计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(5):232−237 [doi: 10.15888/j.cnki.csa.006894]

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

结合对象分割的运动行人检测①

宫法明, 吕轩轩, 宫文娟, 王晓宁

(中国石油大学 (华东) 计算机与通信工程学院, 青岛 266580)
通讯作者: 吕轩轩, E-mail: 2459008120@qq.com

摘　要: 目标检测大量应用于监控系统的行人检测以及人脸识别, 是当前深度学习的研究热点. 监督学习利用人工标注大量数据集训练出针对特定场景的行人检测器. 但是人工标注方法费时费力, 本文针对监督学习需要人工标注数据集的缺点, 研究了一种半自动标注行人的方法. 针对静止的单目摄像机拍摄的监控视频, 利用光流信息提供的初始前景可能性, 以及跨越时间的视觉相似性来迭代地更新初始的前景可能性, 分割出运动的行人, 根据分割的前景对象, 提出了一种半自动标注行人的方法. 实验结果显示, 本文的方法可以为行人检测系统提供大量数据集, 且效率上明显优于传统人工标注的方法.

关键词: 行人检测; 光流; 视频对象分割; 深度学习; 半自动数据集标注

引用格式: 宫法明,吕轩轩,宫文娟,王晓宁.结合对象分割的运动行人检测.计算机系统应用,2019,28(5):232–237. http://www.c-s-a.org.cn/1003-3254/6894.html

Moving Pedestrian Detection Framework with Object Segmentation

GONG Fa-Ming, LYU Xuan-Xuan, GONG Wen-Juan, WANG Xiao-Ning

(Department of Computer and Communication Engineering, China University of Petroleum, Qingdao 266580, China)

Abstract: Object detection is widely used in surveillance systems for pedestrian detection and face recognition. It is a research hotspot of current deep learning. Supervised learning trains pedestrian detectors for specific scenes by manually annotating large datasets. However, the manual labeling method is time-consuming and laborious. In this work, the shortcomings of manual labeling of datasets for supervised learning are studied. A method of semi-automatic labeling of pedestrians is proposed. The surveillance video captured by the stationary monocular camera, using the initial foreground possibilities provided by the optical flow information, and the visual similarity across time, iteratively updates the initial foreground likelihood to segment the moving pedestrians. According to the segmented foreground pedestrians, a method of semi-automatic labeling of pedestrians is proposed. The experimental results show that the proposed method can provide a large number of datasets for the pedestrian detection system, and the efficiency is obviously superior to the traditional manual annotation method.

Key words: pedestrian detection; optical flow; video object segmentation; deep learning; semi-automatic data annotation

1　概述

近年来, 运动目标检测已经成为计算机视觉领域的研究热点, 引起了众多学者的关注, 在视频监控系统、对象跟踪等方面发挥了巨大的作用. 行人检测的研究随着机器学习的巨大发展取得了重大进步, 研究内容为在图像中快速而精确地识别和定位出行人. 基于外观的行人检测器在大规模的数据集上进行训练已经成为主流, 目前流行的训练目标检测器的方法是使

① 收稿时间: 2018-11-21; 修改时间: 2018-12-12; 采用时间: 2018-12-20; csa 在线出版时间: 2019-05-01

用监督算法 (如 AdaBoost[1], 神经网络[2], 支持向量机[3]). 但是这种方法需要大量的人工标注训练数据集, 而且随着检测模型的扩大, 需要标注的数据集也快速增长, 人工标记大型训练集的过程是耗时且乏味的. 因为训练出适应特定场景的检测器需要海量的数据集来覆盖各种视点、分辨率、光照条件、天气环境以及各种复杂的场景, 当训练用于大规模视觉系统的检测器时, 例如在几百个场景中配置摄像机的视频监控网络中, 从每个场景人工收集和标记正面以及负面的训练图像的成本是非常大的. 随着大数据时代的到来, 需要处理海量视频数据, 人工获取数据集的方式已经无法满足实际需要.

目前目标检测图片标注的方式主要以全手工标注为主, 进行重复人工类别标注的成本太高, 效率低下且不可扩展, 尤其是在需要大量标注样本的情况下. 在这种情况下, 数据集的获取方式变得尤为重要. 为了应对大量增长的视频数据, 提高效率. 本文针对静止的单目摄像机拍摄的监控视频图像, 在已有框架上进行步骤改进, 提出了一种融合对象分割的半自动标注方法, 极大减少了人力参与, 降低经济成本.

本文的实验数据来源于海上石油平台的监控视频. 海上作业危险性很高, 为了保证石油工人的安全等问题, 对海上石油平台的监控尤为重要. 海上平台的摄像头数量巨大, 且安装角度各异, 使得视频背景及海上石油平台工作人员在视频中出现的位置复杂, 更增加了训练集制作的难度.

本文的主要贡献如下:

1) 设计一个用于行人检测任务的融合对象分割的半自动标注方法, 能够提供大量的训练样例;

2) 将运动信息与视觉相似性相结合, 更好地分割出前景目标, 并将其应用于数据集的生成.

首先, 为了更加精确地分割前景目标和背景, 本文结合了短期线索的运动信息和跨越大时间圈的视觉相似性, 首先将图像分割成超像素[4], 通过光流[5]大小来提取运动信息, 给出每个超像素的初始显著性分割 (前景或者背景), 通过跨越时间的连续帧在空间区域的外观相似性迭代的纠正每个超像素的分割结果. 在空间区域和时间区域的相似特征将丰富多样的信息传播到整个视频序列, 得到准确的分割结果. 根据提取出的前景目标进行数据集的制作; 其次, 用制作好的数据集学习一个针对特定场景的行人检测器, 最后将其应用于行人检测. 如图 1 所示, 为本文的框架流程图.

文章的组织结构如下: 第 2 部分介绍了行人检测的研究现状; 第 3 部分提出了一种新的数据集标注方法, 将得到的数据集用于行人检测器的学习; 第 4 部分在海上石油平台的监控视频数据集中对本文提出的算法进行了实验验证; 第 5 部分对全文进行了总结.

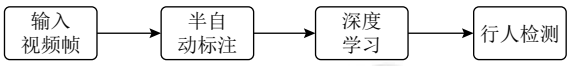


图 1 行人检测框架流程图

2 相关工作

监督学习的行人检测算法一般需要手工标注大量的数据集, 这种方式耗费大量的时间和人力. 目前已经有相关的无监督学习或者半监督学习实现了无需标注数据集的方式来训练行人检测模型. 香港中文大学的王晓刚等[6]提出了一种迁移学习框架, 自动地将通用行人检测器转换成针对特定场景的行人检测器, 过程中无需手工标注数据集. 弗莱堡大学的 Dosovitskiy A 等[7]基于数据增强提出了一种仅使用未标记数据训练卷积神经网络的无监督的目标检测方法, 使用代理类通过一系列基本变换进行数据增强, 如: 旋转、缩放、平移、颜色、对比度等. 卡内基梅隆大学的 Rosenberg C 等[8]使用一组弱标记的示例来增强一小组标记的训练实例去训练出一个目标检测器, 表现优于完全使用标记数据训练出的目标检测器. 巴塞罗那自治大学的 Marin J[9]使用来自虚拟世界的数据集, 免除了手工标注的烦琐过程.

对于监督学习, YOLO[10]和 SSD[11]在目标检测方面基于大量数据集的人工标注取得了良好的效果. 麦吉尔大学的 Nair V 等[12]使用背景减法手动设计了一个自动标注机, 自动标注在办公室走廊上的行人, 免除了手工标注数据集的枯燥乏味, 但是对于复杂场景下, 这种方式可能并不适用.

监督学习和无监督学习是人工神经网络的两种主要的学习方式. 无监督学习不需要带有标签的训练样本, 但是训练过程繁琐且时间冗长. 监督学习从带标记的训练样本中学习特征, 但是需要大量的训练数据. 根据研究现状可知, 以上成果主要基于训练数据集的无监督训练, 数据增强或者虚拟现实来实现无需标注数据集的目的, 基于监督学习表现良好, 本文旨在构建一个应用于监督学习的复杂场景下的数据集类别半自动

标注的方法. 本文的半自动标注方法允许人力参与, 对标注结果进行修正.

3 基于对象分割的数据集半自动标注

在下面的章节中描述了所提出方法的细节, 3.1 节将运动信息与视觉相似性相结合, 提取出前景目标, 为之后的数据集标注做准备; 3.2 节将 3.1 节提取的前景目标应用于数据集的半自动标注, 可以人工调整目标框; 3.3 节使用深度学习方法对本文方法进行了验证.

3.1 前景背景分割

图像前景区域提取有很多应用, 包括对象检测和识别, 视频摘要, 图像压缩等等. 因此, 国内外的学者也在这个方向进行了大量的研究. Mitra NJ 等[13]提出的算法仅仅考虑颜色显著性, 这显然是不够的, 因为一些不同颜色的区域可能是非显著的; 还有算法仅仅考虑检测不同的模式, 如前景对象和背景区域之间的边缘信息, 但这可能导致显著物体的均匀区域缺失; 背景减除法也是是一种有效的对象检测算法, 基本思想是利用背景的参数模型来近似背景图像的像素值, 将当前帧与背景图像进行差分比较实现对运动区域的检测由于监控环境的不同, 真实的背景可能随时会发生变化, 背景模型如果不能及时更新, 则会导致运动目标提取的失败. 针对传统方法存在的不足, 我们既考虑运动显著性, 也考虑跨越大时间圈的视觉相似性. 这些线索都不足以提供良好的分割结果. 我们的方法是以简单而有效的方式融合这两个线索. 本文核心算法流程图如图 2 所示.

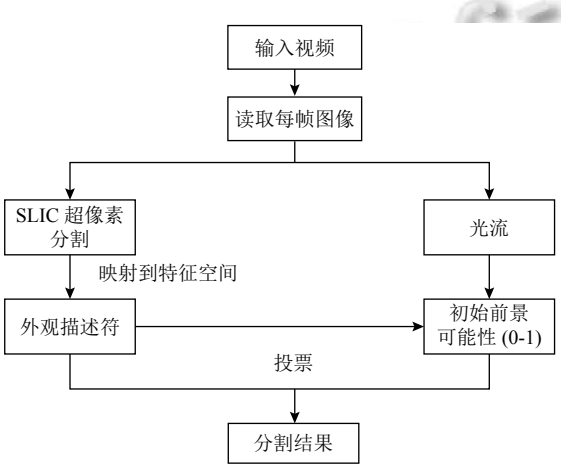


图 2 算法流程图

算法包含了 4 部分: (1) 将视频序列分割成超像素; (2) 根据光流信息获得每个超像素的前景可能性 (0-1), 其中 1 代表前景, 0 代表背景; (3) 将每个超像素用一个外观描述符表示在特征空间中, 特征空间中相邻描述符代表相似的区域, 每一个区域都有一个前景可能性; (4) 根据特征空间中的相邻描述符的前景可能性来更新初始前景可能性.

3.1.1 初始前景可能性结果

最初的光流算法由 Horn 和 Schunck[14]提出, 假定在时刻 t 处于图像坐标 (x, y) 位置的点, 在时刻 $t+d$ 运动到了图像坐标的点 $(x+dx, y+dy)$ 处, 在时刻 t 的像素灰度值为 $I(x, y, t)$, 在时刻 $t+dt$ 的像素灰度值为 $I(x+dx, y+dy, t+dt)$, 时间变化很小的情况下, 认定像素灰度值不变, 那么像素的运动矢量即为运动物体的光流变化. 光流算法根据物体的运动信息来判断物体位置以及显著性特征. 它代表图像中的模式运动的速度, 如果图像中没有运动目标, 那么, 光流矢量在整个图像区域连续变化; 如果物体和图像背景存在着相对运动时, 运动物体形成的速度矢量的大小和方向和邻域背景必然不同, 从而检测出运动物体的位置以及轮廓特征.

本文利用连续帧间差分方法定位运动目标, 获得二值图像, 然后计算二值图像中特征点处的光流. 因为计算二值图像中值为 1 的特征点的光流场分布比计算整个图像的光流场要准确.

本文利用光流信息获取视频帧的初始前景可能性结果, 但是在复杂背景下的视频序列中, 短期的运动显著性只能很好地指示物体位置. 假如前景物体在视频序列中只有一部分移动而另一部分静止, 只用光流的方法是不够的. 这些前景可能性结果很嘈杂, 需要通过视频序列的特征空间的相邻描述符的前景可能性来更新初始结果.

3.1.2 超像素分割

超像素分割技术是指将图像分割为许多小的区域, 这些成为超像素的小区域在颜色和纹理上具有同质性. 由于超像素空间紧凑性高、大小均匀, 并且能够很好地保留了图像中目标的边界结构, 这种过度碎片化使得我们即使在高运动模糊或低分辨率的情况下提取有意义的边界. 本系统采用超像素分割方法为简单线性迭代聚类 (Simple Linear Iterative Clustering, SLIC)[15], 该方法预先设定的超像素个数, 采用 K-mean 聚类方法生成一系列大小一致且保持目标边界的超像素区域 R. 超像素分割示例如图 3 所示.

3.1.3 前景可能性的迭代更新

特征空间中的相邻描述符表示相似的区域, 在视频中可能在空间和时间上相隔很远. 我们的描述符有以下几种类型: RGB 和 LAB 颜色直方图, HOG 描述符. 首先我们需要找到超像素区域 R 在特征空间上的 N 个最邻近区域 (Nearest Neighbors), 计算出区域 R 和它的 N 个最邻近区域的相似性:

$$S\left(R,TNN_n(R)\right)=\exp\left(-\left\|d(R)-d(TNN_n(R))\right\|^2/\sigma^2\right) \tag{1}$$

其中, $d(R)$代表区域 R 的高维特征描述符.

图 3 SLIC 超像素分割图像

然后, 计算图像上所有超像素区域的最邻近距离矩阵 $S(i,j)$, 并归一化; 用 N 个最邻近区域的前景可能性的加权平均值更新每个区域 R 的前景可能性:

$$S(i,j)=\begin{cases} s\left(R_i,R_j\right), & R_j\in TNN(R_i) \\ 1, & i=j \\ 0, & otherwise \end{cases} \tag{2}$$

我们将算法分为两个阶段: 首先限制最邻近区域搜索的视频帧数量, 将每个超像素搜索的范围设置为在 F 等于 10 帧的时间半径之内的包括自身在内的 TNN (即 $2F$+1 帧), 这样的做法保证了算法的效率以及减少了混淆背景和前景区域的机会; 然后放宽对于邻近区域的搜索时间限制, 可以在整个视频序列中搜索, 得到最终的分割结果.

3.2 数据集的半自动标注

我们的半自动标注数据集的方法是根据 3.1 节中提取的前景目标设计而来. 在行人检测任务中, 半自动标注方法允许人力的参与, 当然我们尽力让提取的前景目标更加精确, 以免除或者减少人力参与, 使得标注结果更加精确:

(1) 当目标框不够精确时, 可以手动调整目标框的大小;

(2) 当出现将背景像素标记为前景目标等错误标记情况时, 可以通过删除目标框按钮, 删除已经标记的错误目标框;

(3) 当出现遗漏标记的情况, 可以通过添加目标框按钮来标记遗漏目标.

3.3 深度学习分类器

卷积神经网络[16]作为深度学习模型的一种, 能从数据中自动学习并提取特征, 其泛化能力显著优于传统方法, 已经成功应用于物体检测和识别等领域. 包含输入层、输出层和隐层, 它的隐层由若干个卷积层、池化层和全连接层组成. 简化的神经网络结构图如图 4 所示.

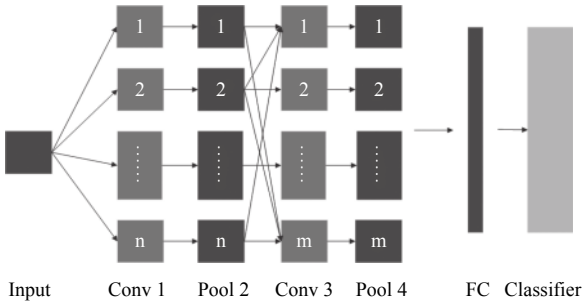


图 4 卷积神经网络结构图

其中, $Conv_1$、$Conv_3$ 是卷积层, $Pool_2$、$Pool_4$ 为池化层, FC 为全连接层. 卷积层中有多个 Feature Map, 每个 Feature Map 对应一种滤波器, 以此提取不同的特征.

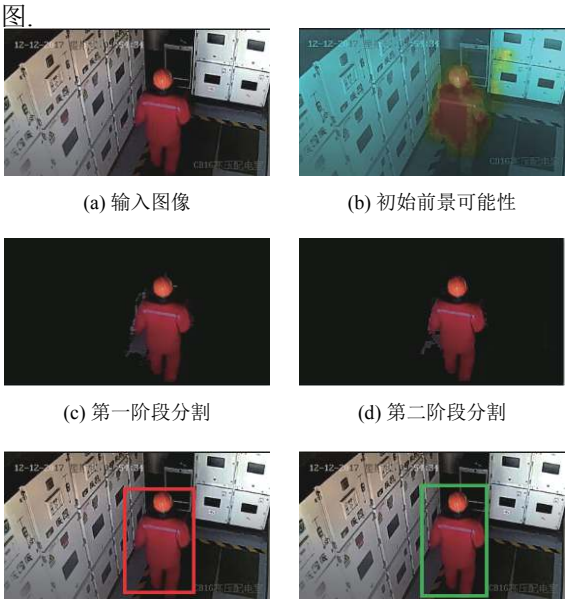
4 实验

基于本文提出的方法, 我们开发了一个半自动标注行人实例的系统. 为了验证本文方法的鲁棒性, 在石油海洋平台的多个场景中进行了实验验证. 本实验采用普通的台式机, 将 CPU 为 Core(TM)i7、主频 3.4 GHz, 内存为 8 G 的台式机作为硬件平台, 搭载英伟达 GTX1060 型号显卡. 软件开发环境为: 64 位 Windows 10 操作系统、MATLAB R2015b 软件开发平台、Visual Studio 2013 平台、Caffe 深度学习框架.

实验所用的视频来自石油海洋平台的静止监控摄像头. 在实验中, 监控设备保持固定不动, 视频序列以海洋工作平台作为背景. 图 5 为方法实现过程.

图 5(a) 为输入视频帧; 图 5(b) 为经过光流算法得

到的初始前景可能性结果; 图 5(c) 为限制邻近区域搜索空间得到的第一阶段分割结果; 图 5(d) 为在整个视频序列中搜索邻近区域得到的第二阶段分割结果; 图 5(e) 为使用本文半自动标注方法得到的训练样例标注结果; 图 5(f) 为本文所提出的目标检测框架所得结果图.



(a) 输入图像 (b) 初始前景可能性

(c) 第一阶段分割 (d) 第二阶段分割

(e) 训练样例 (f) 检测结果图

图 5 本文框架实现过程

如表 1 所示, 为两种标注方法在同一数据集上进行实验的结果. 人工标注与操作人员的标注效率有关, 本实验选用七个标注人员的平均效率作为参考. 其中, 包含目标帧为所选视频中包含行人的帧数, 不包含目标帧为视频帧中没有行人的帧数, 错误标注率为标注错误的视频帧占不包含目标帧的比例, 遗漏标注率为标注遗漏的视频帧占包含目标帧的比例, 标注时长为标注该组数据集所消耗的时间.

表 1 实验结果对比表

标注方法	视频时长 (s)	包含目标帧	不包含目标帧	错误标注率 (%)	遗漏标注率 (%)	标注时长 (s)
人工标注	70	855	863	1.69	0.70	8400
我们的方法	70	855	863	6.95	6.41	4042

由实验可以看出:

(1) 在相同的数据集上进行标注, 本文提出的方法在效率上要优于人工标注数据集.

表 2 为在不同场景下使用本文方法的表现对比结果, 其中, 场景 1 为画质清晰且目标行人尺寸较大的数据集; 场景 2 为画质清晰但目标行人尺寸较小的情况; 场景 3 为背景较为复杂且行人较小的数据集.

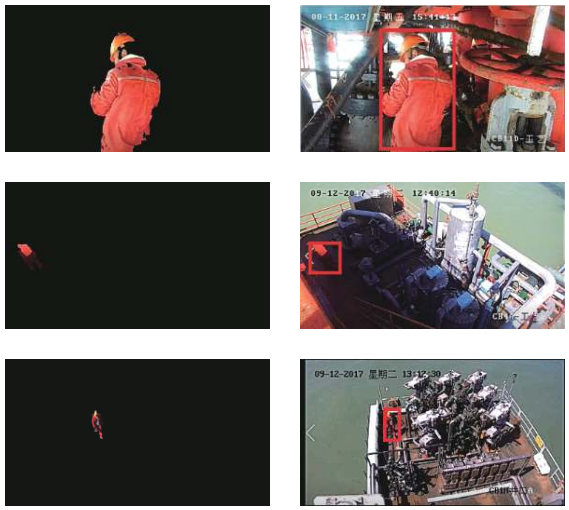
(2) 在相同的数据集上进行标注, 本文提出的方法在标注质量上要低于人工标注, 但在资源消耗上要优于人工标注方法.

(3) 在不同的数据集上进行标注, 本文提出的方法在画质清晰且目标行人尺寸较大的场景中表现明显优于背景模糊场景或者目标行人尺寸较小的场景. 图像质量以及目标尺寸大小对实验结果 (标注效率、标注质量、资源消耗) 有很大影响.

表 2 不同场景下实验结果对比表

标注方法	视频时长 (s)	包含目标帧	不包含目标帧	错误标注率 (%)	遗漏标注率 (%)	标注时长 (s)
场景 1	70	808	910	3.54	2.20	3902
场景 2	70	798	920	6.35	5.41	4582
场景 3	70	832	886	4.22	3.32	4037

图 6 为使用本文的方法针对不同场景下的结果. 实验表明, 本文提出的半自动标注行人实例的方法能够较精确地实现单目标场景中行人训练实例的分割问题, 同时对多场景视频中的复杂环境等有较好的适应性, 提高标注训练实例的效率.



(a) 分割结果 (b) 基于对象分割的半自动标注

图 6 不同场景下的实验结果. 第一列为相应场景的分割结果图, 第二列为对应于第一列的使用本文提出的方法的标注结果图

5 结束语

本文所提出的行人检测框架, 结合了对象分割, 能

较准确的分割出视频中的运动目标, 并将其应用于训练数据集的标注. 一定程度上减轻了人力负担, 提高了数据集标注效率. 但是对于多目标场景,本文还不能很好地解决. 由于海上采油平台远离陆地, 工作环境相对复杂, 我们下一步的工作就是解决多目标复杂场景下的前景分割及标注.

参考文献

1 Rajeshwari J, Karibasappa K, Gopalkrishna MT. Adaboost modular tensor locality preservative projection: Face detection in video using AdaBoost modular-based tensor locality preservative projections. IET Computer Vision, 2016, 10(7): 670–678. [doi: 10.1049/iet-cvi.2015.0406]

2 周志华, 陈世福. 神经网络集成. 计算机学报, 2002, 25(1): 1–8. [doi: 10.3321/j.issn:0254-4164.2002.01.001]

3 张浩然, 韩正之, 李昌刚. 支持向量机. 计算机科学, 2002, 29(12): 135–137, 142. [doi: 10.3969/j.issn.1002-137X.2002.12.038]

4 王春瑶, 陈俊周, 李炜. 超像素分割算法研究综述. 计算机应用研究, 2014, 31(1): 6–12. [doi: 10.3969/j.issn.1001-3695.2014.01.002]

5 Brox T, Malik J. Large displacement optical flow: Descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(3): 500–513. [doi: 10.1109/TPAMI.2010.143]

6 Wang XG, Wang M, Li W. Scene-specific pedestrian detection for static video surveillance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(2): 361–374. [doi: 10.1109/TPAMI.2013.124]

7 Dosovitskiy A, Fischer P, Springenberg JT, *et al*. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(9): 1734–1747. [doi: 10.1109/TPAMI.2015.2496141]

8 Rosenberg C, Hebert M, Schneiderman H. Semi-supervised self-training of object detection models. Proceedings of 2005 IEEE Workshops on Applications of Computer Vision. Breckenridge, CO, USA. 2005. 29–36.

9 Vázquez D, López AM, Marín J, *et al*. Virtual and real world adaptation for pedestrian detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(4): 797–809. [doi: 10.1109/TPAMI.2013.163]

10 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.

11 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.

12 Nair V, Clark JJ. An unsupervised, online learning framework for moving object detection. Proceedings of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC, USA. 2004. 317–325.

13 Cheng MM, Zhang GX, Mitra NJ, *et al*. Global contrast based salient region detection. Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO, USA. 2011. 409–416.

14 Horn BKP, Schunck BG. Determining optical flow. Artificial Intelligence, 1981, 17(1–3): 185–203. [doi: 10.1016/0004-3702(81)90024-2]

15 汪成, 陈文兵. 基于 SLIC 超像素分割显著区域检测方法的研究. 南京邮电大学学报 (自然科学版), 2016, 36(1): 89–93.

16 芮挺, 费建超, 周遊, 等. 基于深度卷积神经网络的行人检测. 计算机工程与应用, 2016, 52(13): 162–166. [doi: 10.3778/j.issn.1002-8331.1502-0122]