,所提出的基于CNN-Transformer的缺陷柑橘视觉分选系统,可直接应用在柑橘加工生产线上实现快速在线分选。

关键词 柑橘; 缺陷检测; 机器视觉; 深度学习; 卷积神经网络; 在线柑橘分选; 轨迹预测; Transformer
中图分类号 TP391.41 **文献标识码** A **文章编号** 1000-2421(2022)04-0158-12

柑橘是世界上最丰富的水果之一,含有大量有益的次生代谢产物,年产量超过1.24亿t,每年约1/3的柑橘被用于后处理加工[1]。水果的后处理过程一般包括清洗、打蜡、分选、分级、包装、运输和储存,其中柑橘的分选和分级有利于鲜果销售利润空间的增加[2]。目前,虽然已经开发了各种柑橘后处理加工机器[3-4],但柑橘表面缺陷检测主要还是由人工进行,耗时且成本昂贵[2]。开发自动化的分拣系统可以更有效、更经济、更准确地对柑橘进行分类,有利于克服柑橘表面缺陷导致的对消费者购买意愿产生的负面影响,以及大幅度地提高其市场价格。

卷积神经网络(convolutional neural networks,CNN)作为深度学习常用技术之一,在机器视觉中显示出了各种应用潜力,如图像分类、目标检测和图像分割等[5-6]。由于它强大的特征提取能力和自动学习能力,与传统的图像处理方法相比,可以收获更好的识别精度[7-8]。在目标检测领域,它也常被应用于农业各种检测任务中,如田间害虫检测、茶叶品质鉴定、水稻产量预测等[8-9]。

Transformer起源于2017年[10],常被应用于自然语言处理,在机器翻译、文本分类等方向表现极佳[11]。最近也被应用于多目标跟踪[12]和轨迹预测领域,均取得良好效果。相比循环神经网络(RNN),Transformer因其并行运算可以避免递归,从而减少由于长期依赖而导致的性能下降[13]。

目前,有关柑橘缺陷检测的研究报道不多。章海亮等[14]使用高光谱成像(HSI)技术检测缺陷柑橘,通过主成分分析提取特征波长实现缺陷柑橘的分类,识别率达到94%,但该方法存在实时应用困难且设备成本较高等问题。龚中良等[15]通过柑橘的颜色特征差异使用边缘分割实现缺陷柑橘的分类,分类识别率达到92%,但这种基于RGB的判别方法受环境、光照等影响较大。针对产线缺陷分选,李善军等[16]提出了一种基于改进SSD深度学习模型的缺陷柑橘实时检测方法,该方法精度达到87.89%,但该方法无法跟踪分类的历史信息,从而无法判断最终柑橘的真实类别。

本研究基于CNN强大的特征提取能力和Transformer的时间序列处理能力,设计了一种基于CNN与Transformer相结合的视觉系统;通过结合检测器和跟踪器提出了一种新的基于检测的跟踪分类策略,检测器检测柑橘的缺陷表面,而跟踪器则沿着它

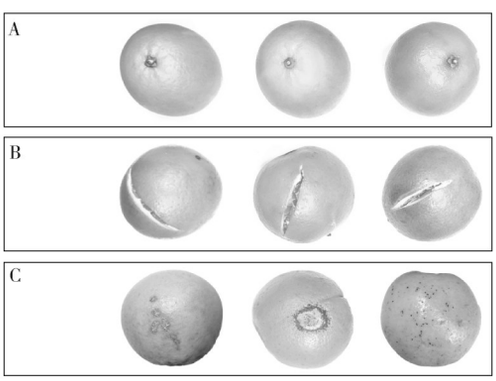
收稿日期: 2021-12-02
基金项目:财政部和农业农村部:国家现代农业产业技术体系、柑橘全程机械化科研基地建设项目(农计发[2017]19号);湖北省农业科技创新行动项目;国家重点研发计划(2020YFD1000101)
安小松,E-mail:279560176@qq.com
通信作者: 李善军,E-mail:shanjunlee@163.com

们的路径记住它们的分类信息，通过历史信息识别它们的真实类别；同时利用轨迹预测算法对缺陷果实的未来路径进行预测，旨在实现缺陷柑橘的分拣并快速地实现在线柑橘分类。

1　材料与方法

1.1　样本与系统配置

样品柑橘为蜜萘夏橙（Midknight Valencia Orange），来源于宜昌市秭归县，其特点是糖酸比适中，在成熟阶段表皮颜色处于绿色到黄色之间。首先通过人工筛选将柑橘分为3类，分别为正常（N，指表面没有任何缺陷并准备好进入新鲜水果市场的柑橘，图1A）、机械损伤（MD，指在收获或收获后处理过程中由于处理不当而受到机械损伤的样本，图1B）和表皮病变（SL，对于被真菌或昆虫感染的果实，表面呈现缺陷的柑橘被归类为SL类，图1C）。随机抽取300个柑橘用于数据收集，其中N、MD、SL类均为100个。



A：正常 Normal；B：机械损伤柑橘 Mechanical damage；C：表皮病变柑橘 Epidermal lesions.

图1　3类柑橘示意图

Fig.1　Examples of 3 types of citrus

在实验室中组装了1条市面上可用的柑橘加工生产线（GJDLX-5），该生产线装备有自动柑橘清洗和打蜡设备（图2A）。传统的柑橘加工生产线上，输送机自由旋转柑橘，以便分拣员检查每个柑橘的所有表面，挑选出缺陷的柑橘；然后将健康的果实输送到清洗机和打蜡机进行再加工。

自动化分拣过程使用网络摄像头（GuceeHD98）安装在输送机上方0.5 m，用于实时传输柑橘视频图像到分类系统，以便检测和跟踪有缺陷的柑橘；网络摄像头的图像分辨率为640像素×480像素，每秒30帧（FPS）；同时使用100 W的LED灯在工作空间内增强和平衡照明条件。

视觉系统包括缺陷柑橘的检测、跟踪和轨迹预测共3步。第1步，传送带不间断地旋转柑橘，让摄像头查看柑橘的不同表面；为检测有缺陷的柑橘，开发了一种基于单阶段神经网络的Mobile-Citrus检测器，用于检测柑橘果实并将其暂时分类为相应的类别。第2步，采用自定义Transformer-One-Step的实时跟踪算法用于跟踪缺陷柑橘，通过存储的历史信息识别柑橘的真实类别。第3步，利用Transformer-Multi-Step轨迹预测算法将未来路径与类别发送到中央控制系统。以上步骤已通过PC端实现，在未来将会通过PC端的计算结果引导机器人手臂分拣缺陷柑橘，实现真正的产业化，如图2B所示。

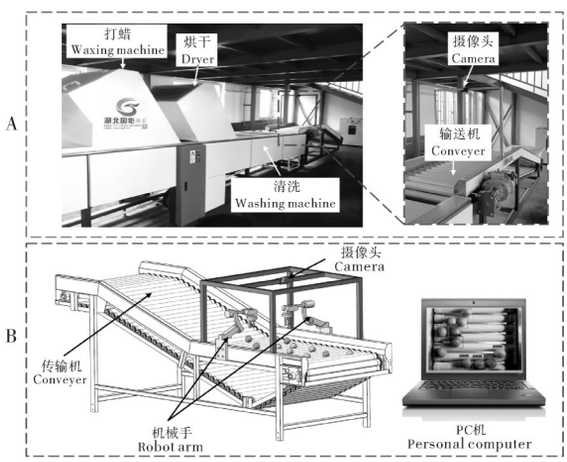


图2　平台设置和计算机视觉系统

Fig.2　Platform setup and computer vision system

在图像采集过程中，柑橘被放置在传送带上，传送速度为0.3 m/s。将300个柑橘放在传送带上拍摄视频，共拍摄6次，每次拍摄之前随机打乱。即共收集了6个帧率为30 FPS的视频，每个视频持续时间为60～70 s。在这些视频中，其中5个被用于所开发的检测器。为避免相邻帧之间的信息重叠，每个视频序列间隔10帧抽取1帧图像，即每秒取3帧图像，经过人工筛选后，总共收集2 400张图像；随机选择1 200幅图像作为网络训练数据，另外500幅图像作为验证数据，其余700幅图像作为测试数据。使用LabelImg工具对采集到的图像进行手动标记，只有当表面损坏或破坏性伤口被捕获时，柑橘被标记为MD或SL类。剩余1个视频用于评估跟踪器的性能。由于跟踪并不需要训练，只需制作一个跟踪相关的测试集即可，最终跟踪测试数据集由1 800个连续序列构成。由于轨迹预测与检测跟踪没有联系，录制的6个视频都可用于轨迹预测，最终收集了8 074个连续序列作为轨迹预测的训练数据，2 525个连续序列作为轨迹预测的测试数据。

1.2 缺陷柑橘检测

提出的检测网络Mobile-Citrus基于YOLOv4[17]架构。轻量级分类网络MobileNet-V2[18]被用于替换YOLOv4中CSPDarknet53主干网络，用于从输入图像中提取多尺度特征图。之后，使用路径-聚合网络（PANet）[19]作为检测分支，从特征图中聚合多尺度信息到P5。最后将检测分支的输出遵循YOLOv4网络设计对YOLO head进行解码获取柑橘的类别信息、置信度、位置信息（图3）。

与原MobileNet-V2模型不同，输入图像尺寸从（224，224，3）变为（416，416，3），只使用了原模型中前面18层网络架构，即18个深度可分离模块（depth-wise dodule），每个模块具有倒残差（inverted residual）结构。最后将第7层C3、第14层C4、第18层C5模块的特征图作为检测分支的输入，以执行缺陷柑橘的检测。

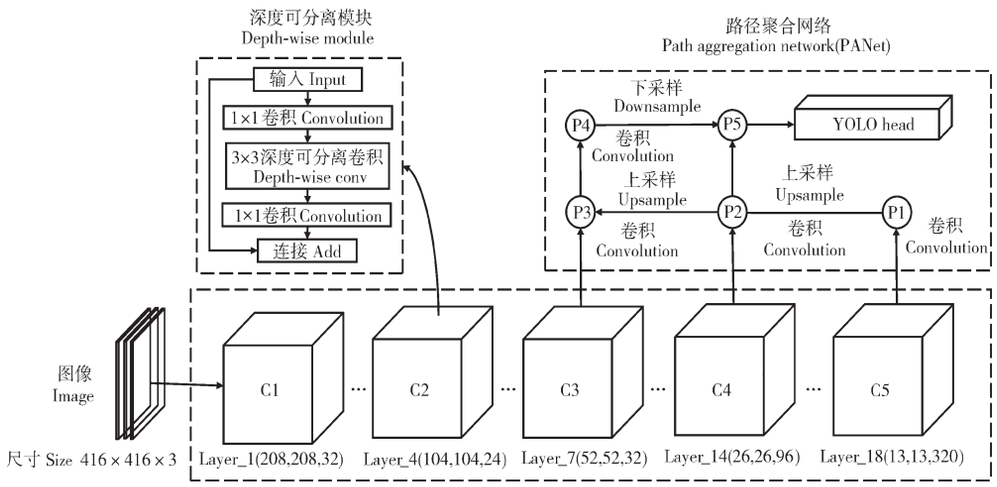


图3 检测器的网络结构

Fig.3 Network architecture of the detector

Mobile-citrus的检测分支从MobileNet网络接收C3、C4和C5的特征图，特征图随后遵循PANet的特定路径到达P5。如C5特征图通过P1路径执行卷积操作（Conv）和上采样（Upsample）操作得到与C4大小一致的特征图，该特征图是P2路径的输入之一。P2路径的另一个输入是C4经过卷积操作后得到的特征图，即P2将会聚合C4特征图与C5特征图信息。同理，P5路径将会聚合由P4经过下采样（Downsample）得到的特征图与P2输出的特征图信息，即P5将会聚合C3、C4、C5模块所有特征。与YOLOv4不同，Mobile-citrus只聚合了1个YOLO head，最终只会对1个YOLO head进行解码获取柑橘的类别信息、置信度、位置信息。

在训练过程中应用多种图像增强方法，让模型学习更多有效特征，使模型鲁棒性更优。包括缩放（0.8～1.2）、翻转（水平和垂直方向）、旋转（±20°）以及HSV颜色空间中饱和度（0.8～1.2）和亮度（0.8～1.2）的调整。网络训练时，Adam优化器用于梯度下降，批次大小为24，训练图像分辨率为416像素×416像素。在训练过程中，冻结骨干网络内的权重，只对检测分支进行了训练。网络训练前80个迭代的学习率为0.001，后40个迭代的学习率为0.000 1。

1.3 缺陷柑橘跟踪及轨迹预测

由于加工生产线上柑橘在做实时运动，机器手需要一定的抓取时间，为了更准确地抓取，可以将未来帧的信息输入到执行器，实现动态抓取。这里采用了基于Transformer算法的预测器对未来的轨迹进行预测。

另外，针对柑橘滚动时会呈现不同的特征面，检测模型难以确定同一柑橘的真实类别的问题，设计一套跟踪算法，可以在不同的特征面中对相同的柑橘进行归类标记。为了获得更好的检测精度，提出了基于Transformer算法的实时多目标跟踪器，用于跟踪和记录工作空间内每个柑橘在其路径上的分类信息。然后，视觉系统可以根据历史分类信息确定每个柑橘的真实类别。

1）Transformer算法。Transformer的特点是编码器和解码器之间的相互注意，以及编码器和解码器内部的自我注意。自我注意的主要优点是它能够将输入序列的任何2个位置联系起来，而不管它们的距离如何，从而允许在广泛的任务上性能显著提高。

与原始架构有所不同，提出的 Transformer 架构主要分为 3 个模块，分别是输入模块（input block）、编码器模块（encoder block）以及输出模块（output block），如图 4 所示。

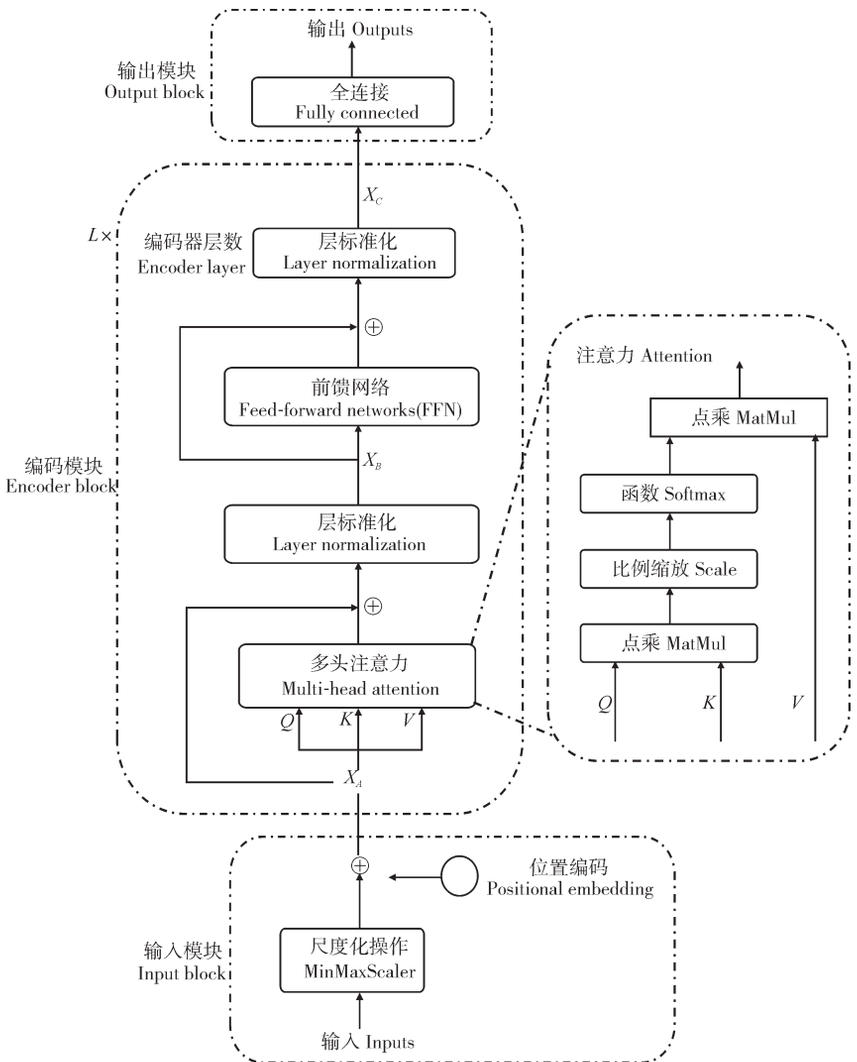


图 4　Transformer 网络架构

Fig.4　Network architecture of the Transformer

输入模块由尺度化（MinMaxScaler）操作和位置编码（Positional embedding）操作组成，MinMaxScaler 操作类似于归一化，具有数据缩放的能力。与归一化的效果相似，可以防止因输入很大使得反向传播时输入层的梯度太大，从而造成越过最优解的情况出现。对于本研究，像素数据被缩放到 $\{-15,15\}$。MinMaxScaler 运算如下：

$$x_{\text{std}}=\left(x-x_{\min}\right)\Big/\left(x_{\max}-x_{\min}\right) \tag{1}$$

$$x_{\text{scaled}}=x_{\text{std}}(\max-\min)+\min \tag{2}$$

假设 $x=\{x^{(1)},\cdots,x^{(n)}\}\in R^{n\times d}$ 由 d 维的 n 个序列组成。上述运算可以将输入值 x 从范围 $\{x_{\min},x_{\max}\}$ 缩放到 $\{\min,\max\}$ 中，其中 x 表示输入值，$x_{\min}$ 表示所有输入值中最小值，$x_{\max}$ 表示所有输入值中最大值，x_{scaled} 为缩放后的值。

Positional embedding 操作可以让特征与特征之间具有相对位置关系，使得模型的学习更加容易。Positional embedding 运算如下：

$$\text{PE}_{(\text{pos},2i)}=\sin(\text{pos}/10000^{2i/d}) \tag{3}$$

$$\text{PE}_{(\text{pos},2i+1)}=\cos(\text{pos}/10000^{2i/d}) \tag{4}$$

其中，pos 代表序列长度所处位置，i 代表每个序列对应的特征维度，如偶数位置的特征将会进行 sin 计算，奇数位置的特征会进行 cos 计算。

编码器模块的输入是输入模块的输出，用 X_A 表示，运算公式如下：

$$X_A=x_{\text{scaled}}+\text{PE}_{(x_{\text{scaled}})} \tag{5}$$

编码器模块由 L 个编码层组成，每个编码层可以

划分为多头注意力模块(multi-head attention)和前馈网络模块(FFN),每个编码器中都有残差连接和分层归一化(LayerNorm)。

注意力模块会通过给定的3个矩阵Q、K、V进行注意力得分运算,如第i个Q与第j个K的注意力得分越大,那么第j个值对第i个值的影响就越大。Q、K、V是输入模块的输出与对应权重矩阵相乘的结果,可以表示为:

$$Q=X_A*W_Q,\ K=X_A*W_K,\ V=X_A*W_V \tag{6}$$

这里的W_Q、W_K、$W_V\in R^{n\times d}$,Q、K、$V\in R^{n\times d}$,将通过梯度下降来更新W_Q、W_K、W_V,最终获得1个合适的权重来拟合真值。

为了获得特征位置上的概率分布,注意力计算使用了Softmax函数;同时为了防止当d值较大时,Q与K^T的点积会变大,从而Softmax值趋于0,因此引入比例因子$\sqrt{d}$。注意力(Attention)的计算如下:

$$\text{Attention}(Q,K,V)=\text{Softmax}(\frac{QK^T}{\sqrt{d}})V \tag{7}$$

为了在多方面、多空间上关注特征之间的注意力,提出了多头注意力(MultiHead)机制,它通过联合多个独立的注意力实例共同决定最终注意力,即将不同空间的头部(head)输出简单连接了起来[10]。多头注意力(MultiHead)计算如下:

$$\text{MultiHead}(Q,K,V)=\text{Concat}(\text{head1};\cdots;\text{head}_h)W^O \tag{8}$$

$$\text{head}_i=\text{Softmax}(\frac{QW_i^Q(KW_i^K)^T}{\sqrt{d_k}})VW_i^V \tag{9}$$

其中,W_i^Q、W_i^K、W_i^V分别是将Q、K和V投影到第i个子空间的矩阵;W^O是计算头部(head)线性变换的矩阵。通常,$d_k=d/h$,其中,h是多头注意力中的head数目,意味着从h个空间去关注特征。

编码器模块中的残差结构可以理解为初始输入与对初始输入进行操作后的结果的叠加处理。随后将残差连接后的结果进行分层归一化(LayerNorm),LayerNorm操作可以一定程度上防止模型的过拟合,以及加快模型的收敛速度[20]。具体操作如下:

$$X_B=\text{LayerNorm}(\text{Attention}(X_A)+X_A) \tag{8}$$

$$X_C=\text{LayerNorm}(\text{FFN}(X_B)+X_B) \tag{9}$$

$$\text{LayerNorm}(x)=g\odot\frac{x-\mu}{\sqrt{\delta^2+\epsilon}}+b \tag{10}$$

$$\mu=\frac{1}{d}\sum_{i=1}^{d}x_i,\ \delta^2=\frac{1}{d}\sum_{i=1}^{d}(x_i-\mu)^2 \tag{11}$$

$$\text{FFN}(x)=\max(0,xW_1+b_1)W_2+b_2 \tag{12}$$

LayerNorm计算中$\odot$表示逐元素乘积,增益g与偏置b是模型训练过程中可学习的参数,FFN中的权重W与偏置b同理,μ和δ^2分别是根据输入x计算的均值与方差,ϵ是一个固定的小常数,可以防止根号下为0的情况出现。

最后输出模块是将编码器模块的输出作为输入,使用全连接(fully connected)计算输出最终的结果。

2)基于Transformer的多柑橘跟踪。本研究提出的跟踪算法实现需要3个步骤,分别为CNN检测当前帧图像,Transformer根据过去柑橘序列位置信息预测当前帧柑橘位置信息,然后将当前帧检测到的柑橘位置信息与Transformer预测到的当前帧柑橘位置信息做数据关联(data association),如图5所示。这里使用Transformer-One-Step模型用于跟踪,可以根据上1帧的柑橘位置信息预测下1帧的柑橘位置信息,即输入与输出是一对一的,因为在没有跟踪之前,如果想要知道该柑橘的历史轨迹,必须做到跟踪,让柑橘对象一一匹配。因为在未做跟踪之前无法获取某个柑橘对象的多个历史轨迹,所以采用了输入与输出为一对一的模式。每帧柑橘图像都会进入CNN模型做进行目标检测,以获取单帧图像的所有柑橘的位置信息和类别信息,同时Transformer-One-Step会根据CNN检测器检测出的上1帧柑橘位置信息预测出当前帧柑橘位置信息,最后将检测器检测到的柑橘与Transformer-One-Step预测器预测的柑橘进行数据匹配,从而实现不同帧对象与对象之间的关联。

如果新检测到的柑橘与现有的跟踪柑橘相匹配,则使用新检测柑橘的包围框来更新现有柑橘的状态,并基于Transformer模型预测下一帧中柑橘的包围框。通过交并比(IOU)计算预测的包围框与新检测包围框之间的相似性,利用匈牙利算法(Hungray algorithm)[21]进行数据关联,当匹配的包围框之间的交并比(IOU)比最小交并比(IOU_{min})低时,则匹配失败,此类柑橘将会放入下一帧继续匹配。最小交并比(IOU_{min})设置为0.3。

3)基于Transformer的缺陷柑橘轨迹预测。由于本研究提出的轨迹预测方案是基于跟踪算法的,因此,对于每个柑橘都会分配1个Tracker-citrus跟踪器,该跟踪器会记录柑橘的历史信息,该历史信息会以信息栈的形式更新,同时信息栈会一直维持最新的40帧历史信息。当信息栈储存到40帧柑橘历史信息时,开始执行轨迹预测。具体预测方法是以连续40帧图像信息作为输入,经过CNN目标检测后,将连续40

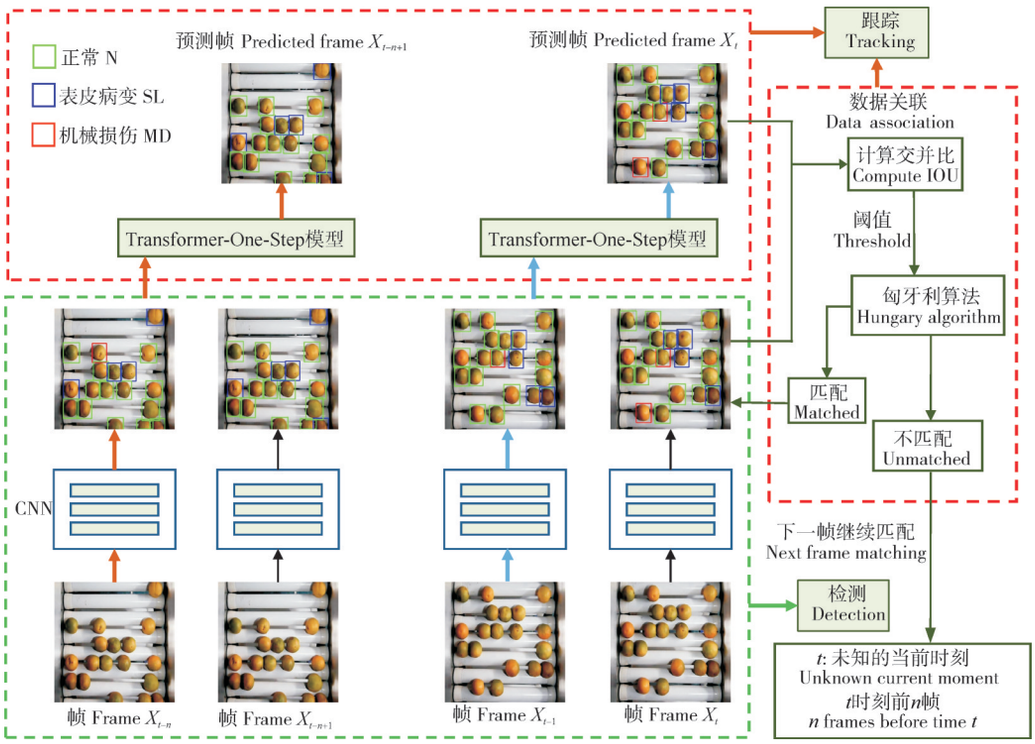


图 5　多目标跟踪工作流程

Fig.5　Workflow of the multi-object tracking

帧的柑橘信息先做跟踪，获取某个柑橘连续 40 帧运动轨迹，然后将过去连续 40 帧运动轨迹输入到基于 Transformer-Multi-Step 轨迹预测器中输出未来 40 帧的轨迹，最后对每个柑橘分别做未来帧的轨迹预测。对于连续的 40 帧未来信息，可以根据机械手抓取柑橘所需的时间动态输出 0～40 帧的柑橘位置（图 6）。

4）从跟踪中分类。在分类过程中，检测器检测工作空间内的所有柑橘，并在每个图像中暂时将它们分为 N、SL、MD 类。然而，缺陷柑橘（SL＋MD）在旋转中，向摄像机呈现其正常（N）表面时，将存在识别误差。在这里，提出的跟踪器可以将检测器的结果记录到一个历史记录表，然后应用一个逻辑来检查每个柑橘的历史列表并确定其真实类别。将相邻 5 帧的图像划分为历史列表，该历史列表一直维持最新的 5 帧，如果历史列表中存在 2 帧及其以上的缺陷柑橘，则当前帧的柑橘将被归类为缺陷，而不去关注当前帧检测结果。如果它们还没有被归类为缺陷，则将被标记为正常。这样的策略可以消除一些随机识别误差，提高检测精度。

如图 7 所示，缺陷柑橘在旋转中，从进入摄像机视野到离开摄像机视野，大约捕获 140～150 帧，期间柑橘大约旋转 540°。可以看出，虽然在 70～85 帧中，柑橘呈现正常表面，但依旧会将其归类为缺陷。

2　结果与分析

2.1　检测性能评估

对于检测性能评估，检测器单独工作在一个单一的图像上，而不考虑在分类过程中连续跟踪的问题。使用 F_1 值衡量检测的整体性能。

$$F_1=\frac{2\times R\times P}{R+P} \tag{13}$$

R（recall）指的是被预测为正例的样本占总的正例样本的比重，P（precision）代表被分类器判定正例中的正样本的比重。同时还引入了推理时间（inference time）来表示检测 1 张图像所花的时间。

Mobile-citrus 检测器对柑橘的检测结果见表 1。Mobile-citrus 获得的总体召回率 R、准确率 A（accuracy）、F_1 值分别为 0.870、0.880、0.871。相比原 YOLOv4 的结果，虽然 F_1 精度略有下降，但检测速度会得到大幅度的提升，约是原 YOLOv4 模型的 4～5 倍左右，这样的速度给与下一阶段的跟踪算法与轨迹预测算法预留了充足的时间。

实际检测结果如图 8 所示。图 8 中检测到的绿色框代表正常（N）柑橘，检测到的蓝色框代表表皮病变（SL）柑橘，检测到的红色框则代表机械损伤（MD）柑橘，图 8 中所示都是同一批柑橘不同时刻的检测结果，可以看出检测算法对于单帧检测情况良好。但

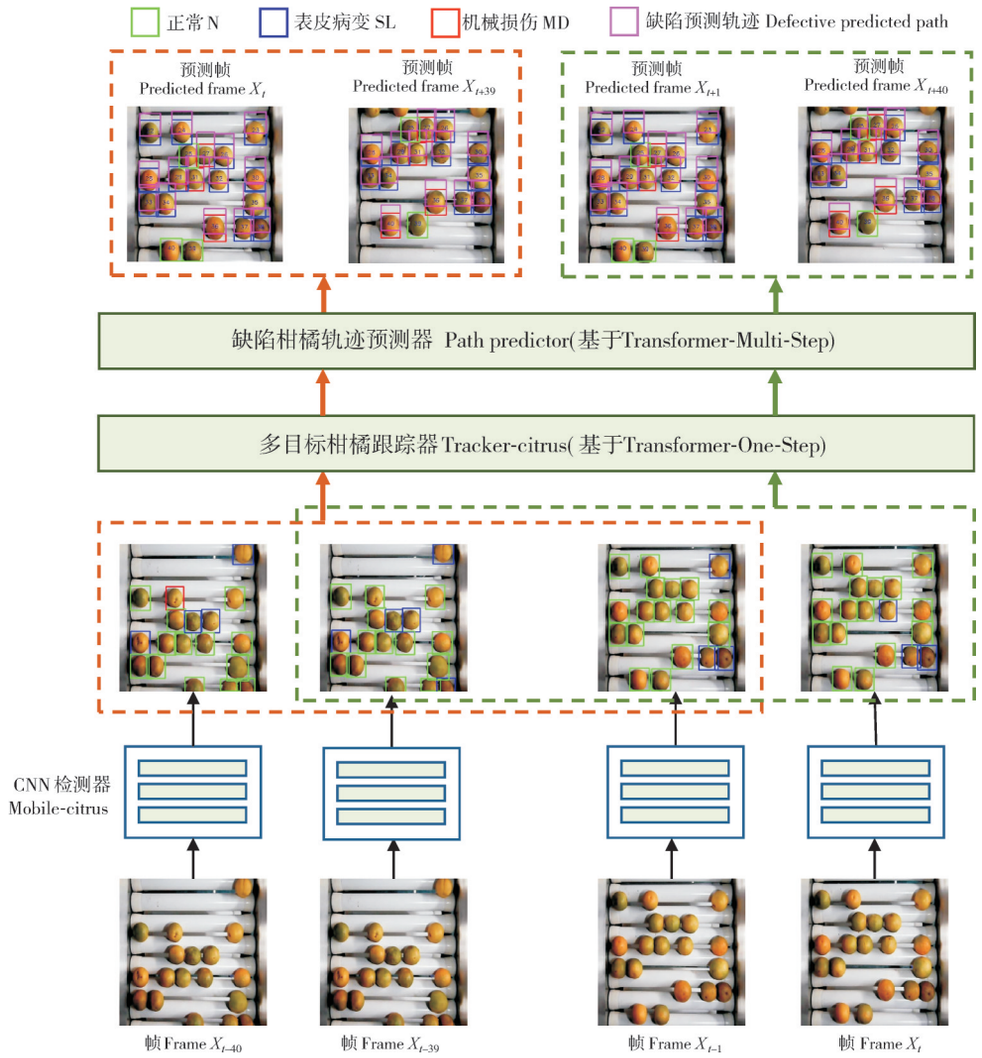


图6 轨迹预测工作流程

Fig.6 Workflow of trajectory prediction

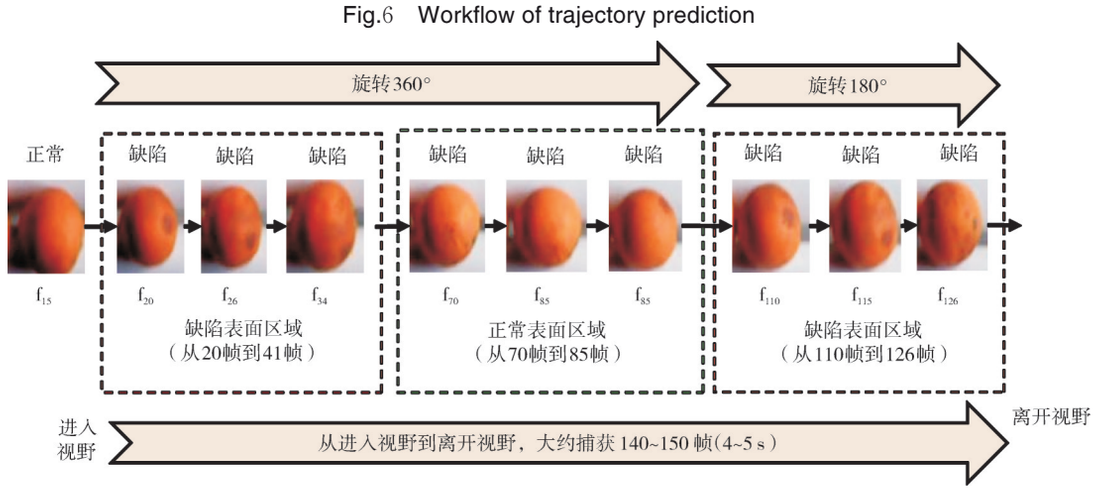


图7 跟踪分类过程

Fig.7 Tracking classification process

不同的帧中同个柑橘可能会出现不同的类型，如第153帧与第202帧中，同一个柑橘会被标记为不同类别。针对这一情况，表明需要开发跟踪算法用于柑橘类别的固定识别。

表 1　检测器的性能评估

Table 1　Performance evaluation of detector

模型 Model	召回值 Recall	准确率 Accuracy	F_1	推理时间/ms Inference time
YOLOv4	0.872	0.895	0.883	53
Mobile-citrus	0.870	0.880	0.871	12

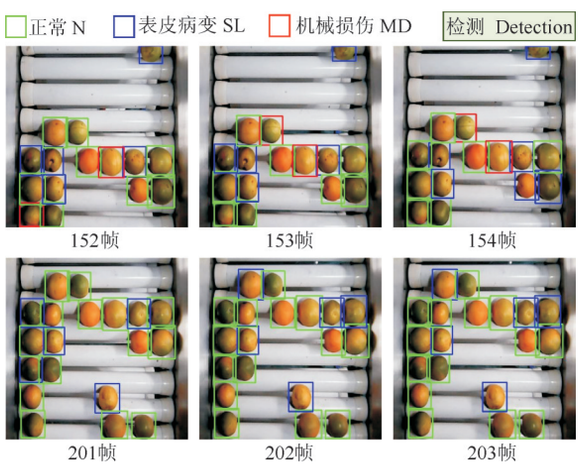


图 8　Mobile-citrus 模型缺陷检测结果

Fig.8　Mobile-citrus model defect detection results

2.2　跟踪性能评估

对于跟踪性能评估，使用多目标跟踪精度（MOTA）和多目标跟踪准确度（MOTP）来评估跟踪的整体性能，测试指标来源于 MOT20 Benchmark[22]。MOTA 规定如下：

$$[\mathrm{MOTA}=[1-\frac{\sum_t(\mathrm{FN}_t+\mathrm{FP}_t+\mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t}]\times 100\% \quad (14)$$

FN_t 表示跟踪过程中 t 时刻目标的漏检数，如图 9A 所示，第 4 帧中真实轨迹（GT）既可以与红色轨迹匹配，又可以与蓝色轨迹匹配，但红色轨迹目标与真实目标更接近，所以最终匹配到了红色轨迹，生成了真阳性（TP），但相对而言，因为真实目标已经被红色轨迹匹配，所以蓝色轨迹的目标丢失，导致了假阴性（FN）的生成。FP_t 表示跟踪过程中 t 时刻目标的误检数。如图 9A 所示，第 3 帧中真实轨迹与红色轨迹实现了匹配，但蓝色轨迹并没有完成匹配，因为轨迹预测目标与真实目标之间的 IOU 太小，不在阈值通道内，导致了假阳性（FP）的生成。IDSW_t 衡量跟踪过程中 t 时刻目标身份发生变动（IDSW）的数量，如图 9A 所示，当红色轨迹切换到蓝色轨迹时会产生 IDSW；如图 9B 所示，当由于目标的重识别处理不当，真实轨迹（GT）产生中断，也会导致 IDSW 的产生。GT_t 是时间 t 时刻物体跟踪的真实目标。

MOTP 用来衡量预测目标与真实目标之间的靠近程度，是一种定位精度的度量，MOTP 的表述如下：

$$\mathrm{MOTP}=\frac{\sum_t^i d_t^i}{\sum_t c_t}\times 100\% \quad (15)$$

其中，d_t^i 是预测的位置之间的交集（IOU）值，c_t 是正确匹配目标的数目。

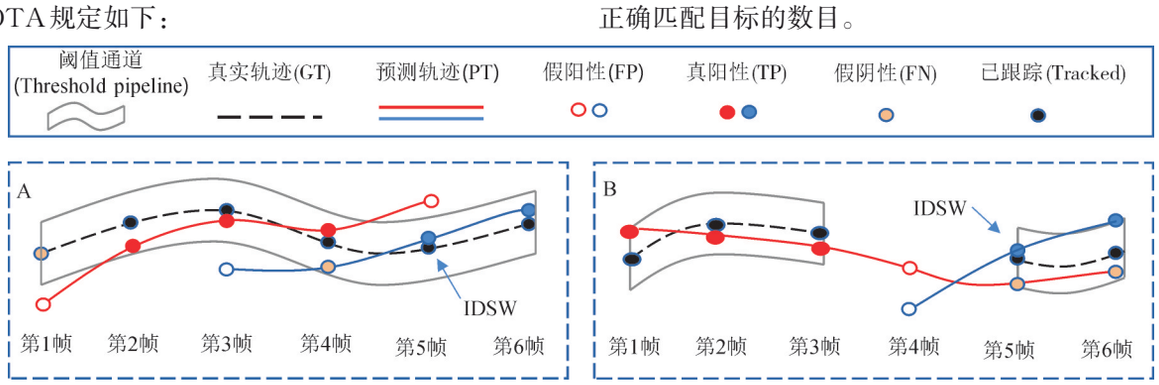


图 9　跟踪指标参数示意图

Fig.9　Schematic diagram of tracking indicator parameters

跟踪器使视觉系统能够记忆历史分类信息，并跟踪每个柑橘的位置。试验采用记录的跟踪列表执行分类，对自定义多目标跟踪数据集进行了评估。结果显示 MOTA＝98.4%，MOTP=81.5%，IDSW=0。MOTA 得到了 98.4% 的高精度，其原因是柑橘与柑橘之间不会出现遮挡，且柑橘一直都是向前运动并不会往复运动，这样就大幅度避免了柑橘对象 ID 的转换，可以从 IDSW 为 0 验证这一点。MOTP 得到了 81.5% 的高精度，可以看出基于 Transformer 的预测模型效果较好，预测的值可以和真实值获取较大的交并比（IOU）。即较高的 MOTA 和 MOTP 表示该跟踪系统具有良好的性能。

如图 10 所示，绿色、蓝色和红色的方框分别表示 N、SL 和 MD 型柑橘的位置。方框左上角的数字表示柑橘从进入视野开始计数的帧号，而不是每个子图对应的视频流的帧数，右下角的数字表示对象号。

可以看出该跟踪系统可以一直确定一个柑橘对象，即对象的ID号不会发生任何改变。例如图10中4号柑橘，64～66帧时分别检测到机械伤口，但在113～115帧呈现正常表面，也能判断出该柑橘为缺陷柑橘，即会被标记为红色框，不会因为检测为正常柑橘而标记为绿色框。

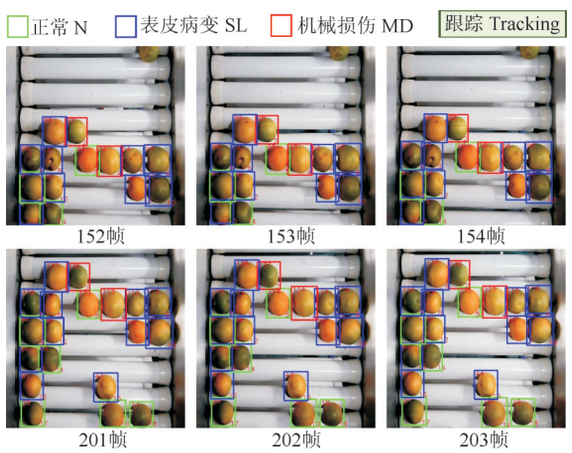


图10 缺陷检测和跟踪结果

Fig.10 Defect detection and tracking results

2.3 轨迹预测评估

对于轨迹预测性能评估，使用平均绝对误差(MAE)表示预测值与真实值的差值，表达如下：

$$e=\frac{|x_1-\hat{x_1}|+|y_1-\hat{y_1}|+|x_2-\hat{x_2}|+|y_2-\hat{y_2}|}{4} \tag{17}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{n}e_n \tag{18}$$

其中，e(error)表示4个像素坐标的误差平均值，每个坐标的误差值表示预测的边界框$(\hat{x_1},\hat{y_1},\hat{x_2},\hat{y_2})$和检测的边界框$(x_1,y_1,x_2,y_2)$之间的绝对坐标差值。如图11所示，图11A中紫色框表示算法预测10帧之后的柑橘边界框，$(\hat{x_1},\hat{y_1})$表示预测框的左上角坐标，$(\hat{x_2},\hat{y_2})$表示预测框的右下角坐标。图11B中绿色框表示10帧之后的真实柑橘边界框，(x_1,y_1)表示真实框的左上角坐标，(x_2,y_2)表示真实框的右下角坐标。最后对所有样本的误差取平均值获取最终误差MAE。

经过轨迹预测测试集验证，Transformer算法无论预测多少帧，准确度始终保持在3个左右像素误差范围，最佳平均绝对误差为2.98个像素。这也在一定程度上体现了Transformer强大的长序列处理能力。

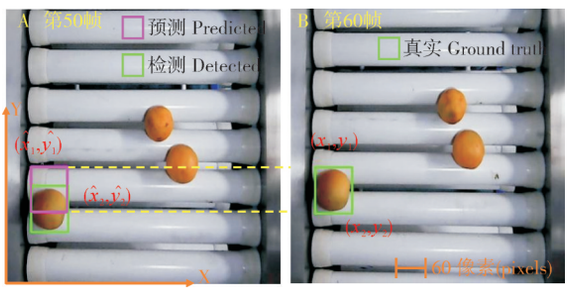


图11 轨迹预测误差计算示意图

Fig.11 Schematic diagram of trajectory prediction error calculation

最终的轨迹预测结果如图12所示，其中紫色框是缺陷柑橘的未来轨迹，图12显示的是未来第10帧的位置。可以看出对于每帧，每个柑橘的预测情况都基本相同，不会存在个别预测误差特别大的情况。

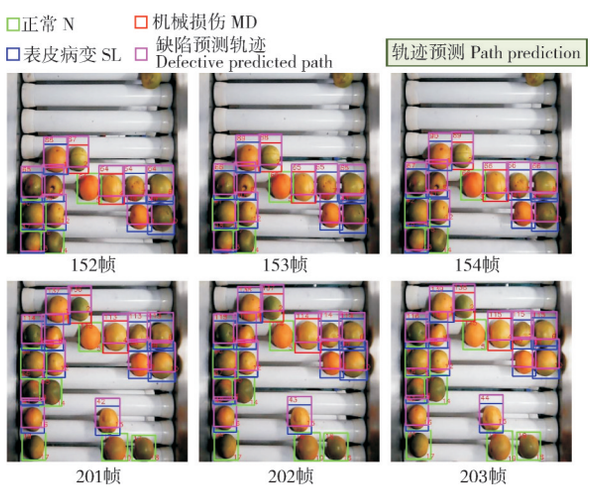


图12 轨迹预测结果

Fig.12 Trajectory prediction results

在模型训练过程中Transformer-Multi-Step模型与Transformer-One-Step模型所用参数一致，优化器使用Adam算法，学习率设置为0.000 1，编码器层数(encoder layer)设置为2，批次大小为5，多头注意力头数设置为2，Transformer-Multi-Step输入输出大小均为40×4的矩阵，Transformer-One-Step输入输出大小均为1×4的矩阵。

2.4 最终分类效果评估

普通分类中常用准确率A(accuracy)来表达模型的分类性能。准确率A被定义为：

$$A=\frac{1}{n}\sum_{i=1}^{n}I(f(x_i)=y_i) \tag{19}$$

以视频序列x为输入经过视觉模型后，预测出结果$f(x_i)$，将$f(x_i)$与真实类别y_i进行对比，如果相同记为1，不同记为0，最后对n个样本求平均。为了评

估视觉系统的分类效果，制作了 270 个柑橘样本，其中 MD、N、SL 类柑橘分别为 43、47、180 个，结果显示 A 值为 89%，如图 13 所示。图 13 中，0、1、2 分别代表对应 N、SL、MD。

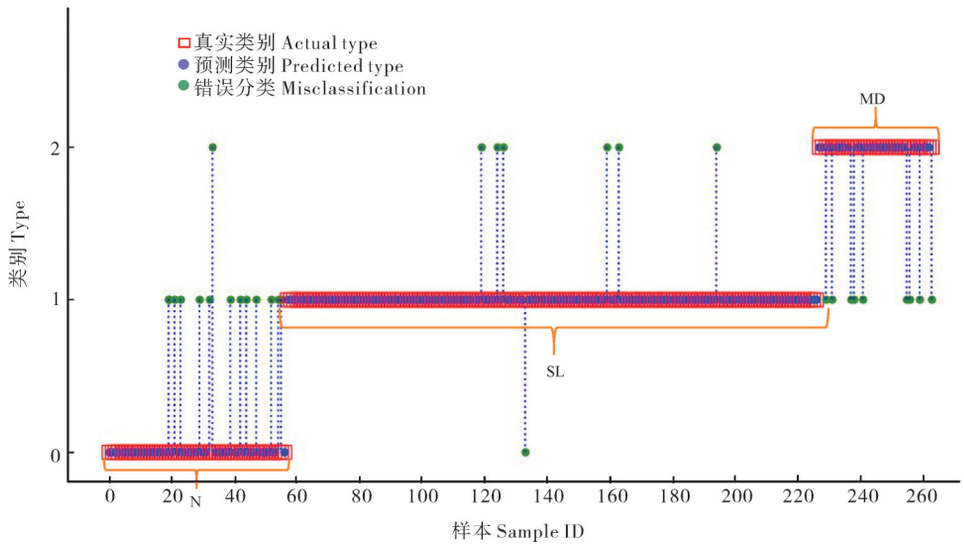


图 13　最终分类结果

Fig.13　Final classification results

由图 13 可知，该视觉系统对 SL 的识别率较高，N、MD 的识别率较差且大部分会识别为 SL，这可能是 MD 的特征表现不明显时，与 SL 的特征具有很多相似的地方；同时由于 SL 柑橘呈现的暗黑和腐烂的外观与正常区域上的深绿色区域相似，因此识别精度可能受到影响。如果只分缺陷（SL＋MD）与正常（N）两类，识别率可达到 92.8%。这说明该方法不仅解决了柑橘在滚动过程中因特征面一直改变而无法确定真实类别的问题，同时运用 Transformer 的方法，结合了历史信息之后，也具有较高的分类精度。

2.5　运行时间评估

在自动柑橘分选中，视觉信息的更新是实时的，分类系统具有实时性能是必不可少的。所提出的视觉系统由 1 个检测器和 1 个跟踪器以及轨迹预测器 3 个部分组成。试验结果显示：检测器的平均处理时间为 12 ms，跟踪器平均处理时间为 20 ms，轨迹预测平均处理时间为 1 ms，总体平均运行时间 34 ms，基本可以实现实时视觉检测。以上处理时间均在摄像头视野中平均存在 20 个左右柑橘时的试验结果。对于视野同时出现 20 个以上柑橘时，由于需要同时检测跟踪和轨迹预测的数量增多将无法达到实时，这是未来将要改进的地方。另外测试时硬件算力相对较低，随着硬件算力的增加，能同时处理的柑橘数量将越多。

3　讨　论

本研究提出了一种基于 CNN-Transformer 的视觉系统，它可以与机器手结合进行实时柑橘分类。Mobile-citrus 算法可以检测到视图中的缺陷柑橘，Tracker-citrus 算法可以在柑橘旋转过程中跟踪它们，并识别出它们的真实类型，Transformer-Multi-Step 算法可以预测柑橘的未来路径以指导机器手的抓取。跟踪获得的总体 MOTA 为 98.4%，表明系统可以跟踪视野中的大部分柑橘，并可以对缺陷柑橘与正常柑橘进行识别分类，识别分类准确率达到 92.8%，路径预测的最低绝对平均误差为 2.98 个像素，约为柑橘直径的 5%，这在机械手抓取可接受的范围内。单帧平均运行时间 34 ms，约为 29.4 FPS，具有良好的实时性能。

参考文献 References

[1]　LIU N，LI X，ZHAO P，et al.A review of chemical constituents and health-promoting effects of Citrus peels[J/OL].Food chemistry，2021，365：130585[2021-12-02].https://doi.org/10.1016/j.foodchem.2021.130585.

[2]　ELKAOUD N S M，ELGLALY A M M.Development of grading machine for citrus fruits（Valencia orange）[J].Journal of soil sciences and agricultural engineering，2019，10(11)：671-677.

[3] SAKUDO A, YAGYU Y. Application of a roller conveyor type plasma disinfection device with fungus-contaminated citrus fruits [J/OL]. AMB express, 2021, 11(1): 16[2021-12-02]. https://doi.org/10.1186/s13568-020-01177-2.

[4] BHATNAGAR A, PATEL R, GUPTA M, et al. Customized sorting and packaging machine[J]. Telecommunication computing electronics and control, 2021, 19(4): 1326-1333.

[5] REDMON J, FARHADI A. YOLOv3: an incremental improvement [EB/OL]. arXiv. 2018.1804.02767 [cs. CV] [2021-12-02]. https://arxiv.org/abs/1804.02767.

[6] KANG H W, CHEN C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards[J/OL]. Computers and electronics in agriculture, 2020, 171: 105302[2021-12-02]. https://doi.org/10.1016/j.compag.2020.105302.

[7] KANG H W, CHEN C. Fast implementation of real-time fruit detection in apple orchards using deep learning[J/OL]. Computers and electronics in agriculture, 2020, 168: 105108[2021-12-02]. https://doi.org/10.1016/j.compag.2019.105108.

[8] WANG Q J, ZHANG S Y, DONG S F, et al. Pest24: a large-scale very small object data set of agricultural pests for multi-target detection[J/OL]. Computers and electronics in agriculture, 2020, 175: 105585[202-12-02]. https://doi.org/10.1016/j.compag.2020.105585.

[9] 杨万里,段凌凤,杨万能. 基于深度学习的水稻表型特征提取和穗质量预测研究[J]. 华中农业大学学报, 2021, 40(1): 227-235. YANG W L, DUAN L F, YANG W N. Deep learning-based extraction of rice phenotypic characteristics and prediction of rice panicle weight[J]. Journal of Huazhong Agricultural University, 2021, 40(1): 227-235(in Chinese with English abstract).

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. arXiv: 1706.03762 [cs. CL] [2021-12-02]. https://doi.org/10.48550/arXiv.1706.03762.

[11] RAGANATO A, SCHERRER Y, TIEDEMANN J. Fixed encoder self-attention patterns in transformer-based machine translation [EB/OL]. arXiv: 2002.10260 [2021-12-02]. https://doi.org/10.48550/arXiv.2002.10260.

[12] SUN P Z, JIANG Y, ZHANG R F, et al. TransTrack: multiple-object tracking with transformer[EB/OL] . arXiv: 2012.15460 [cs. CV] [2021-12-02]. https://doi.org/10.48550/arXiv.2012.15460.

[13] WANG Z W, MA Y, LIU Z T, et al. R-transformer: recurrent neural network enhanced transformer[EB/OL]. 2019: arXiv: 1907.05572 [cs.LG][2021-12-02]. https://arxiv.org/abs/1907.05572.

[14] 章海亮,高俊峰,何勇. 基于高光谱成像技术的柑橘缺陷无损检测[J]. 农业机械学报, 2013, 44(9): 177-181. ZHANG H L, GAO J F, HE Y. Nondestructive detection of citrus defection using hyper-spectra imaging technology [J]. Transactions of the CSAM, 2013, 44(9): 177-181(in Chinese with English abstract).

[15] 龚中良,杨张鹏,梁力,等. 基于机器视觉的柑橘表面缺陷检测[J]. 江苏农业科学, 2019, 47(7): 236-239. GONG Z L, YANG Z P, LIANG L, et al. Detection of citrus surface defects based on machine vision [J]. Jiangsu agricultural sciences, 2019, 47(7): 236-239(in Chinese).

[16] 李善军,胡定一,高淑敏,等. 基于改进SSD的柑橘实时分类检测[J]. 农业工程学报, 2019, 35(24): 307-313. LI S J, HU D Y, GAO S M, et al. Real-time classification and detection of citrus based on improved single short multibox detecter [J]. Transactions of the CSAE, 2019, 35(24): 307-313(in Chinese with English abstract).

[17] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. arXiv: 2004.10934 [cs. CV] [2021-12-02]. https://doi. org/10.48550/arXiv.2004.10934.

[18] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks [EB/OL]. arXiv: 1801.04381 [cs. CV] [2021-12-02]. https://doi. org/10.48550/arXiv.1801.04381.

[19] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. [S.l.]: IEEE, 2018: 8759-8768.

[20] XU J J, SUN X, ZHANG Z Y, et al. Understanding and improving layer normalization [EB/OL]. arXiv: 1911.07013 [cs. LG] [2021-12-02]. https://doi.org/10.48550/arXiv.1911.07013.

[21] BEWLEY A, GE Z Y, OTT L, et al. Simple online and real-time tracking[C]//2016 IEEE International Conference on Image Processing. September 25-28, 2016, Phoenix, AZ, USA. [S. l.]: IEEE, 2016: 3464-3468.

[22] MILAN A, LEAL-TAIXE L, REID I, et al. MOT16: a benchmark for multi-object tracking[EB/OL]. arXiv: 2016.1603.00831 [cs.CV] [2021-12-02]. https://arxiv.org/abs/1603.00831.

A CNN-Transformer-based method for sorting citrus with visual defects

AN Xiaosong[1], SONG Zhuping[1], LIANG Qianyue[1], DU Xuan[1], LI Shanjun[1,2,3]

1.*College of Engineering, Huazhong Agricultural University, Wuhan* 430070, *China*;
2.*National R&D Center for Citrus Preservation, Wuhan* 430070, *China*;
3.*Ministry of Agriculture and Rural Affairs Key Laboratory of Agricultural Equipment in Mid-Lower Yangtze River, Wuhan* 430070, *China*

Abstract Manual sorting of citrus fruit with visual defects on the production line is time-consuming and cost-expensive. This article proposes a sorting solution based on machine vision and CNN-Transformer. The system can be directly implemented on various citrus processing lines for online sorting. For the citrus fruits randomly rotating on the conveyor, a detection algorithm Mobile-citrus based on convolutional neural network (CNN) was developed to detect and temporarily classify the defective one. A tracking algorithm Tracker-citrus was used to record the classification information along the path. The real category of the fruit was identified using the historical information, with tracking accuracy of 98.4% and classification accuracy of 92.8%. In addition, a trajectory prediction algorithm based on Transformer was used to predict the future path of fruit with the average prediction error of 2.98 pixels, which can be used to guide the robot arm to sort defective citrus in real time. The results showed that the method proposed can be applied to citrus production lines for online sorting.

Keywords citrus; defect detection; machine vision; deep learning; convolutional neural network; online citrus sorting; trajectory prediction; Transformer

（责任编辑：陆文昌）