计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(6):197−202 [doi: 10.15888/j.cnki.csa.007507]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

# 基于改进 YOLOv3 的人体行为检测①

李啸天[1], 黄　进[1], 李剑波[2], 杨　旭[1], 秦泽宇[1], 付国栋[1]

[1](西南交通大学 电气工程学院, 成都 611756)
[2](西南交通大学 信息科学与技术学院, 成都 611756)
通讯作者: 黄　进, E-mail: 396341096@qq.com

**摘　要**: 针对人体行为检测中相同行为差异大, 不同行为相似度高, 以及视觉角度、遮挡、不能实时检测等问题, 提出 Hierarchical Bilinear-YOLOv3 人体行为检测网络. 该网络采用 YOLOv3 在 3 个不同尺度上进行预测, 抽取 YOLOv3 金字塔特征提取网络中特定层作为 Hierarchical Bilinear 的输入, 捕获特征图的层间局部特征关系, 并在 3 个不同尺度上进行预测, 最后将 YOLOv3 和 Hierarchical Bilinear 两种预测结果融合. 实验结果显示, 改进后的模型相比于原网络仅增加了少量参数, 在保证检测效率的同时提高原算法的检测精度, 并在一定程度上优于当前行为检测算法.

**关键词**: 人体行为检测; YOLOv3 算法; Hierarchical Bilinear-YOLOv3 网络; 特征提取

引用格式: 李啸天,黄进,李剑波,杨旭,秦泽宇,付国栋.基于改进 YOLOv3 的人体行为检测.计算机系统应用,2021,30(6):197–202. http://www.c-s-a.org.cn/1003-3254/7507.html

## Human Behavior Detection Based on Improved YOLOv3

LI Xiao-Tian[1], HUANG Jin[1], LI Jian-Bo[2], YANG Xu[1], QIN Ze-Yu[1], FU Guo-Dong[1]

[1](School of Electrical Engineering, Southwest Jiaotong University, Chengdu 611756, China)
[2](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

**Abstract**: This study proposes a neural network named Hierarchical Bilinear-YOLOv3 for human behavior detection due to a large disparity in the same behavior and high resemblance between different behaviors in human behavior detection, as well as problems such as visual angle, occlusion, and incapability of continuous real-time monitoring. YOLOv3 is first designed for prediction on three scales, and certain layers in its feature pyramid networks are used as inputs for Hierarchical Bilinear to capture local feature relationships between layers in the feature maps and predict the results on three scales. The integrated results of both YOLOv3 and Hierarchical Bilinear show that the improved network only adds a few parameters compared to the original one. It improves the detection accuracy of the original algorithm without lowering the detection efficiency and thus is superior to the current behavior detection algorithms.

**Key words**: human behavior detection; YOLOv3 algorithm; Hierarchical Bilinear-YOLOv3 network; feature extraction

人体行为检测是计算机视觉领域的热点之一, 其目的是检测图片或者视频中的人体行为. 传统的检测算法可以分为 3 个步骤: 首先采用多尺度、不同长宽比的滑动窗口[1] 选取图片中感兴趣区域. 其次, 从选取区域中提取 SIFT[2]、HOG[3] 以及 Haar-like[4] 等人工特征. 最后, 对选取的特征进行分类. 由于滑动窗口会产生大量冗余窗口计算量大, 人工特征进行分类只能提取物体的部分特征, 鲁棒性较差, 传统的目标检测算法

① 基金项目: 成都市科学技术局项目 (2018-YF05-01424-GX)
Foundation item: Project of Science and Technology Bureau, Chengdu Municipality (2018-YF05-01424-GX)
收稿时间: 2019-12-16; 修改时间: 2020-01-14; 采用时间: 2020-01-21; csa 在线出版时间: 2021-06-01

有待改进.

近几年来, 基于深度学习的目标检测算法得到快速发展, 这些算法主要分为两类: 非端到端检测和端到端检测. 以 Faster-RCNN[5] 为代表的非端到端类算法首先采用区域建议网络 (RPN) 筛选可能含有目标的候选框, 然后通过深度卷积神经网络提取图像特征进行分类. 端到端类算法通过深度卷积网络提取特征, 然后采用回归方式输出图像中目标的位置和类别, 代表性的算法有 SSD[6]、YOLO[7–9].

相比于传统人体行为检测算法, 基于深度学习的行为检测算法使用神经网络自动提取更深层次的图像特征, 避免了人工特征易受干扰的缺陷, 检测效果明显优于传统方法. 在两类深度学习目标检测算法中, 非端到端检测网络产生大量候选框, 然后对每一个候选框进行预测, 检测精度高, 但是比较耗时. 端到端检测网络采用回归方式直接预测, 具有良好的实时性, 但是不能很好的分割图片中的前景区域和背景区域, 容易产生误检和漏检. 因此如何在保证检测效率的前提下提升端到端检测算法的精度具有重要意义.

## 1 行为检测研究现状

目前在行为检测方面主要采用深度卷积神经网络提取特征, 经过特征融合后进行检测. Ji 等[10] 采用三维卷积神经网络, 提出 3-D 卷积神经网络 (3-D Convolutional Neural Networks, 3-D CNN), 提取视频中时空信息. 在 KTH 人体行为数据库上测试, 获得了 90.2% 的识别正确率. Gkioxari 等[11] 利用卷积神经网络对人体姿势和行为进行检测, 在 PASCAL VOC 数据集该方法取得了很好的检测效果, 并对已有的方法进行了对比. Gkioxari 等[12] 通过研究人体部件的动作和属性, 提出了一种基于人体部件的行为检测方法. 实验结果表明, 该方法能够对人体动作较好的分类. Feichtenhofer 等[13] 提出一种时空域上的人体行为检测方法. 该方法将双流卷积神经网络和残差网络 ResNet 进行结合, 采用运动流和外观流进行检测, 在 HMDB51 数据库和 UCF101 数据库取得了较高检测的精度. 莫宏伟等[14] 将 Faster R-CNN 与 OHEM 算法结合, 提出在线难例挖掘算法. 该算法包含两个 RoI 网络, 在 VOC 2012 Action 数据集上实验结果表明, 改进后 Faster R-CNN 算法具有识别精度高的特点. 黄友文等[15] 提出基于卷积神经网络与长短期记忆神经网络的多特征融合人体行为识别算法. 该算法将不同层次的特征进行连接, 通过卷积融合后输入 LSTM 单元, 在 KTH 和 UCF Sports 数据集实验结果表明, 模型有效地提高了行为识别精度.

同时, 朱煜等[16] 对传统行为识别方法和基于深度学习的人体行为识别方法进行了分析总结. 向玉开等[17] 对主流人体行为数据集进行对比, 分析了基于可见光波段、传统方法、深度学习等人体行为检测研究现状及趋势, 并总结面临的挑战.

虽然基于深度学习的行为检测算法在各种数据集上取得了不错的检测效果, 但仍然存在一些问题, 如基于 3D CNN、双流网络、Faster R-CNN 的行为检测算法网络参数量巨大无法实现实时性检测. 由于相同行为差异大, 不同行为相识度高, 检测过程中需要更加注重行为的细粒度特征, 基于人体部件的检测方法虽然能够提取局部和全局特征但额外增加数据标注成本. 端到端目标检测算法 YOLOv3 在 COCO 数据集上的测试结果 mAP 为 57.9% , 比 SSD 算法高出 7.5%, 并且满足实时性检测要求, 因此本文选择 YOLOv3 作为行为检测的基本网络并改进, 在保证检测效率的前提下提高网络对细粒度特征的提取能力, 从而提升检测的精度.

## 2 网络模型介绍

### 2.1 Hierarchical Bilinear Pooling 网络模型

在早期的研究中, 基于 Bilinear CNN 模型的细粒度分类网络[18] 的有效性已经在实验中得到验证. Hierarchical Bilinear Pooling 网络模型[19] 在 Bilinear CNN 模型的基础上提出分层双线性池化结构, 增加不同层之间的交互, 对多个分层双线性池化模块进行集成, 从网络中间的卷积层中提取细粒度互补信息, 其网络框架如图 1 所示.

该模型选取 3 个不同层、大小相同的特征图作为的输入, 如采用 VGG-16[20] 的 relu5_1, relu5_2, relu5_3 层. 然后相互作元素积 (Hadamard product[21]) 进行层间信息互补, 采用和池化操作降维, 经过非线性变换和 $L2$ 正则化提升网络模型表达能力, 最后将 3 个特征图进行维度拼接, 通过全连接层进行分类.

### 2.2 YOLOv3 网络模型

YOLOv3 网络结构可以分为两个部分: Darknet-53 特征提取网络和特征金字塔预测网络. Darknet-53 采用全卷积层和残差结构提取图像特征, 每个卷积层包括

二维卷积、归一化、LeakyReLU 三个操作. 特征金字塔预测网络中高分辨率的特征图通过低分辨率特征图上采样并与 Darknet-53 网络中的特征图拼接得到, 每一个尺度上的特征图都融合了不同分辨率、不同语义强度的特征. YOLOv3 预测过程如图 2 所示.

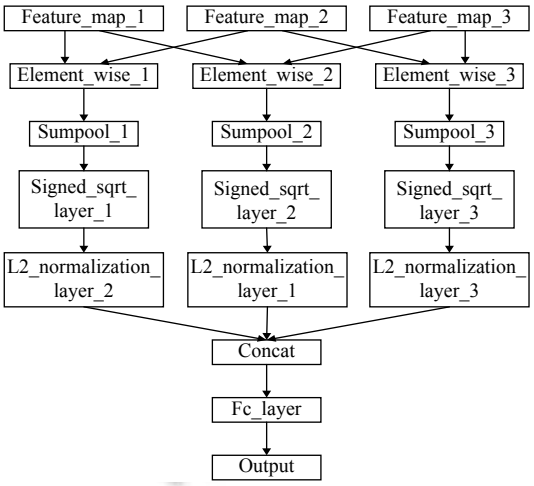



图 1 Hierarchical Bilinear Pooling 网络框架图

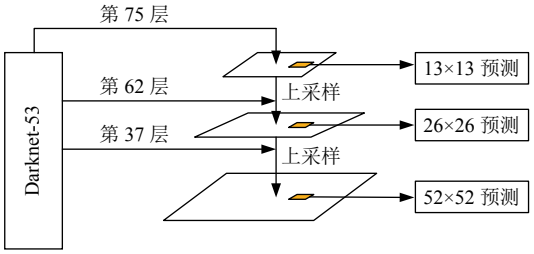



图 2 YOLOv3 预测结构图

416×416 的原始图像经过 YOLOv3 网络后产生 13×13、26×26、52×52 三个尺度上的网格区域, 每个网格区域预测 3 个边框, 每个边框对应四个边框预测值、一个网格区域置信度值和 $n$ 个类别值, 每个预测框输出向量 $y$ 如式 (1) 所示:

$$y=(t_x+t_y+t_w+t_h)+P_0+(P_1+P_2+\cdots+P_n) \quad (1)$$

## 3 Hierarchical Bilinear-YOLOv3 网络

### 3.1 改进 Hierarchical Bilinear Pooling 网络

原 Hierarchical Bilinear Pooling 网络主要用于图片的分类, 即单张图片上只有一个目标的情况. 为了使网络能够检测多个目标, 实现目标定位, 对原网络进行以下两个方面的改进: (1) 省去原网络中的和池化操作, 保留特征图的每一个像素特征; (2) 采用 1×1 卷积分类层代替原网络中的的 L2 归一化层和全连接分类层, 直接输出目标的类别和坐标信息. 改进之后的 Hierarchical Bilinear 网络如图 3 所示.

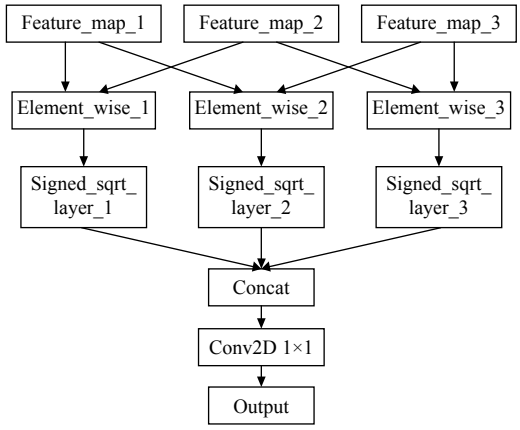



图 3 改进后的 Hierarchical Bilinear 网络结构图

将选取的 3 个大小为 $W\times H\times C$ 的特征图相互作元素积进行层间信息互补得到 3 个大小为 $W\times H\times C$ 的特征图, 其中 $W$、$H$、$C$ 分别为特征图的宽、高、深度. 经过非线性变换, 其表达式为:

$$y=\text{sign}(x)\times\sqrt{|x|+b} \quad (2)$$

其中, $x$ 为输入特征向量, $b$ 为浮点数常量. 将经过非线性变换后的特征图相加, 通过 1×1 卷积分类, 其表达式为:

$$\begin{aligned} Z_{\text{HB}} &= P^{\text{T}}concat(x,y,z) \\ &= \left(t_x+t_y+t_w+t_h\right)+P_0+(P_1+P_2+\cdots+P_n) \end{aligned} \quad (3)$$

其中, $Z_{HB}$ 为分类结果矩阵, $P^{\text{T}}$是分类矩阵, $x$、$y$、$z$ 为特征矩阵, $t_x$、$t_y$、$t_w$、$t_h$ 为目标坐标信息, $P_0$ 置信度值, $P_1, \cdots, P_n$ 为 $n$ 个类别值.

每个边框的预测坐标值计算公式如下:

$$b_x=Sigmoid(t_x)+C_x \quad (4)$$

$$b_y=Sigmoid(t_y)+C_y \quad (5)$$

$$b_w=P_w\times e^{t_w} \quad (6)$$

$$b_h=P_h\times e^{t_h} \quad (7)$$

其中, $t_x$、$t_y$、$t_w$、$t_h$ 为网络预测输出值, $C_x$ 和 $C_y$ 是网格区域相对于图片左上角的偏移量, $P_h$ 和 $P_w$ 表示预设边界框的长和宽, $b_x$ 和 $b_y$ 表示预测边界框的中心坐标, $b_h$ 和 $b_w$ 是预测边界框的长和宽.

置信度 $P_0$ 的计算公式如下:

$$P_0 = Sigmoid(P) \tag{8}$$

其中, $P$ 表示的是物体处于预测框中的输出值.

对预测框所在网格区域进行物体类别得分计算时采用逻辑分类, 计算公式如下:

$$P_i = Sigmoid(x_i) \tag{9}$$

其中, $x_i$ 表示预测该网格区域为某一类别的输出值.

### 3.2 改进后的 YOLOv3 网络

为了增强 YOLOv3 网络层间局部特征交互, 提升网络对细粒度特征的提取能力, 在特征金字塔分类网络中选取 3 个 3×3 卷积特征图作为改进后的 Hierarchical Bilinear 网络的输入, 经过层间信息互补后, 采用回归方式直接在 3 个尺度输出预测结果, 1×1 分类卷积核的深度为 (4+1+类别)×3. 改进后的 YOLOv3 网络如图 4 所示.

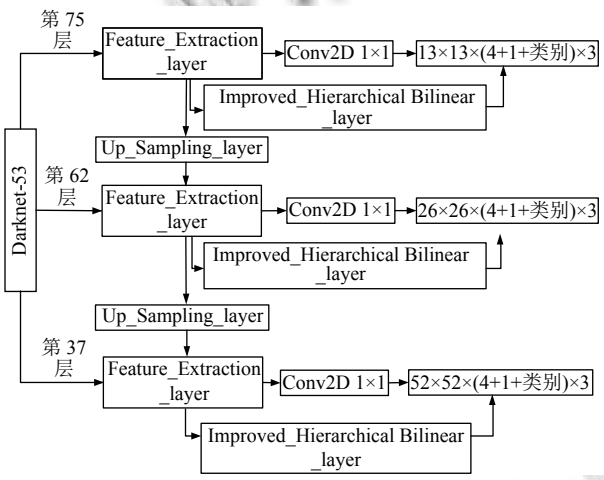



图 4 改进后的 YOLOv3 网络结构图

原网络和改进后的 Hierarchical Bilinear 网络均会在 3 个不同尺度上输出预测结果, 将输出结果进行融合, 计算公式如下:

$$y = \alpha y_{\text{yolo}} + (1-\alpha) y_{\text{hb}} \tag{10}$$

其中, $y$ 为融合结果, $y_{\text{yolo}}$ 指原 YOLOv3 网络输出, $y_{\text{hb}}$ 代表细粒度分类结果, $\alpha$ 为调节参数, 取值为 0.6.

### 3.3 损失函数设计

改进后 YOLOv3 的损失函数计算公式如下:

$$loss = \alpha loss_{\text{yolo}} + (1-\alpha) loss_{\text{hb}} \tag{11}$$

其中, $loss$ 为函数总损失, $loss_{\text{yolo}}$ 为原 YOLOv3 网络的损失, $loss_{\text{hb}}$ 为改进后的 Hierarchical Bilinear 网络损失, $\alpha$ 为权重调节参数, 取值为 0.6.

改进后的 Hierarchical Bilinear 网络损失包括 $xy$ 损失、$wh$ 损失、置信度损失、分类损失, 其中 $wh$ 损失采用误差平方和损失函数, 剩余的使用交叉熵损失函数, 计算公式如下:

$$loss_{xy} = \lambda \sum_{i=1}^{10\,647} I_{ij}^{\text{obj}} (2 - w_{\text{truth}} \times h_{\text{truth}}) \sum_{t \in x,y} binary\_crossentropy(t, \hat{t}) \tag{12}$$

$$loss_{wh} = \lambda_{coord} \sum_{i=1}^{10\,647} I_{ij}^{\text{obj}} (2 - w_{\text{truth}} \times h_{\text{truth}}) \sum_{t \in (w,h)} (t - \hat{t})^2 \tag{13}$$

$$loss_{\text{conf}} = \sum_{i=1}^{10\,647} I_{ij}^{\text{obj}} \times binary\_crossentropy\left(P_0, \hat{P}_0\right) \lambda \sum_{i=1}^{10\,647} \left(1 - I_{ij}^{\text{obj}}\right) \times binary\_crossentropy\left(P_0, \hat{P}_0\right) \tag{14}$$

$$loss_{\text{class}} = \sum_{i=1}^{10\,647} I_{ij}^{\text{obj}} \sum_{c \in class} binary\_crossentropy\left(Pi(c), \hat{P}i(c)\right) \tag{15}$$

其中, $I_{ij}^{\text{obj}}$ 表示该网格中是否存在物体, 如果有目标则为 1, 否则为 0. $\lambda_{\text{coord}}$、$\lambda$ 为权重调节参数, 取值为 0.5. $w_{\text{truth}}$, $h_{\text{truth}}$, $t$, $P_0$, $Pi(c)$ 为真实值, $\hat{t}$, $\hat{P}_0$, $\hat{P}i(c)$ 为预测值.

## 4 实验分析

### 4.1 实验数据集与参数设置

本文选用 PASCAL VOC 2012 action 数据集, 该数据集包含 10 种不同的行为: 跳、打电话、弹奏乐器、阅读、骑车、骑马、跑步、拍照片、使用电脑、走路, 每张图片包含类别信息、位置信息和语义分割信息. 数据集包含 3448 张图像, 分为训练集、验证集、测试集, 三者的比例为 6:2:2, 标签采用类别信息和标注框信息.

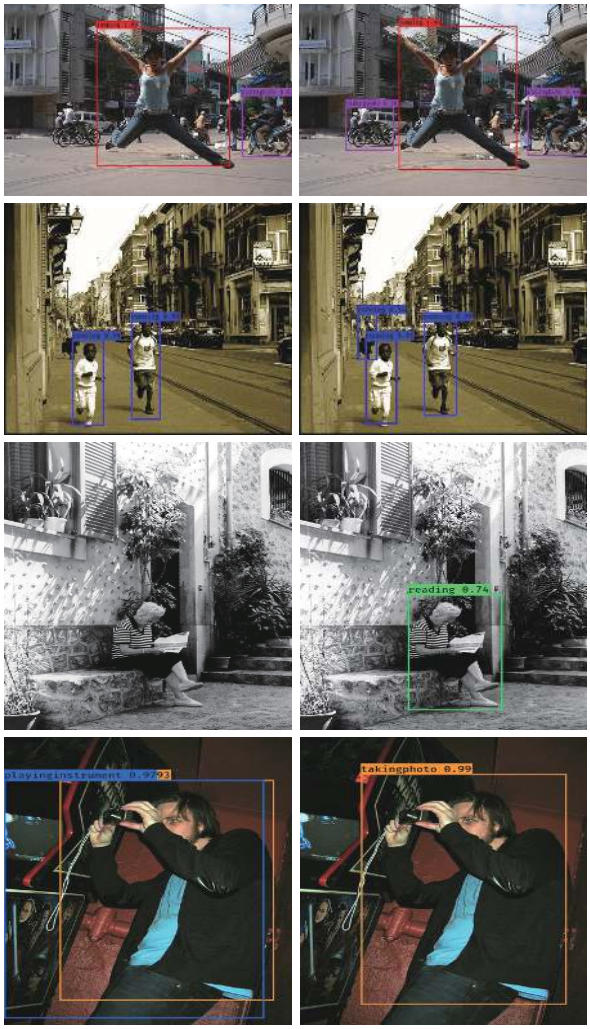
实验平台采用 Ubuntu 16.04 系统, Intel(R) Xeon (R) Silver 4116 CPU, 主频 2.10 GHz, 48 内核, 使用 NVIDIA Tesla K80 GPU 进行加速.

网络输入大小固定为 416×416, 初始化方法采用 He 等[22] 所提出的 MSRA Initialization, 实验训练迭代次数为 200 轮, 参数更新方法采用 Adam, 初始学习率

为 0.001, L2 权重衰减设置为 0.0005.

### 4.2 实验结果分析

本文提出的 Hierarchical Bilinear-YOLOv3 网络模型与原 YOLOv3 模型检测结果对比如图 5 所示.



(a) 原模型检测效果图 (b) 改进后模型检测效果图

图 5 两种模型检测结果对比图

当 IOU=0.5 时, 两种模型在测试集上的 AP (Average Precision) 结果如图 6 所示.

上述实验结果表明, 通过加入改进后的 Hierarchical Bilinear 网络增强特征图的层间交互, 能够提升原网络的细粒度提取能力和小目标检测率, 从而提高行为检测精度. 本文使用平均准确率均值 (mean Average Precision, mAP) 和每秒帧率 (Frame Per Second, FPS) 这两个指标来评价模型的检测效果, 并选择当前行为检测领域比较有代表性的模型进行对比, 实验结果如表 1 所示. 实验结果表明, 本文提出的 Hierarchical Bilinear-YOLOv3 网络模型相比原 YOLOv3 网络、文献 [23]、文献 [24] 在行为检测上的性能指标均有所提升, 改进后的模型虽然在 mAP 指标上没有文献 [12]、文献 [25] 高, 但检测精度已经非常接近, 同时 FPS 性能指标上大幅度优于这些算法, 能够实现实时行为检测.

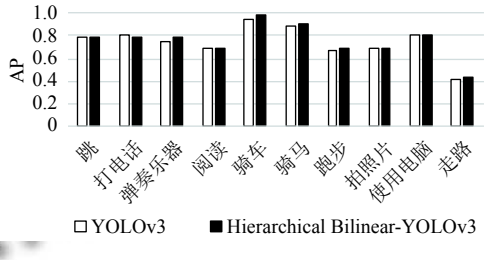



图 6 两种模型 AP 测试结果

表 1 各种行为检测模型实验结果对比数据

| 算法 | mAP | FPS |
|---|---|---|
| YOLOv3 | 0.7437 | 12 |
| Hierarchical Bilinear-YOLOv3 | 0.7583 | 12 |
| 文献[23] | 0.7020 | 1 |
| 文献[24] | 0.7560 | 1 |
| 文献[12] | 0.7700 | 1 |
| 文献[25] | 0.7864 | 1 |

## 5 结论与展望

本文针对 YOLOv3 网络在人体行为检测中精度低等问题, 提出一种基于 Hierarchical Bilinear 模型的 YOLOv3 改进算法. 该模型在 YOLOv3 原特征金字塔分类网络上选取一些特征输出层作为改进后 Hierarchical Bilinear 网络的输入, 增强层间局部信息交互, 进行细粒度分类, 然后与 YOLOv3 网络分类结果进行融合. 实验结果表明改进模型的参数量仅增加了 0.4%, 相比于原 YOLOv3 网络检测精度提升了 1.5%mAP, 在保证检测效率的前提下提高了检测精度.

## 参考文献

1 Sermanet P, Eigen D, Zhang X, *et al*. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv: 1312.6229, 2013.

2 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

3 Wang XY, Han TX, Yan SC. An HOG-LBP human detector

with partial occlusion handling. Proceedings of the 2009 IEEE 12th International Conference on Computer Vision. Kyoto, Japan. 2009. 32–39.
4 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA. 2001. I.
5 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]
6 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 21–37.
7 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.
8 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6517–6525.
9 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.
10 Ji SW, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition. EEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221–231. [doi: 10.1109/TPAMI.2012.59]
11 Gkioxari G, Hariharan B, Girshick R, *et al*. R-CNNs for pose estimation and action detection. arXiv: 1406.5212, 2014.
12 Gkioxari G, Girshick R, Malik J. Actions and attributes from wholes and parts. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 2470–2478.
13 Feichtenhofer C, Pinz A, Wildes RP. Spatiotemporal residual networks for video action recognition. Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain. 2016. 3468–3476.
14 莫宏伟, 汪海波. 基于 Faster R-CNN 的人体行为检测研究. 智能系统学报, 2018, 13(6): 967–973.
15 黄友文, 万超伦, 冯恒. 基于卷积神经网络与长短期记忆神经网络的多特征融合人体行为识别算法. 激光与光电子学进展, 2019, 56(7): 071505.
16 朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述. 自动化学报, 2016, 42(6): 848–857.
17 向玉开, 孙胜利, 雷林建, 等. 基于计算机视觉的人体异常行为识别综述. 红外, 2018, 39(11): 1–6, 33. [doi: 10.3969/j.issn.1672-8785.2018.11.001]
18 Lin TY, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1449–1457.
19 Yu CJ, Zhao XY, Zheng Q, *et al*. Hierarchical bilinear pooling for fine-grained visual recognition. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 595–610.
20 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.
21 Kim JH, On KW, Lim W, *et al*. Hadamard product for low-rank bilinear pooling. Proceedings of the 5th International Conference on Learning Representations. Toulon, France. 2017.
22 He KM, Zhang XY, Ren SQ, *et al*. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1026–1034.
23 Oquab M, Bottou L, Laptev I, *et al*. Learning and transferring mid-level image representations using convolutional neural networks. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1717–1724.
24 Cimpoi M, Maji S, Vedaldi A. Deep filter banks for texture recognition and segmentation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 3828–3836.
25 Zhang Y, Cheng L, Wu JX, *et al*. Action recognition in still images with minimum annotation efforts. IEEE Transactions on Image Processing, 2016, 25(11): 5479–5490. [doi: 10.1109/TIP.2016.2605305]