基于改进的 YOLOv3 网络的实时目标检测

孙佳，郭大波*，杨甜甜，马识途
山西大学物理电子工程学院，山西 太原 030006

摘要 针对 YOLOv3 算法实时目标检测性能不佳的缺陷，提出了一种适应实时目标检测的改进网络结构以及视频目标检测的新方法。首先，提出的 k-means-threshold(k-thresh)方法弥补了 k-means 算法对聚类中心初始位置十分敏感的问题，在包括三个类别的数据集中进行聚类分析选择合适的锚框；然后，将 4 倍下采样和 8 倍下采样特征图拼接融入第三个检测层，以提高对目标的检测精度，将 YOLOv3 算法的平均准确率均值提高了 2%；最后，通过摄像头捕捉图像和前期得到的优秀检测数据来预测新图像的目标以及加入了重新检测阈值，以提高视频检测流畅度。实验结果表明：所提基于改进的 YOLOv3 网络在检测精度上得以提高，实时性也有所提高，在 30 min 的实时检测中最大帧率达到 64.26 frame/s，相比原始 YOLOv3 算法，实时检测速度提高了 4 倍左右。

关键词 机器视觉；图像处理；目标检测；YOLOv3；k-means 算法
中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP57.221505

Real-Time Object Detection Based on Improved YOLOv3 Network

Sun Jia, Guo Dabo*, Yang Tiantian, Ma Shitu
College of Physics and Electronic Engineering, Shanxi University, Taiyuan, Shanxi 030006, *China*

Abstract For the shortcoming of the real-time performance of YOLOv3 algorithm in object detection, we propose an improved network structure and a new method for video object detection adapted to real-time object detection. Firstly, the proposed k-means-threshold (k-thresh) method makes up for the problem of its sensitivities to the initial position of the cluster center, and performs cluster analysis on a data set including three categories to select more appropriate anchor boxes. Then, the 4× down-sampling and 8× down-sampling feature maps are stitched together into the third layer detection layer to improve the detection accuracy of the object and increase the the mean average precision of the YOLOv3 algorithm by 2%. Finally, the camera captures the image and the excellent detection data obtained in the previous period to predict the target of the new image and adds a re-detection threshold to improve the smoothness of video detection. The experimental results show that the proposed improved YOLOv3 network improves the detection accuracy and the real-time performance, the maximum frame rate reaches 64.26 frame/s in 30 min of real-time detection, which is 4 times faster than the original YOLOv3 algorithm.

Key words machine vision; image processing; object detection; YOLOv3; k-means algorithm
OCIS codes 150.0155; 150.1135; 100.4996

1 引 言

目标检测是机器视觉研究的基本课题之一，有较为广泛的应用，例如安保监控、自动驾驶、交通监控和机器人视觉等。一般而言，为了实现目标检测，首先要在图像样本中提取出目标特征，然后在此基础上进行分类判断。在传统的目标检测算法中，前期处理有基于模板检测[1-4]和梯度特征提取[5-8]之分，后期处理有基于统计判断[1,5]和模式学习与分类[2-4,6-8]之分。近年来，随着深度学习技术的不断发展，目标检测技术也向深度学习转化，出现了基于深度卷积神经网络的两阶段模型和一阶段模型[9]。比较流行的两阶段模型有 2014 年 Girshick 等[10]提出的区域卷积神经网络(R-CNN)及改进的 Fast R-

收稿日期：2020-04-02；**修回日期**：2020-04-23；**录用日期**：2020-04-27
基金项目：山西省基础研究项目(201801D121118)
* **E-mail**：dabo_guo@sxu.edu.cn

CNN[11]、Faster R-CNN[12]，之后冯小雨等[13]将Faster R-CNN应用在空中目标检测。2016年由Redmon等[14]提出深度学习时代的一阶段模型YOLO（You Only Look Once）以及后来的YOLOv2[15]和YOLOv3[16]，同年Liu等[17]又提出SSD(Single Shot Multibox Detector)，为了提高精度，Lin等[18]提出RetinaNet算法。比较先进的两阶段模型Fast R-CNN先通过区域候选网络(RPN)获取候选目标的边界框，然后使用感兴趣区域(RoI)池化操作从每个候选框提取特征进行分类和边界框回归任务；然而一阶段模型不需要区域候选网络，例如YOLOv3将特征图划分成$K\times K$个小网格，每个网格会预测3个边界框，同时YOLOv3使用多个独立的逻辑回归分类器对目标进行分类，每个分类器对于目标边框中出现的物体只判断其是否属于当前标签，实现了多标签分类[19]。

评价一个目标检测器的好坏，可以依据目标定位的精确度、检测速度以及密集和遮光条件下的检测效果等[20]，最基本的性能指标是目标检测精度和检测速度。但是它们之间又是相互对立的，如两阶段模型R-CNN系列的优势在于检测精度高，而一阶段模型YOLO系列的主要特点是检测速度快，这也是本文选择YOLOv3算法进行实时目标检测的主要原因。目前已经有很多基于YOLOv3的改进和应用，例如无人机检测识别应用[9,21]和小目标检测的应用[22]等。而视频目标检测的大致原理和图像相同，但是视频中包含的图像数量更多，存在大量冗余信息，主流的视频目标检测方法有D & T (Detect and Track)[23]等方法的目标跟踪、光流法、与循环神经网络(RNN)相结合的方法、非端到端方法等[24]。在NVIDIA Tesla K40 GPU上采用官方Darknet深度学习框架以及yolov3.weight利用YOLOv3直接进行实时检测实验，每秒处理帧数为5.6～6.2 frame/s，所以YOLOv3的实时目标检测还存在很大的提升空间。

本文基于YOLOv3网络，改进了其目标检测的实时性。k-means算法对聚类中心的初始位置很敏感，每次迭代选取不同的初始聚类中心会导致不同的聚类结果[25-26]，本文通过设置阈值选取合适的初始聚类中心获得更好的聚类结果。针对YOLOv3检测器精度不高的问题，改进了第三个YOLO检测层，从而在适当提高精度的基础上满足实时检测的要求。为了提高检测视频的流畅度，提出了一种根据像素阈值及跳帧来预测新图像目标的方法。利用改进的YOLOv3网络实时目标检测在检测精度和检测速度上都有明显提升，30 min的实时检测可达到的最大帧率为64.26 frame/s。

2 YOLOv3网络

YOLO算法用单个神经网络处理整张图像，利用整张图像的上下文信息直接从像素得到可能的类别和边界框。YOLOv3借鉴了特征金字塔(FPN)网络[27]的尺度金字塔结构，将输入的图像缩放到320 pixel × 320 pixel，416 pixel × 416 pixel，608 pixel×608 pixel等32倍数的尺寸，通过上采样实现三个不同尺度的跨层检测。YOLOv3采用了没有全连接层的Darknet-53，网络结构如图1所示。图1(a)为YOLOv3算法网络结构，多尺度检测通过上采样结合了36层、61层和74层的特征图，其中set conv模块和yolo layer模块是由1×1和3×3的卷积层组成，concat表示拼接层，例如74层的8 pixel×8 pixel特征图通过上采样为16 pixel×16 pixel的特征图然后和61层的16 pixel×16 pixel的特征图拼接起来用于检测目标。

目标检测技术不仅可以面向静止的图像，也可以面向视频中某一帧图像，图像可以是可见光的图像，也可以是红外、微波或者其他成像方法获得的数字图像。视频每一帧都是静止的图像，快速连续地显示多帧图像就形成了“运动的假象”，较低帧率视频会出现模糊、移动镜头不流畅和延时等缺点，较高帧率可以得到流畅的画面。与静止图像相比，视频增加了时间维度，检测的目标位置和外观响应应该在时间上与视频保持一致性，否则就会让测试者感到不适。在NVIDIA Tesla K40 GPU上YOLOv3检测器在分辨率为160 pixel时的实时效果较好，但是每帧视频图像的检测精度不高，而在分辨率为416 pixel或者608 pixel时，虽然精度提高了，但是画面流畅度很差。因此，如何利用YOLOv3算法在较高分辨率下提高其实时性就是本文的工作。

3 改进的YOLOv3网络实时检测

3.1 锚框的选取

YOLOv3通过k-means聚类出3×3个锚框用来逻辑回归边界框，这样大大提高了对目标对象的检测，这也意味着选定的锚框大小会影响目标检测器的性能，所以锚框的选择非常重要。为了减小锚框大小对检测的影响，采用矩形框的平均交并比(Avg IOU)对训练集所有目标使用k-means聚类

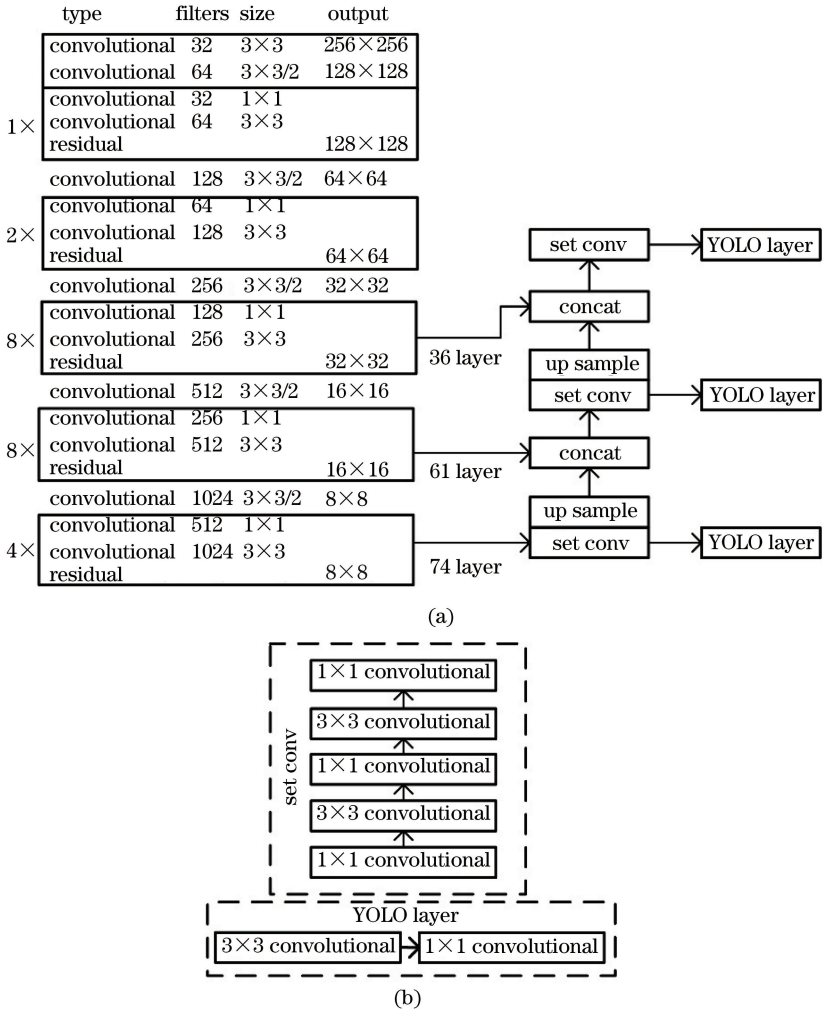

图 1 YOLOv3 网络结构图。(a) YOLOv3 网络整体结构图;(b) set conv 层和 YOLO layer 结构图

Fig. 1 YOLOv3 network structure diagram. (a) Overall structure diagram of YOLOv3 network; (b) structure diagram of set conv layer and YOLO layer

获得锚框的大小,聚类的平均交并比目标函数 p 可表示为

$$I_{\mathrm{IOU}}=\frac{\mathrm{area}(b_{\mathrm{boxTru}}\cap b_{\mathrm{boxPre}})}{\mathrm{area}(b_{\mathrm{boxTru}}\cup b_{\mathrm{boxPre}})},\tag{1}$$

$$d(b_{\mathrm{box}},c_{\mathrm{cen}})=1-I_{\mathrm{IOU}}(b_{\mathrm{box}},c_{\mathrm{cen}}),\tag{2}$$

$$p=\mathrm{argmax}\frac{\sum_{i=1}^{k}\sum_{j=1}^{m_k}I_{\mathrm{IOU}}(b,c)}{m},\tag{3}$$

式中:I_{IOU} 表示真实框和预测框交集面积和并集面积的比值;(2)式为利用 k-means 聚类算法的距离度量;b_{box} 表示矩形框大小;c_{cen} 表示一个簇中心矩形框的大小;b 表示样本;c 表示通过 k-means 算法选择的簇中心;m_k 表示第 k 个聚类中样本的个数,m 表示样本总个数,k 表示聚类中心个数;i 和 j 分别表示样本序号和聚类中的样本序号。

YOLOv3 作者选出(10 13,16 30,33 23,30 61,62 45,59 119,116 90,156 198,373 326)9 个锚框是基于 COCO 数据集聚类出来的,同时由于机械设备以及实际操作需求,本文重新筛选制作了包括 Person、Tvmonitor 和 Chair 三个类别的数据集进行实验,图 2(a)是本文数据集中目标宽高的分布,其中横坐标表示目标的宽,纵坐标表示目标的高;图 2(b)是锚框个数 $k=1\sim9$ 对本文数据集聚类分析得到 k 和平均交并比的关系图。平均交并比随着锚框个数增加而增加,由于本文模型保留三层检测层,最终选取 9 个锚框。

k-means 算法初始聚类中心对聚类结果有很大影响,如果初始中心选取合适,则聚类结果的平均交并比会比较好,但是如果初始中心选择不佳那么最终结果也会很差。张琳等[28]为了提高 k-means 的聚类效果,先使用 Canopy 算法对数据进行聚类得到聚类中心,再使用 k-means 算法进行聚类。本文

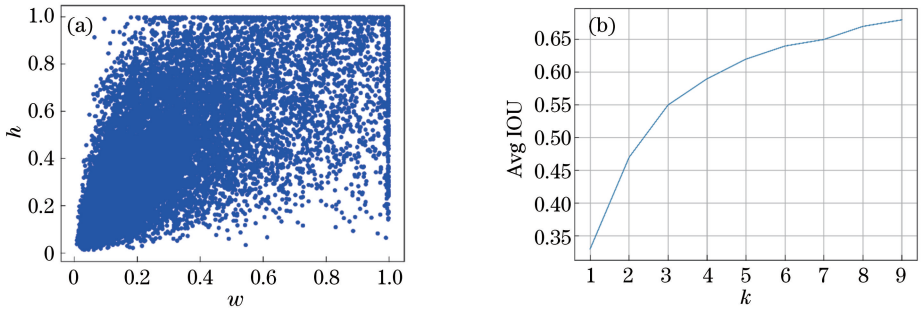

图 2 数据集分析结果。(a)数据集的目标宽高分布图;(b) k-means 聚类分析结果

Fig. 2 Dataset analysis results. (a) Target width and height distribution of the dataset; (b) k-means clustering analysis result

对 k-means 算法增加了阈值来提高聚类效果,命名为 k-means-threshold(k-thresh)方法。首先通过 k-means 算法进行“粗”聚类,此时可能得到一个好的聚类结果也可能是不好的聚类结果,然后通过阈值选择好的聚类结果为新的聚类中心,再次通过 k-means 算法进行“细”聚类得到最终的聚类结果。取每个初始聚类中心与样本的交并比均值作为 Mean IOU,图 3(a)表示均值交并比(Mean IOU)与 Avg IOU 的关系,发现当 Mean IOU 大于 0.3 时,聚类结果的 Avg IOU 表现比较好,所以选择“粗”聚类结果的 Mean IOU 大于 0.3 的结果作为“细”聚类的初始聚类中心,即阈值选取为 0.3。

k-thresh 方法的平均交并比结果在图 3(b)中由 k-thresh 线表示,原 k-means 算法的平均交并比结果在图 3(b)中由 k-means 线表示。从图中可以看出,k-thresh 方法在多次实验的聚类结果平均交并比在 68.60%~68.51%之间波动,而 k-means 算法进行多次实验的平均交并比却在更大范围内波动,所以 k-thresh 方法可以得到更好的聚类效果。通过 k-thresh 方法本文选择 9 个锚框大小为(17 31,27 69,55 61,47 125,99 108,83 187,201 166,143 279,309 334)。

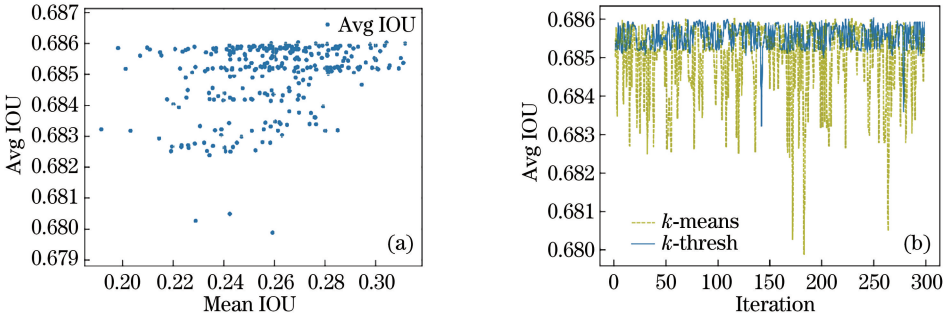

图 3 平均交并比。(a) Avg IOU 与 Mean IOU 关系图;(b) k-means 算法和 k-thresh 方法的 Avg IOU

Fig. 3 Average IOU. (a) Relationship between Avg IOU and Mean IOU; (b) Avg IOU of k-means and k-thresh

3.2 网络结构改进

YOLOv3 网络采用了三个尺度特征图对应不同大小的锚框。尺度越小,感受野越大,分辨率越小,对小目标越不敏感,输入为 416 pixel×416 pixel 时,尺度 13×13 可以用来预测大目标,而对于小目标检测精度将会大幅下降,所以增加了 26×26 和 52×52 尺度特征图提高对中等以及偏小目标的检测精度。如图 4 所示,利用原 YOLOv3 检测器对昏暗图像进行目标检测会出现检测不到的问题。

图 4 目标检测图像

Fig. 4 Image of object detection

针对实时目标检测精度不佳的问题,本文提出了一种新的 video-YOLOv3 网络结构,加强了 YOLOv3 实时目标检测的性能,该网络结构如图 5

所示。深层特征语义信息更加丰富但是目标位置较粗略，而浅层特征虽然语义信息少但是目标位置更准确，YOLOv3 通过多尺度跨层检测结合深层语义信息和浅层语义信息，对不同大小的特征层进行独立预测，更好适应了对小目标的检测[22]。图 6 为两种网络结构的对比图，图 6(b)在保留了图 6(a)三层采样检测的前提下，为了进一步适应对小目标的检测进而提高其检测精度，将 Darketnet-53 网络的第 36 层和 11 层拼接融合到小目标检测层。以输入为 416 pixel×416 pixel 为例，通过上采样将 36 层的 52 pixel×52 pixel 与 11 层的 104 pixel×104 pixel 拼接融入第三个 YOLO 检测层，同时加入两层 3×3 卷积层和三层 1×1 卷积层增加网络深度，这些操作虽然提高了检测精度但是也增加了检测时间。将图 4 分别放入 YOLOv3 和 video-YOLOv3 检测器中的时间为 62.06 ms 和 72.84 ms，即该图像的检测时间大约增加了 10.78 ms 左右，但是 video-YOLOv3 检测器仍然可以满足实时目标检测的需求，所以本文采用 video-YOLOv3 网络结构实现实时目标检测。

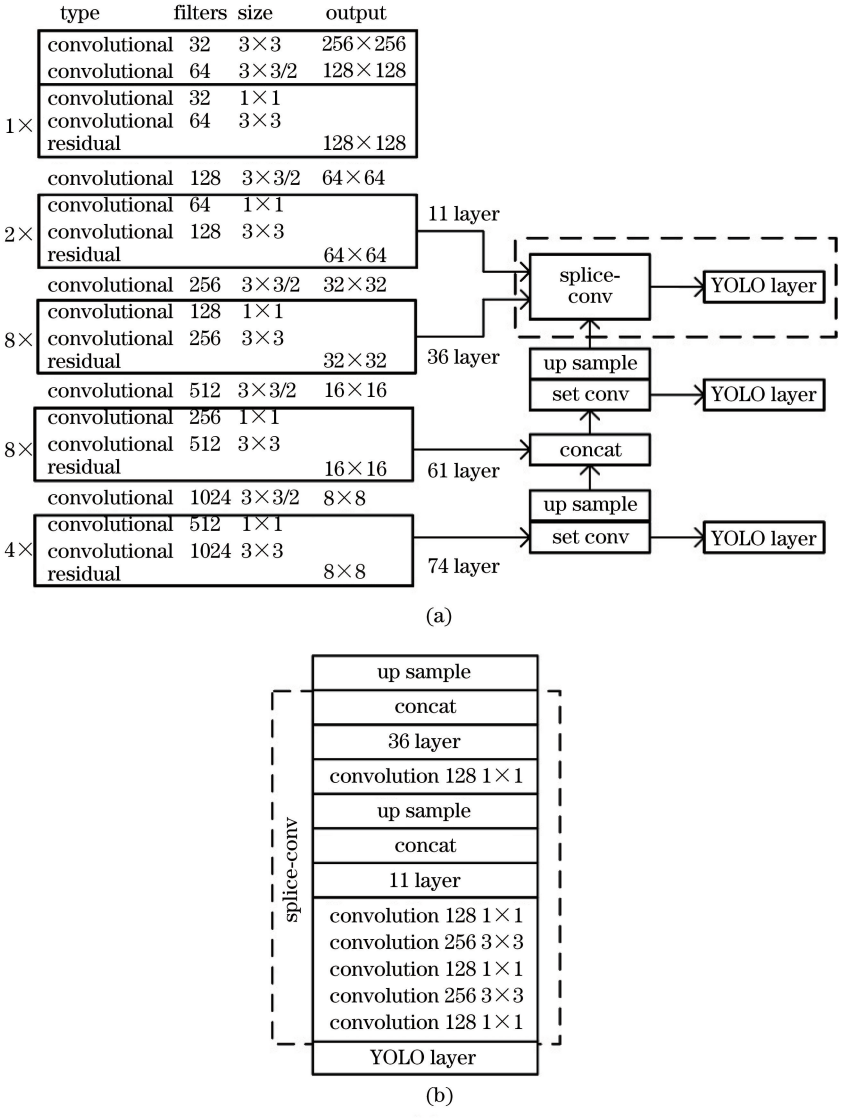

图 5 video-YOLOv3 网络结构。(a) video-YOLOv3 网络整体结构图；(b) splice-conv 模块结构图

Fig. 5 Video-YOLOv3 network structure. (a) Overall structure diagram of video-YOLOv3 network; (b) structure diagram of splice-conv module

3.3 预测新图像目标

视频是由许多帧图像组成，而且相邻每帧图像具有丰富连续相关的上下文信息[29]，所以可以利用视频初始几帧图像的位置以及类别预测连续的后几帧图像的目标。YOLOv3 算法的目标框是以左上角为起始坐标，利用 Darknet-53 网络同时得到多个

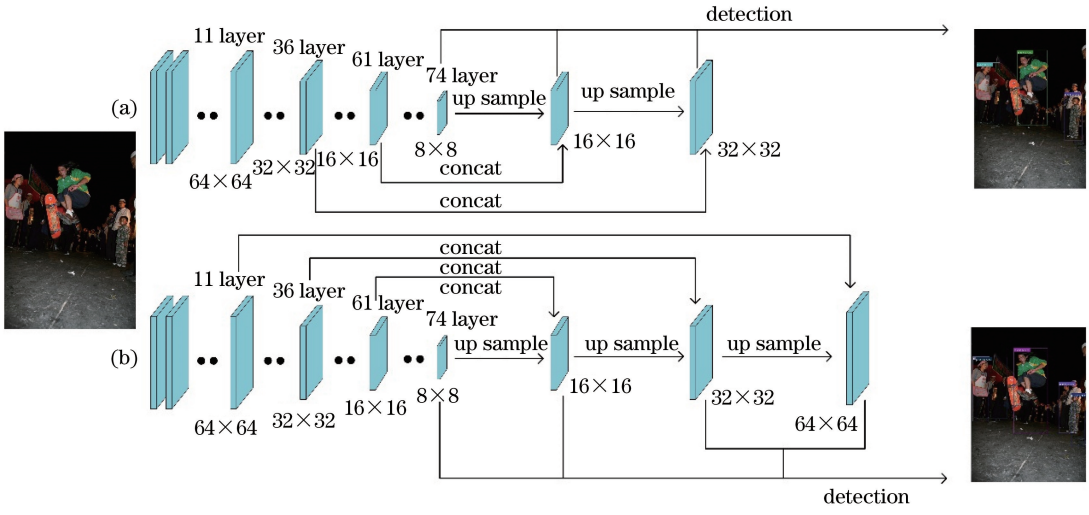

图 6 网络结构对比图。(a) YOLOv3 网络结构;(b) video-YOLOv3 网络结构

Fig. 6 Comparison of network structure. (a) YOLOv3 network structure; (b) video-YOLOv3 network structure

预测的目标位置以及类别,预测的每个目标框有 4 个坐标信息、置信度和类别,然后通过置信度以及非最大抑制筛选出最优的位置和类别。预测新图像目标流程如图 7 所示,对于视频的前几帧图像,我们可以通过 YOLOv3 检测器选择目标位置以及类别表现优秀的参数预测后几帧图像中的目标,从而缩短检测时间,提高目标检测的流畅度和实时性。

视频图像中的目标有三种状态,分别是消失的目标、依旧存在的目标、新出现的目标。对于依旧存在的目标,通过目标检测器选取检测图像中优秀的目标位置和类别参数预测新图像中的目标。对于消失和新出现的目标,本文采用像素阈值判断。视频中连续帧之间像素变化通常比较小,所以将连续图像的每两帧之间的像素点的差值均值化后设定为阈值,随着传入视频图像的变化,该像素阈值也在不断更新变化,当传入的下一帧图像与上一帧图像的像素点差值大于该阈值时,就认为图像有新目标出现或目标消失,此时采用 YOLOv3 检测器重新检测。为了保证视频检测时像素阈值的稳定更新和适应在某些帧图像之间像素阈值的不敏感性,设置预测一定帧数后必须重新检测一次。

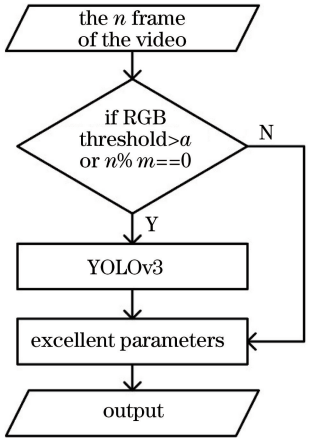

图 7 预测新图像流程图

Fig. 7 Flow chart of predicting new image

4 实验结果与分析

4.1 video-YOLOv3 网络训练和测试结果评价

本文实验操作平台使用 GPU 为 NVIDIA Tesla k40 的 Ubuntu 操作系统和 Pytorch 深度学习框架。PASCAL VOC 数据集格式规范,图像和标注质量高,是目标检测测评常用数据集,为实现实时目标检测充分利用实验环境的物体以及机械设备,本文在数据集 VOC2007 和 VOC2012 的基础上筛选制作了只有 Person、Tvmonitor 和 Chair 三个类别的数据集,这三种物体在实验环境中普遍存在,可以充分利用训练的参数。本文数据集中包含每类目标的图片数量如表 1 所示。

表 1 数据集中包含每类目标的图片数量

Table 1 Number of images with each type of object in the dataset

Class	Person	Tvmonitor	Chair
Total number	2503	1050	1958

在 YOLOv3 网络和 video-YOLOv3 网络的训练中,使用 3.1 节通过 k-thresh 方法选择的 9 个锚框,其初始学习率、衰减系数和动量分别为 0.001、0.0005 和 0.9。训练从 1000 次迭代开始,以初始学习率训练,20000 次迭代后学习率为 0.0001,25000 次迭代后以 0.00001 为学习率,使损失函数进一步

收敛。

video-YOLOv3 网络训练过程的损失值变化曲线如图 8(a)所示，而且在 25500 次迭代后的损失值达到 0.15 左右，其平均交并比曲线如图 8(b)所示。

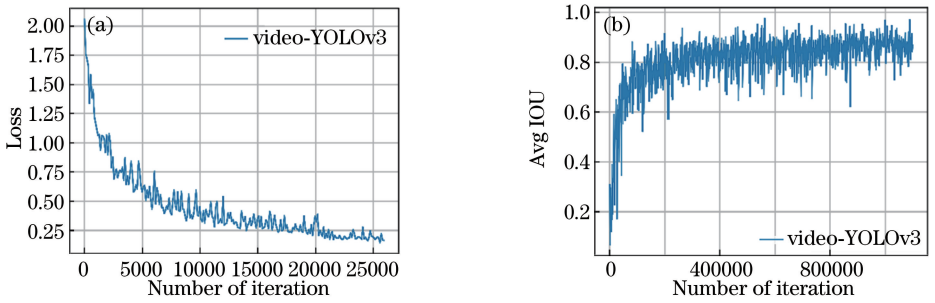

图 8 video-YOLOv3 损失值函数和平均交并比曲线。(a)损失值函数曲线；(b)平均交并比曲线

Fig. 8 Loss function and Avg IOU curve of video-YOLOv3. (a) Loss function curve; (b) Avg IOU curve

表 2 为本文模型和其他目标检测模型准确率对比，相比其他模型，本文模型准确率更高，而且 YOLO 系列整体平均准确率均值(mAP)值都优于 Faster R-CNN 和 SSD。video-YOLOv3 的平均准确率(AP)在较小目标上有所提升，和 YOLOv3 相比，在本文数据集中 Chair 和 Tvmonitor 在 video-YOLOv3 网络中的 AP 值分别增长了 4%和 2%，并且改进网络的 mAP 增长了 2%。图 9 为 video-YOLOv3 网络检测结果，该网络模型可以检测出相对昏暗图像中的更多目标以及比较准确地检测出更多的小目标。

表 2 不同模型在数据集上的 mAP 值

Table 2 mAP values of different models on the dataset

unit: %

Method	AP			mAP
	Person	Tvmonitor	Chair	
Faster R-CNN	70	74	65	69.67
SSD	72	73	58	67.67
YOLOv3	77	74	60	70.33
Video-YOLOv3	77	76	64	72.33

图 9 检测结果对比。(a) YOLOv3 检测结果；(b) video-YOLOv3 检测结果

Fig. 9 Comparison of test results. (a) YOLOv3 test results; (b) video-YOLOv3 test results

4.2 video-YOLOv3 网络实时检测结果评价

为保证在本文实验设备环境下视频检测时像素阈值的稳定更新以及适应在某些帧图像之间像素阈值的不敏感性，设置预测一定帧数自动重新检测。在实时检测实验中，本文选取每 5、6、7 帧进行跳帧检测对比，其检测效果如图 10 所示。在图 10 中，每组图像均为间隔一帧选取的 5 帧图像，可以看出，每 5、6 帧跳帧检测视频检测效果较好，但是每 6 帧检测一次视频更流畅，而每 7 帧检测一次会让观察者感到目标框在追赶目标，所以本文选取每 6 帧自动检测。

本文在 Pytorch 框架下进行 YOLOv3 算法相关实验，实时目标检测的检测帧率会随着时间变化，所以选取实时检测 30 min 的最大帧率以及 30 min 内检测的平均时间做对比。由表 3 CPU 和 GPU 实时检测对比可知，加入预测图像方法时，在 GPU 下加入预测图像方法的实时检测中每张图像的平均检测时间 16.39 ms，最大帧率为 64.26 frame/s，大约是原来未加入预测图像方法的4.09倍；在CPU中

图 10 实时检测对比图。(a)每 5 帧检测;(b)每 6 帧检测;(c)每 7 帧检测

Fig. 10 Comparison of real-time detection. (a) Detection every 5 frames; (b) detection every 6 frames; (c) detection every 7 frames

表 3 CPU 和 GPU 实时检测时间对比

Table 3 Comparison of CPU and GPU real-time detection time

Item	CPU(Intel i7-7700k)		GPU(Tesla K40)	
	Original	Corrected	Original	Corrected
Time /ms	231.22	51.78	67.54	16.39
Max /(frame/s)	4.33	19.45	15.73	64.26

每张图像平均检测时间大约是未加入预测图像方法的 4.49 倍。

5 结 论

为实现实时目标检测对 YOLOv3 网络结构进行了改进,提高了预测新图像目标的检测精度。首先针对 k-means 算法对初始值很敏感的问题,提出了 k-thresh 方法对本文数据集进行聚类选出优秀的锚框。YOLOv3 检测器的检测速度虽然具有很大优势,但是检测精度不高,为了提高检测精度,本文提出了 video-YOLOv3 网络将 4 倍下采样和 8 倍下采样特征图通过上采样拼接融入第三个检测层,该网络在包含三个类别的数据集中 Chair 类的 AP 值提高了 4%,Tvmonitor 类的 AP 值提高了 2%以及 mAP 值也提高了 2%。为了提高检测精度牺牲了一定的检测速度,但是通过本文提出的预测新图像目标方法较好地弥补了检测速度的问题,30 min 的实时目标检测的最大速度达到了 64.26 frame/s,得到了很好的视觉效果。

虽然本文在一定精度下实现了比较好的实时目标检测,但是仍然存在很大的改进空间。实时目标检测容易受到光照强度大小的影响,若捕捉检测的一帧图像受到强光(或者弱光)影响,检测器很难检测到目标,这将会降低检测精度甚至阻断实时目标检测。除此之外,如何利用视频目标检测算法优化本文实时目标检测,这些都是在未来的科研中需要深入研究的内容。

参 考 文 献

[1] Schneiderman H, Kanade T. A statistical method for 3D object detection applied to faces and cars[C]// Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662). June 15-15, 2000, Hilton Head Island, SC, USA. New York: IEEE Press, 2000: 746-751.

[2] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[J]. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.

[3] Felzenszwalb P F, Girshick R B, McAllester D. Cascade object detection with deformable part models[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 2241-2248.

[4] Forsyth D. Object detection with discriminatively trained part-based models[J]. Computer, 2014, 47(2): 6-7.

[5] Lowe D G. Local feature view clustering for 3D object recognition[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. December 8-14, 2001, Kauai, HI, USA. New York: IEEE Press, 2001: I.

[6] Viola P A, Jones M J. Rapid object detection using

a boosted cascade of simple features[C]// Computer Vision and Pattern Recognition, 2001.

[7] Viola P, Jones M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]// 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05). June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

[9] Ma Q, Zhu B, Cheng Z D, et al. Detection and recognition method of fast low-altitude unmanned aerial vehicle based on dual channel[J]. Acta Optica Sinica, 2019, 39(12): 1210002.
马旗, 朱斌, 程正东, 等. 基于双通道的快速低空无人机检测识别方法[J]. 光学学报, 2019, 39(12): 1210002.

[10] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition. June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[11] Girshick R. Fast R-CNN [J]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 1440-1448.

[12] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[13] Feng X Y, Mei W, Hu D S. Aerial target detection based on improved faster R-CNN[J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[15] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[16] Redmon J, Farhadi A. Yolov3: an incremental improvement [EB/OL]. (2018-04-08)[2020-04-02]. https://arxiv.org/abs/1804.02767.

[17] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[J]. Computer Vision - ECCV 2016, 2016.

[18] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[19] Li Y P, Hou L Y, Wang C. Moving objects detection in automatic driving based on YOLOv3 [J]. Computer Engineering and Design, 2019, 40(4): 1139-1144.
李云鹏, 侯凌燕, 王超. 基于 YOLOv3 的自动驾驶中运动目标检测[J]. 计算机工程与设计, 2019, 40(4): 1139-1144.

[20] Zou Z X, Shi Z W, Guo Y H, et al. Object detection in 20 years: a survey[EB/OL]. (2019-05-16)[2020-04-02]. https://arxiv.org/abs/1905.05055.

[21] Ma Q, Zhu B, Zhang H W, et al. Low-altitude UAV detection and recognition method based on optimized YOLOv3[J]. Laser & Optoelectronics Progress, 2019, 56(20): 201006.
马旗, 朱斌, 张宏伟, 等. 基于优化 YOLOv3 的低空无人机检测识别方法[J]. 激光与光电子学进展, 2019, 56(20): 201006.

[22] Ju M R, Luo H B, Wang Z B, et al. Improved YOLO V3 algorithm and its application in small target detection[J]. Acta Optica Sinica, 2019, 39(7): 0715004.
鞠默然, 罗海波, 王仲博, 等. 改进的 YOLO V3 算法及其在小目标检测中的应用[J]. 光学学报, 2019, 39(7): 0715004.

[23] Feichtenhofer C, Pinz A, Zisserman A. Detect to track and track to detect [C]// 2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 3057-3065.

[24] Liu R. Video object detection based on deep learning [D]. Guangzhou: South China University of Technology, 2019: 7-8.
刘荣. 基于深度学习的视频目标检测研究[D]. 广州: 华南理工大学, 2019: 7-8.

[25] Celebi M E, Kingravi H A, Vela P A. A comparative study of efficient initialization methods

for the k-means clustering algorithm [J]. Expert Systems with Applications, 2013, 40(1): 200-210.

[26] Lei J S, Jiang T, Wu K, et al. Robust K-means algorithm with automatically splitting and merging clusters and its applications for surveillance data[J]. Multimedia Tools and Applications, 2016, 75(19): 12043-12059.

[27] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [EB/OL]. (2017-04-19)[2020-04-02]. https://arxiv.org/abs/1612.03144.

[28] Zhang L, Mou X W. Chinese text clustering algorithm based on Canopy+K-means[J]. Library Tribune, 2018, 38(6): 113-119.
张琳, 牟向伟. 基于 Canopy+K-means 的中文文本聚类算法[J]. 图书馆论坛, 2018, 38(6): 113-119.

[29] Kang K, Li H S, Yan J J, et al. T-CNN: tubelets with convolutional neural networks for object detection from videos [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 2896-2907.