文章编号：1005-6734(2020)01-0129-05 doi.10.13695/j.cnki.12-1222/o3.2020.01.020

一种基于 YOLO-V3 算法的水下目标识别跟踪方法

徐建华，豆毅庚，郑亚山
（北京理工大学，北京 100081）

摘要：为协助水下平台完成自主拍摄任务，针对水中成像模糊，物体多自由度运动的特点，提出一种基于 YOLO-V3 算法的目标识别模型。通过降采样重组，多级融合、优化聚类候选框、重新定义损失函数等方式优化网络结构，提高了目标识别的准确率，同时提升算法的计算速度。将具有旋转不变性的特征描述应用于跟踪水中多自由度运动的物体，通过评价结果修正跟踪状态。实验表明，该方法能够自主识别和跟踪目标，具有自适应能力，对输入像素为 416*416 的图片，处理速度达到 15 帧/秒以上，置信度为 0.5 时的平均准确度值达到 75.1，满足实时性和准确性要求。

关 键 词：水下平台；目标识别；目标跟踪；YOLO-V3 算法；多自由度
中图分类号：TP275 **文献标志码**：A

Underwater target recognition and tracking method based on YOLO-V3 algorithm

XU Jianhua, DOU Yigeng, ZHENG Yashan
(Institute of Technology, Beijing 100081, China)

Abstract: In order to assist the underwater platform to complete the autonomous shooting task, a target recognition model based on the YOLO-V3 algorithm is proposed. The network structure is optimized by the means of down-sampling recombination, multi-stage fusion, optimization of clustering candidate box, and redefinition of loss functions, etc, which improves the accuracy of target recognition and the calculation speed of the algorithm. The feature description method with rotational invariance is applied to track the motion of multi-degree of freedom objects in water, and the tracking state is corrected by evaluation results. Experiments show that the method can identify and track the target autonomously and has adaptive ability. For the image with input pixel of 416*416, the processing speed reaches more than 15 frames per second and the mean Average Precision (mAP) reaches 75.1 when the confidence degree is 0.5, which meets the real-time and accuracy requirements.

Key words: underwater platform; target recognition; tracking; YOLO-V3 algorithm; multiple degrees of freedom

近年来，随着水下活动需求的丰富，水下移动平台的控制方式也由最初的手动遥控模式逐渐转变为自主移动模式，例如自主跟踪拍摄模式。采用图像信息跟踪的方法在小型浅水平台有较好的实用性。常用电磁跟踪信号在传播过程中会快速衰减，声波信号则受限于平台的体积。相比之下，采用图像信息跟踪的方法表现出更强的实用性。

目标识别与跟踪是视觉应用领域重要的分支。基于视觉的目标识别方式主要包括传统的机器学习算法和深度学习算法[1-4]。传统方式如 Haar 特征与 AdaBoost 分类器组合而成的识别算法，对于背景简单，目标区域清晰，轮廓特征明显的场景效果良好[5]。但由于水对光线的吸收、散射、漫射等作用，水下拍摄的图像往往存在清晰度差，边缘锐度低，整体亮度低，局部折射光照强度过高等问题。目标的颜色强度会随着水深不断减弱，轮廓信息也会受到漂浮物、波

收稿日期：2019-11-05；**修回日期**：2020-02-15
基金项目：装备发展预研项目（151741417030103）
作者简介：徐建华（1975—），男，讲师，从事组合导航研究。E-mail： xujianhua@bit.edu.cn

纹、气泡的影响，传统的识别目标方式会造成误识别的情况。深度学习网络方法则是通过多次卷积计算提取图像的特征，过程中充分利用像素信息，以此来提高检测器的性能[6]，能够应用于水下目标的识别。

基于深度学习的 YOLO-V3 算法网络将图像分割成 S×S 的网格，不同网格只负责其对应区域的物体的识别，减少重叠识别，提高检测速度，该算法因其快速和准确而近年来被广泛使用[7,8]。本文在 YOLO-V3 原算法的基础上使用重组成与多级融合的方法进行特征提取，使得对于水下图像的检测效果明显提升。在此基础上，还针对检测结果使用基于旋转不变性的特征跟踪目标[9-11]，并通过评价结果判断是否重新进行识别，以增强算法的抗干扰能力。

1 目标检测

基于 YOLO-V3 网络的检测方法将候选框提取、特征提取、目标分类、目标定位统一于一个神经网络中。神经网络可直接从图像中提取候选区域，通过整幅图像特征来预测目标位置和置信度。

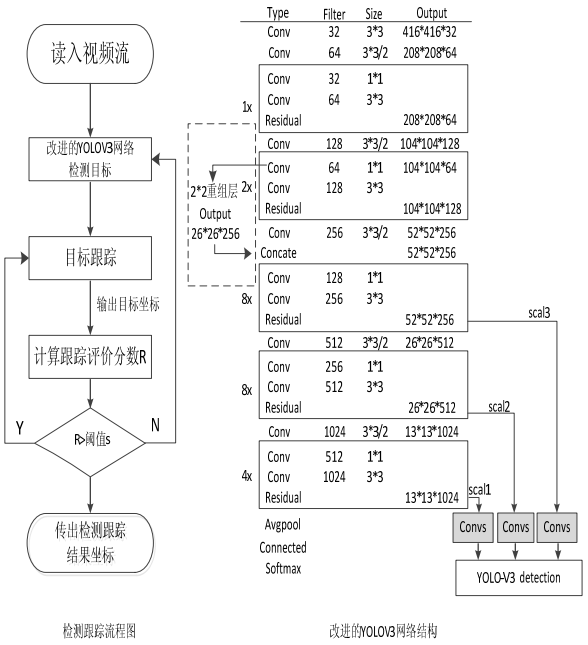

图 1 算法流程图

Fig.1 Architecture of the algorithm

1.1 重组层与多级融合

YOLO-V3 网络主要通过卷积和池化操作实现对图像特征的提取，本文提出的目标检测跟踪模型的流程图和改进的 YOLO-V3 目标检测算法流程图如图 1 所示。通常，水下拍摄画面模糊性较高，水下目标的局部性特征比较明显，为了增强 YOLO-V3 检测网络对于水下场景的适用性，提高网络对水下目标检测的精确度，本文使用重组层代替传统的卷积和池化操作进行特征提取；此外，在网络结构中还加入多级融合的思想，使得网络充分利用水下目标的局部特征，以增加算法的鲁棒性。重组层的结构示意图如图 2 所示，其将每个通道上的 2*2 图像块中的 4 个像素点拆解排列成 4 通道的 1*1 图像块，此方法与传统的池化相比，极大程度地保留了像素中的局部细节，并且实现了图像特征降采样的过程。本文还将重组层的输出特征图与同步进行的卷积池化的输出特征图进行多级融合，生成叠加后的特征图。

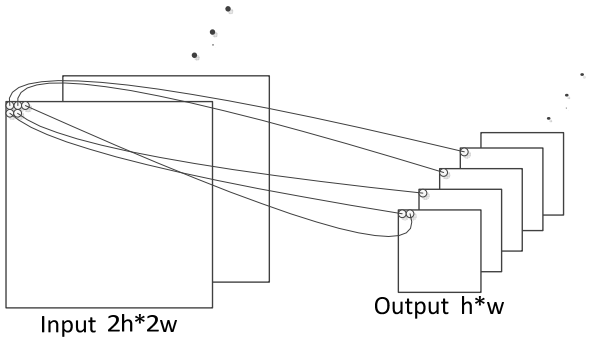

图 2 重组层示意图

Fig. 2 Restructuring layer diagram

1.2 损失函数

由于拍摄时的角度变化，目标在水中旋转等原因，同一类物体会出现不常见的长宽比。识别过程中，原始的 YOLO-V3 网络的泛化能力偏弱，会出现识别不到目标或错误识别的情况。为弥补网络无法检测物体较大角度变化的缺点，文章将网络原型中长宽的损失，转化为区域框对角线的损失，重新定义模型的损失函数：

$$
\begin{aligned}
loss = {} & \lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[\sqrt{\left|\left(w^2+h^2\right)-\left(w'^2+h'^2\right)\right|}\right]+\\
& \lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[\sqrt{\left|\left(x_i-x_i'\right)^2+\left(y_i-y_i'\right)^2\right|}\right]+\\
& \lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{obj}\left(c_i-c_i'\right)^2+\lambda_{noobj}\sum_{i=0}^{s^2}\sum_{j=0}^{B}1_{ij}^{nobj}\left(c_i-c_i'\right)^2+\\
& \lambda_{coord}\sum_{i=0}^{s^2}1_{ij}^{obj}\sum_{c\in classes}\left(p_i-p_i'\right)^2
\end{aligned}
\tag{1}
$$

式中，C_i 表示目标分类，p_i 表示类别概率。对于待检测区域，存在目标的置信度设置为 1；不存在目标的区域置信度为 0。训练时权重 $\lambda_{coord}=\lambda_{noobj}=0.5$。

1.3 网络训练

选取泳池中于不同距离、不同角度采集人的不同

泳姿图片作为训练集。根据拍摄角度将训练集进行分类：前向(front)、侧向(side)、后向(back)，并利用 LabelImg 工具进行标注。训练集图像共包含 3 类目标共 2000 张；

网络训练过程是通过不断调整预测框，使其接近真实框的过程。在训练开始前，需要设定初始的候选框的大小及数量。合适的初始框不仅能加快网络的训练过程，还能增加识别算法的准确率。文章针对水下目标的特点，引入 K 均值聚类的算法，生成各类目标的最相近初始框。聚类的目标函数：

$$\min \sum_{i=0} \sum_{j=0} \left[1 - IOU\left(Box[i], Box[j]\right)\right] \tag{2}$$

式中，IOU 表示聚类得到的结果 $Box[i]$ 与真实值 $Box[j]$ 之间的交并比。

2　目标跟踪

2.1　特征描述

常用的图像描述特征如 HOG 特征、SIFT 特征对旋转后目标等的描述都不够准确，易造成误匹配的情况。本文设计了一种具有旋转不变性的特征描述用于水下目标的跟踪，较好解决了上述问题。

如图 3 所示，目标中心点 (x, y)，由中心向外扩展半径 r，以圆心水平方向为 x 轴，与 x 轴夹角 θ 的灰度值表示为：

$$I(r,\theta),\ r \in [0, R],\ \theta \in [0, 360] \tag{3}$$

其中 R 表示目标区域最大半径。

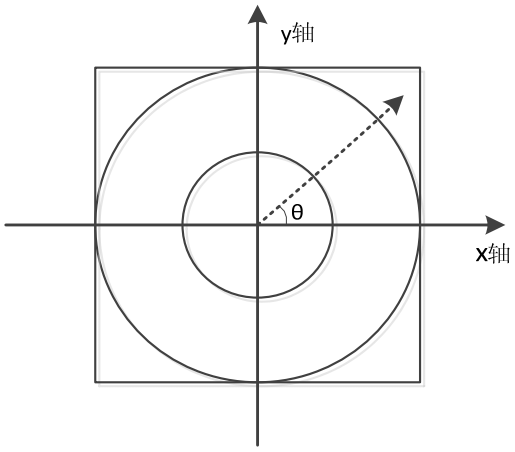

图 3　旋转不变性特征

Fig. 3　Rotation invariant feature

为减小物体运动时产生的波纹影响，增强算法的抗干扰能力，本文采用沿半径方向计算特征描述梯度的方法：

$$I_k = I(r+1,\theta) - I(r,\theta), k = 0, 1 \cdots r-1 \tag{4}$$

综合得到旋转不变性特征描述：

$$S_k = \sum_{i=0}^{\theta} I_k \tag{5}$$

2.2　跟踪实现

通过 2.1 节所述的特征描述方式，可以将当前帧中的目标表示为 $1\times(r-1)$ 的梯度行向量 $S=[S_0, S_1, .. S_{r-1}]$。在获取下一帧图像后，以上一帧中目标中心点为基准，在目标的原始区域 2.5 倍范围内进行目标检索，计算目标的梯度行向量 S 与待检区域梯度行向量 S' 的余弦相关性。检索结果：

$$Q(x', y') = \frac{s \times s'}{|s||s'|} = \frac{\sum_{k=0}^{R-1} S_k \times S'_k}{\sqrt{\sum_{k=0}^{R} (S_k)^2} \times \sqrt{\sum_{k=0}^{R} (S'_k)^2}} \tag{6}$$

其中 (x', y') 表示待检区域中可能是目标中心点的坐标。遍历待检区域后，$Q(x', y')$ 极大值位置即新的一帧中目标中心所在，记录 $Q_{\max}$ 并更新目标模板 S_Q，继续采用相同的搜索策略以实现连续帧中的目标跟踪。

2.3　跟踪结果评价

跟踪过程中，对跟踪结果准确性的判断主要包括两个方面：相邻帧中目标的相似程度和移动距离。前者保证识别的准确性，后者保证目标在视频流中的连续性。由此确定跟踪结果评价函数：

$$f(Q_{\max}) = \lambda\left(1 - \frac{\sqrt{(x-x')^2 + (y-y')^2} \times \tan\frac{a}{2} \times d \times F}{\frac{T}{2} \times v}\right) + \mu Q_{\max} \tag{7}$$

式中，α 表示相机视角，d 表示目标与相机焦点之间的距离，T 表示相机分辨率，F 表示视频帧率，v 表示物体速度，λ 和 μ 表示权重。设定一个阈值 s，当 $f(Q_{\max}) > s$，跟踪结果准确，否则重新进行目标识别。本文中 $\lambda = \mu = 0.5$，阈值 s 取为 0.7。

3　实验及结果分析

将本文设计的算法布署到研扬科技 UP Squared 主板上进行验证试验。主板配置：CPU Intel Pentium TM N4200，4GB RAM 缓存，64GB eMMC 内存；Ubuntu16.04，64 位操作系统；AI Core 深度学习网络

加速模块，通过 Mini-PCIe 接口连接 UP Squared 主板；高清摄像头，分辨率 1980×1080。试验场地为 20×50 m^2 的泳池，运动员在泳池内随机运动。

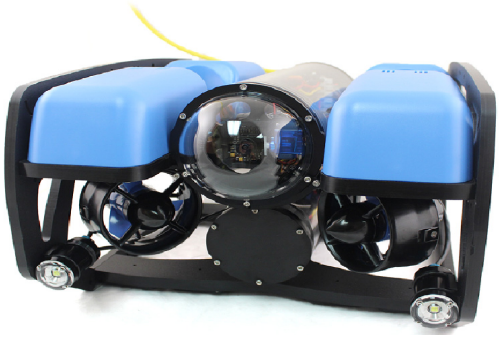

图 4　水下移动平台

Fig. 4　Underwater mobile platform

3.1　目标检测跟踪效果评价

使用改进的 YOLO-V3 的目标检测算法以及利用具有旋转不变特性的特征描述对水下目标进行识别跟踪，取得了良好的效果。效果如图 5 所示，各图中黄色框表示识别与跟踪到的游泳运动员，改进后算法的鲁棒性高。此外，为了更直观比较本文改进 YOLO-V3 算法的性能，使用控制变量的对比方法，分别使用输入图像尺寸为 608*608 和 416*416 像素的图片，采用 YOLO-V3、YOLO-V3-tiny 和改进的 YOLO-V3 算法进行训练和检测，训练图片为具有标注的水下运动员运动图像 2000 张（其中 60%训练集，30%测试集，10%验证集），训练 20000 个 step 后使用研扬科技 UP Squared 嵌入式主板进行检测与评价。表 1 所示为使用置信度 0.5 进行筛选后的结果的性能指标。可见，改进的 YOLO-V3 的方法相比于其他方法在平均准确度值（mean Average Precision，mAP）上均有提升，并且其速度可以达到 15 帧/秒，能够满足水下机器人识别跟踪任务的需求。

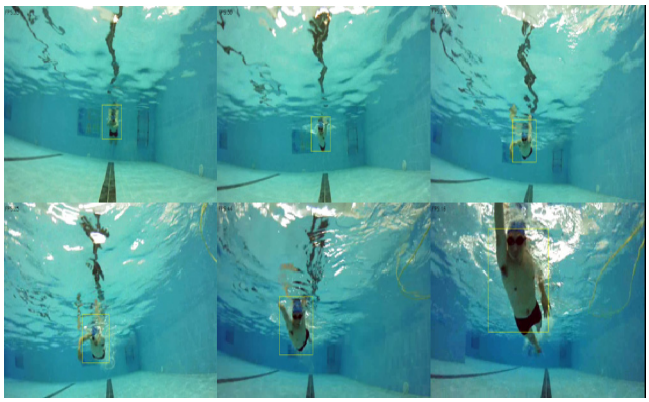

图 5　改进 YOLO-V3 目标检测跟踪效果图

Fig.5　Improved YOLO-V3 t detection and tracking result

表 1　置信度 0.5 时各算法检测结果

Tab.1 Detection results of each algorithm at a confidence level of 0.5

算法-输入图像像素	mAP（%）	平均速度 （帧/秒）
YOLO-V3-608	70.3	11.7
YOLO-V3-416	68.5	14.1
YOLO-V3-tiny-608	61.2	20.7
YOLO-V3-tiny-416	56.4	25.2
改进 YOLO-V3-608	76.7	10.5
改进 YOLO-V3-416	75.1	15.2

由表 1 可知，在一定范围内利用重组与多级融合的方式能够降低识别过程中的误差率，提高算法的准确性。但被替代的卷积层数较多时，虽然识别速度有所提升，提取的深度学习特征不够明确，准确性下降明显，在实际应用中，需要进行取舍，因此本文采用了 2 次重组层与多级融合的机制来修改原网络结构。

3.2　平台跟踪实验

平台跟踪路线结果如图 6 所示，‘.’状线条表示运动员运动路径，‘+’状线条表示移动平台跟踪路径。由图 6 可以看出，运动平台初始在运动员后方约 2 米，当运动员开始以‘S’形路线运动时，移动平台在本文所使用的检测及跟踪算法的驱动下能够与运动员始终保持在 2 米距离内，完成了跟踪拍摄任务。

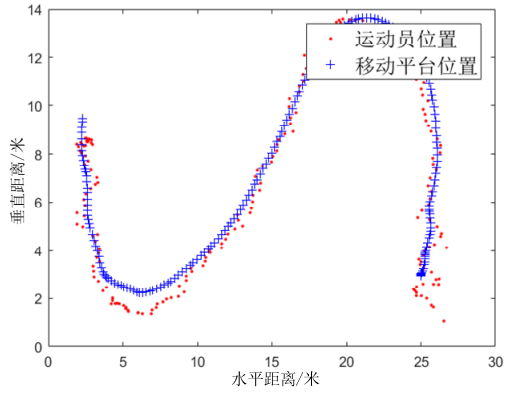

图 6　跟踪结果

Fig.6　Tracking result

跟踪结果评价曲线如图 7 所示，在跟踪过程中，使用本文所提出的跟踪结果评价方法对 60 米不规则运动的跟踪结果进行评价，由图 7 可得跟踪结果评价在大多数时间都是大于阈值 0.7，表现出了本文提出的跟踪算法的稳定性。当跟踪结果评价小于阈值 0.7 时，

算法会自动做出调整，重新对目标进行识别定位，以保证平台对目标的跟踪效果。总体看来，本文提出的跟踪算法对于运动员的不规则运动表现出一定的适应能力。

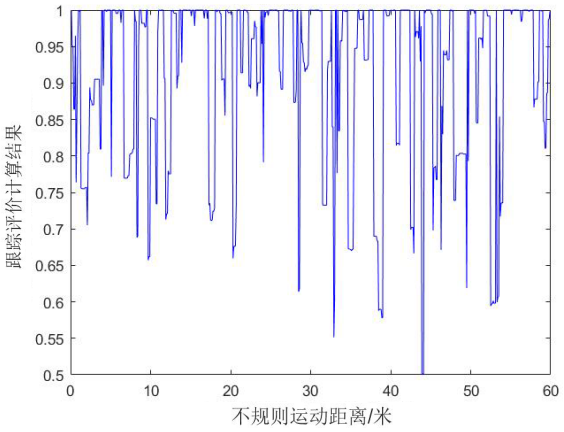

图 7　跟踪评价曲线

Fig.7　Tracking evaluation curve

4　结 论

针对水中图像成像模糊、水下物体运动多自由度的特点，本文利用深度学习方法和基于旋转不变性的特征描述，提出了一种基于 YOLO-V3 算法的水下目标识别跟踪方法，实现了对水中目标的识别、定位与跟踪。

该方法在 YOLO-V3 原算法的基础上使用重组成与多级融合的方法进行特征提取，其目标检测模型在嵌入式平台上的检测速度达到 15 帧/秒；当置信度为 0.5 时，mAP 值达到 75.1，水下图像的检测效果明显提升。

针对检测结果使用基于旋转不变性的特征跟踪目标，对跟踪情况做出实时评判，当跟踪情况不佳时自适应地调用检测算法进行辅助，增强了算法的抗干扰能力。实验结果表明其跟踪评价分数稳定在 0.7 以上，取得了较好的跟踪效果。因此，该方法针对水下作业，尤其是水下运动员跟拍等任务具有较好的工程应用价值。

参考文献（References）：

[1] 尹宏鹏，陈波，柴毅，等．基于视觉的目标检测与跟踪综述[J]．自动化学报，2016，42(10)：1466-1489．
Yin H, Chen B, Chai Y, et al. Vision-based object detection and tracking: a review[J]. Acta Automatica Sinica, 2016, 42(10): 1466-1489.

[2] Han B. Learning Multi-Domain Convolutional neural networks for visual tracking[J]. CVPR, 2016(7): 4293－4302.

[3] Diaz R, Hallman S, Fowlkers C. Detecting dynamic objects with multi-view background subtraction[J]. Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 273－280.

[4] 马娟娟，潘全，梁彦，等. 基于深度优先随机森林分类器的目标检测[J]. 中国惯性技术学报，2018，26(4)：518-523.
Ma J, Pan Q, Liang Y, et al. Object detection for depth-first random forest classifier[J]. Journal of Chinese Inertial Technology, 2018, 26(4): 518-523.

[5] 朱文青，刘艳，卞乐，等．基于生成式模型的目标跟踪方法综述[J]．微处理机，2017，38(1)：41-47．
ZHU W, LIU Y, BIAN L, et al. Survey on Object Tracking Method Base on Generative Model[J]. Microprocessors, 2017, 38(1): 41-47.

[6] Karpathy A, Toderoco G, Shetty．Large sacal video classification with convolutional neural networks[C]// IEEE Conference on CVPR, 2014: 1725－1732．

[7] Redmon, Divvala, Girshick, You only look once: Unified, real-time object detection[J]. Proceedings of IEEE Conference on CVPR. Washington D.C.USA: IEEE Computer Society, 2016: 779－788.

[8] Redmon, Farhadi. YOLO9000: Better, Faster, Stronger[C]//Proceedings of IEEE Conference on CVPR. Washington D.C, USA: IEEE Computer Society, 2017: 6517－6525．

[9] Simoserra E, Trulls E, Ferraz L, et al. Discriminative learning of deep convolutional feature point descriptors[C]//IEEE International Conference on Computer Vision, 2016: 118-126．

[10] Li Y, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration[J]．European Conference on Computer Vision, 2014(9): 254-265．

[11] Henriques J.F, Rui C, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 37: 583-596.