DOI: 10.11992/tis.201710019
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180416.1401.010.html

多层卷积特征的真实场景下行人检测研究

伍鹏瑛[1,2], 张建明[1,2], 彭建[1,2], 陆朝铨[1,2]
（1. 长沙理工大学 综合交通运输大数据智能处理湖南省重点实验室, 湖南 长沙 410114; 2. 长沙理工大学 计算机与通信工程学院, 湖南 长沙 410114）

摘　要: 针对真实场景下的行人检测方法存在漏检、误检率高, 以及小尺寸目标检测精度低等问题, 提出了一种基于改进 SSD 网络的行人检测模型 (PDIS)。PDIS 通过引出更底层的输出特征图改进了原始 SSD 网络模型, 并采用卷积神经网络不同层输出的抽象特征对行人目标分别做检测, 融合多层检测结果, 提升了小目标行人的检测性能。此外, 针对数据集样本多样性能有效地提升检测算法的泛化能力, 本文采集了不同光照、姿态、遮挡等复杂场景下的行人图像, 对背景比较复杂的 INRIA 行人数据集进行了扩充, 在扩增的行人数据集上训练的 PDIS 模型, 提高了在真实场景下的行人检测精度。实验表明: PDIS 在 INRIA 测试集上测试结果达到 93.8% 的准确率, 漏检率低至 7.4%。

关键词: 行人检测; 卷积神经网络; SSD; 真实场景; 多尺度特征; 目标检测; 小目标行人; 行人数据集
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2019)02−0306−10

中文引用格式: 伍鹏瑛, 张建明, 彭建, 等. 多层卷积特征的真实场景下行人检测研究[J]. 智能系统学报, 2019, 14(2): 306–315.
英文引用格式: WU Pengying, ZHANG Jianming, PENG Jian, et al. Research on pedestrian detection based on multi-layer convolution feature in real scene[J]. CAAI transactions on intelligent systems, 2019, 14(2): 306–315.

Research on pedestrian detection based on multi-layer convolution feature in real scene

WU Pengying[1,2], ZHANG Jianming[1,2], PENG Jian[1,2], LU Chaoquan[1,2]
(1. Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation, Changsha University of Science and Technology, Changsha 410114, China; 2. School of Computer and Communication Engineering, Changsha University of Science and Technology, Changsha 410114, China)

Abstract: Pedestrian detection methods in real scenes face some problems due to the high miss detection and false detection as well as the low detection accuracy of small size objects. To solve these problems, a pedestrian detection model based on improved SSD (PDIS) is proposed. The PDIS method improves the original SSD network model by extracting the lower-level output feature maps. It employs the abstract features of different convolutional neural network layers to detect pedestrians respectively, and then integrates the detection results of multi layers to increase the pedestrian detection performance for small sizes. Considering that the diversity of dataset can effectively enhance the generalization ability of detection algorithm, the paper expands the INRIA pedestrian dataset with complex background by collecting pedestrian images with different illumination, pose and occlusion. The PDIS method trained on expanded pedestrian dataset increases the precision rate of pedestrian detection in real scenes. The experiment results on INRIA test set indicate that the precision rate of PDIS algorithm is up to 93.8% and the miss rate is as low as 7.4%.

Keywords: pedestrian detection; CNN; single shot multibox detector; real scene; multi-scale features; object detection; small target pedestrians; Pedestrian dataset

收稿日期: 2017−10−31.　**网络出版日期**: 2018−04−16.
基金项目: 国家自然科学基金项目 (61402053); 湖南省教育厅科研重点项目 (16A008); 湖南省交通厅科技项目 (201446); 长沙理工大学研究生科研创新项目 (CX2017SS19); 长沙理工大学研究生课程建设项目 (KC201611).
通信作者: 张建明. E-mail: jmzhang@csust.edu.cn.

行人检测是判断输入的图像或视频中是否含有行人, 并准确的找出行人的具体位置。行人检测作为目标检测的一个子方向, 在视频监控、行人识别[1]、图像检索以及先进的驾驶员辅助系统

等领域有着广泛的应用[2]。由于行人具有非刚性属性, 决定了行人检测不同于普通的目标检测, 另外存在着许多制约行人检测的因素, 如现实场景中背景的复杂多样性、光照变化、行人遮挡、姿态变化、拍摄角度多样化、实时性要求、小目标行人等。这些因素给行人检测带来了巨大的挑战, 因此行人检测一直是计算机视觉领域中的研究热点和难点。

传统的行人检测的效果依赖于特征的选取以及分类器的学习。一个好的特征即使结合简单的分类器仍能够取得不错的检测效果, 所以传统的行人检测研究重点在于行人的特征提取及分类。尽管传统的行人检测算法取得了不少的研究成果, 但是在实际生活场景的检测效果依然不理想。近年来深度学习的方法在目标检测、语音识别、图像分类等方面取得了突破性的进展, 与传统检测算法相比, 卷积神经网络 (CNN) 通过权值共享, 大大减少了网络的参数, 进而降低了算法复杂度。CNN 的卷积运算以及下采样能很好的学习到图像的颜色、纹理等特征, 使之对图像的缩放、平移具有很好的鲁棒性。因此, 深度学习算法在行人检测领域里的检测精度以及实时性都优于传统算法。

针对真实场景下的行人检测精度不高, 小目标行人的漏检率较高的问题, 本文对目前优秀的深度模型进行了改进。通过引出 SSD[3]网络模型中更底层特征做检测以及增加输入图像大小来增加深度模型的分辨率, 提高了对小目标行人的检测性能。卷积网络中的底层特征能检测到尺寸较小的目标, 而深层特征可以检测到尺寸较大的目标, 因此引出 SSD 网络中多层输出特征图, 将检测结果综合后确定目标位置。此外, 训练数据集的数量跟数据集样本的多样性也是深度学习算法取得优秀成果的主要原因。因此本文采用车载摄像头拍摄了各种场合、光照、遮挡、姿态等复杂的背景下的行人视频, 在 INRIA[4]行人数据集上, 扩增了一个复杂场景下的行人数据集 CSUSTPD。

1　相关工作

传统的行人检测流程主要由行人图像输入、行人的特征提取、分类与定位、检测结果等几个模块组成[5], 其研究重点在于行人的特征提取及分类, 比较常见的特征提取算子有 SIFT[6]、Haar[7]、梯度方向直方图 HOG[4]等; 代表性的分类器有神经网络、Adaboost[8]、支持向量机 SVM[9]、随机森林 RF[10]等。基于 HOG 特征的提取极大地推动了行人检测的发展, 并随后出现了在 HOG 特征上融合颜色特征、纹理特征等诸多算法; 2005 年 Dalal 等[4]提出了 HOG 结合分类器 SVM 的算法, 取得了较好的效果, 并陆续提出的 ACF(aggregated channel features)[11]、LDCF[12]等算法都具有很好的检测效果。2015 年 Zhang 等[13]把 HOG 特征结合光流特征进一步提高了行人检测性能。针对在同一张图像有不同尺寸的目标时, 传统方法主要有两种处理方法: 1) 将原始图像转换成不同尺寸大小的图像输入固定尺寸的滑动窗口分别提取特征, 该方法的检测精度较好, 但是计算复杂, 其流程如图 1 所示; 2) 用固定大小不变的图像输入多尺度缩放的滑动窗进行特征提取。方法 2) 避免了测试图像的多尺度计算, 检测速度较快但其精度比较差。

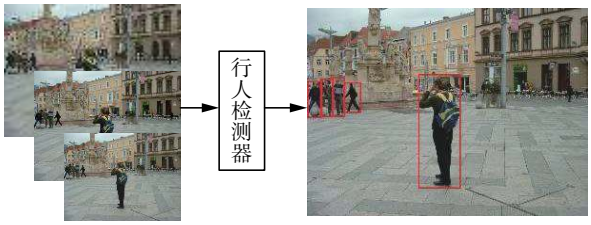

图 1　多尺度输入图像检测流程

Fig. 1　Object detection with multi-scale input

2012 年 Krizhevsky 等[14]在 ImageNet 竞赛上训练出一个 8 层的卷积神经网络, 取得了视觉领域竞赛 ILSVRC 2012 的冠军。在此之后, 深度学习算法层出不穷, 典型的算法有 VGG-Net[15]、R-CNN[16]、Faster R-CNN[17]; Redmon 等[18]提出的 YOLO 直接在输出层回归目标位置与类别, 加快了目标检测速度及精度; Liu 等[3]提出了 SSD 融合卷积层的多层输出特征做检测, 进一步提高了目标检测精度。由于 CNN 提取的特征比传统特征更具鲁棒性, 其良好的特征表达能力提高了行人检测性能, 因此许多研究学者将深度学习算法应用于行人检测领域。文献[19]引入级联 CNN 网络在拥挤场景中准确地检测行人; Hosang 等[20]利用 SquaresChnFtrs[21]产生行人候选窗口用于训练 AlexNet[14]。文献[22]利用光流卷积神经网络对光流图序列中提取行人特征, 该特征具有较强的全局描述能力; Tian 等[23]利用深度学习结合部件模型解决行人检测中的遮挡问题。文献[24]采用级联的 Adaboost 检测器对行人目标进行初步筛选, 再用迁移学习技术训练卷积神经网络来提高检测精度; Zhang 等[25]利用级联的决策森林来分类 RPN 网络 (region proposal network) 产生的行人候选窗口。

训练深度 CNN 模型时, 数据集的数量跟数据

集样本的多样性能增强算法检测的泛化能力。现有的行人数据集如 Daimler 行人数据集[26]含训练样本集有正样本大小为 18×36 和 48×96 的图像。较早公开的 MIT 行人数据集[27]含 924 张宽高为 64×128 行人图像，肩到脚的距离约 80 像素。该数据库只含正面和背面两个视角，无负样本，并且未区分训练集和测试集。NICTA 行人数据集[28]标注要求行人高度至少要大于 40 个像素。这些数据集训练样本存在从大图像中剪切出的单个行人图像、分辨率偏低、对小目标行人无标注的问题，且行人数据集训练样本背景单一。因此，这些数据集不适合用于训练深度卷积网络模型。

2　SSD 网络

SSD 算法是一种直接预测目标边界框的坐标和类别的检测算法，整个网络没有生成候选窗口的过程。SSD 算法的骨干网络结构是 VGG16[15]，将 VGG16 最后两个全连接层改成卷积层再增加 4 个卷积层构造网络结构。表 1 展示了整个 SSD 网络中每个卷积层中卷积核的大小、数目，卷积的步长，特征图有无填充以及每层输出特征图的大小。图 2 为 SSD 算法的目标检测流程图，SSD 检测算法分别把 conv4_3、fc7、conv6_2、conv7_2、conv8_2 和 conv9_2 等 6 个不同卷积层的特征图引出做检测，其特征图与两个 3×3 的卷积核卷积后得到两个输出，分别作为分类时使用的置信度以及回归时使用的位置信息。将每层计算结果合并后传递给损失层，该层对所有层的检测结果进行综合，通过非极大值抑制输出目标的检测结果。

表 1　SSD 网络参数表

Table 1　Parameters of SSD Network

卷积层	卷积核	卷积核数量	步长	填充	输出特征图像素大小
Conv1_1	3×3	64	1	1	300×300
Conv1_2	3×3	64	1	1	300×300
Maxpool1	2×2	1	2	0	150×150
Conv2_1	3×3	128	1	1	150×150
Conv2_2	3×3	128	1	1	150×150
Maxpool2	2×2	1	2	0	75×75
Conv3_1	3×3	256	1	1	75×75
Conv3_2	3×3	256	1	1	75×75
Conv3_3	3×3	256	1	1	75×75
Maxpool3	2×2	1	2	0	38×38
Conv4_1	3×3	512	1	1	38×38
Conv4_2	3×3	512	1	1	38×38
Conv4_3	3×3	512	1	1	38×38
Maxpool4	2×2	1	2	0	19×19
Conv5_1	3×3	512	1	1	19×19
Conv5_2	3×3	512	1	1	19×19
Conv5_3	3×3	512	1	1	19×19
Maxpool5	3×3	1	1	1	19×19
Fc6	3×3	1 024	1	1	19×19
Fc7	1×1	1 024	1	0	19×19
Conv6_1	1×1	256	1	0	19×19
Conv6_2	3×3	512	2	1	10×10
Conv7_1	1×1	128	1	0	10×10
Conv7_2	3×3	256	2	1	5×5
Conv8_1	1×1	128	1	0	5×5
Conv8_2	3×3	256	1	0	3×3
Conv9_1	1×1	128	1	0	3×3
Conv9_2	3×3	256	1	0	1×1

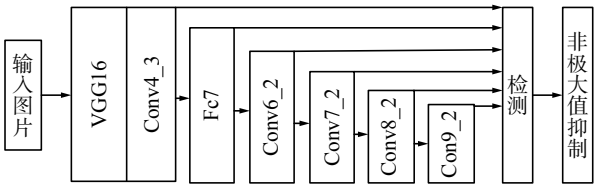

图 2　SSD 目标检测流程图

Fig. 2　Detection flowchart for SSD object algorithm

SSD 算法关键核心点是同时采用底层和顶层的特征图做检测。在不同层卷积输出的若干特征图中的每个位置处评估不同纵横比的默认框。默认框是指在特征图的每个网格上都有一系列固定大小的框。每个网格有 k 个默认框, 每个默认框预测 c 个目标类别的分数和 4 个偏移。若特征图的大小为 mn, 即有 mn 个特征图网格, 则该特征图共有 $(c+4)\times k\times m\times n$ 个输出。在训练阶段, 将默认框和真实框 (ground truth box) 进行匹配, 一旦匹配成功, 则默认框即为正样本, 反之则为负样本。根据置信度的损失值对负样本进行排序, 得到靠前的负训练样本, 使得正负样本的比例保持在 3:1。在预测阶段, 得到默认框的偏移及目标类别相应的置信度。

SSD 网络的目标损失函数表示为

$$L(x,c,l,g)=\frac{1}{N}(L_{\text{conf}}(x,c)+\alpha L_{\text{loc}}(x,l,g)) \tag{1}$$

式中: N 为匹配到的默认框个数; l 为预测框; g 为真实框; c 为多类别目标的置信度; L_{loc} 为位置损失; L_{conf} 为置信度损失; α 通过交叉验证设为 1。

位置损失是预测框 l 和真实框 g 之间的 smooth_{L1} 损失[29], 如式 (2) 所示, 通过对边界框的坐标中心点 (x, y) 以及宽度 w 和高度 h 的偏移进行回归, 使得位置损失最小。

$$L_{\text{loc}}(x,l,g)=\sum_{i\in(\text{Pos})}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}\text{smooth}_{L1}(L_i^m-\hat{g}_j^m) \tag{2}$$

式中: $\hat{g}^{cx}=(g_j^{cx}-d_i^{cx})/d_i^w$, $\hat{g}_j^{cy}=(g_j^{cy}-d_i^{cy})/d_i^h$, $\hat{g}_j^w=\log(g_j^w/d_i^w)$, $\hat{g}_j^h=\log(g_j^h/d_i^h)$; g_j^{cx}、g_j^{cy} 分别表示第 j 个真实框中心点 (x, y); d_i^{cx}、d_i^{cy} 分别表示第 i 个默认框的中心点 (x, y); g_j^w、g_j^h 分别表示第 j 个真实框宽度跟高度; d_i^w、d_i^h 分别表示第 i 个默认框的宽度跟高度。

置信度损失是多类别置信度 c 的 softmax 损失如式 (3) 所示。

$$L_{\text{conf}}(x,c)=-\sum_{i\in\text{Pos}}^{N}x_{ij}^{p}\log(\hat{c}_i^p)-\sum_{i\in\text{Neg}}\log(\hat{c}_i^0) \tag{3}$$

式中: $\hat{c}_i^p=\exp(c_i^p)/\sum_p\exp(c_i^p)$, $\hat{c}_i^p$ 表示第 i 个默认框的类别的置信度, p 表示目标的类别, 0 表示目标外的背景, x_{ij}^p 表示第 i 个默认框与类别 p 匹配的第 j 个真实框相。

相比现有的目标检测方法, SSD 算法不管是在检测速度还是检测精度上都取得了非常优秀的效果, 但受卷积神经网络中特定特征层感受野大小限制, 单独一层的特征无法应对多姿态多尺度的行人[30]。因此, 本文提出了改进的 SSD 模型用于行人检测。

3　多层卷积特征的行人检测算法

3.1　基于改进 SSD 的行人检测算法

随着深度学习的快速发展, CNN 已经广泛地应用于目标检测中, 在实时性和准确性上都优于传统算法的性能。SSD 算法是以 VGG16 网络模型为基础的前向传播的深度卷积网络模型, 对卷积后得到的特征图分别预判目标位置跟类别置信度, 实现快速且精准目标检测效果。但原始 SSD 算法对同一张图像中小尺寸目标检测效果较差, 主要原因有两点: 1) 输入图像在深度卷积神经网络中经过网络的卷积、池化后特征图变小, 原始 SSD 算法 conv4_3 输出的特征图与原始输入图像相比缩小至原来的 1/8, 特征图的变小导致检测的目标丢失了大部分的细节信息, 在训练阶段严重的影响了算法对各项参数的学习, 且后续的 fc7、conv6_2 等层输出的特征图缩小更多, 对算法的训练影响更大; 2) 输入图像分辨率的大小对 SSD 算法的影响。训练的图像较小, 卷积池化后得到的特征图会对应地减小, 使得训练阶段 SSD 网络参数的学习不完全造成过拟合; 若输入图像较大, 网络学习的参数大量增加, 使得算法计算复杂度增加, 速度减慢。

本文在权衡算法的精准度及实时性的基础上, 对 SSD 模型更底层的输出特征图进行特征提取, 获取更多特征图的纹理、边缘等细节信息, 增强了 SSD 模型对行人目标的检测性能, 提升对小目标行人的检测能力。图 3 为本文基于改进 SSD 模型的行人检测 (pedestrian detection based on improved SSD, PDIS) 框架, 行人图像通过改进 SSD 卷积网络中的各卷积层输出多层次特征图, 并在多层次的特征图上提取特征做检测, 将多层特征图的检测结果进行综合实现行人检测。由图 3 的特征图可视化结果可知, 底层卷积 conv3_3 输出的特征图比较大, 且纹理、轮廓信息明确, 因此该层的特征图可以提取到小目标行人的细节信息。conv9_2 卷积层输出的特征图变得很小, 原始图像的大部分信息丢失, 尤其小物体信息丢失严重, 因此该层只能获取较大目标的行人信息。随着网络层数增加, 原始图像的信息会随着输出特征的尺寸变小而减少。底层输出特征图可以检测较小的行人目标, 深层输出的特征可以检测较大的行人目标, 因此 PDIS 通过结合多层特征图检测结果, 提升了多尺寸行人的检测性能。

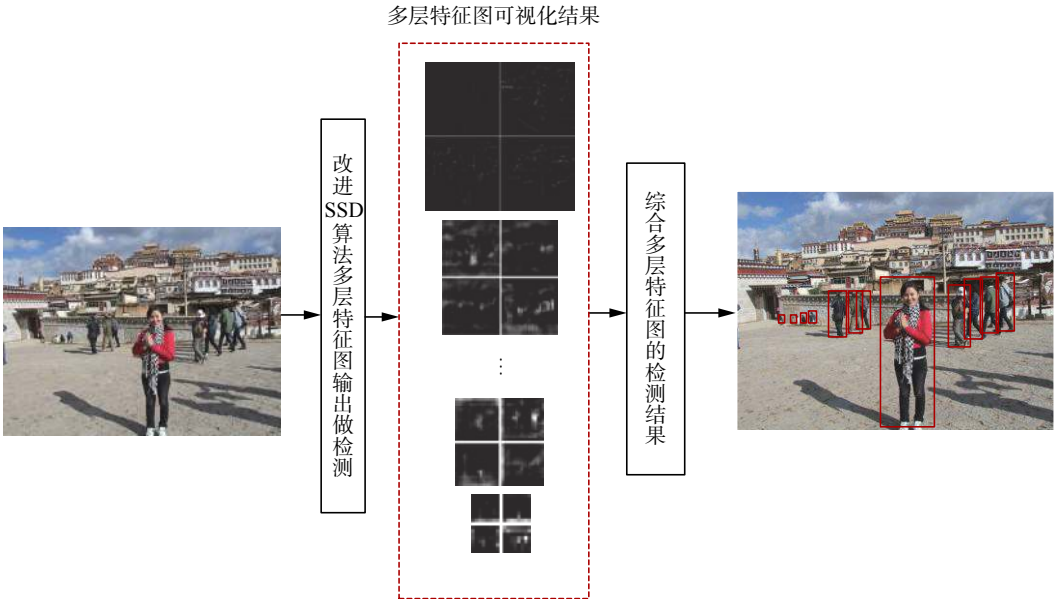

图 3　PDIS 框架

Fig. 3　PDIS framework

PDIS 模型通过引出 SSD 卷积网络中更底层 conv3_3 输出的特征图做检测。图 4 为本文改进之一的 PDIS 流程图，应用 SSD 算法卷积层 conv3_3、conv4_3、conv7、conv6_2、conv7_2、conv8_2 和 conv9_2 等 7 个输出层的特征图做检测，图中可以看出不同卷积层输出的特征图可以检测图像中不同尺度的行人，conv3_3 输出的特征图能检测到图像中尺度很小的行人，但对尺寸大的行人检测效果很差；卷积层 conv9_2 输出的特征图可以检测图像中的尺寸较大的行人，但对小尺寸行人检测效果不理想。因此 PDIS 模型把网络中各层次输出的特征图由底层到深层依次引出做检测，检测到行人目标尺寸越来越大。尽管每一层对图像整体的检测效果不理想，但综合所有层的检测达到了精准的行人检测结果。因此，本文通过修改后的 SSD 网络，应用多个卷积层输出的特征图做检测，实现了图像中多尺度的行人检测问题，增加了算法的行人检测分辨率，提升了对图像中尺寸相对较小行人的检测效果。

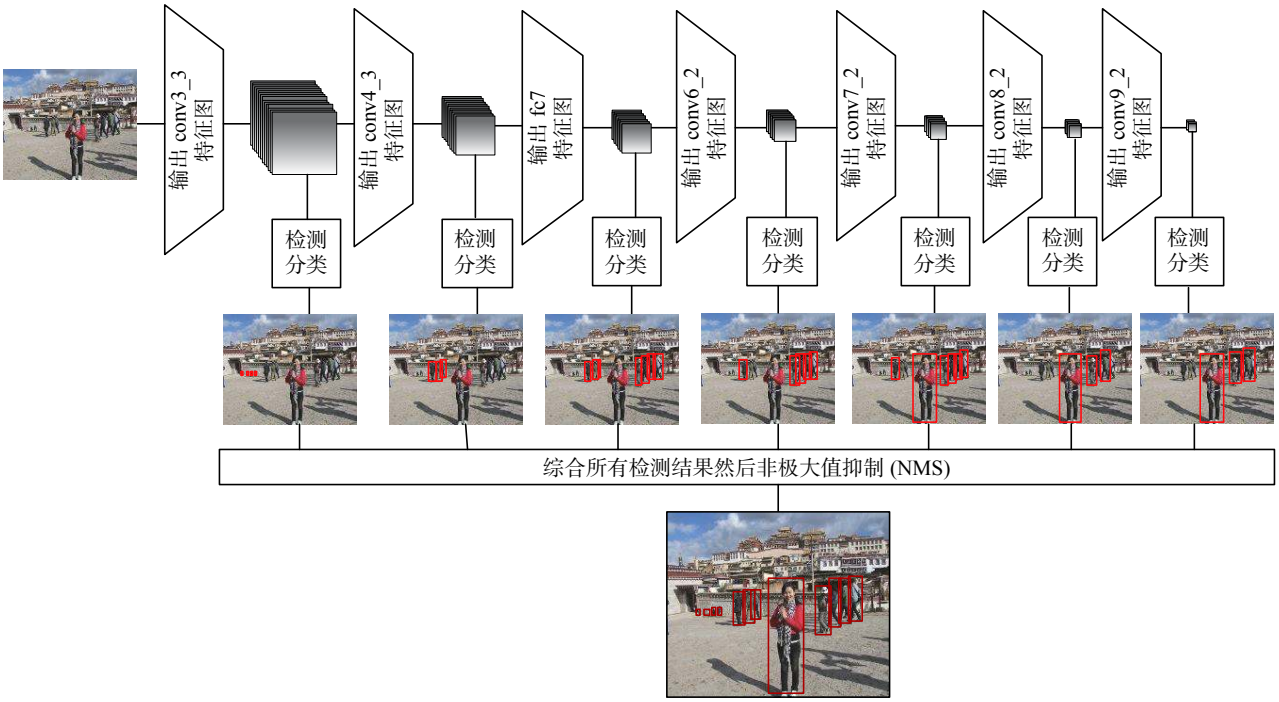

图 4　PDIS 流程

Fig. 4　PDIS flowchart

PDIS 在每层输出特征图上的每个特征图网格预设有 4 个默认框, 在不同输出层的特征图上有不同尺寸大小的默认框, 在同一个特征图上又有不同纵横比的默认框, 实现了图像中各种形状和尺寸大小的目标检测。行人的高度跟宽度之比一般在 1∶1、2∶1、3∶1、1∶2 这些比例内, 不同于原始 SSD 算法, 本文将默认框的纵横比 a_r 设置为符合行人的 4 种比例分别为 1∶1、2∶1、3∶1、1∶2, 加速了行人区域定位。且默认框的尺度定义为

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), k \in [1, m] \tag{4}$$

式中: $w_k^a = s_k\sqrt{a_r}$, $h_k^a = s_k/\sqrt{a_r}$, 当 $a_r = 1$ 时 $s_k' = \sqrt{s_k s_k + 1}$; s_k、s_k' 表示默认框的尺寸, $s_{\min}$、$s_{\max}$ 分别表示 0.1 和 0.9, m 表示 PDIS 中间层输出做预测特征图层数, w_k^a、h_k^a 分别表示每一个默认框的宽度跟高度。

PDIS 模型融合多层特征图的特征做检测, 解决了同一张图像中不同尺度的目标检测问题。通过研究用不同层输出特征图以及网络输出层的数量对算法行人性能的影响, 本文在原始 SSD 的基础上引出更底层卷积 conv3_3 的特征图做检测, 在该层的特征图上提取特征获得了原始输入行人图像的更多细节信息。实验表明引出卷积 conv3_3 的输出特征图做检测, 特征维度的计算复杂度相应的增加, 实时性相比原始 SSD 算法有所下降, 但依然能满足行人检测实时性的要求, 并相比原始 SSD 算法对小尺寸行人检测性能提升很高。同时, 研究了融合不同卷层基输出的特征图对算法的影响。在引出底层 conv3_3 的输出特征图做检测的基础上, 训练了多个组合不同输出特征图的网络模型: 模型 2 引出 conv3_3、conv4_3、fc7、conv6_2、conv7_2、conv8_2 和 conv9_2 等 7 个卷积层的特征图做检测, 模型 3 引出 conv3_3、fc7、conv6_2、conv7_2、conv8_2 和 conv9_2 等 6 个卷积层的特征图做检测, 模型 4 引出 conv3_3、conv5_3、conv6_2、conv7_2、conv8_2 和 conv9_2 等 6 个卷积层的特征图做检测, 模型 5 引出 conv3_3、conv5_3、conv6_2、conv8_2 和 conv9_2 等 5 个卷积层的特征图做检测。在扩增的行人数据集分别训练各个 PDIS 网络模型, 利用 INRIA 行人数据集的测试集分别对模型进行测试。实验表明: 不同的网络模型在引出不同的特征层以及引出不同层的数目直接影响网络模型的检测效果, 改进的模型 2 取得了最好的检测性能。

此外, 为了进一步提升 PDIS 模型对小目标的检测能力, 通过增加输入图像的分辨率提升算法检测性能的鲁棒性。原始的 SSD 算法输入图像大小为 300×300。在 CNN 中经过卷积、池化特征图不断减小, 原始 SSD 算法最底层 conv4_3 引出特征图大小为 38×38,相比原始图像缩小至原来的 1/8, 在原始图像中一个 8×8 的目标在 conv4 输出的特征图表现为一个像素点, 该目标的细节信息完全丢失。卷积输出的特征图会随网络层数的增加而减小。导致训练阶几乎无法学习到小目标物体的信息。因此, 数据集的训练图像分辨率大小很大程度影响了卷积神经网络的学习, 训练图像分辨率太小, 训练时模型很难收敛, 检测精度低。本文把训练预设输入图像尺寸大小从 300×300 变大到 512×512, 增加卷积后输出特征图的分辨率, 能获得原图像中更加丰富、更加细节的信息。测试结果表明: 用放大的行人图像训练 PDIS 模型能够检测到同一张图像中更小尺寸的行人, 进一步提升了行人小目标检测效果。

3.2　数据集扩增

增加数据集的多样性来训练 PDIS 模型可以增强算法检测的泛化能力。一般使用单一的行人数据集训练卷积网络模型时, 在其本身数据集上测试的效果会很理想, 然而在其他数据集上测试时效果往往不好。因此, 行人数据集所包含的样本的数量、样本背景的多样性以及样本中有无对小尺寸行人目标的标注等因素, 在训练 CNN 的过程中会严重影响算法的学习。在训练卷积神经网络时, 深度模型学习的参数往往比较多, 用于训练的样本数据量太少, 容易造成网络过拟合。此外, 现有的行人数据集公布时间较早, 而且训练样本基本是从较大图像中剪切出的单个行人图像, 背景单一, 像素分辨率普遍偏低, 因此不适合用于训练深度卷积网络模型。

为了增强 PDIS 模型在行人检测领域的泛化能力, 本文对已有的 INRIA 行人数据集进行了扩增。首先, INRIA 行人数据集的选取: INRIA 行人数据集是目前使用最多的静态行人检测数据集。其中包含沙滩、机场、城市、山等复杂的场景, 且拍摄条件多样, 存在光线变化、人体遮挡等情形, 符合本文所需求的行人样本的背景多样性。其次, 扩增 INRIA 行人数据集: INRIA 行人数据集中训练集的正样本只包含 614 张图像 (包含 2 416 个行人), 用于训练 CNN 模型的数量远远不够。本文在各种天气、场景、光照下采集了数万张图像, 并对图像进行人工筛选标注, 目前已有 5 000 多张图像用于训练。部分数据如图 5 所示, 扩增的行人数据集中包含学校、街道、车站等不同场景下的样本, 组合成一个复杂背景下的真实场景行人数据集, 并对训练样本中姿态变化、遮挡、小

目标的行人都进行了标注，如：骑自行车、打伞、拥挤，图像中像素很小的行人等。扩增的数据集图像使得行人数据集样本背景复杂化、多样化，并大大增加了对小目标行人标注数目。实验表明：采用本文扩充的行人数据集训练 PDIS 模型，不管在真实场景下还是小目标行人检测 PDIS 都取得了非常优秀的效果。

图 5　真实场景下的训练样本

Fig. 5　Training samples in real scenes

4　实验结果与分析

4.1　性能评价指标

本文应用漏检率、准确率来衡量检测算法的性能，通过在 INRIA 行人数据集的测试集上测试训练好的模型，记录每张图像检测窗口，计算检测框跟真实框的 IOU 值。假设检测框为 BB_{dt}，真实框为 BB_{gt}，若 IOU 值大于阈值时，则 BB_{dt} 与 BB_{gt} 是匹配的。本文 IOU 设定的阈值为 0.5，如 (5) 式所示：

$$\mathrm{IOU}=\frac{\mathrm{area}(BB_{dt}\cap BB_{gt})}{\mathrm{aera}(BB_{dt}\cup BB_{gt})}>0.5 \tag{5}$$

在 BB_{dt} 与 BB_{gt} 匹配过程中，未匹配到的 BB_{dt} 是误检的行人框 (false positive,FP)，未匹配的 BB_{gt} 是漏检的行人框 (false negative,FN),漏检率统计用到的标准如表 2 所示

表 2　行人统计量

Table 2　Pedestrian statistics

分类结果	真实值	
	行人 (Positive)	非行人 (Negative)
行人 (Positive)	True Positive(TP)	False Positive(FP)
非行人 (Negative)	False Negative(FN)	True Negative(TN)

漏检率 R_M (Miss Rate) 定义为

$$R_M=\frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} \tag{6}$$

准确率 R_P (Precision Rate) 定义为

$$R_P=\frac{\mathrm{TP}}{\mathrm{TN}+\mathrm{FP}} \tag{7}$$

式中：TP、FP、FN 分别表示将行人样本分类成行人样本数、将非行人样本分类为行人样本数、将行人样本分类成非行人样本数。

4.2　实验环境与模型对比

本文的实验环境为 Ubuntu14.04 系统，处理器型号为 Intel® Xeon(R) CPU E5-2670 v3 @ 2.30 GHz×24，显卡型号为 GeForce GTX TITAN X，显存 12 GB，内存 32 GB。

本文在 INRIA 行人数据集上，扩增成一个 5 000 多张图像的数据集，在该数据集上训练了的 6 个不同的模型，如表 3 所示，输入图像的大小会直接影响算法的精度和实时性，SSD 300×300 比 SSD 512×512 输入图像小，在 INRIA 的测试集上测试，一张图像的平均测试时间快了一倍多，但检测精度有所下降；将原始的 SSD 网络模型 conv3_3 的特征图引出做检测，并在此基础上融合多个卷基层的特征图做检测，提出了表 3 的 4 个检测模型。实验表明引出 conv3_3 的特征图的模型，相比原始 SSD 模型在每张图像的平均检测速度有所下降，但测试一张图像最慢速度依然达到 0.16s，满足行人检测的实时性要求，并相比原来的 SSD 算法，本文算法准确率达到了 93.8%，漏检率下降至 7.4%。

表 3　不同模型的检测率

Table 3　Detection rates of different models

模型	conv 3_3	conv 4_3	conv 5_3	fc 7	conv 6_2	conv 7_2	conv 8_2	conv 9_2	漏检率/%	时间/s
SSD 300×300		√		√	√	√	√	√	12.1	0.02
SSD 512×512		√		√	√	√	√	√	10.7	0.07
模型 2	√	√		√	√	√	√	√	7.4	0.16
模型 3	√			√	√	√	√	√	10.9	0.11
模型 4	√		√		√	√	√	√	8.7	0.14
模型 5	√		√		√		√	√	9.0	0.14

注：打钩的表示该层卷积输出的特征图被引出

6 种模型在 INRIA 的测试集上 R_M-FPPI 曲线如图 6 所示，模型 2 在 INRIA 的测试集上取得了最好的检测效果。

在训练过程中采用本文扩增的数据集，分别用 300×300 与 512×512 的图像训练原始的 SSD 模型及本文改进的模型 2，如图 7 所示，使用 300×300 的图像训练原始模型时，无法学习到扩增数据集

中的小尺寸行人，导致训练 loss 曲线收敛效果最差，而增加输入图像大小能有效提高收敛效果，利用大小相同的图像分别训练 SSD 模型与 PDIS 模型，PDIS 的 Loss 曲线收敛效果较好。因此，本文增加图像大小来训练 PDIS 模型，能够得到最好的检测模型。

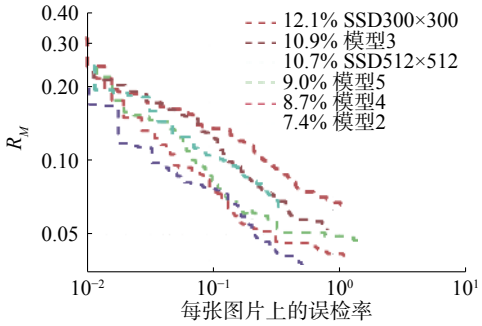

图 6　6 种模型在 INRIA 数据集上 R_M-FPPI 曲线

Fig. 6　Miss Rate-FPPI curves of 6 models on INRIA dataset

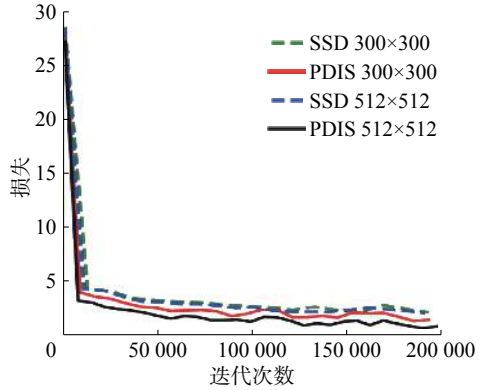

图 7　不同算法的 loss 下降曲线

Fig. 7　Loss curves of different algorithms

4.3　与现有流行算法性能对比

利用当前流行的多个算法在 INRIA 的测试集上进行测试，实验表明：在扩增的 INRIA 行人数据集训练 PDIS 模型的测试漏检率比现有的比较流行的算法都要低，如表 4 所示，本文的算法取得了最好的效果。

表 4　不同算法在 INRIA 行人数据集的漏检率

Table 4　Miss rates of different algorithms in the INRIA pedestrian dataset　%

算法	漏检率
HOG[4]	45.8
ACF[8]	17.2
LDCF[9]	13.8
R-CNN[13]	12.77
Faster R-CNN[14]	17.6
本文算法	7.4

不同算法在 INRIA 的测试集上 Miss Rate-FPPI 曲线如图 8 所示，可以看出本文算法在 INRIA 的测试集上取得了最好的检测效果。

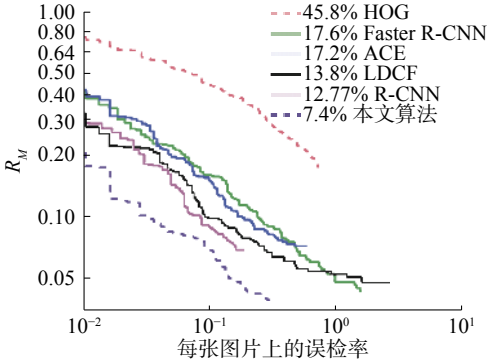

图 8　不同算法在 INRIA 数据集上 R_M-FPPI 曲线

Fig. 8　Miss Rate-FPPI curves of different algorithms on INRIA dataset

本文对真实场景中拍摄的 200 张图像进行了测试，从中挑选了 3 张代表性的图像在 2 个不同的模型上分别检测，其结果如图 9 所示，可以看出，针对图像中的大目标行人，PDIS 与 SSD 相比具有同等的检测能力甚至更优；而对图像中的小目标行人，SSD 检测性能很差，但 PDIS 在小目标检测上达到了非常好的性能，漏检率更低。

(a) SSD 算法　　(b) 本文算法

图 9　不同算法的行人检测结果

Fig. 9　Pedestrian detection results for different algorithms

5　结束语

本文通过引出 SSD 网络模型中更底层特征图做检测以及增加输入图像大小来增加深度模型的分辨率，相比原始的 SSD 算法，改进的 SSD 模型提高了对小目标行人的检测性能。另外增加 INRIA 数据集的数量跟多样性也是本文算法检测性能提升的主要原因。尽管通过扩增的 INRIA 数据集训练改进的 SSD 模型取得较好的检测效果，但检测性能还有待优化。下一步研究工

作主要针对两点：1）应用本文算法在多个基准行人数据集（如 Caltech 行人数据集等）上进行实验，针对每个数据集的测试结果进行统计分析，优化本文算法的检测性能；2）继续扩充行人数据集的数量跟多样性能够进一步的提升算法的检测性能。

参考文献：

[1] 宋婉茹, 赵晴晴, 陈昌红, 等. 行人重识别研究综述[J]. 智能系统学报, 2017, 12(6): 770–780.
SONG Wanru, ZHAO Qingqing, CHEN Changhong, et al. Survey on pedestrian re-identification research[J]. CAAI transactions on intelligent systems, 2017, 12(6): 770–780.

[2] YE Qixiang, LIANG Jixiang, JIAO Jianbin. Pedestrian detection in video images via error correcting output code classification of manifold subclasses[J]. IEEE transactions on intelligent transportation systems, 2012, 13(1): 193–202.

[3] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Proceedings of 2016 European Conference on Computer Vision. Cham, Germany, 2016: 21–37.

[4] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886–893

[5] 苏松志, 李绍滋, 陈淑媛, 等. 行人检测技术综述[J]. 电子学报, 2012, 40(4): 814–820.
SU Songhi, LI Shaozi, CHEN Shuyuan, et al. A survey on pedestrian detection[J]. Acta electronica sinica, 2012, 40(4): 814–820.

[6] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91–110.

[7] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference Computer Vision and Pattern Recognition. Kauai, USA, 2001: 511–518.

[8] FERREIRA A J, FIGUEIREDO M A T. Boosting algorithms: a review of methods, theory, and applications[M]. New York, USA: Springer, 2012: 35–85.

[9] VAPNIK V. The nature of statistical learning theory[M]. 2nd eds. New York: Springer-Verlag, 2000.

[10] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1): 5–32.

[11] DOLLÁR P, APPEL R, BELONGIE S, et al. Fast feature pyramids for object detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(8): 1532–1545.

[12] NAM W, DOLLÁR P, HAN J H. Local decorrelation for improved detection[J]. Advances in neural information processing systems, 2014, 1: 424–432.

[13] ZHANG Shanshan, BENENSON R, SCHIELE B. Filtered channel features for pedestrian detection[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1751–1760.

[14] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25(2): 1097–1105.

[15] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2014.

[16] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 580–587.

[17] REN Shaoqing, HE Kaiming, 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.

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

[19] 王梦来, 李想, 陈奇, 等. 基于 CNN 的监控视频事件检测[J]. 自动化学报, 2016, 42(6): 892–903.
WANG Menglai, LI Xiang, CHEN Qi, et al. Surveillance event detection based on CNN[J]. Acta automatica sinica, 2016, 42(6): 892–903.

[20] HOSANG J, OMRAN M, BENENSON R, et al. Taking a deeper look at pedestrians[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 4073–4082.

[21] BENENSON R, OMRAN M, HOSANG J, et al. Ten years of pedestrian detection, what have we learned?[C]//Proceedings of 2014 European Conference on Computer Vision. Cham, Germany, 2015: 613–627.

[22] 吕静, 高陈强, 杜银和, 等. 基于双通道特征自适应融合的红外行为识别方法[J]. 重庆邮电大学学报(自然科学版), 2017, 29(3): 389–395.
LYU Jing, GAO Chenqiang, DU Yinhe, et al. Infrared action recognition method based on adaptive fusion of dual channel features[J]. Journal of Chongqing university of posts and telecommunications (natural science edition), 2017, 29(3): 389–395.

[23] TIAN Yonglong, LUO Ping, WANG Xiaogang, et al.

Deep learning strong parts for pedestrian detection[C]// Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1904–1912.

[24] 张雅俊, 高陈强, 李佩, 等. 基于卷积神经网络的人流量统计[J]. 重庆邮电大学学报(自然科学版), 2017, 29(2): 265–271.
ZHANG Yajun, GAO Chenqiang, LI Pei, et al. Pedestrian counting based on convolutional neural network[J]. Journal of Chongqing university of posts and telecommunications (natural science edition), 2017, 29(2): 265–271.

[25] ZHANG Liliang, LIN Liang, LIANG Xiaodan, et al. Is faster r-cnn doing well for pedestrian detection?[C]//Proceeding of 2016 European Conference on Computer Vision. Cham, Germany, 2016: 443–457.

[26] ENZWEILER M, GAVRILA D M. Monocular pedestrian detection: survey and experiments[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(12): 2179–2195.

[27] MOHAN A, PAPAGEORGIOU C, POGGIO T. Example-based object detection in images by components[J]. IEEE transactions on pattern analysis and machine intelligence, 2001, 23(4): 349–361.

[28] OVERETT G, PETERSSON L, BREWER N, et al. A new pedestrian dataset for supervised learning[C]//Proceedings of 2008 IEEE Intelligent Vehicles Symposium. Eindhoven, Netherlands, 2008: 373–378.

[29] GIRSHICK R. Fast R-CNN[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440–1448.

[30] 王成济, 罗志明, 钟准, 等. 一种多层特征融合的人脸检测方法[J]. 智能系统学报, 2018, 13(1): 138–146.
WANG Chengji, LUO Zhiming, ZHONG Zhun, et al. Face detection method fusing multi-layer features[J]. CAAI transactions on intelligent systems, 2018, 13(1): 138–146.

作者简介:

伍鹏瑛, 男, 1990 年生, 硕士研究生, 主要研究方向为计算机视觉、模式识别。

张建明, 男, 1976 年生, 副教授, 博士, 主要研究方向为计算机视觉、智能交通系统。发表学术论文 50 余篇, 其中 EI 收录 26 篇, SCI 收录 9 篇。

彭建, 男, 1971 年生, 副教授, 主要研究方向为目标检测、计算机视觉。发表学术论文 20 余篇。