doi:10.19562/j.chinasae.qcgc.2024.01.001

基于多尺度骨架图和局部视觉上下文融合的驾驶员行为识别方法*

胡宏宇,黎烨宸,张争光,曲 优,何 磊,高镇海
(吉林大学,汽车仿真与控制国家重点实验室,长春 130022)

[摘要] 识别非驾驶行为是提高驾驶安全性的重要手段之一。目前基于骨架序列和图像的融合识别方法具有计算量大和特征融合困难的问题。针对上述问题,本文提出一种基于多尺度骨架图和局部视觉上下文融合的驾驶员行为识别模型(skeleton-image based behavior recognition network,SIBBR-Net)。SIBBR-Net通过基于多尺度图的图卷积网络和基于局部视觉及注意力机制的卷积神经网络,充分提取运动和外观特征,较好地平衡了模型表征能力和计算量间的关系。基于手部运动的特征双向引导学习策略、自适应特征融合模块和静态特征空间上的辅助损失,使运动和外观特征间互相引导更新并实现自适应融合。最终在 Drive&Act 数据集进行算法测试,SIBBR-Net在动态标签和静态标签条件下的平均正确率分别为 61.78% 和 80.42%,每秒浮点运算次数为 25.92G,较最优方法降低了 76.96%。

关键词:驾驶员行为识别;多尺度骨架图;局部视觉上下文;多模态数据自适应融合

Driver Behavior Recognition Based on Multi-scale Skeleton Graph and Local Visual Context Method

Hu Hongyu, Li Yechen, Zhang Zhengguang, Qu You, He Lei & Gao Zhenhai
Jilin University, State Key Laboratory of Automotive Simulation and Control, Changchun 130022

[**Abstract**] Non-driving behavior identification is one of the important ways to improve the safety of driving. The current recognition method based on skeleton sequence and image fusion has the problems of large model calculation and the difficulty of feature fusion. To address the above problems, the skeleton-image based behavior recognition network (SIBBR-Net) is proposed in this paper, which is based on the multi-scale skeleton graph and the local visual context. SIBBR-Net fully extracts motion and appearance features through a graph convolution network based on multi-scale skeleton graphs and a convolutional neural network based on local vision and attention mechanisms, and better balances the relationship between model representation capabilities and model calculation. The feature bi-directional guided learning strategy based on hand motion, an adaptive feature fusion module and an auxiliary loss on the static feature space can guide mutual guidance and updating between motion and appearance features to achieve adaptive fusion. SIBBR-Net is finally tested on the Drive & Act dataset, and the average accuracy is 61.78% for dynamic labels and 80.42% for static labels. The Floating-point Operations per Second (FLOPS) of SIBBR-Net is 25.92G, which is 76.96% lower than that of the optimal method.

Keywords: driver behavior recognition; multi-scale skeleton graph; local visual context; multi-model data adaptive fusion

* 吉林省自然科学基金(20210101064JC)、国家自然科学基金(52272417)、新能源智能汽车关键技术研发及产业化项目(TC210H02S)和大学生创新创业训练计划项目(X202310183158)资助。
原稿收到日期为 2023 年 07 月 26 日,修改稿收到日期为 2023 年 09 月 09 日。
通信作者:何磊,教授,博士,E-mail: jlu_helei@jlu.edu.cn。

前言

在驾驶车辆时,从事非驾驶行为会降低驾驶员对车辆的操控能力及对周围驾驶环境的感知能力[1]。为了保障驾驶安全性,识别驾驶员行为并提醒驾驶员保证对车辆的正常操控是十分重要的。

目前基于相机传感器的驾驶员行为识别方法是主流的研究方法。该类方法主要包括基于图像或骨架序列的识别方法。根据相机位置,基于图像的识别方法可被细分为基于头面部或手部图像的识别方法。对于头面部图像,研究人员通过追踪识别头部或双眼的运动,确定驾驶员的注视方向,从而识别驾驶员行为。Yang等[2]基于面部图像,利用条件神经过程算法提取面部关键点,构建非线性模型获取驾驶员视线方向的热力图,从而确定行为类别。对于手部图像,研究人员主要识别双手的位置,以及手部交互动作,进而识别驾驶员行为。Zheng等[3]基于注意力机制提出CornerNet-Saccade模型,用于识别手机等交互物体。对于由图像提取的骨架序列,研究人员需要对其进行预处理,如构造辅助关键点[4],进而构建时空运动特征提取模型,确定驾驶员行为类别。Holzbock等[5]将连续的24帧骨骼关键点坐标序列输入多层感知机(multilayer perceptron,MLP),通过关键点之间的时空关系识别驾驶员行为。Li等[6]基于图卷积神经网络和遗传算法,选择时空运动变化显著的骨架关键点。进而,研究人员对骨架关键点特征和驾驶员行为类别间进行相关性分析。然而,目前基于图像的方法占用大量计算资源且难以获取具体的动态运动信息;基于骨架序列的方法虽然计算资源开销较低,但由于忽略了所有外观信息,运动情况相似的行为间易发生混淆。为了更好地识别驾驶员行为,基于图像和骨架序列融合的识别方法成为研究趋势。

对于图像与骨架序列融合的识别方法,主要采用双流网络结构分别提取骨架运动信息和图像外观信息,再通过融合分类模块融合上述特征并输出驾驶员行为类别[7-8]。Weyers等[9]从卷积神经网络提取双手邻近区域的图像外观特征,然后与骨架序列特征进行拼接。拼接后的特征向量被输入到长短期记忆网络中,进而对驾驶员行为进行分类。Tan等[10]提出姿态和外观交互网络(bidirectional posture-appearance interaction network ,BPAI-Net)。BPAI-Net通过时空图卷积神经网络(spatial temporal graph convolutional networks,ST-GCN)[11]和3D卷积神经网络(inflated 3D convNet,I3D)[12]分别提取运动信息和外观信息,学习两者之间的交互特征,提高识别驾驶员行为的准确性。然而,该类方法的骨架和图像分支网络均对时序信息进行建模,存在冗余计算的问题。同时运动和外观信息分别处于动态和静态特征空间,存在不同特征空间信息间融合困难的问题。

因此,考虑到驾驶员行为识别任务的实时性要求高,同时为了解决跨特征空间信息融合难的问题,本文中提出基于多尺度骨架图和局部视觉上下文融合的驾驶员行为识别模型:SIBBR-Net。主要贡献如下:

(1)SIBBR-Net采用基于多尺度骨架图的图卷积网络,和以单帧为输入的基于局部视觉和注意力机制的卷积神经网络,保证模型表征能力的同时,减少了所需的计算开销;

(2)SIBBR-Net构建特征间双向交流模块(bidirectional exchange module,BEM)、自适应特征融合模块(adaptively feature fusion module, AFFM)和辅助损失,使骨架与图像信息间互相引导更新并自适应融合。

1 研究方法

1.1 问题定义

驾驶员行为识别任务中,给定原始样本$S=\{I_t\in\mathbb{R}^{H\times W\times 3},G_t\in\mathbb{R}^{V\times C}\}_{t=1}^{T}$,本文选取第$t$帧图像$I_t\in\mathbb{R}^{H\times W\times 3}$与长度为$T'$的骨架序列$G\in\mathbb{R}^{T'\times V\times C}$,组成样本$X=\{I_t,G\}$,其中$T$为原始样本长度;$H\times W$为分辨率;$V$为骨架关键点数量;$C$为坐标维数。给定样本$X$,本文的任务是提出行为识别模型$f$,并输出类别分数$Y=f(X,\theta)\in\mathbb{R}^{N\times 1}$,其中$N$为驾驶员行为类别总数,$\theta$为可训练参数集合。

1.2 模型整体框架

首先,为了获取骨架序列在不同语义层次上的运动信息,SIBBR-Net的骨架分支网络选用基于多尺度图的图卷积神经网络。为了获取局部视觉上下文特征,图像分支网络以单帧图像为输入,通过基于注意力机制的卷积神经网络,提取驾驶员行为中的判别性外观信息;其次,考虑到驾驶员的手部活动较

为活跃,本文中提出一种基于手部运动信息的BEM,使运动和外观特征互相引导更新;最后,通过AFFM和辅助损失,实现骨架和图像特征的互补融合。SIBBR-Net框架如图1所示。

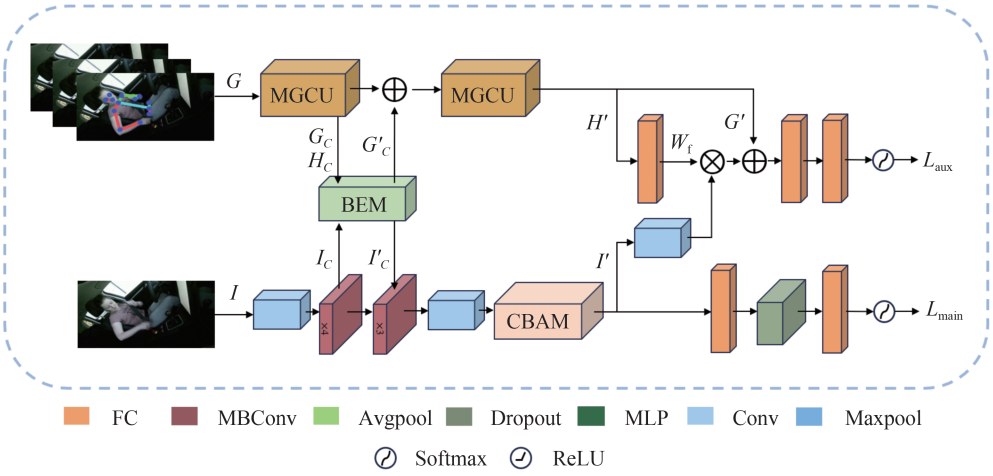


图1　SIBBR-Net的结构示意图

1.3　骨架分支网络

考虑到非驾驶行为的发生过程中骨架关键点间存在实际的物理连接关系,具有分组运动的特点,使用基于动态多尺度图神经网络(dynamic multiscale graph neural networks,DMGNN)[13]的骨架分支网络。本文将多尺度图定义为$\{\boldsymbol{G}_{\mathrm{g}}\in\mathbb{R}^{T'\times V_{\mathrm{g}}\times C_{\mathrm{g}}};\boldsymbol{G}_{\mathrm{p}}\in\mathbb{R}^{T'\times V_{\mathrm{p}}\times C_{\mathrm{p}}};\boldsymbol{G}_{\mathrm{d}}\in\mathbb{R}^{T'\times V_{\mathrm{d}}\times C_{\mathrm{d}}}\}$,其中,$\boldsymbol{G}_{\mathrm{g}}$为全局尺度图,$\boldsymbol{G}_{\mathrm{p}}$为局部尺度图,$\boldsymbol{G}_{\mathrm{d}}$为动态尺度图。DMGNN通过两个多尺度图计算模块(multiscale graph computational unit,MGCU)捕获不同尺度图上的空间和时序判别性特征,并实现跨尺度图间的特征融合,整体结构如图2所示。

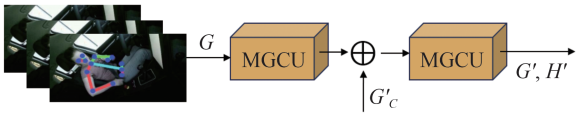


图2　骨架分支网络的结构示意图

MGCU模块包含单尺度图卷积模块(single-scale graph convolution block,SS-GCB)和跨尺度图融合模块(cross-scale fusion block,CS-FB)。MGCU结构如图3(a)所示

SS-GCB按照分组关系,将单尺度图内的关键点进行拼接。接着,SS-GCB将初始化后的单尺度图$\boldsymbol{G}_{\mathrm{in}}$分别输入时空图卷积模块ST-GCN,获得单尺度图的运动信息$\boldsymbol{G}_{\mathrm{out}}$。ST-GCN的计算公式为

$$\boldsymbol{G}_{\mathrm{out}}=\hat{\boldsymbol{A}}\boldsymbol{G}_{\mathrm{in}}\boldsymbol{W} \tag{1}$$

$$\hat{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}} \tag{2}$$

$$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N \tag{3}$$

式中:$\boldsymbol{A}$为邻接矩阵;$\boldsymbol{W}$为可训练参数;$\boldsymbol{D}$为度矩阵;$\boldsymbol{I}_N$为单位矩阵。

CS-FB能够充分地利用骨架序列间的内部关系,捕获骨架序列在空间和时序上的依赖性运动特征。CS-FB首先在时间通道上对跨尺度图S_1、S_2上各节点进行卷积;然后将单尺度图上的特征进行拼接;最后通过带有残差结构的MLP,实现跨尺度图间的特征融合,整体结构如图3(b)所示。

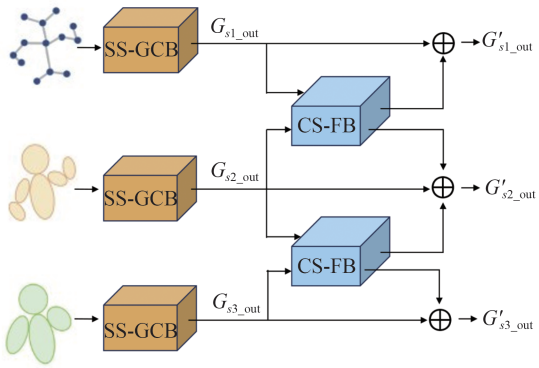
1.4　图像分支网络

图像分支网络以单帧图像I_t为输入,选取EfficientNet[14]作为骨干网络,结合注意力机制,提高外观信息的提取能力,整体结构如图4所示。

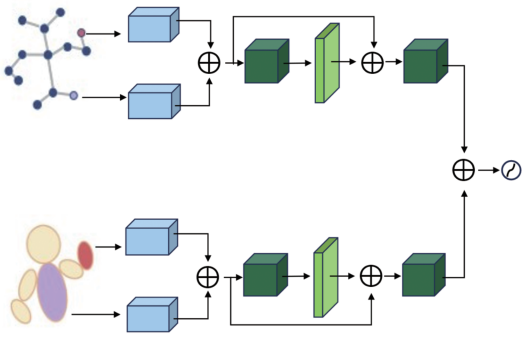
EfficientNet由若干移动可翻转卷积块(mobile inverted residual bottleneck block,MBConv)堆叠构成。MBConv由普通卷积、深度可分离卷积、压缩与激发模块(squeeze and excitation,SE)和普通卷积组成,其中SE由平均池化层和全连接层组成。MBConv模块结构如图5所示。

卷积注意力机制(convolutional block attention module,CBAM)[15]集成了通道注意力机制(channel attention,CA)和空间注意力机制(spatial attention,SA),整体结构如图6所示。

CA和SA分别用于捕捉特征图在不同通道和位置之间的依赖性特征。CA首先将输入的特征图并



(a) MGCU模块结构示意图



(b) CS-FB模块结构示意图

图 3　MGCU模块结构和CS-FB模块结构示意图

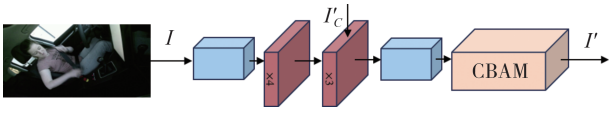


图 4　图像分支网络结构示意图

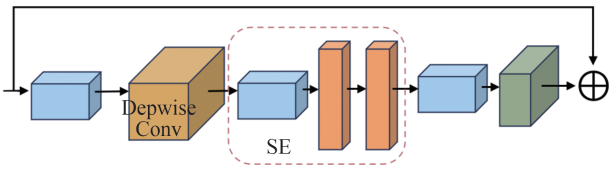


图 5　MBConv模块结构示意图

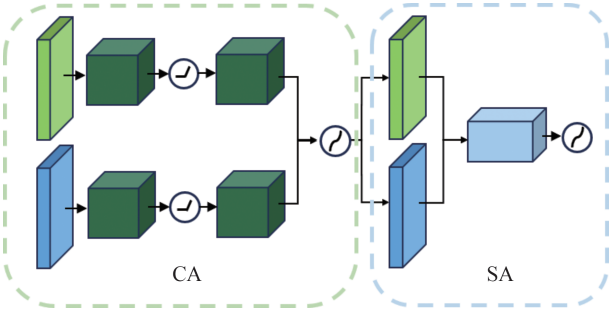


图 6　CBAM模块结构示意图

行通过平均和最大池化层,特征图的尺寸由 $\mathbb{R}^{C\times H\times 3}$ 变为 $\mathbb{R}^{C\times 1\times 1}$;然后经过共享权重的MLP,特征图的通道数被压缩为原来的 1/r,接着通过MLP将ReLU激活函数的输出特征图扩张到原通道数;最后相加两个输出结果并通过sigmoid激活函数得到CA权重。SA将CA权重作为输入,首先在通道维度并行通过平均和最大池化层;然后将二者输出进行拼接;最后通过卷积和sigmoid激活函数,获得尺寸为 $\mathbb{R}^{1\times H\times W}$ 的SA权重。CBAM将原始特征图与SA权重相乘得到图像分支网络的最终特征图 $I'_t\in\mathbb{R}^{H\times W\times 3}$。

1.5　融合模块

融合模块由BEM和AFFM组成,能够实现运动和外观特征间的相互引导更新和自适应融合。本文通过拼接各尺度图中手部关键点的特征,获取手部运动信息 H_C 和 H',获取方法如图 7 所示。

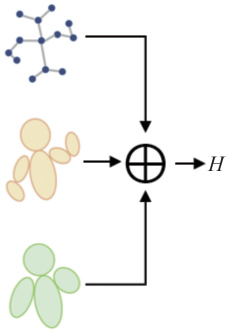


图 7　获取手部运动信息示意图

BEM由骨架更新模块(skeleton update,SU)和图像位置注意力模块(image situation attention,ISA)组成,能够促使运动和外观特征间相互引导更新,整体结构如图 8 所示。SU以手部运动信息 H_C、图像特征图 I_C 及骨架特征图 G_C 为输入,使外观特征引导运动特征的提取过程。通过全连接层后的 H_C 和通过卷积后的 I_C 相乘,BEM得到骨架更新权重 $W_G=G(H_C)\otimes H(I_C)$,其中,G 代表全连接层,H 代表卷积层。更新后的骨架特征图 $G'_C=w\otimes G_C+G_C$。SA通过图像特征的位置注意力,使图像分支网络进一步关注与运动相关的外观特征。该模块的输入是手部运动信息 H_C 和图像特征图 I_C,位置注意力 $W_I=G(G(H_C))$,更新后的图像特征图 $I'_C=W_I\otimes I_C$。

考虑手部运动信息能够判别性地反映模型对外观辅助性特征的需求程度,本文采取高维特征间自

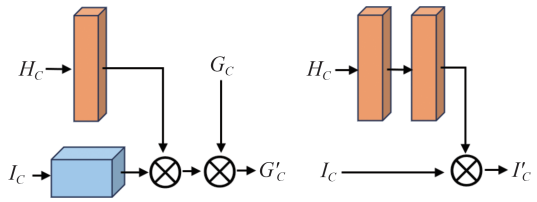


图 8　BEM结构示意图

适应融合的设计思路。AFFM首先从手部运动特征获得自适应权重 $W_f = G(H')$;接着将外观特征 $W_f \otimes H(I')$与运动特征 G'进行结合,使外观特征能够自适应地补充运动特征;最后通过由全连接层和Softmax激活函数组成的分类器获得分类结果。AFFM结构如图9所示。

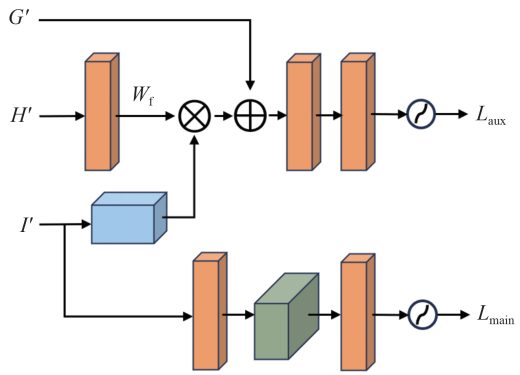


图 9　AFFM结构示意图

1.6　损失函数

为了确保各分支网络都能在各自特征空间中充分地提取判别性特征,本文以动态特征空间的交叉熵损失为主损失 L_{main},以静态特征空间的交叉熵损失为辅助损失 L_{aux},总损失 L 为两者的加权之和。L_{aux} 为图像分支网络在静态标签上的分类损失;L_{main} 为SIBBR-Net在动态标签上的分类损失。总损失计算公式为

$$L = \alpha L_{main} + (1 - \alpha) L_{aux} \tag{4}$$

$$L_{main} = -\sum_{i=1}^{N} y_i^{d} \log\left(\widehat{y_i^{d}}\right) \tag{5}$$

$$L_{aux} = -\sum_{i=1}^{N} y_i^{s} \log\left(\widehat{y_i^{s}}\right) \tag{6}$$

式中:α 为权重系数;N 为样本总数;y^d 表示动态标签样本真实分布的概率;$\widehat{y^d}$ 表示动态标签样本预测概率;y^s 表示静态标签样本真实分布的概率;$\widehat{y^s}$ 表示静态标签样本预测概率。

2　实验部分

2.1　数据集

本文使用Drive&Act[16]数据集验证算法的有效性。该数据集在真实场景下,使用带有5个视角的多视角相机系统,采集了约12 h的驾驶员行为视频数据和骨架关键点3D坐标。骨架序列以“Front-top”角度拍摄的红外视频为蓝本,以MSCOCO的标准格式对骨架关键点进行标注。由于舱内驾驶员的下肢被遮挡,本文选取驾驶员头部及上肢共13个关键点组成骨架序列。同时,数据集的样本标签按照整体行为、交互物体和动态行为被分为3个层次,即task-level、object-level和mid-level。本文根据模型的训练要求,以mid-level标签作为动态标签,如表1所示;以静态交互物体类别为分类标准,本文将动态标签自行重新划分为静态标签,如表2所示。Drive&Act数据集的数据标注示例如图10所示。

表1　动态标签类别

C0	Closing-door-outside	C17	Drinking
C1	Opening-door-outside	C18	Closing-bottle
C2	Entering-car	C19	Looking-or-moving-around
C3	Closing-door-inside	C20	Preparing-food
C4	Fastening-seat-belt	C21	Eating
C5	Using-multimedia-display	C22	Taking-off-sunglasses
C6	Sitting-still	C23	Putting-on-sunglasses
C7	Pressing-automation-button	C24	Reading-newspaper
C8	Fetching-an-object	C25	Writing
C9	Opening-laptop	C26	Talking-on-phone
C10	Working-on-laptop	C27	Reading-magazine
C11	Interacting-with-phone	C28	Taking-off-jacket
C12	Closing-laptop	C29	Opening-door-inside
C13	Placing-an-object	C30	Exiting-car
C14	Unfastening-seat-belt	C31	Opening-backpack
C15	Putting-on-jacket	C32	Putting-laptop-into-backpack
C16	Opening-bottle	C33	Taking-laptop-from-backpack

2.2　实验设置

(1)数据划分及数据标签:本文与Drive&Act数据集的原始数据划分方式保持一致,根据驾驶员id划分训练集、测试集和验证集。在驾驶员行为领域中,目前本文调研到的领先水平在动态行为标签上的平均正确率为67.83%[10]。在标注动态行为标签时,Drive&Act数据集细分了驾驶员行为的发生过程。考虑到细分行为间存在连贯性,以及缺乏明确

表 2 静态标签类别

C0	Default	C9	Interact-with-jacket
C1	Interact-with-seat-belt	C10	Drinking
C2	Interact-with-multimedia-display	C11	Looking-or-moving-around
C3	Sitting-still	C12	Eating
C4	Interact-with-automation-button	C13	Interact-with-sunglasses
C5	Fetching-an-object	C14	Reading
C6	Interact-with-laptop	C15	Writing
C7	Interact-with-phone	C16	Interact-with-backpack
C8	Placing-an-object		

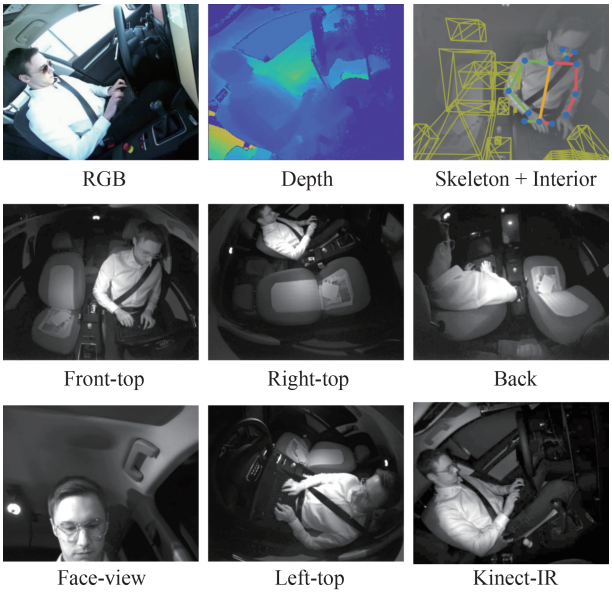


图 10 以 C10“Working-on-laptop”类别为例的图像和多模态数据标注示例

的行为边界,本文认为该数据集的正确率主要受标签类别的影响。同时动态行为标签包含 C0(closing-door-outside)和 C1(opening-door-outside)等发生于舱外的驾驶员行为。该类行为无法通过舱内相机传感器采集图像和骨架序列数据,进而识别驾驶员行为类别,这也将对识别效果产生不良影响。因此使用动态行为标签来识别驾驶员行为是具有挑战性的。

(2)数据预处理:在数据预处理环节中,本文在原始样本 S 中选取样本 $X=\{I_t, G\}$,选取方式如图 11 所示。预处理后单帧图像 $I_t \in \mathbb{R}^{H\times W\times 3}$,其中,$H=224$,$W=224$;骨架序列 $G \in \mathbb{R}^{T'\times V\times C}$,其中,$T'=8$ 为样本长度,$V=13$ 为骨架关键点数量,$C=3$ 为坐标维数。

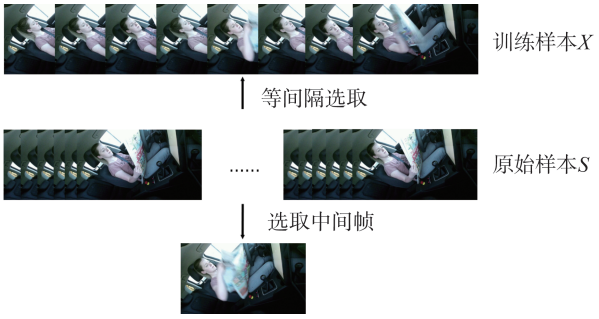


图 11 数据预处理过程示意图

(3)实验平台及训练参数:在 SIBBR-Net 的训练阶段,本文首先对骨架分支网络和图像分支网络分别进行训练;然后剔除各分支网络训练权重的分类权重,获得预训练权重;最后利用该预训练权重对 SIBBR-Net 进行联合训练。硬件平台及训练参数的设置情况,如表 3 所示。

表 3 训练参数配置情况

参数名称	参数
操作系统	Window11
测试框架	Pytorch1. 11. 0
显卡	GeForce RTX 3070ti
优化器	SGD
批次大小	32
迭代轮次	150
初始学习率	0. 001

2.3 评价指标

本文选取每类平均正确率(Accuracy)作为评价指标;选取混淆矩阵对分类性能进行评价;选取每秒浮点运算次数(floating-point operations per second, FLOPs)对模型计算量进行评价。

$$Accuracy=\frac{1}{N}\sum_{i=1}^{N}\frac{TP_i+TN_i}{TP_i+FN_i+FP_i+TN_i} \tag{7}$$

式中:N 为样本总数;TP 为正样本被正确识别的数量;FP 为误报的负样本数量;TN 为负样本被正确识别的数量;FN 为漏报的正样本数量。

2.4 实验超参数设置

对 SIBBR-Net 中数据预处理和损失函数中的超参数设置进行实验对比,进而消除实验超参数设置对后续实验的影响。

(1)数据预处理:在数据预处理环节中,本文需要从原始样本 S 中获取长度为 T' 的训练样本 X 和第 t 帧图像 I_t。在长度 T' 的选取方式中,本文在 S 中等间距选取 $T'=8$ 与 $T'=16$,并在动态行为标签下进

行对比实验;在单帧图像的选取方式中,本文主要对比“选取中间帧”和“随机选取”两种方式,并在静态标签下进行对比实验。从表4实验结果可见,当$T'=8$且单帧图像以“选取中间帧”方式选取时,骨架和图像分支网络的识别效果最佳,因此本文以此方式对样本进行预处理。

表4　不同数据预处理方式的实验结果

模型分类	*Accuracy*/%
DMGNN($T'=8$)	**50.98**
DMGNN($T'=16$)	46.83
EfficientNet+CBAM(选取中间帧)	**75.12**
EfficientNet+CBAM(随机选取)	72.95

(2)损失函数的权重系数:为了验证权重系数α对模型性能的影响,本文对权重系数α的取值进行实验对比,实验结果如表5所示。由于当$\alpha=0.6$时模型性能最佳,因此本文选择$\alpha=0.6$进行后续对比实验和消融实验。

表5　选取不同权重系数的实验结果

权重系数	*Accuracy*/%
$\alpha=0.2$	59.62
$\alpha=0.3$	60.63
$\alpha=0.4$	60.08
$\alpha=0.5$	59.53
$\alpha=0.6$	**61.78**
$\alpha=0.7$	61.67
$\alpha=0.8$	61.54

2.5　对比实验

对SIBBR-Net与其他行为识别模型在Drive&Act数据集上进行实验对比。由于Drive&Act数据集并没有可用于选取对比模型的排行榜,本文与其他研究人员的对比做法保持一致,即与数据集的基线模型进行对比,实验结果如表6所示。基线模型包括基于骨架序列的模型:Pose、Two-stream和ST-GCN,以及基于图像的模型:P3D[17]、C3D[18]与I3D。在本文调研的非基线模型中,达到最佳识别效果的模型为BPAI-Net,其平均正确率为67.83%。

随着局部视觉的研究方法、CBAM、ISA和辅助损失的引入,SIBBR-Net的平均正确率由50.98%(DMGNN)提升至61.78%。融合模型SIBBR-Net的识别效果均大幅度优于基于骨架序列的模型和基于图像序列的C3D和P3D。虽然SIBBR-Net的平均正确率仍低于I3D和BPAI-Net,但其计算量较最优方法降低了76.96%。I3D和BPAI-Net的FLOPs分别为111.3G和112.5G,而SIBBR-Net的FLOPs为25.92G。因此,SIBBR-Net保证了准确性的同时,减少了计算开销,在实时性上更具优势。同时,在静态标签的平均正确率为80.42%,达到实际应用场景所需的识别精度,具有一定的实际应用价值。

表6　各方法在Drive&Act数据集上的对比实验结果

模型分类	模型	*Accuracy*/%
纯骨架模型	Pose	44.36
	Two-stream	45.39
	ST-GCN	45.34
	DMGNN	50.98
纯图像模型	C3D	43.41
	P3D	45.32
	I3D	**63.64**
融合模型	SIBBR-Net	61.78

2.6　消融实验

为了验证CBAM、BEM,以及辅助损失的有效性,分别对上述模块进行消融实验。

本文中共设置了6组消融实验:(a)保留所有模块,其结构如图1所示;(b)去除CBAM;(c)去除BEM;(d)去除ISA,保留SU;(e)去除SU,保留ISA;(f)去除辅助损失(AL),总损失$L=L_{main}$。消融实验设置情况和实验结果如表7所示。

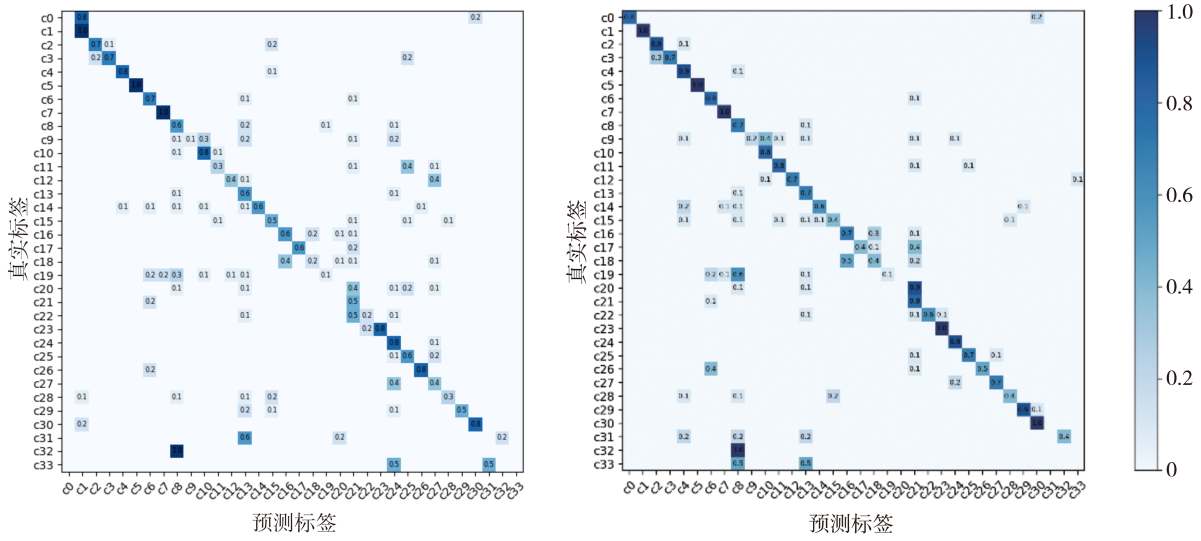
通过对比(a)和(b)消融实验结果可见,引入CBAM后,平均正确率提升了1.63%。由此可得CBAM能够促进SIBBR-Net进一步提取判别性外观特征;通过对比(a)和(c)消融实验结果可见,引入BEM后,平均正确率提升了2.76%。由此可得,BEM的运动和外观信息互相引导更新策略是具有有效性的;通过对比(a)和(f)消融实验结果可见,引入辅助损失后,平均正确率提升了2.3%。由此可得添加辅助损失能确保SIBBR-Net在动静态特征空间分别提取运动和外观特征,有助于提升识别效果。

为了验证运动和外观信息互补关系对识别驾驶员行为的有效性,本文对比并分析骨架分支和整体

表7　消融实验设置及结果

编号	CBAM	ISA	SU	AL	*Accuracy*/%
(a)	√	√	√	√	**61.78**
(b)		√	√	√	60.15
(c)	√			√	59.02
(d)	√		√	√	61.19
(e)	√	√		√	60.22
(f)	√	√	√		59.48

网络分类结果的混淆矩阵,混淆矩阵如图 12 所示。在动态行为标签下,融合模型 SIBBR-Net 的平均正确率为 61.78%,比骨架分支网络提升了 10.8%,可见融合运动和外观信息后,模型的整体识别能力得到大幅提升。通过分析 C12(closing-laptop)、C25(writing)和 C27(reading-magazine)的平均正确率可得,由于骨架分支网络忽略所有外观信息,当运动信息具有相似性时,行为间将存在混淆现象。然而当 SIBBR-Net 引入补偿性的外观特征后,这种混淆情况得到缓解。



(a) 骨架分支网络的混淆矩阵 (b) SIBBR-Net的混淆矩阵

图 12 骨架分支网络的混淆矩阵和 SIBBR-Net 的混淆矩阵

3 结论

在驾驶员行为识别任务中,本文提出基于多尺度骨架图和局部视觉上下文融合的驾驶员行为识别模型 SIBBR-Net,实现对运动信息的多层次性表达,保留所需的外观信息并有效减少模型参数。本模型构建 BEM、AFFM 和静态特征空间的辅助损失,实现图像和骨架序列信息间的高效融合。在 Drive&Act 数据集中,SIBBR-Net 于测试集的 FLOPs 为 25.92G,动态标签的平均正确率为 61.78%,静态标签的平均正确率为 80.42%。在未来的研究中,将继续探索如何将驾驶员的生理信息融合到现有的融合模型中,进一步提高识别驾驶员行为方法的性能。

参考文献

[1] MARBERGER C, MIELENZ H, NAUJOKS F, et al. Understanding and applying the concept of "driver availability" in automated driving[C]. International Conference on Applied Human Factors & Ergonomics. Springer, Cham, 2017.

[2] YANG L, DONG K, DMITRUK A J, et al. A dual-cameras-based driver gaze mapping system with an application on non-driving activities monitoring[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(10): 4318-4327.

[3] ZHENG W, ZHANG Q Q, NI Z H, et al. Distracted driving behavior detection and identification based on improved cornernet-saccade[C].2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). IEEE, 2020: 1150-1155.

[4] JIAO S J, LIU L Y, LIU Q. A hybrid deep learning model for recognizing actions of distracted drivers [J]. Sensors, 2021, 21(21): 7424.

[5] HOLZBOCK A, TSAREGORODTSEV A, DAWOUD Y, et al. A spatio-temporal multilayer perceptron for gesture recognition [C]. 2022 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2022: 1099-1106.

[6] LI P, LU M, ZHANG Z, et al. A novel spatial-temporal graph for skeleton-based driver action recognition[C].2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019: 3243-3248.

[7] XU Q, ZHENG W, SONG Y, et al. Scene image and human skeleton-based dual-stream human action recognition [J]. Pattern Recognition Letters, 2021, 148: 136-145.

(下转第 28 页)

mind? a mental and perceptual load estimation framework towards adaptive in-vehicle interaction while driving[C]. Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. New York, NY, USA: Association for Computing Machinery, 2022: 215-225.

[11] 文晗. 基于情境感知的汽车人机交互界面设计研究[D]. 长沙: 湖南大学, 2016.
WEN H. Study on automotive human machine interface design based on context awareness[D]. Changsha: Hunan University, 2016.

[12] 李坤刚. 基于情境感知的车载HUD安全提示信息界面设计研究[D]. 北京:中国矿业大学, 2020.
LI K G. Research on interface design of vehicle hud safety information based on context awareness[D]. Beijing: China University of Mining and Technology, 2020.

[13] 柳冠中, D M. 事理学论纲——概述[J]. 设计, 2013(9): 114-115.
LIU G Z, D M. An outline of science of affairs[J]. Design, 2013(9): 114-115.

[14] FREES S. Context-driven interaction in immersive virtual environments[J]. Virtual Reality, 2010, 14(4): 277-290.

[15] TAYLOR R M. Situational awareness rating technique (SART): the development of a tool for aircrew systems design[M]//Situational Awareness. Routledge, 2011.

[16] ENDSLEY M R. Toward a theory of situation awareness in dynamic systems[J]. Human Factors, 1995, 37: 32-64.

[17] BROOKE J. SUS: a quick and dirty usability scale[J]. Usability Eval. Ind., 1995, 189.

(上接第8页)

[8] BRUCE X B, LIU Y, ZHANG X, et al. Mmnet: a model-based multimodal network for human action recognition in rgb-d videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3): 3522-3538.

[9] WEYERS P, SCHIEBENER D, KUMMERT A. Action and object interaction recognition for driver activity classification[C]. 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019: 4336-4341.

[10] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]. International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[11] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C].Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1).

[12] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.

[13] LI M, CHEN S, ZHAO Y, et al. Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 214-223.

[14] TAN M, LE Q. Efficientnet: rethinking model scaling for convolutional neural networks[C]. International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[15] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[16] MARTIN M, ROITBERG A, HAURILET M, et al. Drive&act: a multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles[C].Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 2801-2810.

[17] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 5533-5541.

[18] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C].Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

(上接第17页)

[18] 杨龙海, 张春, 仇晓赟, 等. 冰雪条件下中国驾驶员跟驰行为及模型研究[J]. 交通运输系统工程与信息, 2020, 20(6): 145-155.
YANG L H, ZHANG C, QIU X Y, et al. Car-following behavior and model of Chinese drivers under snow and Ice conditions[J]. Journal of Transportation Systems Engineering and Information Technology, 2020, 20(6): 145-155.

[19] 马小龙, 余强, 刘建蓓, 等. 基于无人机视频拍摄的高速公路小型车换道行为特性[J]. 中国公路学报, 2020, 33(6): 95-105.
MA X L, YU Q, LIU J B, et al. Analysis of lane change behavior of passenger cars on the freeway using UAVs[J]. China Journal of Highway and Transport, 2020, 33(6): 95-105.