JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
01
VOL.26

ISSN1006-8961
CN11-3758/TB

“自动驾驶技术与应用”专刊

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第1期（总第297期）
2021年1月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

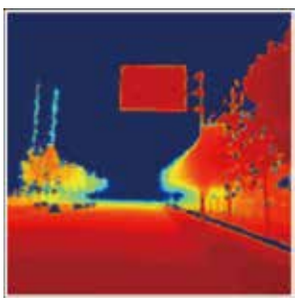
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2021年1月 第26卷 第1期（总第297期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

目次

面向智能驾驶的平行视觉感知：基本概念、框架与应用(第0067页)



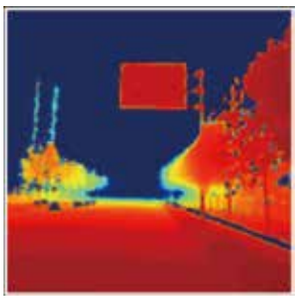
Edge-guided GAN：边界信息引导的深度图像修复(第0186页)

利用边缘计算的多车协同激光雷达SLAM(第0218页)

Volume 26, Number 1
Published January 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Parallel visual perception for intelligent driving: basic concept, framework and application (P0067)



Edge-guided GAN: a depth image inpainting approach guided by edge information (P0186)

Cooperative LiDAR SLAM for multi-vehicles based on edge computing (P0218)

中图法分类号:TP183; TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2021)01-0143-11

论文引用格式: Zhuang Y and Qi Y. 2021. Driving fatigue detection based on pseudo 3D convolutional neural network and attention mechanisms. Journal of Image and Graphics,26(01):0143-0153(庄员,戚湧. 2021. 伪 3D 卷积神经网络与注意力机制结合的疲劳驾驶检测. 中国图象图形学报,26(01):0143-0153)[DOI:10.11834/jig.200079]

伪 3D 卷积神经网络与注意力机制结合的疲劳驾驶检测

庄员,戚湧

南京理工大学计算机科学与工程学院, 南京　210094

摘　要: **目的**　复杂环境下的疲劳驾驶检测是一个具有挑战性的技术问题。为了充分利用驾驶员面部特征信息与时间特征,提出一种基于伪 3D(Pseudo-3D,P3D)卷积神经网络(convolutional neural network,CNN)与注意力机制的驾驶疲劳检测方法。**方法**　采用伪 3D 卷积模块进行时空特征学习;提出 P3D-Attention 模块,利用 P3D 的结构融合双通道注意力模块和适应的空间注意力模块,提高对重要通道特征的相关度,增加特征图的全局相关性,将多层深度卷积特征进行融合。利用双通道注意力模块分别在视频帧之间和每一帧的通道上施加关注,去除背景和噪声对识别的干扰,使用自适应空间注意模块使模型训练更快、收敛更好;使用 2D 全局平均池化层替代 3D 全局平均池化层获得更具表达能力的特征,进而提高网络收敛速度;运用 softmax 分类层进行分类。**结果**　在公共数据集 YawDD(a yawning detection dataset)上开展对比实验,本文方法在测试集上的 F1-score 检测准确率达到 99.89%,在打哈欠类别上召回率达到 100%;在数据集 UTA-RLDD(University of Texas at Arlington real-life drowsiness dataset)上,本文方法在测试集上的 F1-score 检测准确率达到 99.64%,在困倦类别上召回率达到 100%;与 Inception-V3 融合 LSTM(long short-term memory)的方法相比,本文方法模型大小为 42.5 MB,是其模型大小的 1/9,本文方法预测时间约 660 ms,是其 11% 左右。**结论**　提出一种基于伪 3D 卷积神经网络与注意力机制的驾驶疲劳检测方法,利用注意力机制进一步分析哈欠、眨眼和头部特征运动,将哈欠行为与说话行为动作很好地区分开来。

关键词:3D 卷积神经网络;伪 3D 卷积;全局平均池化;注意力机制;疲劳驾驶

Driving fatigue detection based on pseudo 3D convolutional neural network and attention mechanisms

Zhuang Yuan, Qi Yong

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing　210094, China

Abstract: **Objective** Fatigue driving is one of the main causes of traffic accidents. Drivers in fatigue state will have reduced alertness, weakened ability to deal with abnormal events, and inability to react traffic control and dangerous events, which will lead to accidents. The current technology for detecting fatigue driving behavior can be divided into three methods based on physiological parameters, vehicle behavior, and facial feature analysis. Detection methods based on physiological parameters require various sensors. These sensors use physiological signals to detect the driver's drowsiness, but they need

收稿日期:2020-03-20;**修回日期**:2020-05-13;**预印本日期**:2020-05-20

基金项目:国家重点研发计划政府间国际科技创新合作重点专项(2019YFE0123800,2016YFE0108000);江苏省重点研发计划(产业前瞻与共性关键技术)项目(BE2017163)

Supported by: National Key Research and Development Program Key Special Projects of International Intergovernmental Science and Technology Innovation Cooperation of China(2019YFE0123800,2016YFE0108000)

contact with the driver's body, rely on expensive equipment, and are invasive. Detection methods based on vehicle behavior use vehicle behavior parameters, such as lane departure detection, steering wheel angle, and yaw angle information, to detect driving fatigue behavior, but they also depend on external factors such as road conditions. Detection methods based on facial feature analysis need to extract feature points from the driver's facial features and compare the driver's performance in fatigue or normal conditions by detecting fatigue behavior characteristics such as eye state, blinking, and yawning. Compared with the two earlier methods, this method has the advantages of noninvasiveness and easy implementation. In several current methods, spatiotemporal features cannot be well integrated, and interference of background and noise on recognition is not removed. This paper proposes a driving fatigue detection method based on pseudo 3D (P3D) convolutional neural network (CNN) and attention mechanism to solve these problems. **Method** The dataset is cropped into small videos of around 5 s each. The training video data interval is 90 video frames, and the picture resolution is set to $80 \times 80 \times 3$. First, the feature map of each frame is fully extracted through the P3D module to generate a fixed-size feature set. Second, the P3D structure uses a $1 \times 3 \times 3$ convolution kernel and a $3 \times 1 \times 1$ convolution kernel to simulate $3 \times 3 \times 3$ convolution in the spatial and time domains, decoupling $3 \times 3 \times 3$ convolutions in time and space. Based on the feature that P3D decouples $3 \times 3 \times 3$ convolutions in time and space, a module named P3D-Attention is proposed. The 3D convolutional neural network and attention mechanism are integrated to improve the correlation of important channel features, increase the global correlation of feature graphs, and remove the interference of background and noise on recognition by translating 3D temporal and spatial features into 2D features and embedding them in the dual-channel and spatial attention modules. The dual-channel attention module is used to apply attention on video frames and channels of each frame, which removes the interference of background and noise on recognition. For driving scenarios, this paper selects convolution kernels of different sizes to adapt to convolution features of different depths and uses the adaptive spatial attention module to make the model training converge faster and better. Afterward, 2D global average pooling layer is used instead of 3D global average pooling layer to obtain more expressive features, improving network convergence speed. Finally, the softmax classification layer is used for classification. **Result** A comparative test is performed on the public dataset——a yawning detection dataset (YawDD). The detection accuracy of the method in this paper reaches 98.75%, and the recall rate of the yawning category reaches 100%. On the University of Texas at Arlington real-life drowsiness dataset (UTA-RLDD), the F1-score detection accuracy of the method in this paper reaches 99.64% on the test set, and the recall rate reaches 100% in the drowsy category. In terms of running time and model size, experimental results show that compared with the long short-term memory (LSTM) fusion method using ImageNet-trained Inception_v3 model, the algorithm in this paper has evident advantages in terms of running time and predicts that a 5 s video will take 660 ms on average, which is 11% of it. In terms of the size of the unpruned model, the method of Inception-v3 plus LSTM has 396.15 MB, and the model size in this paper is 42.5 MB, which is 1/9 of it. **Conclusion** A driving fatigue detection method based on P3D convolutional neural network and attention mechanism is proposed. Attention mechanisms are used to remove background and noise from recognition interference, improve the accuracy of driving fatigue detection, distinguish yawning behavior from mouth opening and mouth closing behaviors such as talking, and analyze yawning behavior, blinking, and head characteristic movements. The further work of this paper will 1) verify whether features can be extracted through a smaller network structure, design a more efficient network structure, and further reduce the size of the model. 2) Future work will also focus on using 3D convolution to distinguish more complicated driving behaviors because distracted driving behavior not only needs to focus on predicting the driver's fatigue status.

Key words: 3D convolutional neural network (CNN); pseudo-3D (P3D) convolutional; global average pooling; attention mechanisms; fatigue driving

0 引 言

疲劳驾驶是交通事故的主要原因之一。处于疲劳状态的驾驶员经常感到困倦、短暂意识丧失,降低警觉性与应对异常事件能力,造成对交通管制和危险反应时间变慢,从而导致事故的发生。美国汽车协会报告指出,驾驶疲劳状况占道路交通事故比例最大,所有事故的7%和致命交通事故的21%是由疲倦驾驶员引起的(Tefft,2018)。现有技术用于检

测疲劳驾驶行为可分为基于生理参数、基于车辆行为和基于面部特征分析 3 类方法。

基于生理参数的检测方法要求传感器与驾驶员身体接触,利用生理信号进行驾驶员睡意检测,如心电图(electrocardiography,ECG)(Chui 等,2016)和脑电图(electroencephalogram,EEG)等(Balandong 等,2018;Wang 等,2018)。利用多种传感器组合测量不同的生理参数,如融合测量肌电图(electromyography,EDA)、呼吸和心电图检测驾驶员睡眠状态的系统(Forsman 等,2013),该方法虽然能够获得较高的疲劳驾驶检测精度,但是高昂的实验成本与侵入性特征限制了其应用范围。基于车辆行为的检测方法利用车辆行为参数检测驾驶疲劳行为(Sahayadhas 等,2012),如车道偏离检测、方向盘转角(steering wheel angle,SWA)和偏航角(yaw angle,YA)信息(Li 等,2017),但是该方法也依赖于道路状况(Caffier 等,2003)等外部因素。

基于面部特征分析的检测方法,通过对驾驶员面部特征提取特征点,比较驾驶员在疲劳状态和正常状态的表现,检测驾驶员头部移动姿态、眼睛状态、眨眼和哈欠等疲劳行为特征。与上述两种方法相比,该方法具有非入侵、易于实现等优点。Caffier 等人(2003)通过红外传感器连续记录眼睑的运动并研究自发眨眼参数的有效性,考察眨眼持续时间的子成分,即关闭时间、重新打开时间和关闭时间,研究表明,闪烁持续时间和重新打开时间这两个参数随着睡意的增加而可靠地变化。Friedrichs 和 Yang(2010)评估最新的基于眼睛追踪的车内疲劳预测措施的性能,研究基于摄像机的驾驶员睡意检测方法,将符合标准(最小/最大持续时间、形状和最小振幅)的候选眨眼标记为有效的眨眼。Clavijo 等人(2015)采用面部识别的算法,应用一种基于边缘检测和纹理测量的技术分割眼睛并计算随时间变化的眼睛特征,在高照度下获得 95.83% 的有效性,在中等照度下获得 87.5% 的有效性。Alioua 等人(2014)提出一种基于支持向量机的人脸提取系统,使用基于面部提取的支持向量机和一种基于圆形霍夫变换的新嘴部检测方法,应用于嘴部提取区域,并通过嘴部的张开大小判断疲劳状态。这些方法基于手工制作的特征,无法彻底探索不同视觉线索之间的复杂关系,忽略了眼部和嘴部遮挡问题,各人打哈欠时间、嘴巴张开大小存在明显差异,也没有考虑面部表情的特征变化和头部移动姿态等问题。

与基于手工制作特征的面部特征分析检测方法相比,Zhang 和 Su(2017)提出一种基于卷积神经网络的图像空间特征提取系统和一种基于 LSTM(long short-term memory)的图像时间特征分析系统,利用 LSTM 在时间序列上集成信息以获得最佳的判断性能,将帧级卷积神经网络(convolutional neural network, CNN)特征输出聚合为视频级特征进行预测,研究结果表明哈欠检测应该在连续视频上进行,准确率高达 87% 以上。Xie 等人(2019)对训练好的 Inception-v3 模块迁移学习,利用 Inception-v3 模块进行空间特征提取,之后将提取到的空间特征输入到 LSTM 层,融合时间特征预测疲劳状态。Xing 等人(2019)提出一种基于多种 CNN 的驾驶员活动检测模型,利用高斯混合模型分割原始图像从背景中提取驾驶员身体,有效地区分驾驶员是否分心,准确率为 91.4%。这些方法比基于手工特征的方法具有更强的鲁棒性,能够更好地捕捉不同特征间的关系。但是,由于空间特征提取使用 GoogLeNet 和Inception-v3 模型,导致预测模型参数量巨大,包含大量冗余的空间数据,卷积空间特征转换为 1 维向量输入到时序模型中,并没有考虑到空间上的相关性,且没有去除背景和噪声对识别的干扰,导致时空特征不能很好地融合。自 3D-CNN (three-dimensional convolutional neural network)(Ji 等,2013)模型提出以来,其在动作识别领域表现出优越的时空特征学习能力(Tran 等,2015)。Qiu 等人(2017)提出伪 3D 残差网络结构(Pseudo-3D-ResNet,P3D-ResNet)降低昂贵的计算成本和内存需求,并在 Sports-1M 视频分类数据集相对于 3D-CNN 和基于帧的 2D-CNN 分别提高 5.3% 和 1.8%。

本文提出一种基于伪 3D 卷积神经网络与注意力机制的疲劳驾驶检测模型,设计 P3D-Attention 模块,其以 P3D 模块将空间与时间卷积解耦为基础,分别与适应的空间注意力模块和双通道注意力模块融合,充分融合时空特征,提高重要通道特征的相关度,增加特征图的全局相关性,以提高对疲劳驾驶行为的预测性能。

1 基于伪 3D 卷积神经网络与注意力机制的驾驶疲劳检测模型

基于伪 3D 卷积神经网络与注意力机制的驾驶

疲劳检测模型如图 1 所示。该模型由 P3D 卷积模块、P3D-Attention 模块、池化模块与 softmax 分类模块组成。根据 Tran 等人(2015)的结论,3×3×3 卷积核的设置是 3D 卷积网络的最佳选择。P3D 结构(Qiu 等,2017)利用 1×3×3 卷积核和 3×1×1 卷积核在空间域和时间域上模拟 3×3×3 卷积,在时间和空间上将 3×3×3 卷积进行解耦。为了从图像数据中对 2D-CNN 进行预训练,本文利用 P3D 结构来获取图像特征。

通过 P3D 模块充分提取每一帧的特征图,生成固定大小的特征集合;再通过 P3D-Attention 提高重要通道特征的相关度,增加特征图的全局相关性;并使用全局平均池化层(global average pooling, GAP)(Lin 等,2014)代替全连接层。GAP 与全连接层不同,不需要训练大量参数,其在整个网络结构上通过正则化防止过拟合,使模型更加健壮。为了在没有完全折叠时间信号的时候获取更多的时间特征信息,选择将特征转置后输入到 2D GAP 中,而不是使用 3D GAP。

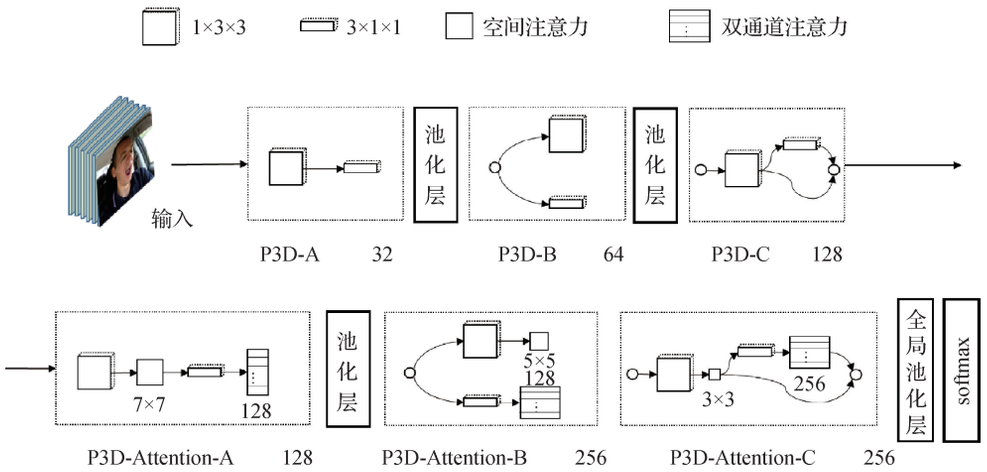


图 1 模型结构

Fig. 1 Model structure

1.1 伪 3D 卷积模块

P3D 卷积网络利用 1×3×3 卷积核在空间域(相当于 2D-CNN)上模拟 3×3×3 卷积和 3×1×1 卷积,在时间上构造相邻特征图上的时间连接,设计多个不同的瓶颈构建块(Qiu 等,2017)。如何结合这两个卷积运算,第 1 个要考虑的问题是空间维数 S 的 2D 卷积核和时间维 T 的 1D 卷积核具有直接作用还是间接作用。直接作用是空间卷积结果直接作为时间卷积的输入,而间接作用意味着空间卷积运算和时间卷积运算以并行方式执行。第 2 个问题是这两个卷积核是否都应该直接影响最终的输出。P3D 模块基于这两个设计问题,分别设计 3 个不同的 P3D 块,命名为 P3D-A、P3D-B 和 P3D-C。

1.2 注意力机制

注意力机制的使用使得各种计算机视觉任务的模型性能得到改善,包括图像识别和目标检测等任务。注意力机制通常应用于通过卷积神经网络(CNN)提取的特征。在预测类之前,本文在特征层上应用通道注意力(Hu 等,2018)和空间注意力(Woo 等,2018)。CBAM(convolutional block attention module)(Woo 等,2018)是通道注意力模块级联空间注意力模块的顺序排列结构。CBAM 的两个子模块都使用平均池化和最大池化操作来聚合特征图的时空信息,提高网络的表达能力。两个注意力模块计算互补,分别关注“什么”和“哪里”,并通过实验验证了两者优于仅使用通道注意力(Hu 等,2018)的方法。

本文通过将 3D 的时间与空间特征转置为 2D 特征嵌入到通道与空间注意力模块中,实现 3D 卷积神经网络与注意力机制的融合。通道注意力模块用于关注帧与帧之间、通道与通道之间的权重,获取对关键帧的注意。空间注意是对通道注意的补充,可以用于关注疲劳检测中关键特征,如眼睛、嘴巴和整个面部的特征。注意力机制可以表达为

$$\boldsymbol{F}' = M(\boldsymbol{F}) \otimes \boldsymbol{F} \tag{1}$$

式中,M 代表注意力模块,$\boldsymbol{F}$ 为特征图,⊗代表矩阵元素依次相乘。

1.2.1 通道注意力模块

本文以 CBAM(Woo 等,2018)模块的通道注意

力为基础,为适应 3D 卷积构建了如图 2 所示的通道注意力模块,称为双通道注意力模块(dual-channel attention model)。为了在 3D 卷积上使用注意力机制,本文的双通道注意力模块分别在视频帧之间和每一帧的通道上施加关注,而不是只在时间帧的层面上进行。以特征图 $\boldsymbol{F} \in \mathbf{R}^{(F\times H\times W\times C)}$ 例,其中 F 代表帧,C 代表每一帧下的通道数,H 和 W 代表不同通道下的特征,但各通道对最终检测结果的贡献并不相等。双通道注意力模块学习 $\boldsymbol{M} \in \mathbf{R}^{(F\times 1\times 1\times C)}$ 的权重来决定每个通道的重要性,需要将 (F,H,W,C) 的特征图转置为 $(H,W,F\times C)$,并嵌入到 2D 空间注意力模块中,分别学习 $\boldsymbol{M} \in \mathbf{R}^{(F\times 1\times 1\times 1)}$ 与 $\boldsymbol{M} \in \mathbf{R}^{(1\times 1\times 1\times C)}$ 的权重来分别表达对帧与通道的关注。

图 2 中 F、W、H 和 C 分别代表每一次预测传入视频帧池化后的数量、特征图的宽、特征图的高和特征图的通道数,⊕代表求和操作,⊕后级联的㋡型符号代表 sigmoid 激活函数。

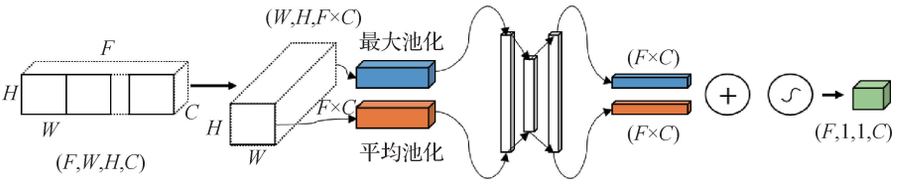


图 2　通道注意力模块结构

Fig. 2　The modular structure of channel attention

1.2.2　空间注意力模块

相对于背景,为了获取关键信息,人类的视觉机制更关注主要目标。因此,本文通过空间注意力模块来获取特征层在空间维度上的权重图。以特征图 $\boldsymbol{F} \in \mathbf{R}^{(F\times H\times W\times C)}$ 为例,空间注意力模块学习 $\boldsymbol{M}_{\mathrm{S}} \in \mathbf{R}^{(1\times W\times H\times 1)}$ 权重来确定每个特征图的重要性。空间注意力机制主要通过一个 2D 卷积核来获取空间特征权重。由于在驾驶过程,车内场景几乎不会变化,与其他需要考虑多尺度的任务不同,针对疲劳检测这一特殊场景,可以选用不同大小的卷积核来适应不同深度的卷积特征,模块结构如图 3 所示。

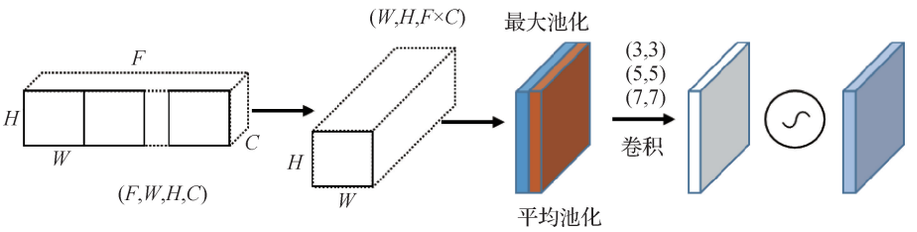


图 3　空间注意力模块结构

Fig. 3　The modular structure of spatial attention

1.2.3　注意力模块放置位置

注意力机制通常放置在:1)CNN 网络的末端,在深层 CNN 网络的末端应用视觉注意力是合并视觉注意力的最直接方法;2)迭代在 CNN 网络的不同层上,有多种方法可以在多个 CNN 特征层上进行操作,从而反复施加视觉注意力(Khandelwal 和 Sigal,2019)。注意力机制的差异在于其可以是全局的,即使用所有可用的图像上下文来共同预测注意掩码中的值;注意力机制也可以是局部的,其中每个变量的关注度为单独生成或使用相应的本地图像区域生成的图像。本文使用全局信息计算注意掩码迭代地在网络的后 3 层卷积模块添加注意力。

1.3　P3D-Attention 模块

在本文驾驶疲劳行为检测的任务中,模型输入的数据是一个视频的连续帧。在 P3D 模块将 3D 卷积解耦成空间与时间卷积的基础上,本文融合注意力模块设计 3 个不同的 P3D-Attention 块来实现网络模型,如图 4 所示,即 P3D-Attention-A 至 P3D-Attention-C。该模块使用双通道注意力模块使关键帧在分类中起更大的作用,同时添加空间注意力模块,依据特征图自动给不同关节点分配不同的注意力,关注眼睛部位和嘴巴部位这些先验知识所提到

的位置,去除背景和噪声对识别的干扰。

P3D-Attention-A:原有的 P3D-A 模块是在残差单元(residual unit)的基础上通过将时间 1D 卷积核(T)级联到空间 2D 卷积核(S)之后来考虑堆叠架构。因此,这两种卷积核可以在同一路径上直接相互影响,并且只有时间 1D 卷积核直接连接到最终输出。考虑到疲劳检测任务不需要太深的卷积层,在去掉残差单元只用 P3D-A 的基础上,通过在 S 后级联空间注意力模块(spatial attention, SA),并接着在 T 后级联通道注意力模块(channel attention, CA),实现 P3D-Attention-A 结构,即

$$(CA \cdot T \cdot SA \cdot S) \cdot \boldsymbol{X}_t = CA(T(SA(S(\boldsymbol{X}_t)))) = \boldsymbol{X}_{t+1} \tag{2}$$

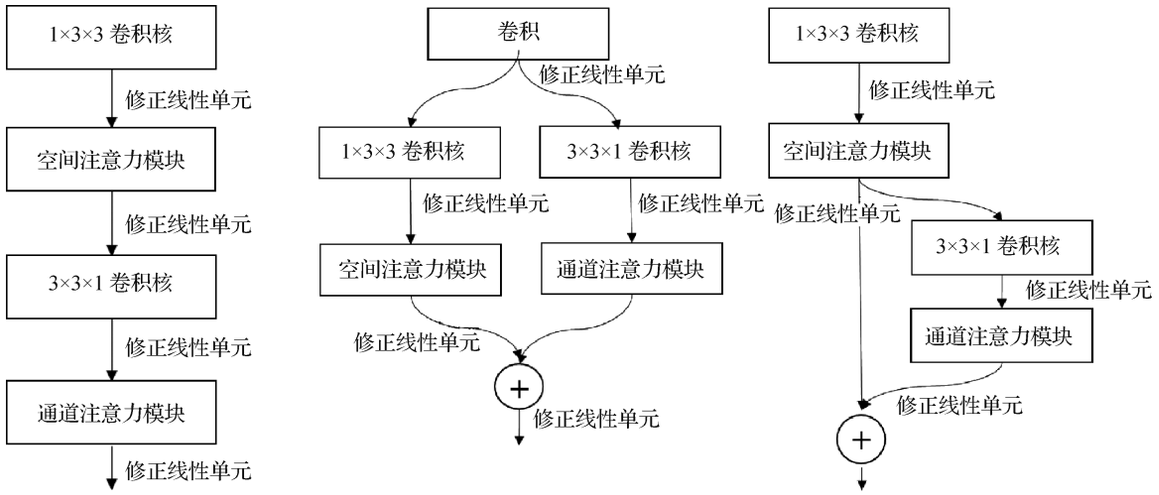
式中,$\boldsymbol{X}_t$ 表示输入特征图,$\boldsymbol{X}_{t+1}$表示施加注意力机制后的输出,$\boldsymbol{X}_t$与 $\boldsymbol{X}_{t+1}$具有相同的特征维度,T 表示时间 1D 卷积核,S 表示空间 2D 卷积核,SA 表示空间注意力模块,CA 表示通道注意力模块。

P3D-Attention-B:原有的 P3D-B 采用两个卷积核之间的间接影响,使得两个卷积核以并行方式处理卷积特征。在去掉残差单元基础上,在 S 位置后级联空间注意力模块(SA),并接着在 T 位置后级联通道注意力模块(CA),可以表示为

$$(SA \cdot S + CA \cdot T) \cdot \boldsymbol{X}_t = SA(S(\boldsymbol{X}_t) + CA(T(\boldsymbol{X}_t))) = \boldsymbol{X}_{t+1} \tag{3}$$

P3D-Attention-C:原有的 P3D-C 模块是 P3D-A 和 P3D-B 之间的折中,同时建立 S 与最终输出之间的直接影响和 T 与最终输出之间的直接影响。具体来说,为了实现基于级联 P3D-A 架构的 S 和最终输出之间的直接连接,通过添加注意力模块构建 P3D-Attention-C,可以表示为

$$(SA \cdot S + CA \cdot T \cdot SA \cdot S) \cdot \boldsymbol{X}_t = SA(S(X_t) + CA(T(SA(S(\boldsymbol{X}_t))))) = \boldsymbol{X}_{t+1} \tag{4}$$



(a) P3D-Attention-A (b) P3D-Attention-B (c) P3D-Attention-C

图 4 P3D-Attention 结构

Fig. 4 The structure of P3D-Attention((a) P3D-Attention-A; (b) P3D-Attention-B; (c) P3D-Attention-C)

2 实 验

2.1 数据集

2.1.1 YawDD

YawDD(a yawning detection dataset)(Abtahi 等,2014)数据集包含两组具有不同面部特征的驾驶员视频数据,主要用于检测算法模型,也可用于人脸和嘴巴的识别和跟踪,视频是在真实和变化的照明条件下拍摄的。在第 1 个数据集中,摄像头安装在汽车前视镜下方,每个参与者有 3 ~ 4 个视频,每个视频包含不同的口腔状况,如正常、说话/唱歌和打哈欠。该数据集提供了 322 个视频,包括来自不同种族的男女司机、戴或不戴近视镜/太阳镜,以及在 3 种不同情况下的表现(1 表示正常驾驶(不说话);2 表示开车时说话或唱歌;3 表示开车时打哈欠)。

2.1.2 UTA-RLDD

UTA-RLDD (University of Texas at Arlington real-life drowsiness dataset)(Ghoddoosian 等,2019)是包含 60 名受试者的大型公共真实数据集,指示受

试者通过手机/网络摄像头以 3 种不同的睡意状态拍摄自己的 3 个视频,每个视频大约 10 min。这个数据集由大约 30 h 的视频组成,内容从轻微的睡意到更明显的睡意。视频片段标记为警戒(alert)、低警惕(low vigilant)或昏昏欲睡(drowsy)。

1)警戒。受试者被告知警戒意味着他们完全有能力长时间开车。

2)低警惕性。此状态对应于一些微妙的情况,即出现一些嗜睡迹象或存在嗜睡现象,但无需努力保持警觉。尽管受试者可能会在这种状态下驾驶,但不鼓励驾驶。

3)困倦。此状态表示受试者需要积极尝试不入睡,不可在此状态进行驾驶行为。

2.1.3　数据集处理

为了满足实验的需要,在保证训练与验证结果一致性的情况下,对数据集进行裁剪。通过固定位置裁剪将 YawDD 数据集进行数据增强以获得更大的数据集。打哈欠是一个动态连续的事件,通常持续时间 3～5 s,因此设置裁剪后每个视频为 5 s 左右,每一个视频通过固定位置裁剪出 5 个不同的视频。如表 1 所示,新的数据集共有 8 660 个视频,其中正常驾驶类别有 2 865 个视频,说话类别的视频有 4 299 个,打哈欠类别视频有 1 495 个。表 1 给出了数据集按 6∶2∶2 的比例划分后的训练、验证和测试集。

表 1　裁剪后的 YawDD 数据集的视频数量

Table 1　Number of videos in the cropped YawDD

	说话	正常	打哈欠
训练集	2 579	1 719	897
验证集	860	573	299
测试集	860	573	299

由于 UTA-RLDD 数据集是受试者通过手机/网络摄像头自己拍摄的,所以每组视频的像素大小及头部位置都不同。本文通过裁剪每个视频的头部位置附近的图像作为数据集,裁剪后每个视频为 5 s 左右。如表 2 所示,新的数据集共有 21 518 个视频,其中警戒状态的类别有 7 252 个视频,低警惕性类别的视频有 7 182 个,困倦类别的视频有 7 084 个。数据集按 8.5∶0.5∶1 的比例划分为训练、验证和测试集。

表 2　裁剪后的 UTA-RLDD 数据集的视频数量

Table 2　Number of videos in the cropped UTA-RLDD

	警戒	低警惕性	困倦
训练集	6 201	6 091	6 029
验证集	364	358	355
测试集	686	733	700

2.1.4　评估

由于数据集类别比例不均衡,为了减少模型的误判,本文采用 F1 得分(F1-score)而不是准确率用于评估模型的性能。F1-score 又称平衡 F 分数(balanced F score),定义为查准率 P 和召回率 R 的调和平均数。

P(precision)是指被分类器判定正例中的正样本的比率,即

$$P = \frac{TP}{FP + TP} \tag{5}$$

式中,TP 表示分类器判定为正例,且判定正确;FP 表示分类器判定为正例,但是判定错误。

R(recall)是指被预测为正例的样本占总的正例的比率,即

$$R = \frac{TP}{TP + FN} \tag{6}$$

式中,FN 表示分类器判定为负例,但是判定错误。

F1 得分是准确率和召回率的调和平均值,即

$$F1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

2.2　模型训练结果

实验以主频 3.60 GHz、内存 32 GB 的 Intel(R) Xeon(R) W-2123 处理器和 Quadro GP100 GPU 为实验平台,采用 TensorFlow/Keras 平台建立卷积神经网络模型,并选择 TensorFlow 框架作为 Keras 平台的后端。

在准备好训练数据集后,下一步是验证网络模型的性能。本文在深度学习框架上建立神经网络模型,利用 TensorFlow 提供的多线程输入数据框架,将训练数据(批)组合起来,每一个 epoch 通过打乱数据序列,以提高模型训练的效率。

训练模型时,每段视频分为 5 s,间隔选取视频 90 帧,训练数据每一批次 24 样本,并设置图像分辨率 80×80 像素。训练时占用 GPU 内存约为 16 GB,使用全局池化层(GAP)代替 softmax 之前的全连接

层,以此来降低空间参数且防止过拟。使用 Adam 优化算法用于反向传播,初始学习率设为 0.000 1,设置学习率衰减为 0.000 01,一个 epoch 是指所有的训练集都训练一次。训练中使用早停法(early stopping),设置 loss 在 10 个 epoch 不下降即可停止训练。

2.2.1 YawDD 上的测试结果

图 5 是在 YawDD 上,模型训练 F1-score 结果折线图。可以看出,训练集和验证集在 27 个 epoch 的 F1 分数稳定在 99.06% 。

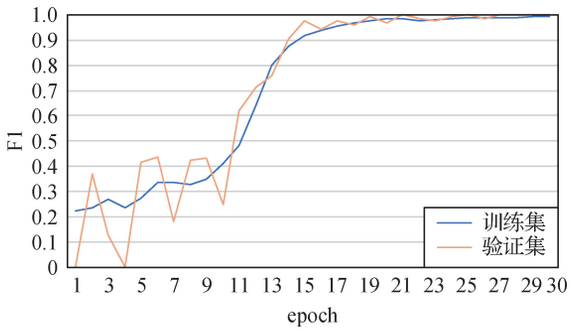


图 5 训练集与验证集的 F1-分数折线图

Fig. 5 A line graph of F1-scores for training sets and test sets

本文采用 F1-Socre 评估模型的性能,如表 3 所示。

表 3 YawDD 数据集上模型测试结果

Table 3 Test results of the model on YawDD

	正常	说话	打哈欠
R	0.996 3	**1**	**1**
P	**1**	0.997 4	**1**
F1	0.998 1	0.998 7	**1**

注:加粗字体为每行最优值。

无论是戴着太阳镜或近视镜,或是不同的面部特征,本文方法检测车辆中打哈欠(yawning)的行为达到了召回率 100% 和精确度 100% 的结果,没有漏检和误检,将哈欠行为与其非常相似的说话行为很好地区分开来,在 F1 分数上达到了 1。

值得注意的是,正常类别(normal)的召回率是 0.996 3,说话行为类别(talking)的精确度是 0.997 4,全部是由于出现说话行为类别误报为正常行为类别导致的。分析了误报的原因,具体的情况是说话行为通常包含在正常行为中,说话行为在 5 s 的视频中小于 0.5 s,没有足够的张口时间,导致了说话行为误报为正常行为。

2.2.2 UTA-RLDD 上的测试结果

通过对在 YawDD 上的训练模型的迁移,快速在 UTA-RLDD 上训练模型。删除 YawDD 上训练模型的 softmax 层,添加一个新的 softmax 层,新的 softmax 层作为分类层。在 UTA-RLDD 数据集上本文方法检测在困倦疲劳状态上达到了召回率 100% 和精确度 100% 的结果,没有漏检和误检,在 F1 分数上达到了 1。如表 4 所示,警戒(alert)的召回率和低警惕性(low vigilant)的精确率都达到了 100% ,这表明模型所有的错误来自于把低警惕性状态预测为警戒状态。分析了误报的原因,具体的情况是低警惕性状态由警戒状态过度而来,由于一些样本并没有太过频繁的眨眼行为和表情变化,特征介于两者之间,这种情况下的低警惕性状态样本被误报为警戒状态。

表 4 UTA-RLDD 数据集上模型测试结果

Table 4 Test results of the model on UTA-RLDD

	警戒	低警惕性	困倦
R	**1**	0.989 3	**1**
P	0.989 5	**1**	**1**
F1	0.994 7	0.994 6	**1**

注:加粗字体为每行最优值。

2.3 方法比较与分析

本文基于伪 3D 卷积神经网络与注意力机制的驾驶疲劳检测效果,与通过 LSTM 学习时间特征的模型进行对比试验。Inception-v3 融合 LSTM 的模型以 Inception-v3 模型作为特征提取程序,删除架构的 softmax 层,添加一个 LSTM 层和一个新的 softmax 层,新的 softmax 层作为分类层。两种方法都在数据增强后的 8 660 个视频的数据集上进行,学习率、衰减参数和 batch 等参数均保持一致。

从预测 F1 分数、预测帧频、模型大小和首次达到相对极值点的 epoch 等 4 个方面进行比较,如表 5 所示。首先,LSTM 学习模型的 F1 分数为 0.975 8,本文方法 F1 分数达到 0.998 9。

在预测时间方面,本文计算视频预处理与模型预测时间的累加。LSTM 学习模型使用 Inception-v3 模型进行空间特征提取需要消耗大量时间,整个模

表 5　不同模型测试结果

Table 5　Test results of different models

方法	F1	预测时间/ms	模型大小/MB
Inception-v3 + LSTM	0.977 2	5 900	396.15
本文	**0.998 9**	**660**	**42.50**

注:加粗字体为每列最优值。

型预测时间平均约为 5 900 ms,本文方法预测一次 5 s 的视频需要 660 ms,本文方法预测时间约为 LSTM 学习模型的 11% 左右。

Inception-v3 融合 LSTM 的方法需要使用 Inception-v3 模型进行空间特征提取,Inception-v3 模型参数大小为 84.81 MB,Inception-v3 结合一层 LSTM 的融合模型大小为 396.15 MB。本文方法的模型参数包含 6 个注意力模块参数、2D GAP 参数和 6 个 P3D 结构的卷积神经网络参数,未剪枝的模型大小为 42.5 MB。本文方法为 LSTM 学习模型大小的 1/9。

2.4　注意力模块比较实验

注意力机制为不同的通道与特征赋予不同的权重,在经过几次卷积后,90 帧图像时空特征信息已经融合为 7 帧,图 6 是同一通道经过通道注意力机制与空间注意力机制后的特征对比图,可以明显看出,脸部的特征以及更重要的眼部与嘴部的特征更为明显。

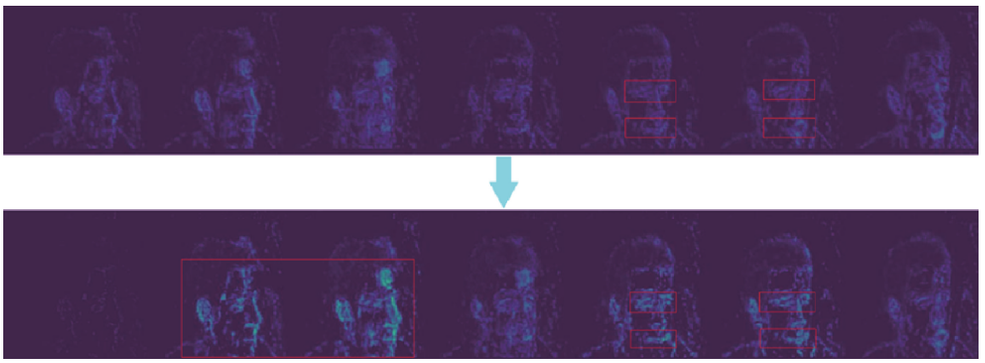


图 6　注意力机制处理后的特征对比图

Fig. 6　Feature comparison diagram after attention mechanism processing

图 7 是不同注意力模块在训练集上的 F1 分数对比图。本文方法在后 3 层 P3D 卷积中根据 P3D 具体结构添加两种注意力模块,具体内容已在 1.3 节中论述。在融合注意力模块方面,比较 4 种不同的结构:1)所有层加入注意力模块;2)后 3 层卷积模块直接级联 CBAM 注意力模块;3)后 3 层卷积模块使用 7 × 7 大小卷积核的空间注意力;4)本文方法。

所有层加入注意力模块的方法与本文方法相比,目的是为了验证在 CNN 网络的不同层上反复施加的视觉注意力是否越多越好。在提出的模型结构中,其后 3 个 P3D-Attention 模块与本文模型结构相同,前 3 个空间注意力模块的卷积核都为 7 × 7 大小。从图 7 中关于两者折线图之间的比较来看,本文方法在后续训练中 F1 分数保持领先优势。

使用 7 × 7 大小卷积核的空间注意力模块与本文方法相比,是为了验证使用不同大小卷积核的空间注意力模块是否可以满足不同大小的特征图对固定尺度的关注。在提出的模型结构中,最后 3 个卷积模块中分别添加 7 × 7、5 × 5、3 × 3 大小的空间注意力模块。如图 7 所示,在第 13 个 epoch 后本文方法在后续训练中 F1 分数快速上升,并一直保持领先优势。

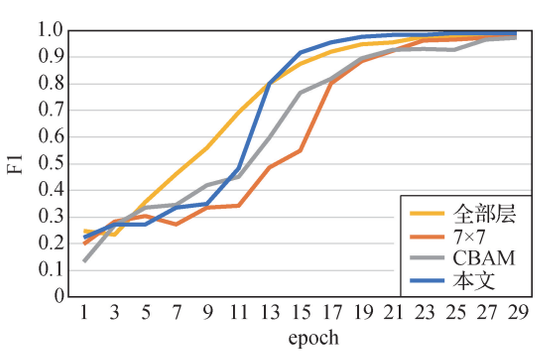


图 7　注意力模块的训练 F1 分数的对比

Fig. 7　Comparison of F1 score in training sets of different attention modules

直接级联 CBAM 模块与本文方法相比,本文方法更有优势。CBAM 模块具有普适性,可以通过直接级联的方法,在 CNN 卷积层后添加两种注意力模

块。为了单独比较级联 CBAM 模块与本文方法的差异,级联 CBAM 模块的方法在注意力模块放置位置上与本文方法相同,其后 3 个 P3D 模块直接级联 CBAM 模块。如图 7 所示,级联 CBAM 模块方法的训练集 F1 分数在第 30 个 epoch 时只有 0.97,其训练的速度与精度在 4 种方法中最差。

2.5 全局平均池化比较实验

使用 GAP 的目的是进行全连接的替换,减少参数的数量,防止过拟合。本文实验中 GAP 替换全连接层,并采用 2D GAP 而不是 3D GAP 以保留更多的时间特征。

图 8 是不同维度全局平均池化层的两种模型在训练集上 F1 分数的对比图。通过对比实验表明,训练过程中使用 2D GAP 的模型更快地收敛,F1 分数相比较而言更高。使用 2D GAP 的模型在第 27 个 epoch 时,F1 分数即达到 99.68%,并一直保持领先优势。比较而言,使用 3D GAP 的模型在 82 个 epoch 后才会达到 2D GAP 的训练预测效果,且后续训练波动较大并不稳定。

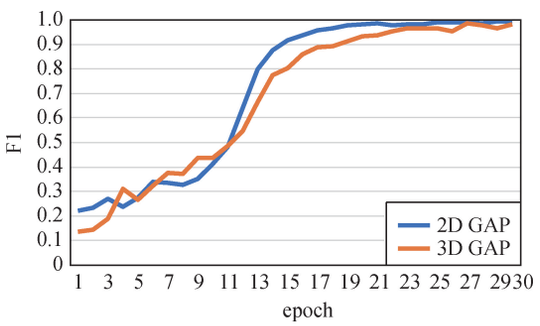


图 8 2D GAP 和 3D GAP 的训练 F1 分数的对比

Fig. 8 Comparison of F1-score of the training set for the global average pooling layer of different dimensions

3 结 论

本文为加大对时间特征的关注,提出一种基于伪 3D 卷积神经网络与注意力机制的驾驶疲劳检测方法。以 P3D 模块时间卷积与空间卷积解耦的特性融合通道和空间注意力模块,设计出 P3D-Attention 模块,对关键帧的时空特征建模,使用通道注意力模块提高重要通道特征的相关度,通过适应的空间注意力模块增加特征图的全局相关性。使用 2D 全局平均池化层来获取更丰富、更精确的特征。实验结果表明,本文模型可以有效利用时间序列上的集成信息,提高对驾驶疲劳行为的预测性能。

本文进一步工作为:1)验证是否可以通过更小的网络结构来提取特征,设计更高效的网络结构,进一步缩减模型大小;2)由于分心驾驶行为不仅仅只需要关注预测驾驶员的疲劳状态,未来工作将集中在利用 3D 卷积区分更多、更复杂的驾驶行为上。

参考文献(References)

Abtahi S, Omidyeganeh M, Shirmohammadi S and Hariri B. 2014. YawDD: a yawning detection dataset//Proceedings of the 5th ACM Multimedia Systems Conference. Singapore, Singapore: Association for Computing Machinery: 24-28 [DOI: 10.1145/2557642.2563678]

Alioua N, Amine A and Rziza M. 2014. Driver's fatigue detection based on yawning extraction. International Journal of Vehicular Technology, 2014: #678786 [DOI: 10.1155/2014/678786]

Balandong R P, Ahmad R F, Saad M N M and Malik A S. 2018. A review on EEG-based automatic sleepiness detection systems for driver. IEEE Access, 6: 22908-22919 [DOI: 10.1109/ACCESS.2018.2811723]

Caffier P P, Erdmann U and Ullsperger P. 2003. Experimental evaluation of eye-blink parameters as a drowsiness measure. European Journal of Applied Physiology, 89(3): 319-325 [DOI: 10.1007/s00421-003-0807-5]

Chui K T, Tsang K F, Chi H R, Ling B W K and Wu C K. 2016. An accurate ECG-Based transportation safety drowsiness detection scheme. IEEE Transactions on Industrial Informatics, 12(4): 1438-1452 [DOI: 10.1109/TII.2016.2573259]

Clavijo G L R, Patiño J O and León D M. 2015. Detection of visual fatigue by analyzing the blink rate//Proceedings of the 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA). Bogota, Colombia: IEEE: 1-5 [DOI: 10.1109/STSIVA.2015.7330398]

Forsman P M, Vila B J, Short R A, Mott C G and Van Dongen H P A. 2013. Efficient driver drowsiness detection at moderate levels of drowsiness. Accident Analysis and Prevention, 50: 341-350 [DOI: 10.1016/j.aap.2012.05.005]

Friedrichs F and Yang B. 2010. Camera-based drowsiness reference for driver state classification under real driving conditions//2010 IEEE Intelligent Vehicles Symposium. San Diego, USA: IEEE: 101-106 [DOI: 10.1109/IVS.2010.5548039]

Ghoddoosian R, Galib M and Athitsos V. 2019. A realistic dataset and baseline temporal model for early drowsiness detection [EB/OL]. [2020-02-20]. https://arxiv.org/pdf/1904.07312.pdf

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Pro-

ceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141 [DOI: 10.1109/CVPR.2018.00745]

Ji S W, Xu W, Yang M and Yu K. 2013. 3D Convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231 [DOI: 10.1109/TPAMI.2012.59]

Khandelwal S and Sigal L. 2019. AttentionRNN: a structured spatial attention mechanism[EB/OL]. [2020-03-20]. http://arxiv.org/pdf/1905.09400.pdf

Li Z J, Chen L K, Peng J and Wu Y. 2017. Automatic detection of driver fatigue using driving operation information for transportation safety. Sensors, 17(6): #1212 [DOI: 10.3390/s17061212]

Lin M, Chen Q and Yan S C. 2014. Network in network//Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014-Conference Track Proceedings [EB/OL]. [2020-03-20]. https://arxiv.org/pdf/1312.4400.pdf

Qiu Z F, Yao T and Mei T. 2017. Learning spatio-temporal representation with pseudo-3D residual networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5534-5542 [DOI: 10.1109/ICCV.2017.590]

Sahayadhas A, Sundaraj K and Murugappan M. 2012. Detecting driver drowsiness based on sensors: a review. Sensors, 12(12): 16937-16953 [DOI: 10.3390/s121216937]

Tefft B C. 2018. Acute sleep deprivation and culpable motor vehicle crash involvement. Sleep, 41(10): #zsy144 [DOI:/10.1093/sleep/zsy144]

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]

Wang P, Min J L and Hu J F. 2018. Ensemble classifier for driver's fatigue detection based on a single EEG channel. IET Intelligent Transport Systems, 12(10): 1322-1328 [DOI: 10.1049/iet-its.2018.5290]

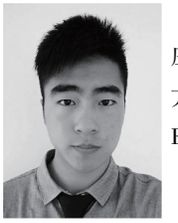
Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Xie Y Q, Chen K X and Murphey Y L. 2019. Real-time and robust driver yawning detection with deep neural networks//2018 IEEE Symposium Series on Computational Intelligence (SSCI). Bangalore, India: IEEE: 532-538 [DOI: 10.1109/SSCI.2018.8628881]

Xing Y, Lv C, Wang H J, Cao D P, Velenis E and Wang F Y. 2019. Driver activity recognition for intelligent vehicles: a deep learning approach. IEEE Transactions on Vehicular Technology, 68(6): 5379-5390 [DOI: 10.1109/TVT.2019.2908425]

Zhang W W and Su J Y. 2017. Driver yawning detection based on long short term memory networks//2017 IEEE Symposium Series on Computational Intelligence (SSCI). Honolulu, USA: IEEE: 1-5 [DOI: 10.1109/ SSCI.2017.8285343]

作者简介



庄员,1996 年生,男,硕士研究生,主要研究方向为视频理解、目标检测与计算机视觉。
E-mail:2681844336@qq.com

戚湧,通信作者,男,教授,博士生导师,主要研究方向为交通大数据。
E-mail: 790815561@ qq.com