doi:10.3969/j.issn.1001-4616.2020.01.019

基于类加权 YOLO 网络的水下目标检测

朱世伟,杭仁龙,刘青山

(南京信息工程大学自动化学院,江苏 南京 210044)

[**摘要**] 由于水下目标检测面临着图像模糊、尺度多样化、复杂背景等问题,给水下目标检测应用带来很多挑战. 本文提出了一种基于类加权 YOLO 网络的水下目标检测方法,主要思想是在深度网络 YOLO 的基础上,构造了类加权损失函数,来平衡样本难易程度以获得更好的效果,并引入了目标框自适应维度聚类方法,进一步提升了检测性能. 实验结果表明,本文算法与传统的 YOLO 网络模型相比,在每幅图片包含近 20 个目标的密集目标检测任务中,能够将平均准确率从 71.2%提升至 74.1%,召回率由 71.1%提升到 78.3%.

[**关键词**] 水下目标,YOLO,类加权损失,自适应维度聚类

[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2020)01-0129-07

Underwater Object Detection Based on the Class-Weighted YOLO Net

Zhu Shiwei, Hang Renlong, Liu Qingshan

(College of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Underwater objects detection exist many issues, such as blur image, various object scales, complex background and so on. In this paper, we propose a class-weighted YOLO net for underwater object detection, in which a class-weighted loss is designed to balance sample of difficulty so as to acquire better effect. Moreover, a dimension adaptive clustering of object box is introduced to promote the detection performance. The experimental results show that the proposed method outperforms to the traditional YOLO net, with the increasing of the mAP from 71.2% to 74.1% and the recall from 71.1% to 78.3%, in the task of dense object detection which every image nearly contained 20 objects.

Key words: underwater object, YOLO, class-weighted loss, dimension adaptive clustering

水下目标检测任务是指对水下图像中包含的目标进行检测. 与传统目标检测任务不同,对于水下目标而言,由于光照、摄像机的抖动、复杂背景干扰、目标类型多样化等因素,目标检测的效果会受到影响. 此外海水对光线的吸收和散射特性,会造成水下图像颜色特征的缺失,使得目标检测任务变得更加困难.

由于水下目标检测任务的特殊性,已有的检测算法大多依赖图像的灰度信息. 例如,2002 年,Adriana 等人提出了一种基于无约束水下视频的人造目标检测算法[1],该算法主要利用轮廓信息完成目标检测,但检测速度较慢. 2005 年,上海交通大学图像处理与模式识别研究所和哈尔滨工程大学潜器与水下机器人实验室的王猛、杨杰等人提出了一种基于区域分割的水下目标检测算法[2],该算法不仅鲁棒性较好,检测速度也有了一定的提升. 但是当物体发生旋转或缩放时,检测效果会受到明显的影响. 2006 年,Rizon M 等人将 1961 年 Hu 提出的不变矩[3]应用于目标检测,较好地解决了平移和缩放问题,但在离散情况下,检测结果易受缩放因子的影响. 针对以上问题,2009 年,哈尔滨工程大学的张铭钧等人提出了一种基于不变矩的水下目标检测算法[4],该算法是利用最小交叉熵来确定阈值,可确保灰度信息的完整,从而实现图像的模糊增强;针对水下图像存在光照不均等问题,该算法利用灰度-梯度不变矩实现水下图像的分割,鲁棒性较好,并且召回率较高[5]. 2012 年,Gracias N 等人提出了利用视觉显著性进行水下视频检测和跟踪,将单帧图像与其相邻帧间的均值作差,再采用显著性算法进行目标检测,在速度上取得了比较大的突破,

收稿日期:2018-09-21.

基金项目:江苏省高校自然科学研究面上项目(18KJB520032)、江苏省青年基金项目(BK20180786).

通讯作者:刘青山,教授,博士生导师,研究方向:模式识别与计算机视觉. E-mail:qsliu@nuist.edu.cn

但精度上仍旧不能达到预期要求.

近年来,通过逐层学习的方式,深度学习,特别是卷积神经网络(convolutional neural networks,CNN),在目标检测领域取得了突破性进展. 2014 年,Ross Girshick 等人在 CVPR 大会上提出了 R-CNN[6](region based convolutional neural networks)网络,在 VOC2012 数据集上,将目标检测的平均准确率(mean average precision,mAP)提升了 30%,达到 53.3%. 在此基础上,Ross Girshick 等人提出了 Faster R-CNN[7]网络,让不同的"感兴趣"区域(region of interest,ROI)尽可能地共享了计算,并用基于锚点框(anchor box)的 RPN 网络(region proposal network)来产生 ROI,在优化了准确率的同时,检测速度也得到了很大的提升,但仍然没有达到实时目标检测任务的要求. 2016 年 Joseph Redmon 等人提出了基于回归的目标检测算法 YOLO[8](you only look once),同时回归目标的类别和边框,并于同年对网络进行了改进,提出了 YOLO v2[9],在 VOC2007 数据集上检测速度达到了 67fps,同时平均准确率达到了 76.8%,使得目标检测任务真正意义上达到了实时的速度. 但由于用来预测边框和类别的最后一层的特征图(feature map)的空间信息有限,YOLO v2 在小目标检测任务上表现较差. 2017 年,Tsung-Yi Lin 等人提出了 FPN[10](feature pyramid networks),同时利用低层特征的空间信息和高层特征的语义信息用于检测任务,取得了突出的成绩. 受此启发,Joseph Redmon 等人于 2018 年提出了 YOLO v3 网络,将不同层的特征图所提取的特征进行融合,明显提升了目标物体(尤其是小目标物体)的检测性能.

无论是传统的模式识别与图像检测算法,还是基于 YOLO 等深度学习的目标检测算法,对于背景模糊复杂、紧凑密集且高度重叠的目标的检测性能通常较差. 基于深度学习的水下目标检测算法,虽然在精度和速度上具有一定的优越性,但面对复杂的水下图像也有很多不可忽略的局限性. 由于水下目标形状尺度的多样性,基于锚点框的深度学习算法很难获得较高的召回率. 此外,水下目标形态差异较大,不同类型样本的特征的学习难度差异较大,这也会影响目标检测效果,增加模型的不稳定性. 为此,本文在深度学习算法 YOLO 的基础上,借鉴了传统的水下目标检测的方法,提出了目标框自适应维度聚类算法来提升召回率,并用类加权损失来权衡不同样本的难易程度,在平均准确率上取得了比较理想的提升.

1　类加权 YOLO 网络

如图 1 所示,本文基于 YOLO 网络的基本框架提出了类加权 YOLO 网络,用以权衡不同类别的难易程度. 为了增加模型的召回率,本文新定义了一个距离,用 *K*-means 算法对锚点框的形状与维度进行聚类,使其更加接近目标框的形状大小. 相对于原始的 YOLO 网络,本文的类加权 YOLO 网络在召回率和平均准确率上面都有了相应的提升.

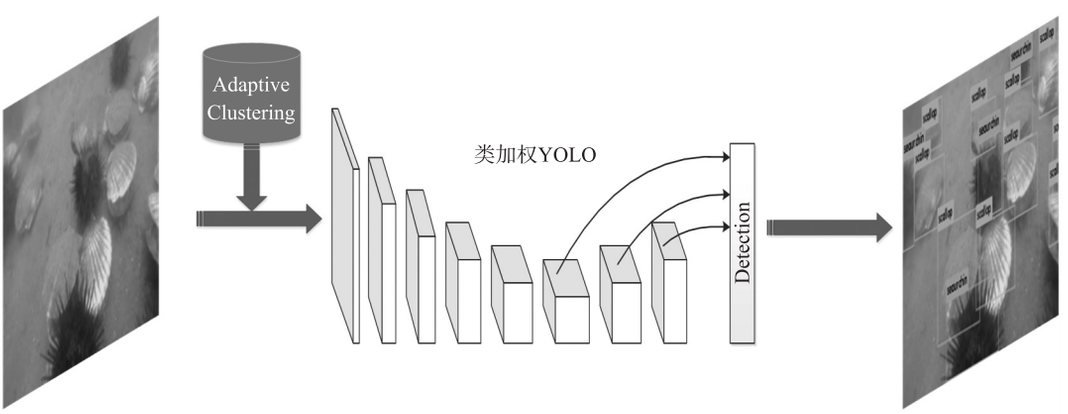

图 1　本文方法

Fig. 1　The method of this paper

1.1　网络结构

本文的基础网络模型和 YOLO v3 一样,是参考 ResNet[11]结构的一个新的分类网络 Darknet-53,Darknet-53 的结构以及准确率与 ResNet-101 相近,但在速度上有一定的优势. 表 1 是本文的基础网络模型 Darknet-53 的网络结构,其中 R E 表示一个残差单元,包含两个卷积层和一个残差层. Darknet-53 和

ResNet 一样,特征提取器是一个残差模型. 它包含 52 个卷积层和一个全连接层.

表 1　Darknet-53 网络参数

Table 1　Darknrt-53 parameters

数量	类型	滤波器个数	卷积核	特征图大小
	conv	32	3 * 3	256 * 256
	conv	64	3 * 3/2	128 * 128
1	R E			128 * 128
	conv	128	3 * 3/2	64 * 64
2	R E			64 * 64
	conv	256	3 * 3/2	32 * 32
8	R E			32 * 32
	conv	512	3 * 3/2	16 * 16
8	R E			16 * 16
	conv	1024	3 * 3/2	8 * 8
4	R E			8 * 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

1.2　目标框自适应维度聚类

2015 年 Ross Girshick 等人在 Faster R-CNN 算法中提出了锚点框的思想,成功地将寻找 ROI 这一步骤放入了卷积神经网络中进行,很大程度地减少了计算时间,之后的目标检测算法几乎都借鉴了这种思想. 当锚点框在特征图上滑动时,若它与人为标定的目标框(ground truth,GT)间的重叠率(intersection over union,IOU)大于一个阈值,该锚点框所覆盖的区域即为我们“感兴趣”的区域.

假设规定 IOU>0.5 时,锚点框覆盖的区域为“感兴趣”的区域,若形状大小差异太大,则锚点框在特征图上无论怎样滑动都不可能与目标框的重叠率满足这个要求,这对模型的召回率有很大的影响. 此外,如图 2 所示,锚点框的形状大小还影响着目标检测的速度和位置回归的精度,锚点框的形状大小与目标框相近时,回归网络的最终结果会更加准确.

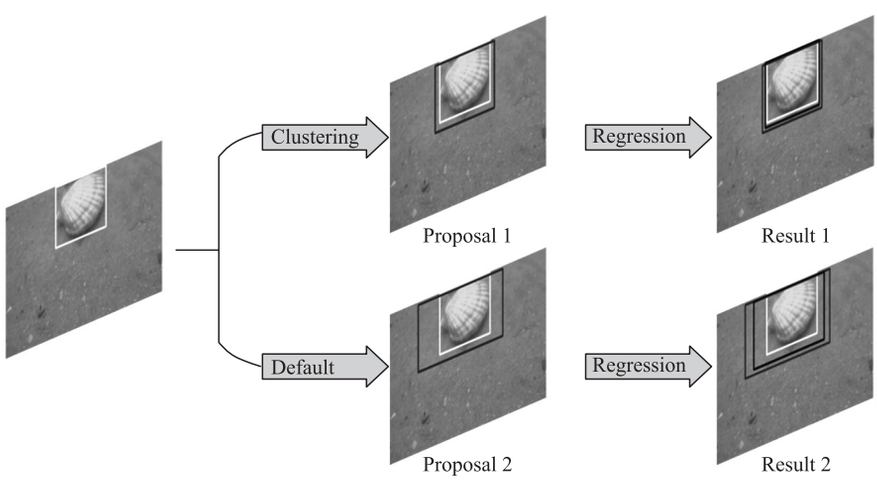

图 2　先验框对结果的影响

Fig. 2　The effect of prior boxes on results

Faster R-CNN 中锚点框的形状和大小是人为设置的,Joseph Redmon 等人在此基础上加以改进,对目标框进行了维度聚类,使得锚点框更好地适应相应数据集的目标框. 对目标框的聚类是为了让锚点框与目标框间的重叠率尽可能地大,YOLO v2 中对目标框聚类采用的是 *K*-means[12] 算法,利用欧式距离进行聚类,使得聚类的中心点与目标框归一化后所对应的点之间距离的均方差最小. 但是欧式距离并不能直接反映它们重叠率的大小. 对于较大的目标框,锚点框相对目标框的偏移对重叠率的影响不明显,而对于较小的目标框,锚点框即使偏移很小,重叠率也会发生很大的变化. 由此,本文提出了目标框自适应维度聚类,对于不同大小的目标框,在聚类的时候有着自适应的偏移. 要使得目标框的偏移量与其本身大小的自

适应,欧式距离是无法做到的,因此本文新定义了一个距离:

$$A_i=(x_i,y_i), \tag{1}$$

$$B_j=(m_j,n_j). \tag{2}$$

设 A_i 为第 i 个聚类中心,B_j 为第 j 个目标框宽高归一化后的值,D_{ij}为新定义的距离计算方法,用于计算第 j 个目标框与第 i 个聚类中心的距离.

$$D_{ij}=\sqrt{(\sqrt{m_j}-\sqrt{x_i})^2+(\sqrt{n_j}-\sqrt{y_i})^2}. \tag{3}$$

新定义的距离 D_{ij}能够根据目标框的大小自适应的调整偏移量,从而提高目标框与聚类后的锚点框间的平均重叠率.

1.3　类加权损失

目标检测的两个子任务分别是边框预测和类别预测. 原始 YOLO 算法为了完成这两个子任务,其损失函数设计包含了 3 个部分,分别是坐标预测、重叠率预测以及类别预测. 坐标预测和重叠率预测的损失函数都是在保证边框回归的准确度,在经过了锚点框的自适应维度聚类后,边框回归的准确度得到了相应的提升,另一个子任务,类别预测的准确度则变得更加重要. 类别预测的损失函数公式如下:

$$C_{\text{loss}}=\sum_{i=0}^{s^2}1_i^{\text{obj}}\sum_{c\in cls}(p_i(c)-\hat{p}_i(c))^2, \tag{4}$$

其中 S^2 是最后的特征图的大小,1_i^{obj} 是判断是否有目标的中心落在网格 i 中,若网格中包含目标的中心,就负责预测该目标的类别概率,cls 代表类别. 本文经过实验发现,水下目标数据集中 3 类目标的平均准确率相差很大. 扇贝和海胆的平均准确率比海参高出 20%.

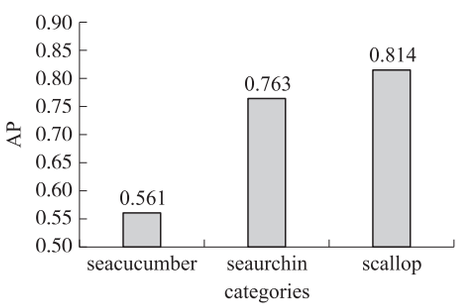

图 3　YOLO v3 下各类 AP 值

Fig. 3　The AP score of each categories with YOLO v3

由于海参形态多变,相互聚集重叠时会呈现多种姿态,因此,与扇贝和海胆不同,海参的特征学习难度较大. 为了缩小它们的差距,本文人为地给每一个类别加上一个权重,权衡了不同类别间样本的难易程度,修改后损失函数如下:

$$C_{\text{loss}}=\sum_{i=0}^{s^2}1_i^{\text{obj}}w_{c\in cls}\sum_{c\in cls}(p_i(c)-\hat{p}_i(c))^2. \tag{5}$$

通过调整 $\omega_{c\in cls}$,可以使得模型在边框预测和类别预测这两个子任务间寻找一个最佳的平衡点,使得算法获得最佳的检测效果.

2　实验数据及结果分析

2.1　实验数据及其预处理

本文所用的数据集是 2017 年水下机器人目标抓取大赛官方所提供的数据集,共有 3 类目标,分别是海参、海胆和扇贝. 该数据集由 5 段视频组成,每段视频拆成单帧图片后放到对应的文件夹中. 在训练的时候,将所有图片合并到一起. 共 17 655 张图片,334 366 个样本,除去部分无效数据,平均每张图片含有近 20 个样本,较为密集. 其中海参、扇贝、海胆的样本数据比大约是 1∶1∶1.5. 由于生活习性的不同,数据集中大部分都是单个类别聚集的图片,3 类同时出现在一张图片时,扇贝的样本数量要远远大于海参和海胆. 样本的不均匀分布同样给模型的训练增加了难度.

视频数据包含图片数据所不具备的时间信息和空间信息,时间信息对应的是每一张图片所属的帧的序号,空间信息则反映了相邻帧的图片变化较小,本文借助视频数据的时空信息,用 OpenCV 对数据进行了均值去燥. 在一段视频中,设第一帧各像素点的灰度值为 $A_1(x,y)$,第 i 帧各像素点的灰度值为 $A_i(x,y)$,均值去燥即截取视频中相邻帧求均值,这样可以抑制单帧出现异常噪声. 公式如下:

$$A_i(x,y)=\frac{1}{5}\sum_{l=i-2}^{i+2}A_l(x,y), \tag{6}$$

在一段连续的图片序列中,不改变起始两帧和结束两帧,对其它任意帧图片取其本身以及该帧的前后各两帧图片,对这 5 帧图片的像素点求均值,由于前后两帧间待检目标的变化忽略不计,而背景噪声却不

可控制,经过这样的处理后得到的图片对随机的噪声有了很好的抑制,每张图片结合了相邻帧的帧间信息,也使得训练的模型泛化能力更强,在测试集上未对数据进行处理,模型也能够很好地完成检测任务,同时 mAP 相较于未处理过的数据集的检测结果提升了 0.6%.

图 4 是随机选取的均值去噪的对比图,左边是原图,由于水下图像有较多不可控的随机噪声,这些噪声产生的因素包括水中的悬浮物,光源的位置及强弱等,会对水下目标造成不可忽略的干扰. 右边是处理后的图片,由于结合了前后 5 帧的信息,随机性的噪声得到了一定的抑制,图片相对于原图要更加清晰.

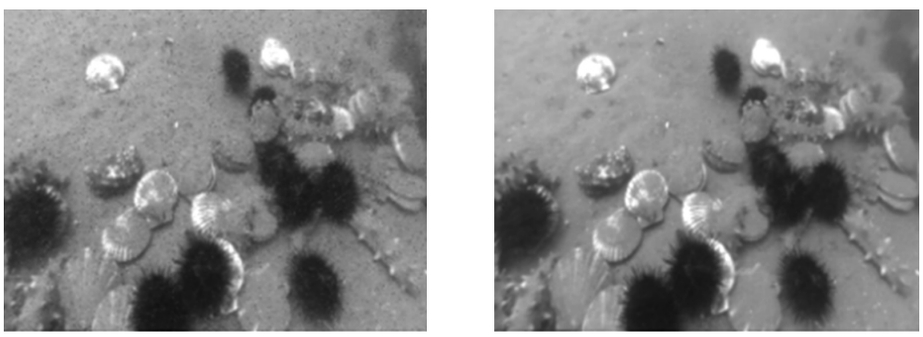

图 4　去噪效果

Fig. 4　The effect of noise reduction

2.2　网络训练

分类网络的预训练是目标检测中的重要环节,分类网络提取特征的能力和速度直接影响着目标检测的效果,主流的网络都会在 ImageNet 上进行预训练,但由于全连接的限制,输入尺寸会被统一调整为固定大小,这会使得模型对不同分辨率的泛化能力减弱. 本文针对预训练阶段的不足做了一些改进,先用 ImageNet 数据集对 Darknet-53 进行预训练,再从数据集中选取一些比较清晰且目标较少的数据微调 Darknet-53,使网络适应数据集的特征,当这一轮训练逐渐收敛时,再选用数据集中分配好的训练集进行训练. 本文采用的类加权 YOLO 网络不包含全连接层,因此可以任意改变输入图像的尺寸,为了让模型适应不同分辨率的输入,本文借鉴多尺度输入的方法对检测网络进行训练,在训练过程中,每隔 10 个批次随机改变网络的输入尺寸. 由于类加权 YOLO 网络的下采样因子为 32,因此相应调整输入图片的尺寸为 32 的倍数,随机调整输入图片大小的范围如下所示:

$$S=32(10+x), \tag{7}$$

其中 x 是 0 到 9 之间的任意整数.

这种多尺度训练会使得模型对不同分辨率的图片有着良好的适应能力,对于高分辨率的输入,检测速度相对变慢,但准确率较高,对于低分辨率的输入,准确率有所下降,但检测速度更快.

2.3　实验结果分析

为了验证目标框自适应维度聚类对结果的影响,本文从召回率和平均准确率两个角度对比分析,结果如图 5.

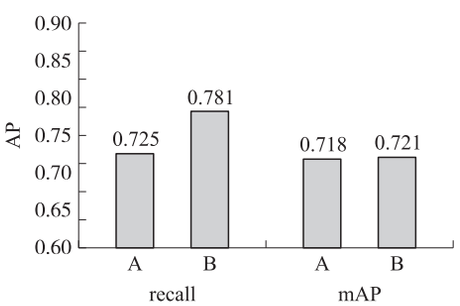

图 5　加权维度聚类的效果

Fig. 5　The effect after weighted dimension clustering

如图 5,其中角标 A 代表用普通的聚类方法,角标 B 是用本文定义的距离进行目标框维度聚类的方法. 可以看出,由于新定义的距离能够依据目标框的大小自适应的调整对应的锚点框的偏移量,聚类后的锚点框与目标框的形状和大小更加相近,这使得训练的正样本数量相应地有所增加,模型的 mAP 也因此有了一定的提升,此外,目标框的自适应维度聚类也让测试时模型的召回率得到了很大的提升. 最终的实验结果显示,B 方法相对于 A 方法,模型的 mAP 提升了 0.3%,召回率提升了 7.6%.

本文在实验过程中将类加权 YOLO 网络和 YOLO v3 网络做了详细的对比,在训练参数控制一致的情况下,YOLO v3 模型测试中海参的表现与海胆和扇贝的差距很大,并且在默认的 IOU>0.5 视为召回的情

况下,模型的召回率只有 71.1%. 在经过数据预处理(均值去燥)后,mAP 和召回率都有了些许的上升. 在这基础上加了对目标框自适应维度聚类,模型的召回率上升了 5.6%,并且 mAP 值也有所上升. 由于现有的大部分目标检测算法都是利用锚点框来寻找感兴趣的区域,这在理论上证明了该方法的通用性. 此外,在加权维度聚类的基础上,本文根据实验结果的异常表现,在类别间 AP 值相差较大的前提下,提出的类加权损失来权衡不同类别样本的难易程度,实验结果显示,经过类加权损失处理后,虽然海胆和扇贝的 AP 值略微下降,但海参的 AP 值有了很大的提升,最终本文的类加权 YOLO 网络与 YOLO v3 网络相比,mAP 提升了 2.9%,召回率提升了 7.2%. 表 2 是详细的实验结果对比.

表 2　本文方法的实验结果

Table 2　Experimental results of our method

	YOLO v3			类加权 YOLO
均值去燥		√	√	√
自适应维度聚类			√	√
类加权损失				√
海参(AP)	56.1	57.2	58.3	67.2
海胆(AP)	76.3	76.7	76.4	73.7
扇贝(AP)	81.4	81.7	81.7	81.5
MAP	71.2	71.8	72.1	74.1
RECALL	71.1	72.5	78.1	78.3

为了验证本文提出的算法的有效性,本文还在其他几个经典的目标检测算法上训练和测试了该数据集,分别在召回率和平均准确率上与本文算法做了对比,如表 3.

表 3　不同模型的实验结果

Table 3　Experimental results of different models

	Faster RCNN	SSD	YOLOv3	类加权 YOLO
平均准确率	64.3	68.7	71.2	74.1
召回率	72.2	66.2	71.1	78.3

可以看出本文提出的算法无论是在召回率还是平均准确率上,都是要优于其他当前顶尖的基于深度学习的目标检测算法. 下图为测试集上随机选取的部分图片的测试结果.

如图 6 所示,本文的算法可以检测出大部分的目标物体,但由于经过了目标框自适应维度聚类,对于个别目标框形状特殊的目标物体(如图 6 中下部分的海参),本文的算法并不能很好地将其召回. 图 7 中可以看出,在背景比较模糊,目标高度密集的情况下本文的方法仍存在一定的误检,这也是当前目标检测算法共同面临的难题.

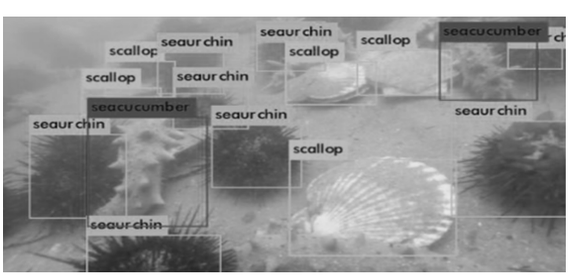

图 6　部分测试结果 1

Fig. 6　Partial test results 1

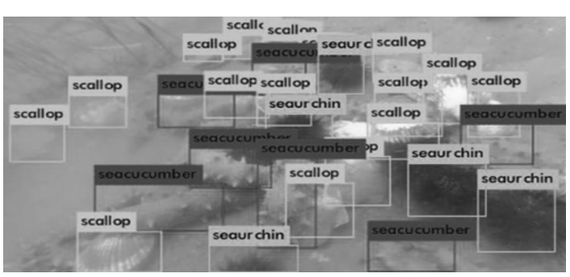

图 7　部分测试结果 2

Fig. 7　Partial test results 2

3　结论

目标检测是计算机视觉领域的重要研究方向,水下目标检测由于其特殊的环境因素干扰,现有的目标检测算法难以得到满意的效果,本文借鉴了传统的水下目标检测算法,首先对数据进行了均值去燥,接着对目标框进行了自适应维度聚类,并提出了类加权 YOLO 网络用来训练和测试目标数据集,最终平均准确率与 YOLO v3 模型相比提升了近 3%,召回率提升了 7%. 但由于每张图片包含的目标数目太多,在实际的测试中速度仅仅达到 12 fps 左右,这还不能达到实时目标检测的要求,下一步研究将借鉴 TPN[13] 的思想,重点放在利用视频数据的帧间时空信息来加快检测速度.

[参考文献]

[1] OLMOS A,TRUCCO E. Detecting man-made objects in unconstrained subsea videos[C]//IEEE International Conference on British Machine Vision. British Publisher:BMVA Press,2002:517-526.

[2] 王猛,杨杰,白洪亮. 基于区域分割的水下目标实时识别系统[J]. 计算机仿真,2005,22(8):101-105.

[3] RIZON M,HANIZA Y,PUTEH S,et al. Object detedtion using geometric invariant moment[J]. American journal of applied sciences,2006,3(6):1876-1878.

[4] 张铭钧,尚云超,杨杰. 基于灰度-梯度不变矩的水下目标识别系统[J]. 哈尔滨工程大学学报,2009,30(6):653-657.

[5] SHIHAVUDDIN A S M,GRACIAS N,GARCIA R. Online sunflicker removal using dynamic texture prediction[C]//International Conference on Computer Vision Theory and Applications. Barcelona,Spain,2012:161-167.

[6] GIRSHICK R,DONAHUE J,DARREL1 T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. IEEE transactions pattern analysis and machine intelligence,2015,38(1):142-158.

[7] REN S,HE K,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,2015,39(6):1137-1149.

[8] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified,real-time object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,America,2016:779-788.

[9] REDMON J,FARHADI A. YOLO9000:Better,Faster,Stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition. Hawaii,America,2017:7263-7271.

[10] LIN T Y,DOLLAR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Hawaii,America,2016:2117-2125.

[11] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//IEEE Conference Computer Vision and Pattern Recognition. Las Vegas,America,2016:770-778.

[12] 乔小妮,张明新,史变霞. 一种基于密度的 *K*-means 算法[J]. 电脑开发与应用,2008,21(10):9-11.

[13] KANG K,LI H,XIAO T,et al. Object detection in video with tubelet proposal networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Hawaii,America,2017:727-735.

[责任编辑:陆炳新]