基于深度学习的车载视频图像动态识别算法与应用研究

郑爱玲[1] 许艳艳[1]
ZHENG Ailing XU Yanyan

摘 要 在道路环境错综复杂且干扰源众多的场景下，车辆、行人、交通标志等物体之间的空间关系变得极为复杂。传统方法过度依赖基本的特征提取，未能充分挖掘图像中物体间的空间关系信息，难以精准捕捉不同物体间的相对位置、距离等空间特征。这使得模型在图像轮廓识别上精度欠佳，难以准确把握目标的动态变化。鉴于此，文章提出一种基于深度学习的动态识别算法研究。该方法先采用深度卷积神经网络（CNN）与循环神经网络（RNN）构建模型，提取车载视频图像的空间与时序特征。然后融合深层语义与浅层细节信息，增强检测性能，实现实时目标检测与跟踪。最后引入图与通道注意力模块，提升复杂环境下空间特征提取和动作特征表示能力，实现视频图像动态识别。对比实验表明，新算法在车载视频图像动态识别方面表现出较高的准确性，显著提升了车辆行驶的安全性和智能化水平，为车载视频图像动态识别提供了有效的技术手段，具有重要的理论意义和实际应用价值。

关键词 深度学习；车载视频；视频图像；特征融合；动态识别

doi：10.3969/j.issn.1672-9528.2025.03.011

0 引言

随着科技的飞速发展，自动驾驶技术已经成为汽车工业和人工智能领域的研究热点。而车载视频图像动态识别技术是自动驾驶技术的核心，通过摄像头捕捉路况，用算法处理视频图像，为自动驾驶提供环境感知和决策支持。高效准确地提取视频信息，辅助安全驾驶，同时助力智能交通建设，实现车辆信息共享和协同控制，提高交通效率。同时，该技术也应用于车辆防盗和追踪系统，从而保障车辆安全。

目前，已有多种车载视频图像动态识别算法被提出并应用于实际场景中，为自动驾驶技术的发展提供了有力支持。例如，张宗等人[1]探讨了基于时空图卷积网络（STGCN）的视频图像人体动作轮廓动态识别方法。该研究利用 STGCN 算法对视频图像中人体动作的时空特征进行建模和识别，取得了良好的效果。然而，虽然 STGCN 能够处理人体动作的空间特征，但道路场景中的空间关系更为复杂多样。STGCN 难以适应这种复杂的空间关系变化，因为空间关系处理能力是基于人体动作相关的图结构设计的，与道路场景中的物体空间关系结构不完全匹配，进而导致在识别物体间空间特征时精度欠佳。费春国等人[2]则提出了一种基于动态衰减网络的图像识别方法。该方法通过引入动态衰减机制，提高了模型对图像特征的提取能力和鲁棒性。但在处理道路环境复杂且干扰源众多的场景时，无法精准地调整其识别过程，从而难以持续准确地捕捉交通标志与其他物体之间的空间关系，最终对识别精度产生影响。刘小溪等人[3]提出针对路牌检测类别少的问题，对 YOLOv3 进行轻量化改造，提出了简化的 YOLOv3（simplified YOLOv3, S-YOLOv3）；为了提高字符分类精度，对 YOLOv3 进行特征融合策略改进，提出高精度的字符分类 YOLOv3（high precision character classification YOLOv3, HPCC-YOLOv3）；分别对 S-YOLOv3 与 HPCC-YOLOv3 进行训练与测试；按照字符检测结果所处的位置进行字符聚类，实现字符序列识别；设计了由车载大恒水星相机，计算机组成的图像采集，车牌检测与字符识别，但该方法主严重受到使用环境的影响。

基于上述分析，本文提出一种基于深度学习的车载视频图像动态识别算法，旨在克服传统方法在复杂道路环境下空间特征提取受限、人体动作特征表征能力不足及缺乏高效特征融合策略等挑战，显著提升视频图像的动态识别性能。

1 基于深度学习的车载视频图像动态特征提取

车载视频图像由连续图像帧构成，动态特征往往更能反映车辆和行人的运动状态、行为模式等关键信息。因此，提

1. 郑州工业应用技术学院信息工程学院 河南郑州 451100
[基金项目] 教育部供需对接就业育人项目（2023122163922）；郑州工业应用技术学院 2024 年度省级工程技术研究中心开放课题支持项目（ETRC-240103）

取动态特征有助于识别模型更好地理解视频中的动态场景，提高识别和检测的准确性[4]。为此，本文提出构建了一个结合 CNN 与 RNN 的深度学习模型。其中，CNN 擅长图像特征提取，RNN 能捕捉时序关系[5]。结合两者，可全面提取车载视频图像的静态空间特征和帧间动态变化，为精准目标识别和动态跟踪奠定基础。

车载视频图像的本质是动态的，包含了丰富的时序信息，这些信息对于理解视频中的动态变化至关重要。CNN 的特征提取过程可以表示为：

$$F_{\mathrm{CNN}}(x)=\sigma(W_{\mathrm{conv}}\cdot x+b_{\mathrm{conv}}) \tag{1}$$

式中：x 表示车载视频中经预处理后的某一帧图像；W_{conv} 表示卷积层权重；b_{conv} 表示卷积层偏置值；σ 表示激活函数。在 CNN 特征提取过程中，卷积操作的数学公式为：

$$z_{i,j,k}=\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}x_{i+m,j+n}\cdot w_{m,n,k}+b_k \tag{2}$$

式中：$z_{i,j,k}$ 表示第 k 个卷积核在位置 (i,j) 的输出；$x_{i+m,j+n}$ 表示输入图像在位置 $(i+m,j+n)$ 的像素值；$w_{m,n,k}$ 表示第 k 个卷积核在位置 (m,n) 的权重；b_k 表示第 k 个卷积核的偏置。通过多层卷积和可能的池化操作，CNN 能够从图像中提取出高级的空间特征。

RNN 擅长处理时序数据，通过循环连接能够捕捉视频图像序列中的时间依赖性。RNN 的时序特征提取过程可以表示为：

$$h_t=\sigma(W_{\mathrm{rnn}}\cdot[h_{t-1},F_{\mathrm{CNN}}(x_t)]+b_{\mathrm{rnn}}) \tag{3}$$

式中：h_t 表示第 t 时刻的隐藏状态；W_{rnn} 表示 RNN 的权重；b_{rnn} 表示 RNN 的偏置值；$F_{\mathrm{CNN}}(x_t)$ 表示第 t 时刻 CNN 提取的空间特征。在处理每一帧图像时，RNN 会接收 CNN 提取的空间特征作为输入，并结合上一时刻的隐藏状态来更新自己的隐藏状态。这一过程使得 RNN 能够提取出视频帧之间的动态特征。

在道路环境复杂且干扰源众多的情境下，采用 CNN 与 RNN 相结合的架构进行车载视频图像的动态特征提取。这一过程使得模型能够同时利用空间和时间信息，对复杂道路环境下的车载视频图像进行全面深入的理解，为后续更精准的目标识别和动态跟踪等任务奠定基础[6]。动态特征 y 通过对 RNN 最后一个隐藏状态 h_T 进行变换得到：

$$y=\mathrm{softmax}(W_{\mathrm{fc}}\cdot h_T+b_{\mathrm{fc}}) \tag{4}$$

式中：W_{fc} 表示全连接层的权重；b_{fc} 表示全连接层的偏置值；h_T 表示 RNN 输出的最后一个隐藏状态。

2 特征融合与动态检测

车载视频图像动态识别通常涉及对视频图像序列数据的实时分析和处理，以捕捉其中的动态变化。动态特征融合可以提供更全面、更丰富的特征表示，有助于模型更准确地捕捉序列数据中的动态变化，实现动态检测。为提升模型的检测性能，特征融合成为一个重要的技术手段。深层特征通常包含更高级的语义信息，对于理解图像中的复杂场景至关重要；而浅层特征则更多地保留了图像中的细节信息，如边缘和纹理等，对于精确捕捉目标的形态和轮廓非常有用。通过融合这些不同层次的特征，能够为动态检测构建一个更加丰富的特征空间，使模型能够更准确地捕捉视频图像序列中的动态变化，进而实现对目标的实时检测和跟踪。

以尺寸为 5 px×5 px 的输入图像为例，通过卷积神经网络提取其动态特征。利用 3×3 的卷积核进行反卷积操作（步长 S =1，扩充 P =0，并在特征单元间插入 0 进行上采样），输出特征图大小可通过公式计算：

$$\mathrm{output}=S\times(i-1)+K-2\times P \tag{5}$$

式中： output 表示输出特征图大小；i 表示输入图像大小；K 表示卷积核。

在动态检测中，深层特征图（如 Conv11 的 19 px×19 px 和 Conv13 的 10 px×10 px）蕴含丰富的动态特征，这些特征是通过深层卷积网络逐步提取得到的[7]。为提升检测性能，采取特征融合策略，将深层特征（Conv11 和 Conv13 层提取的动态特征）与浅层特征（Conv5 层提取的动态特征）进行融合。具体来说，Conv13 层会输出 1 024 个 10 px×10 px 的动态特征图。这些特征图首先经过一个 3×3 卷积核、步长为 2、扩充为 1 的反卷积操作，上采样到 19 px×19 px 的尺寸。接着，通过一个 1×1 卷积核进行降维处理，以便与 Conv11 层的 19 px×19 px 动态特征图进行融合。融合后的特征图采用 3×3 卷积来消除混叠效应，并用于后续的分类和预测任务。

同样，Conv11 层融合后的特征图（512 个 19 px×19 px 的动态特征图）会经过一个 2×2 卷积核、步长为 2、扩充为 0 的反卷积操作，上采样到 38 px×38 px 的尺寸。然后，通过一个 1×1×256 的卷积核进行降维，以便与 Conv5 层的 38 px×38 px 动态特征图进行融合。融合后的特征图同样采用 3×3 卷积来消除混叠效应，并用于分类和预测。

特征融合结合不同层或来源的特征图，提升模型检测性能。在车载视频识别中，融合深层语义与浅层细节信息，构建全面特征空间，对捕捉动态变化至关重要。特征融合的特征图呈现出较高的精度且鲁棒性，是实现动态检测的关键。通过特征融合策略，能够有效地将深层特征与浅层特征相结

合，为车载视频图像动态识别任务提供更加准确和可靠的特征表示，进而提升模型的检测性能[8]。

3 视频图像轮廓动态识别

复杂道路环境中，传统方法难精确捕捉物体空间关系，影响图像识别和目标动态把握。因此，本文引入图注意力模块和通道注意力模块，弥补此不足。图注意力模块构建物体图结构关系，精确描绘人体动作轮廓和物体空间联系。通道注意力模块聚焦特征图通道维度，重新分配特征权重，减少无用特征影响，增强有用特征表达能力[9]。这两个模块结合，提升网络空间特征提取能力和人体动作特征表示能力，实现更精准的视频图像动态识别。

采用两个卷积层将输入特征图f转换为矢量$\boldsymbol{R}$和$\boldsymbol{Q}$。接着，利用归一化技术生成不同动作骨架的时空图样本。通过学习两个随机关节点之间的权值关系，能够更精确地实现人体动作轮廓的识别。

此外，通道注意力模块的加入进一步增强了 STGCN 网络在人体动作特征表示方面的能力。通过全连接层进行特征升维，并使用 Sigmoid 激活函数获取权值。最后，将这些权值与输入特征图相乘，实现特征权重的重新分配[10-12]。这一过程有效减少了无用特征的影响，同时增强了有用特征的表达能力[13]。将权值矢量与输入特征图f相乘，实现特征权重的重新分配，可以表示为：

$$f' = w \odot f \tag{6}$$

式中：⊙表示逐元素相乘操作；f'表示重新分配权重后的特征图；w表示权值。通过上述内容，实现对视频图像轮廓的动态识别。

4 对比实验

4.1 实验环境

本文基于深度学习概念，创新性地提出了一种车载视频图像动态识别算法。为了全面且准确地评估该算法的实际效能，设计并实施了一项对比实验。实验中，新算法被用于对车载视频图像进行动态识别，并设为实验组。为确保实验的可比性和公正性，实验同时设置了两个对照组。对照 A 组采用基于时空图卷积网络（STGCN）的识别方法，对相同的车载视频图像进行识别；而对照 B 组则运用基于动态衰减网络和特定算法的图像识别技术，对同一批车载视频图像进行识别。

这三组实验均在相同的实验环境中进行，以保障实验结果的准确性和可靠性。实验环境配置如下：实验平台选用 Ubuntu14.04 操作系统，搭载英特尔酷睿 i5 处理器和 GTX1060 显卡（显存容量 6 GB）。开发环境则使用 Anaconda2 作为管理工具，结合 caffe 深度学习框架及 Python2.7 编程语言，用于构建和运行算法模型。

通过上述实验环境和配置，能够全面客观地评估新提出的车载视频图像动态识别算法的性能，并与现有的识别方法进行对比分析，为后续研究和应用提供有力支持。

4.2 实验数据集

目前，车载视频图像识别中应用较多的数据集包括 INRIA、ETH、TUD、Cityscapes 等。表 1 记录了各个数据集的基本情况。

表 1 数据集基本情况记录表

序号	名称	拍摄场景	数据集分布	图像分辨率/(px×px)
（1）	INRIA	复杂且多姿态动态图像	训练集 1 834 张 测试集 645 张	>100
（2）	ETH	城市白天车载摄像机	训练集 22 356 张 测试集 21 625 张	640×480
（3）	TUD	城市白天车载摄像机	训练集 12 800 张 测试集 12 100 张	640×480
（4）	Cityscapes	城市白天车载摄像机	训练集 7 560 张 测试集 7 512 张	2 048×1 024

满足本次实验需求，选定 Cityscapes 数据集作为实验数据集。Cityscapes 数据集由车载摄像头拍摄自大约 50 个城市街道的视频帧组成，主要覆盖了春夏秋 3 个季节。该数据集因其详尽的标注而适用于道路场景的识别分析，信息被细致划分为 8 大类、30 小类（例如车辆、行人等）。数据集包含 20 000 张粗标注图像和 5 000 张细标注图像，适用于本文的研究要求。图 1 展示了该数据集中行人尺度的分布情况。

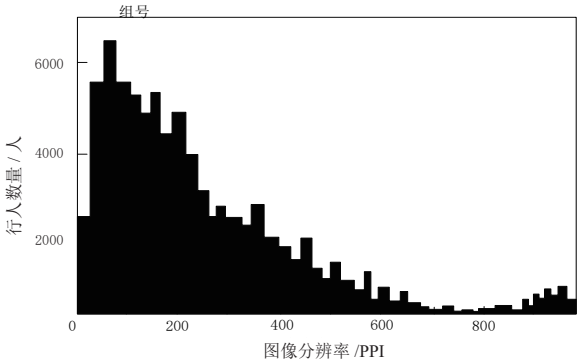
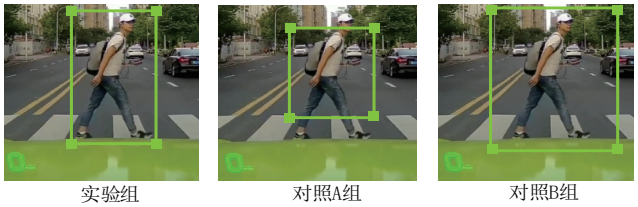


图 1 Cityscapes 数据集行人尺度分布

从图 1 可以看出，在该数据集中行人尺度分布极不均匀，大部分行人尺度集中在 40~60 PPI 的位置。根据图 1 所示，将行人尺度图像分辨率小于 40 PPI 的定义为小目标；将图像分辨率 40~80 PPI 的定义为中目标；将图像分辨率 80 PPI 以上定义为大目标。

4.3 实验结果

图 2 所示为三组方法在实验过程中识别到的行人动态目标。



实验组　对照A组　对照B组

图 2 三组方法识别结果示意图

由图 2 可知，实验组在行人动态目标识别上表现出卓越性能。实验组能精确捕捉行人动态特征并准确框定，体现在位置判断和标记框大小的匹配上，确保了识别结果的可靠性。对照 A 组覆盖不全，易导致识别误差，且标记框放置缺乏统一标准，降低了可靠性。对照 B 组标记框过大，可能纳入无关物体，增加识别复杂性和运行负担。本文方法优势在于引入图注意力模块和通道注意力模块，强化了模型在复杂场景中的空间特征提取能力，提升了模型对人体动作特征的敏感度及识别精度，提高了识别准确性。

在此基础上，记录了三组方法针对不同目标的识别准确率，如表 2 所示。

表 2 三组方法识别结果对比表

类别	实验组		对照 A 组		对照 B 组	
	准确识别数量 / 个	准确率 /%	准确识别数量 / 个	准确率 /%	准确识别数量 / 个	准确率 /%
小目标	49	98.0	26	52.0	18	36.0
中目标	48	96.0	37	74.0	22	44.0
大目标	49	98.0	26	52.0	16	32.0

从表 2 结果可知，实验组对三种目标图像识别准确率高至 95%，准确识别数超过 45 个，远超对照 A、B 组。由此证明本文识别算法实际应用中更准确。优势在于引入的图注意力模块提升了空间特征提取能力，通道注意力模块增强了人体动作特征表示，使模型在复杂场景下鲁棒性、适应性更强，提高了识别准确性。

5 结语

本文聚焦于基于深度学习的车载视频图像动态识别算法，创新性地设计了一种算法框架，并在真实车载视频数据集上完成了实验验证。实验结果显示，该算法在识别精度与运算速度上均实现了显著提升，有效克服了现有技术的局限性。本研究不仅成功解决了这些技术难题，更为车载视频图像动态识别技术的发展注入了新的动力。展望未来，将持续深化研究，致力于开发更加高效、智能的算法，为智能交通系统的广泛应用提供更加可靠的技术保障。

参考文献：

[1] 张宗，石林．基于 STGCN 算法的视频图像人体动作轮廓动态识别 [J]. 现代电子技术，2024, 47 (18): 144-148.

[2] 费春国，刘启轩．基于动态衰减网络和算法的图像识别 [J]. 电子测量与仪器学报，2022, 36 (7): 230-238.

[3] 刘小溪，程佳诚，程咏梅，顾一凡，雷鑫华，汪波．车载视频图像路牌检测与字符序列识别方法 [J]. 计算机工程与应用，2023,59(8):175-181.

[4] 代勇，罗四海，秦政阳，等．基于动态图像识别叶轮气动不平衡诊断方法的应用研究 [J]. 中国设备工程，2024(13): 172-174.

[5] 李进，岳华峰，程生博，等．供应链质量追溯的烟草叶片图像帧特征动态识别方法 [J]. 计算技术与自动化，2024, 43 (1): 111-116.

[6] 肖帅，郭嘉，芦天罡，等．基于图像识别技术的北京市设施温室动态监测研究 [J]. 农业工程，2023, 13 (10): 21-26.

[7] 薛薇，张锋，凡静，等．基于边缘检测及 RBF 神经网络的遥感图像帧特征动态识别技术 [J]. 计算机测量与控制，2023, 31 (7): 163-168.

[8] 陈佳鑫，赵国贞，程伟，等．基于动态称量和图像识别技术的井下煤矸石智能分选装置研发 [J]. 矿业研究与开发，2023, 43 (2): 178-183.

[9] 赵琰，赵凌君，张思乾，等．自监督解耦动态分类器的小样本类增量 SAR 图像目标识别 [J]. 电子与信息学报，2024, 46 (10): 3936-3948.

[10] 叶明，苏海涛，高洪章，等．基于车牌动态图像识别的电子汽车衡自动称重系统的研究与应用 [J]. 衡器，2022, 51 (11): 33-37.

[11] 王硕，孙梦轩，杨志晓，等．基于涡旋电磁波雷达回波时频图像的动态手势识别 [J]. 火力与指挥控制，2022, 47 (8): 109-115.

[12] 朱木清，邹欢．深度学习下动态目标识别算法优化仿真 [J]. 计算机仿真，2023, 40(12):321-324.

[13] 王书朋，贺瑞，王瑜婧，等．中值直方图均衡的动态场景多曝光图像融合算法 [J]. 计算机工程，2022,48(10):224-229.

【作者简介】

郑爱玲（1995—），女，河南信阳人，硕士研究生，助教，研究方向：深度学习、目标检测、模式识别。

许艳艳（1995—），女，河南三门峡人，硕士研究生，助教，研究方向：区块链技术、Web 开发、数据分析。

（收稿日期：2024-11-14）