基于多阶段提议稀疏区域卷积网络的城市交通目标检测

柳长源[1],张玉亮[1],毕晓君[2]
(1. 哈尔滨理工大学测控技术与通信工程学院,黑龙江哈尔滨 150080;2. 中央民族大学信息工程学院,北京 100081)

摘 要: 针对城市交通场景多目标检测算法检测速度慢,检测精度低等问题,本文提出多阶段提议稀疏区域卷积网络算法(Multi-stage Proposal Sparse Region-based Convolutional Neural Network,MPS R-CNN). 算法主要有以下特点:提出了一种多阶段提议框过滤更新机制,提高算法检测精度;提出了一种双向并联特征金字塔网络(Bidirectional Parallel Feature Pyramid Network,BPFPN),增强了模型的特征融合能力;针对城市交通场景目标检测问题引入了Copy-Paste数据增强方法和CIoU损失函数. 实验结果显示,MPS R-CNN算法在Urban Object Dataset数据集上mAP达到了77%,算法检测速度保持在37 fps,优于目前其他城市交通场景目标检测算法.

关键词: 目标检测;城市交通;提议过滤;特征金字塔;数据增强
基金项目: 国家自然科学基金(No.51779050);黑龙江省自然科学基金(No.F2016022)
中图分类号: TP391.4;TP181 **文献标识码:** A **文章编号:** 0372-2112(2023)01-0026-06
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211648

Urban Traffic Object Detection Based on Multi-Stage Proposal Sparse R-CNN

LIU Chang-yuan[1], ZHANG Yu-liang[1], BI Xiao-jun[2]
(1. *College of Measurement and Control Technology and Communication Engineering*, *Harbin University of Science and Technology*, *Harbin*, *Heilongjiang* 150080, *China*;
2. *School of Information Engineering*, *Minzu University of China*, *Beijing* 100081, *China*)

Abstract: Aiming at the slow speed and low accuracy of multi-object detection algorithms in urban traffic scenes, this paper proposes a multi-stage proposal sparse region-based convolutional neural network algorithm (MPS R-CNN). The algorithm mainly has the following characteristics: a multi-stage proposal box filtering update mechanism is proposed to improve the detection accuracy of the algorithm; a bidirectional parallel feature pyramid network (BPFPN) is proposed to enhance the model feature fusion capability; for the problem of object detection in urban traffic scenes, the Copy-Paste data augmentation method and CIoU loss function are introduced. The experimental results show that the MPS R-CNN algorithm achieves 77% mAP on the urban object dataset, and the algorithm detection speed remains at 37 fps, which is better than other current urban traffic object detection algorithms.

Key words: object detection; urban traffic; proposal filtering; feature pyramid; data augmentation
Foundation Item(s): National Natural Science Foundation of China (No.51779050); Natural Science Foundation of Heilongjiang Province (No.F2016022)

1 引言

目标检测的发展为自动驾驶辅助系统、无人驾驶汽车自动导航等未来智能交通系统奠定了基础. 经典目标检测主要利用滑动窗口或图像分割方法来产生大量候选区域,对候选区域进行特征提取,常用方法如SIFT[1]等,最后将局部特征传递给分类器进行识别. 传统方法检测效果难以满足实际应用需求. 近年来,大量基于深度学习的目标检测算法被提出. 这些算法通常分为两大类:(1)基于候选区域(Region Proposal)的二阶段目标检测算法,如R-CNN[2]等;(2)基于回归的一阶段目标检测算法,如YOLO[3]等. 前者使用区域提议网络生成候选区域,后者直接在特征图上生成候选框.

收稿日期:2021-12-12;修回日期:2022-07-01;责任编辑:李勇锋

目前,两类方法都有着大量的应用. 李宝奇[4]等人采用并行附加特征的SSD网络检测地面小目标. 伍锡如等人改进MaskR-CNN[5]算法用于交通场景的多目标检测和分割.

目前针对城市交通场景的目标检测方法如RetinaNet[6]等,仍存在检测精度低,检测速度慢的问题. Sparse R-CNN[7]是2020年Sun等人提出的目标检测算法,使用ResNet-50[8],FPN[9]模型在MSCOCO数据集上检测速度为40 fps,检测精度mAP50为63.4%,是目前最优秀的目标检测算法之一.

2　Sparse R-CNN算法简介

Sparse R-CNN由骨干网络、动态实例交互头和两个预测层组成,结构简洁. 算法采用稀疏提议结构,提议框数量和交互特征远少于传统方法[10],降低了计算量. 其采用ResNet与FPN作为主干网络,使用可学习提议框在特征图上提取感兴趣区域特征,然后通过动态交互头与可学习提议特征交互得到特征向量用于分类和回归,并重复多次. 算法使用Focal Loss[6]函数计算分类损失,使用GIoU损失函数计算定位损失. 同时,算法使用基于固定数量的集合预测损失来进行分类和回归. 在预测结果与真实结果之间进行最优二分图匹配,匹配代价如式(1)所示:

$$L=\lambda_{\text{cls}}\cdot\mathcal{L}_{\text{cls}}+\lambda_{L1}\cdot\mathcal{L}_{L1}+\lambda_{\text{giou}}\cdot\mathcal{L}_{\text{giou}} \tag{1}$$

其中,$\mathcal{L}_{\text{cls}}$是预测类别与真实类别的焦点损失(Focal Loss),$\mathcal{L}_{L1}$和$\mathcal{L}_{\text{giou}}$是预测框与真实框的归一化的中心坐标、宽度和高度之间的L1损失和GIoU损失,$\lambda_{\text{cls}}=2$,$\lambda_{L1}=5$和$\lambda_{\text{giou}}=2$是各部分损失的比例系数.

模型采用基于集合的损失,避免了R-CNN系列算法中多对一匹配的问题,但存在随检测框数量增加检测精度提高而检测速度降低的问题.

3　MPS R-CNN算法

本文提出的MPS R-CNN算法,继承Sparse R-CNN算法稀疏、简洁的特点,并进一步提高. 为此,本文提出多阶段提议过滤机制,BPFPN结构,引入Copy-Paste[11]数据增强和CIoU[12]损失函数,MPS R-CNN算法整体结构如图1所示.

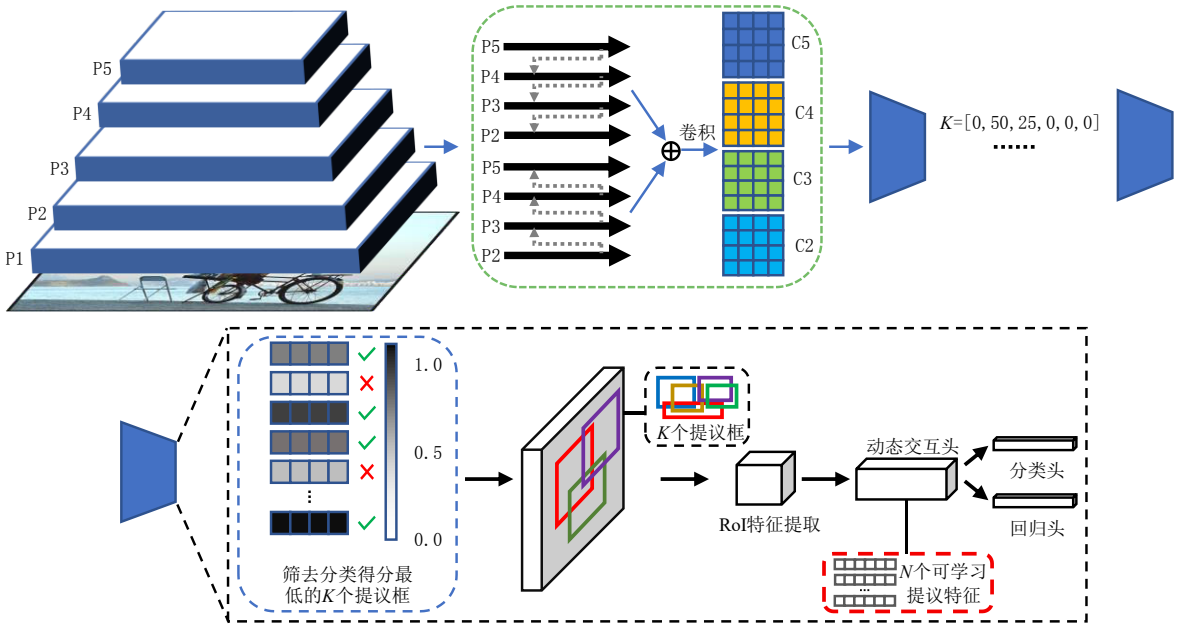

图1　MPS R-CNN算法整体结构

3.1　双向并行特征金字塔

城市交通状况复杂,物体检测易受干扰,需结合多层次特征提高精度. 受PAFPN[13](Path Aggregation Feature Pyramid Network)启发,本文提出了图2所示双向并行特征金字塔网络(Bidirectional Parallel Feature Pyramid Network,BPFPN). 其中P2~P5是ResNet50提取的4种含不同层次信息的特征图,向上的箭头表示上采样,反之表示下采样. BPFPN采用并行双向特征融合,精度与PAFPN相当,高于自顶向下融合的FPN,同时在检测速度上快于PAFPN.

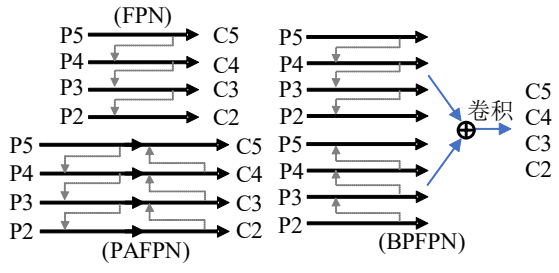

图2　特征金字塔网络结构对比图

3.2 多级提议过滤机制

SparseR-CNN算法在开始阶段引入可学习提议框，若初始N个提议框未覆盖图中所有目标，后续环节也将丢失目标，且随检测框数量增加检测速度会下降. 本文提出图1蓝色虚线框所示的提议过滤机制(即算法1)，对前三个环节提议框进行过滤筛选，将分类得分最低的K个提议框参数重置. 算法检测精度提高7%mAP，速度仅降低1 fps，检测效果超过使用$N+K$个可学习提议框的Sparse R-CNN模型.

算法1 多级提议过滤

输入:提议框个数N=100

上级预测框集合$A=\{(x_i,y_i,w_i,h_i,\text{score}_i)\mid i\in[1,\cdots,N]\}$

输出:新的提议框集合B

$b\leftarrow\{0.5,0.5,1,1,0.5\}$

$K\in[0,50,25,0,0,0]$

$G\leftarrow$ sort A by score

FOR $i\leftarrow 1$ to N DO

  FOR $j\leftarrow 1$ to K DO

    IF $A[i]=G[j]$ THEN

      $A[i]\leftarrow b$

    END IF

    $j\leftarrow j+1$

  END FOR

  $i\leftarrow i+1$

END FOR

$B\leftarrow A$

3.3 CIoU

本文使用CIoU损失替换GIoU损失进行目标边界框回归. 解决了预测框与真实框为垂直或包含关系时GIoU退化成普通IoU，优化能力严重下降的问题. 式(2)所示的CIoU损失函数能直接最小化中心点间的距离，将边框纵横比也考虑在内，提升检测框回归的合理性和检测精度.

$$\mathcal{L}_{\text{CIoU}}=1-\text{IoU}+\frac{\rho^2(b,b^{\text{gt}})}{c^2}+\alpha v$$
$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{\text{gt}}}{h^{\text{gt}}}-\arctan\frac{w}{h}\right)^2 \tag{2}$$
$$\alpha=\frac{v}{(1-\text{IoU})+v}$$

其中，b,b^{gt}分别为预测框和真实框中心点，ρ表示两个中心点间的欧氏距离，IoU为两框交并比，c为同时包含预测框和真实框的最小闭包区域的对角线距离，v用来衡量长宽比的相似性，α用来平衡各项之间的重要性，$w,h,w^{\text{gt}},h^{\text{gt}}$分别表示预测框的宽高和真实框的宽高.

3.4 数据增强

在本文使用的城市目标数据集[14](Urban Object Dataset)中，存在大量小尺度目标，如交通标志、交通灯、远处的车辆行人等，本文使用Copy-Paste数据增强提高模型对小目标的检测能力. 采用随机选择图中目标复制后再随机粘贴到图中任意位置的方式进行样本扩充，操作简单且效果良好. 为减少目标背景对待检测目标的干扰，本文还进行了目标背景抠除的Copy-Paste matting数据增强实验进行对比. 其中，Copy-Paste与Copy-Paste matting数据增强使用如图3所示的方式完成.

图3 Copy-Paste(中)与Copy-Paste matting(右)数据增强效果

4 实验及结果分析

4.1 实验数据集

本文采用RoViT(Robotics & Tridimensional Vision Research Group)组织公布的城市目标检测数据集[14]进行实验，数据集包含7类常见城市目标共106 917张，分别为自行车、公交车、汽车、摩托车、人、交通灯、交通标志，随机划分训练集、验证集、测试集比例约为4:4:2. 样本分布如图4所示，较多小尺度目标和遮挡目标，部分类别样本量较少，符合真实情况.

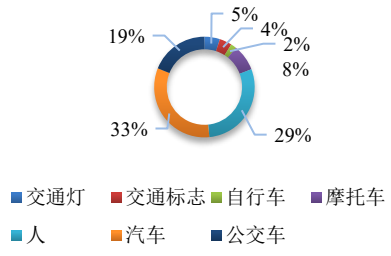

图4 各类样本比例分布图

4.2 实验环境

实验硬件环境采用IntelCore i9-9900K处理器，GTX 1080Ti显卡，软件环境为CUDA10.2和cuDNN9.2，使用Pytorch1.6深度学习框架. 在训练中使用AdamW优化器进行参数优化，Batchsize为32，学习率0.000 25，共训练40轮.

4.3 实验结果分析

目标检测中，使用预测目标框与真实框的交并比(Intersection over Union, IoU)来评价模型检测效果，IoU>0.5即为成功预测目标位置. 精确率(Precision)与召回率(Recall)可以多角度地评价模型性能. 平均精确度(Aver-

age Precision,AP)表示不同召回率下精确率的均值,用于评价单类别检测效果,所有类别检测精确度取平均可得到评价目标检测算法整体性能的平均精确度均值(Mean Average Precision, mAP)指标. 每秒检测图片数量(Frames Per Second,fps)用于评价模型检测速度.

表1是本文改进点实验结果,带Δ标记表示采用该改进结构,实验1是SparseR-CNN模型的检测效果,实验2至8是本文各改进点消融实验结果,实验9是本文提出的MPS R-CNN模型的检测效果. 接下来本文将根据实验结果逐一分析. 实验1、2、3显示,BPFPN结构与PAFPN结构检测精度相近且均高于原模型,mAP相差仅0.01%,比PAFPN结构快0.6 fps. 这表明BPFPN结构具有较强的特征融合能力,能很好地融合高层语义信息与低层位置信息. 实验1、4、5显示,增加提议框数量能显著提升算法检测精度,但对模型检测速度影响较大,降低了6.4 fps. 本文提出的多级提议过滤机制比直接采用N+K个提议框的单级多框提议模型在检测精度上提高5.69%mAP,在检测速度上提高4.8 fps,性能十分出色.

图5显示,多级提议过滤机制对遮挡目标(自行车、人)和小目标(交通灯、交通标志)的检测能力提升显著. 提高了模型对这类难检样本的捕获能力,使模型后续阶段重新检测到遗漏目标,对低品质目标提议框的过滤进一步提升检测精度,优于单纯增加提议框和特征数量的Sparse R-CNN单级多框提议模型.

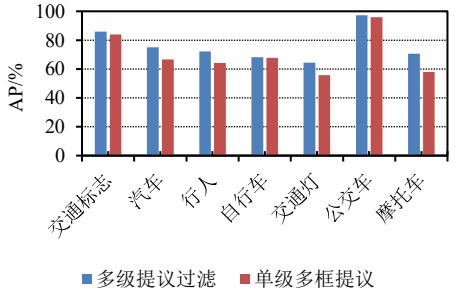

图5　多级提议过滤与单级多框提议模型对比

实验1、6显示,CIoU损失增强了模型对数据集的拟合能力,提升算法泛化能力. CIoU损失相对于Sparse R-CNN采用的GIoU损失在检测精度上提升了0.45%mAP,这表明CIoU损失设计更合理,有利于模型得到更优的解.

实验1、7、8显示,Copy-Paste方法提升了算法检测精度0.32%mAP,Copy-Pastematting方法使得检测精度降低了0.28%. 两种方法都扩充了中小尺度样本的数量,后者还剔除了目标背景的干扰,但受抠图效果影响,Copy-Paste matting方法未能带来提升. 实验发现,现有抠图算法在交通场景下易造成图6所示中小尺度目标的缺损或不合理赘余,对训练过程产生了负面影响. 因而本文选用Copy-Paste方法作为最终方法.

实验1、9显示,MPS R-CNN算法检测精度较原模型提升7.41%mAP,速度下降3 fps,满足实时性需求. 如图7,改进前模型对遮挡以及小尺度目标存在较多漏检、误检,改进后显著减少.

图6　抠图算法带来的目标缺损和不合理赘余

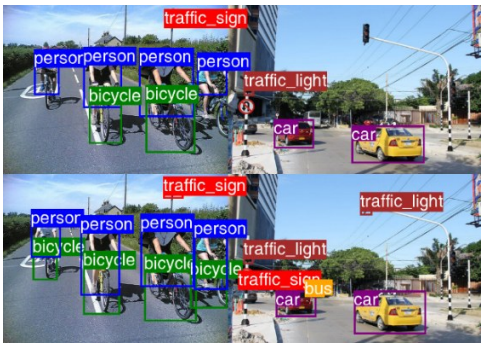

图7　模型改进前(上)与改进后(下)检测效果对比

表1　模型改进实验结果

实验编号	PAFPN	BPFPN	多级提议	单级多框提议	CIoU	Copy-Paste	Copy-Paste matting	mAP/%	fps
1								70.32	40.0
2	Δ							70.51	35.8
3		Δ						70.52	36.4
4			Δ					76.23	38.4
5				Δ				70.54	33.6
6					Δ			70.77	40.1
7						Δ		70.64	39.9
8							Δ	70.04	40.0
9		Δ	Δ		Δ	Δ		77.73	37.1

本文还与YOLOv5[15]，EfficientDet[16]，FlowR-CNN[17]，MaskR-CNN[5]，YOLOv2[3]，Faster R-CNN[AutoUpdated][14]等先进目标检测算法进行了比较. 实验采用的图片长宽均为320像素，本文选择模型输入尺寸最接近的YOLOv5-S和EfficientDet-D0模型作为对照，表2实验结果显示：MPS R-CNN算法与SparseR-CNN算法检测速度相当，检测精度上大幅提升7.41%mAP，比当下优秀的YOLOv5-S算法高2.43%mAP；本文提出的MPS R-CNN算法在检测性能处于前列，优于以往城市交通场景多目标检测算法.

表2 各目标检测算法性能对比

检测算法	mAP/%	fps
Sparse R-CNN	70.32	40
MPS R-CNN(本文提出)	77.73	37
YOLOv5-S	75.3	85
EfficientDet-D0	63.46	24
Flow R-CNN	71.4	<20
Mask R-CNN	71.2	
Faster R-CNN[AutoUpdated]	74.2	
YOLOv2	62	

5 结论

针对城市交通场景多目标检测问题，本文提出MPS R-CNN目标检测算法，提出BPFPN结构取代原FPN结构，提出多级提议过滤机制，引入CIoU损失和Copy-Paste数据增强方法进行训练，得到了更优的算法模型. 本文通过实验得到以下结论：MPS R-CNN算法在城市目标检测数据集上进行测试，检测精度为77.73%mAP，检测速度达到37 fps，优于目前主流城市交通多目标检测方法，实现了对城市交通场景多目标的端到端快速准确检测.

本文提出的MPS R-CNN算法有着出色的检测速度和检测精度. 但算法在检测速度上受到多级提纯结构的制约，后续将考虑采用更高效的动态交互头缩减提纯结构，进一步提升检测速度.

参考文献

[1] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

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

[3] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 7263-7271.

[4] 李宝奇, 贺昱曜, 强伟, 何灵蛟. 基于并行附加特征提取网络的SSD地面小目标检测模型[J]. 电子学报, 2020, 48(1): 84-91.
LI Baoqi, HE Yuyao, QIANG Wei, HE Lingjiao. SSD small target detection model based on parallel additional feature extraction network[J]. Acta Electronica Sinica, 2020, 48(1): 84-91. (in Chinese)

[5] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2961-2969.

[6] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2980-2988.

[7] SUN P, ZHANG R, JIANG Y, et al. Sparse r-cnn: End-to-end object detection with learnable proposals[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 14454-14463.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

[9] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2117-2125.

[10] 侯庆山, 邢进生. 基于Grad-CAM与KL损失的SSD目标检测算法[J]. 电子学报, 2020, 48(12): 2409-2416.
HOU Qingshan, XING Jinsheng. SSD target detection algorithm based on Grad-CAM and KL loss[J]. Acta Electronica Sinica, 2020, 48(12): 2409-2416. (in Chinese)

[11] GHIASI G, CUI Y, SRINIVAS A, et al. Simple copy-paste is a strong data augmentation method for instance segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 2918-2928.
[12] ZHENG Z , WANG P , REN D , et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021,62(01): 1-13.
[13] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8759-8768.
[14] DOMINGUEZ-SANCHEZ A, CAZORLA M, ORTS-ESCOLANO S. A new dataset and performance evaluation of a region-based cnn for urban object detection[J]. Electronics, 2018, 7(11): 301.
[15] GLENN J. ultralytics/yolov5[EB/OL]. (2021-07-08) [2021-07-08]. https://github.com/ultralytics/yolov5.
[16] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10781-10790.
[17] PSALTIS A, DIMOU A, ALVAREZ F, et al. Flow R-CNN: Flow-enhanced object detection[C]//ICPR International Workshops and Challenges. Berlin: Springer,2021: 685-700.

作者简介

柳长源　男. 1970年10月出生,黑龙江肇东人. 副教授、硕导. 1993年、2005年、2013年分别在吉林大学、哈尔滨理工大学、哈尔滨工程大学获理学学士、工学硕士和工学博士学位,现为哈尔滨理工大学测控技术与通信工程学院教师,主要从事模式识别、机器学习、图像处理等方面的研究工作.
E-mail: liuchangyuan@hrbust.edu.cn

张玉亮　男. 1998年1月出生于安徽省阜阳市. 哈尔滨理工大学测控技术与通信工程学院硕士研究生. 研究方向为模式识别、目标检测.
E-mail: 2497484650@qq.com

毕晓君　女. 1964年11月生于黑龙江哈尔滨. 教授、博士生导师. 1987年、1990年、2006年于哈尔滨工程大学、哈尔滨工业大学、哈尔滨工程大学获工学学士、工学硕士和工学博士学位,现为中央民族大学信息工程学院教师,主要研究进化计算、数据挖掘.