

# 基于信息瓶颈理论的驾驶员分心行为识别

张 戟, 白亚坤, 韩双庆, 刘家栋

(同济大学 汽车学院, 上海 201804)

**摘要:** 针对驾驶员分心行为识别问题, 将信息瓶颈理论与图卷积网络相结合, 提出一个基于二维姿态估计的动作识别网络, 增加神经网络对有效信息的保留程度, 从而弥补输入信息量的不足。基于通道级拓扑细化图卷积网络, 在有限输入信息下实现了准确的动作识别。

**关键词:** 深度学习; 分心行为识别; 信息瓶颈; 图卷积

中图分类号: TP311; TP391

文献标志码: A

## Driver Distracted Behavior Recognition Based on Information Bottleneck Theory

ZHANG Ji, BAI Yakun, HAN Shuangqing, LIU Jiadong

(School of Automotive Studies, Tongji University, Shanghai 201804, China)

**Abstract:** Aiming at the problem of driver distracted behavior recognition, the information bottleneck theory and the graph convolutional network were combined to realize the action recognition based on the 2D pose estimation, which effectively increases the retention degree of neural network for effective information, so as to make up for the lack of input information. The accurate action recognition was achieved with the limited input information in combination with CTR-GCN.

**Keywords:** deep learning; distracted behavior recognition; information bottleneck; graph convolution

分心驾驶是导致交通事故的主要原因之一。据美国国家公路交通安全管理局<sup>[1]</sup>、世界卫生组织<sup>[2]</sup>及交通部门统计数据显示, 全球每年因危险驾驶行为导致的伤亡事故频发, 因此驾驶员异常行为监测对提升道路安全至关重要。目前, 驾驶员行为识别主要分为基于可穿戴设备和基于视觉 2 种方式。相比于依赖传感器的接触式检测, 基于计算机视觉的方

案具有非侵入性、易部署等优势。

近年来, 基于骨架的动作识别方法因轻量化和抗干扰性而受到广泛关注。Yan 等<sup>[3]</sup>提出的时空图卷积网络开启了图卷积网络在骨架建模中的应用。此后, 动作-结构图卷积网络<sup>[4]</sup>、双流自适应图卷积网络<sup>[5]</sup>及高效图卷积网络<sup>[6]</sup>等方法通过自适应拓扑结构和轻量化设计不断提升性能。Chen 等<sup>[7]</sup>提出的通道级拓扑细化图卷积网络进一步通过通道级拓扑细化捕捉关节间的内在联系, 成为当前的主流方法。Chi 等<sup>[8]</sup>提出的注意力图卷积网络采用基于注意力的图卷积来捕获人类动作的内在拓扑结构。然而, 在驾驶场景下, 受限于车内空间导致的严重遮挡和二维姿态信息的深度缺失, 现有方法仍面临挑战。

本文提出一种基于信息瓶颈 (information bottleneck, IB) 理论与通道级拓扑细化图卷积网络相结合的驾驶员行为识别网络。通过引入信息瓶颈理论, 在训练过程中压缩冗余信息并提取关键特征, 有效提升了基于二维姿态数据的识别准确率。

## 1 信息瓶颈图卷积动作识别网络

基于视觉信号的驾驶员行为识别方法仅依赖于手工提取特征和浅层机器学习模型, 因此在复杂的驾驶环境中显得力不从心<sup>[9]</sup>。结合已有的人体姿态估计研究, 本文通过深度学习方式, 实现对驾驶员的行为识别。

基于骨架的动作识别依托图卷积网络 (GCN) 对人体骨架进行建模, 从而提取出驾驶员的关键姿态特征。然而, 图卷积网络在处理三维 (3D) 姿态数据时效果较好, 但对于仅包含二维 (2D) 信息的姿态数据, 在识别精度上存在不足<sup>[10]</sup>。实际应用中, 获取 3D 姿态数据往往需要较高的成本和复杂的设备支持<sup>[11]</sup>, 而 2D 姿态数据则可以通过普通摄像头捕获, 更适合大规模部署。因此, 如何提高基于 2D 姿态数

收稿日期: 2024-09-11

第一作者: 张 戟, 副教授, 工学博士, 主要研究方向为智能驾驶及汽车电子控制技术。

E-mail: jizhang@tongji.edu.cn



论  
文  
拓  
展  
介  
绍

据的驾驶员行为识别性能,是本文的研究重点。

信息瓶颈理论能够在特征提取过程中有效权衡数据的表示能力和对任务的预测能力<sup>[12]</sup>。该理论在深度学习应用中展现出了提升模型泛化能力的潜力。本文将信息瓶颈理论与图卷积网络相融合,旨在提炼出紧凑且有效的特征表示,以解决基于2D姿态数据的驾驶员行为识别问题。

通道级拓扑细化图卷积网络(channel-wise topology refinement graph convolutional network, CTR-GCN)是从时序人体姿态数据中学习行为特征的网络基础,本文沿用CTR-GCN进行姿态估计。随后,基于信息瓶颈理论,定义了信息瓶颈目标,并重新定义了网络损失函数。最后,在NTU-RGB+D60数据集<sup>[13]</sup>上验证了模型性能,并使用自制的D1-DDB数据集和KIT数据集,实现了对21种驾驶行为的准确识别。

将信息瓶颈理论和图卷积网络相结合,提出基于时序2D姿态估计的动作识别网络,如图1所示。动作识别网络主要分为3个部分:时序姿态编码、采

样以及解码。在时序姿态编码阶段,采用图卷积网络作为基础模型,充分利用其在图结构数据处理方面的强大能力<sup>[14]</sup>。通过图卷积网络,有效学习人体姿态在时间和空间上的复杂交互和依赖关系,从而捕捉到动作的本质特征。在此过程中,不仅关注姿态的空间排列和变化,还重视姿态随时间的演变,确保了动作特征识别的全面性和准确性。在特征采样阶段,基于信息瓶颈理论,利用编码阶段获得的隐藏空间分布特征进行采样。这一步骤的关键在于如何从大量可能的状态中选择最具信息量的表示,以此为基础进行后续动作判断。通过这种方式,动作识别网络不仅能够降低过拟合风险,还能够提高模型对动作细节的敏感度和判别能力。最后,在解码预测阶段,动作识别网络将采样所得的隐藏样本映射回标签分布空间,执行最终的动作类别预测。这一步骤确保了从抽象的隐藏特征空间到具体的动作类别之间的有效转换,使得动作识别网络能够准确识别和分类各种动作。

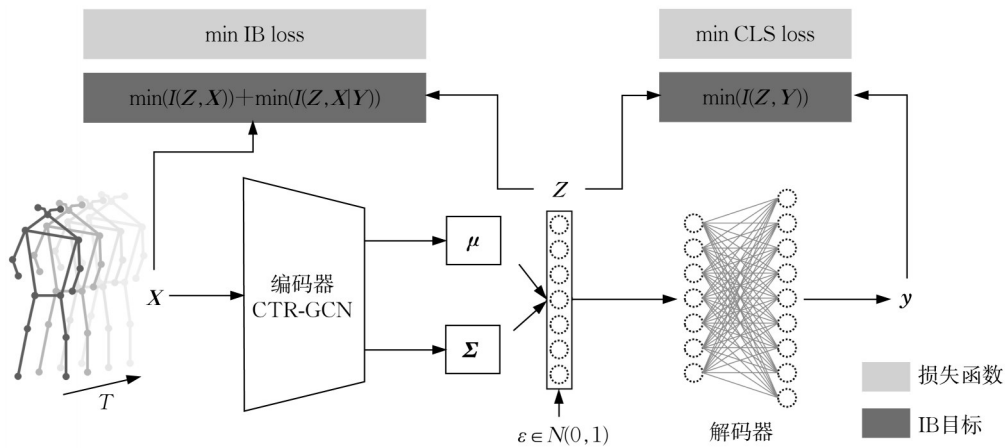


图1 动作识别网络结构

Fig.1 Action recognition network architecture

网络输入  $X \in \mathbf{R}^{T \times N \times C}$  为时序人体姿态。其中,  $T$  表示时序长度,  $N$  表示关节数目,  $C$  表示姿态维度。假设与输入  $X$  对应的隐藏样本  $Z \sim N(z|\mu, \Sigma)$  服从高斯分布 ( $Z \in \mathbf{R}^c$ ),  $Z$  的平均值与对数标准差向量分别为编码阶段的输出结果  $\mu$  和  $\Sigma$ 。图1中,  $\min$  IB loss 表示最小化信息瓶颈损失,用于约束隐藏表示的信息压缩程度,通过最小化互信息  $I(Z, X)$  与  $I(Z, X|Y)$ ,实现对隐藏表示的压缩与类内紧凑;  $\min$  CLS loss 表示最小化分类损失(交叉熵损失),通过分类损失(等价于最大化  $Z$  与标签  $Y$  的互信息),提升动作分类精度。编码阶段的输出结果为

$$p(z|x) = N(z|f_{e,\mu}(x), f_{e,\Sigma}(x)) \quad (1)$$

式中,  $f_e$  表示编码器,也就是 CTR-GCN。

隐藏空间采样过程中,为保证模型在训练过程中的正常反向传播与参数更新,使用自编分编码器(variational autoencoders, VAEs)中的重参数法进行采样<sup>[12]</sup>。首先通过指数运算得到实际标准差,然后引入服从标准正态分布的随机数  $\epsilon \sim N(0, 1)$ ,将其映射到目标分布空间,得到

$$Z = \mu + \epsilon \exp \Sigma \quad (2)$$

在反向传播过程中,随机数  $\epsilon$  仅被当作一个系数看待,保证网络能够正常训练。在解码器中,使用全连接网络,将隐藏样本映射到动作标签空间中,并且使用 softmax 函数进行最后的动作分类,得到

$$\hat{y} = \text{softmax}(\text{Linear}(\text{ReLU}(Z))) \quad (3)$$

### 1.1 CTR-GCN

Chen等<sup>[7]</sup>所提出的CTR-GCN是通过学习的方式共享一个共同的拓扑结构,每个通道动态学习不同的拓扑结构。CTR-GCN能够有效聚合不同通道的关节特征,从而实现基于骨架的行为识别。CTR-GCN在NTU-RGB+D、NTU-RGB+D120和NW-UCLA数据集上均拥有很好的表现。

CTR-GC是一种动态图卷积,根据不同的输入样本自适应地学习卷积拓扑结构。每个CTR-GCN基本模块都由空间特征提取以及时间特征提取两部分构成。在空间特征提取部分,通过3个并列的CTR-GC模块动态学习骨骼拓扑结构,结合残差网络输出空间姿态特征。在时间特征提取部分,通过不同扩张率的时间卷积实现多尺度的时间特征学习。

### 1.2 信息瓶颈目标

本节的目的是基于信息瓶颈定义一个目标,用于从一系列姿态数据中隐藏特征分布学习,并推导变分上界以及可用于模型训练的损失函数。

如图1所示,输入信息 $X$ (时序下的人体2D姿态)后,基于信息瓶颈理论,动作识别网络利用编码阶段获得的隐藏空间分布特征进行采样,生成包含与动作类别有高度相关性信息的随机隐藏变量 $Z$ ,并且滤除输入信息中的干扰部分。这一目标保证了模型的可解释性以及泛化能力。同时,隐藏变量 $Z$ 中需要包含尽可能多的与标签(动作类型) $Y$ 相关的信息,保证模型预测的准确性。该受限优化问题可以通过引入拉格朗日乘子转化为一个无约束优化问题,计算式为

$$O(Z) = I(Z, Y) - \beta_1 I(Z, X) - \beta_2 I(Z, X|Y) \quad (4)$$

式中: $\beta_1, \beta_2$ 为拉格朗日乘子,是用于控制信息压缩程度的因子;互信息 $I(Z, Y)$ 保证了隐藏样本 $Z$ 中拥有足够多的与实际标签 $Y$ 相关的信息,用于准确预测动作类别;互信息 $I(Z, X)$ 保证了从输入 $X$ 到隐藏样本 $Z$ 的信息压缩率<sup>[15]</sup>;互信息 $I(Z, X|Y)$ 表示在给定输入 $X$ 对应标签 $Y$ 的情况下,尽可能地对 $X$ 进行压缩,这一项带限定条件的信息瓶颈目标保证了隐藏样本 $Z$ 中不丢失与标签 $Y$ 相关的重要信息<sup>[16]</sup>。

后续的推导过程均基于输入 $X$ 、隐藏样本 $Z$ 以及标签 $Y$ 符合标准马尔科夫链的关系( $Z \leftrightarrow X \leftrightarrow Y$ )。联合分布符合

$$p(X, Y, Z) = p(Z|X, Y) p(Y|X) p(X) = p(Z|X) p(Y|X) p(X) \quad (5)$$

除了联合数据分布 $p(X, Y)$ 中的结构之外,唯一能够在模型中获取并且定义的是先验分布 $p(Z|X)$ 的模型,所有其他分布都由马尔科夫链约束确定。对式(4)中等号右边的3项分别推导变分界。结合标准马尔科夫链的假设,互信息 $I(Z, Y)$ 的定义为

$$I(Z, Y) = \int dy dz p(y, z) \lg \frac{p(y, z)}{p(y)p(z)} = \int dy dz p(y, z) \lg \frac{p(y|z)}{p(y)} \quad (6)$$

$p(y, z)$ 由解码器以及马尔科夫链完全定义为

$$p(y|z) = \int dx p(x, y|z) = \int dx p(y|x) p(x|z) = \int dx \frac{p(y|x) p(z|x) p(x)}{p(z)} \quad (7)$$

由于 $p(y|z)$ 无法直接获取,因此使用 $q(y|z)$ 作为 $p(y|z)$ 的变分近似。 $q(y|z)$ 对应图1中的全连接层解码器。利用库尔贝-莱布勒(Kullback-Leibler, KL)散度为正的的性质,可以得到 $I(Z, Y)$ 的变分下界为

$$\text{KL}[p, q] \geq 0 \Rightarrow$$

$$\int dy p(y|z) \lg p(y|z) \geq \int dy p(y|z) \lg q(y|z) \quad (8)$$

$$I(Z, Y) \geq \int dy dz p(y, z) \lg \frac{q(y|z)}{p(y)} =$$

$$\int dy dz p(y, z) \lg q(y|z) + H(Y) \quad (9)$$

式中, $H(Y)$ 表示标签 $Y$ 的熵。由于 $H(Y)$ 与图1中的网络结构相互独立,可以忽略,因此式(9)简化为

$$I(Z, Y) \geq \int dx dy dz p(x) p(y|x) p(z|x) \lg q(y|z) \quad (10)$$

推导式(4)等号右边的第2部分 $I(Z, X)$ 。 $I(Z, X)$ 可以扩展为

$$I(Z, X) = \int dz dx p(x, z) \lg p(z|x) - \int dz p(z) \lg p(z) \quad (11)$$

由于隐藏样本 $Z$ 的边缘分布计算复杂,因此使用先验分布 $r(Z)$ 代替 $p(Z)$ 。设定先验分布 $r(Z)$ 为标准正态分布<sup>[15]</sup>。

同理,基于KL散度的性质,可以推导出 $I(Z, Y)$ 的变分上界为

$$I(Z, Y) \leq \int dx dz p(x) p(z|x) \lg \frac{p(z|x)}{r(z)} \quad (12)$$

由式(12)类比出式(4)等号右边的第3部分 $I(Z, X|Y)$ 的变分上界为

$$I(Z, X|Y) \leq \int dx dz p(x) p(z|x) p(y|x) \lg \frac{p(z|x)}{r(z|y)} \quad (13)$$

将式(10)、(12)、(13)代入式(4),得到信息瓶颈目标的下界为

$$O(Z) \geq \int dx dy dz p(x) p(y|x) p(z|x) \lg q(y|z) - \beta_1 \int dx dz p(x) p(z|x) \lg \frac{p(z|x)}{r(z)} - \beta_2 \int dx dz p(x) p(z|x) p(y|x) \lg \frac{p(z|x)}{r(z|y)} = -L \quad (14)$$

式中, $L$ 为动作识别网络的损失函数,是训练过程中梯度下降的对象。

为计算在训练过程中的损失,采用经验数据分布来近似,得到

$$p(x, y) = p(x) p(y|x) \approx \frac{1}{N} \sum_{n=1}^N \delta_{x_n}(x) \delta_{y_n}(y) \quad (15)$$

式中: $N$ 表示实际样本数量; $(x_n, y_n)$ 表示第 $n$ 个观测样本对; $\delta_{x_n}(x)$ 与 $\delta_{y_n}(y)$ 分别表示以 $x_n$ 和 $y_n$ 为中心的狄拉克函数。

将式(15)代入式(14)中信息瓶颈目标变分下界的第1部分,可以得到第1部分损失函数为

$$L_1 = - \int dx dy dz p(x) p(y|x) p(z|x) \lg q(y|z) \approx - \frac{1}{N} \sum_{n=1}^N \int dz p(z|x_n) \lg q(y_n|z) \quad (16)$$

式(16)表示在训练过程中模型预测结果的负对数似然损失。进一步可以得出第2部分损失函数,简化为

$$L_2 = \int dx dz p(x) p(z|x) \lg \frac{p(z|x)}{r(z)} = I(Z, X) + D_{\text{KL}}(p(z)||r(z)) \quad (17)$$

式中, $D_{\text{KL}}$ 表示分布之间的KL散度。损失函数关注的是隐藏样本 $Z$ 中的信息压缩程度,而不是与输入 $X$ 的相关性。因此,对式(17)进行简化,丢弃第1部分 $I(Z, X)$ 。采用最大均值差异(maximum mean discrepancy, MMD)代替式(17)中的KL散度。Rezende等<sup>[17]</sup>在其研究中证明了这种代替方法的可行性。最终,第2部分损失函数可以表示为

$$L_2 = D_M^2[p(z), r(z)] = \left| \int \phi(z) p(z) dz - \int \phi(z) r(z) dz \right|^2 \quad (18)$$

式中: $D_M$ 表示最大均值差异; $\phi(z)$ 表示对于隐藏样本 $Z$ 的映射,此处保持原始值不变。使用相同方法可以推导出第3部分损失函数为

$$L_3 = D_M^2[p(z|x), r(z|x)] = \left| \int \phi(z) p(z|x) dz - \int \phi(z) r(z|x) dz \right|^2 \quad (19)$$

综上,完整的损失函数可表示为

$$L = L_1 + \beta_1 L_2 + \beta_2 L_3 \quad (20)$$

## 2 RGB行为数据集与数据预处理

### 2.1 公开数据集

本文选用NTU-RGB+D数据集<sup>[13]</sup>进行算法基准验证。该数据集包含60类骨骼动作序列,是动作识别领域的主流基准。此外,为验证驾驶场景下的泛化性,引入KIT Drive&Act数据集<sup>[14]</sup>中的RGB数据。该数据集包含12h的多视角驾驶行为视频,涵盖15名驾驶员的多种动作,本文仅使用RGB模态进行训练与测试。

### 2.2 D1-DDB驾驶行为数据集

D1-DDB驾驶行为数据集是自制的RGB数据集,用于分心驾驶行为模型识别与验证。数据集使用720p网络摄像头拍摄,拍摄帧率为15帧·s<sup>-1</sup>,与Drive&Act数据集保持一致。如图2所示,D1-DDB数据集采用双视角拍摄,分别为中控台上方以及A柱顶部,坐标横轴表示时间轴,图中数字表示各段驾驶行为视频样本采集的持续时长,纵轴表示不同的驾驶状态。

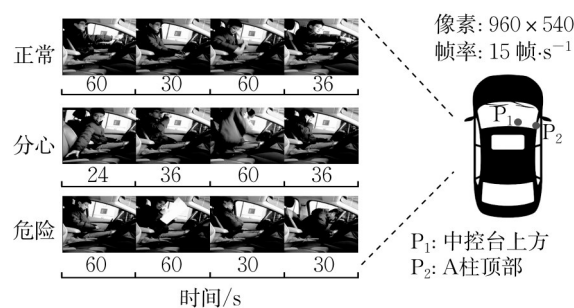


图2 D1-DDB驾驶行为数据集

Fig.2 D1-DDB driving behavior dataset

数据集由5名志愿者共同拍摄完成,每人拍摄5段视频,总拍摄时长约为12h。本文定义了3种驾驶行为状态,分别为:正常驾驶(安全状态)、分心驾驶(警告状态)以及危险驾驶(报警状态)。根据这3种状态定义了21种驾驶行为,分别为:解安全带、系安全带、使用多媒体、正常驾驶、取东西、放东西、使用电脑、玩手机、打电话、穿外套、脱外套、打开水瓶、喝水、关水瓶、四处张望、吃东西、读报纸、阅读杂志、写

字、睡觉以及晕厥。拍摄时间包括白天和夜间。拍摄载具为福克斯2012款以及大众高尔夫2023款。按照5:1的比例划分训练集与数据集。

### 2.3 数据预处理

本文网络输入为2D姿态序列,然而KIT Drive&Act数据集与D1-DDB驾驶行为数据集均为RGB数据集,且KIT Drive&Act数据集仅进行3D姿态标注,因此动作识别网络训练与验证之前需要进行数据提取。首先,使用视频标注数据对原始视频进行分割,按照动作类别进行归类;然后,使用多尺度特征融合姿态估计模型逐帧对视频数据进行2D姿态估计,设置置信度阈值为0.6,即丢弃置信度低于阈值的姿态估计结果。

为保证模型收敛,需要对输入姿态序列进行归一化,如图3所示。图3中, $T$ 表示时序长度,即输入视频片段所包含的帧数。

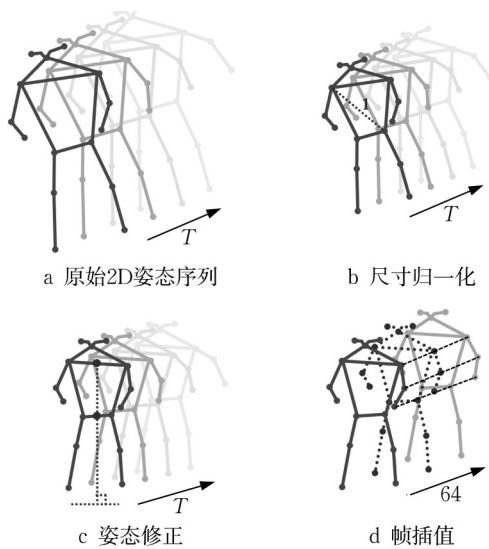


图3 数据预处理

Fig.3 Data preprocessing

每个视频拍摄者存在体型差异,直接导致姿态估计尺度不一致。如图3b所示,首先在空间维度上进行姿态尺寸归一化。选取每段视频中存在有效姿态估计的第1帧作为参考帧,连接参考帧中人物右肩膀到左胯的像素坐标。将连线的中点与视频图像中心对齐,并归一化至单位长度,随后计算转移矩阵。采用转移矩阵对后续各帧中的姿态估计结果进行平移与缩放。如图3c所示,连接第1帧姿态估计中左右肩膀中点与左右胯部中点,旋转连线与地面垂直,并计算旋转矩阵。采用旋转矩阵对各帧中的姿态估计结果进行旋转。网络输入帧数为64帧,因此需要对姿态估计序列进行插值,如图3d所示。在

需要插值的时序点,连接连续的2帧中对应关键点,并采用线性插值方式产生插值姿态。

## 3 评价指标与实验设置

### 3.1 评价指标

本文使用的评价指标除了模型在训练集与数据集上的损失以外,还有预测结果的准确率与召回率。

### 3.2 实验设置

首先使用NTU-RGB+D数据集集中的2D姿态序列与3D姿态序列进行动作识别网络训练与算法验证。为保证拥有足够的数据量,避免模型过拟合,同时提高模型的泛化性,使用KIT Drive&Act数据集与D1-DDB驾驶行为数据集一起训练神经网络,实现驾驶员分心行为检测任务。

使用Pytorch建立上述基于图卷积与信息瓶颈理论的动作识别网络模型,并且使用随机梯度下降优化器最小化损失函数,动量系数为0.9,其余超参数见表1。

表1 动作识别网络超参数

Tab.1 Action recognition network hyperparameters

超参数	数值
时序长度	64
初始学习率	0.05
训练里程碑	90
学习率降低率	0.1
丢弃率	0.3
批次大小	128

在训练过程中,使用每个训练批次的隐藏样本均值来近似计算损失函数 $L_2$ 中的 $\int \phi(z)r(z)dz$ ;使用条件隐藏样本均值来近似计算损失函数 $L_3$ 中的 $\int \phi(z)r(z|x)dz$ 。不同拉格朗日乘子 $\beta_1, \beta_2$ 下在NTU-RGB+D 2D姿态数据集上的训练结果见图4。

图4中,虚线表示当 $\beta_1$ 取0.10时,不同 $\beta_2$ 取值对

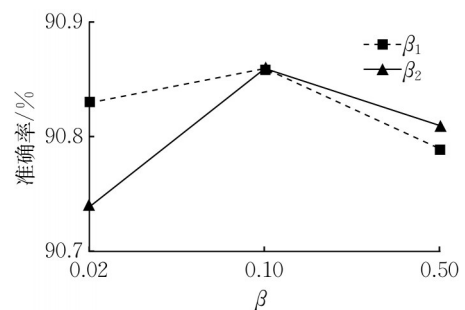


图4  $\beta_1$ 与 $\beta_2$ 对模型训练结果的影响

Fig.4 Effect of  $\beta_1$  and  $\beta_2$  on model training results

于动作识别准确率的影响;实线表示当 $\beta_2$ 取0.10时,不同 $\beta_1$ 取值对于动作识别准确率的影响。可以看到,当两者均取值0.10时,模型预测准确率达到最高值,为90.86%。在后续模型训练中,均设 $\beta_1$ 与 $\beta_2$ 为0.10。此外,改变 $\beta_2$ 时预测结果准确率的影响大于改变 $\beta_1$ 所导致的准确率波动,这说明损失函数第3部分 $L_3$ 相较于 $L_2$ 对训练结果有更大影响。在训练过程中,将 $\beta_1$ 与 $\beta_2$ 均设置为零时,损失函数仅保留第1部分,此时损失函数退化为负对数似然损失,模型退化为一类的多分类任务。

为了定量评估信息瓶颈目标在提取与输出标签高度相关信息中的效果,在不同数据集上比较了引入信息瓶颈目标与未引入信息瓶颈目标的模型表现差异。

## 4 实验结果与讨论

### 4.1 NTU-RGB+D数据集不同姿态维度输入比较

使用NTU-RGB+D60经过预处理后的姿态序列进行训练。根据不同的验证集划分方式、输入姿态维度、是否使用信息瓶颈目标分别对比网络在验证集上的准确率,Top-1准确率指模型预测概率最高的类别与真实类别完全一致的比例,Top-5准确率指模型预测概率最高的前5个类别中真实类别的比例,结果见表2。

表2 不同姿态维度输入实验结果

Tab.2 Input experimental results for different pose dimensions

分类方法	姿态维度	是否引入信息瓶颈目标	准确率/%	
			Top-1	Top-5
X-Sub	2D	是	90.86	98.13
		否	90.28	98.23
	3D	是	93.31	98.40
		否	93.06	98.59
X-View	2D	是	93.59	98.98
		否	93.37	98.64
	3D	是	96.53	99.86
		否	96.68	99.92

表2显示,在相同的姿态维度下,跨视角(X-View)的识别准确率普遍优于跨主体(X-Sub),这主要归因于模型对视角变化的泛化能力强于对个体动作差异的适应性。此外,由于缺乏深度信息,因此2D姿态输入的准确率略低于3D输入。在引入信息瓶颈目标后,2D输入的Top-1准确率在不同划分下均有提升(0.2%~0.6%)。

图5的混淆矩阵中最深的颜色表示概率 $\geq 5\%$ 。

从图5可以看出,2种训练方式获得的混淆矩阵整体分布相似,但是在部分非对角线元素中,使用信息瓶颈目标的训练方式拥有更淡的颜色,即更低的错误预测概率,这表明信息瓶颈目标会使得部分类别的预测结果更加准确。在整体上,不同动作类别的预测结果分布保持了相似性。

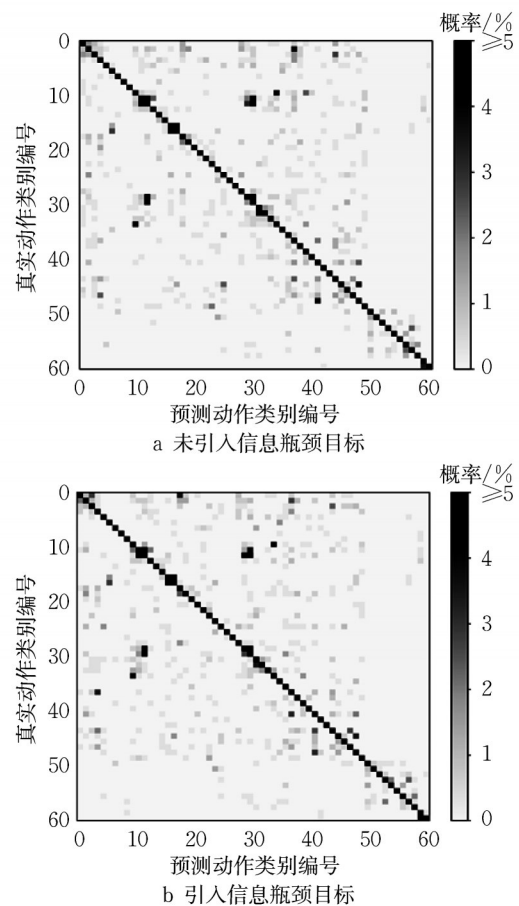


图5 NTU-RGB+D数据集动作分类混淆矩阵

Fig.5 NTU-RGB+D dataset action classification confusion matrix

### 4.2 NTU-RGB+D数据集不同信息瓶颈目标比较

为了探究信息瓶颈目标引入对于模型的影响,随机选取8种动作类别,对隐藏样本进行主成分分析(principal component analysis, PCA)<sup>[18]</sup>。在NTU-RGB+D数据集上使用X-Sub方法进行分类验证,输入姿态维度均为2D,唯一区别在于是否使用信息瓶颈目标。隐藏层样本维度为256,使用PCA降维后的维度为3,可视化结果见图6、7。

结合三维主视图与投影视图,可以看到使用信息瓶颈目标后,8种动作类别的隐藏层样本在空间中的分布边界清晰,且相互之间没有重叠区域。没有使用信息瓶颈目标时,动作类别7、8拥有明显的分布区域重

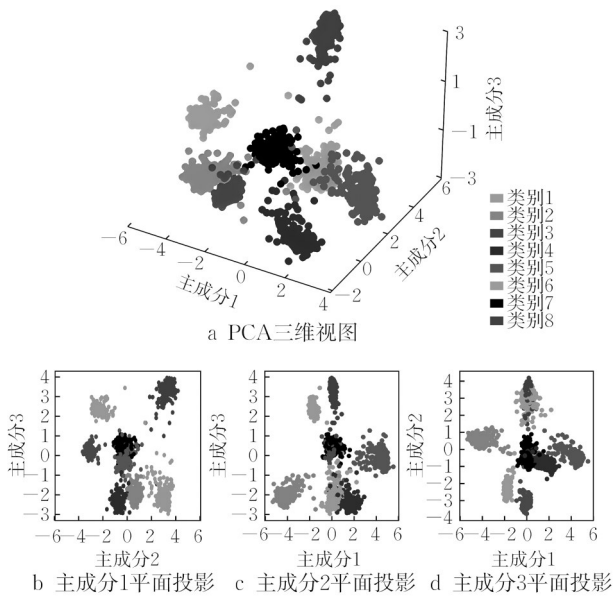


图6 NTU-RGB+D数据集使用信息瓶颈目标时PCA降维结果

Fig.6 PCA dimensionality reduction results on NTU-RGB+D dataset using the IB objective

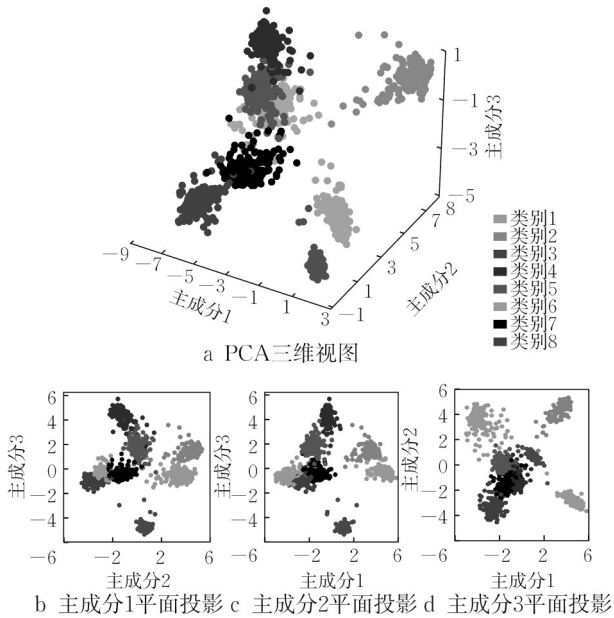


图7 NTU-RGB+D数据集不使用信息瓶颈目标时PCA降维结果

Fig.7 PCA dimensionality reduction results on NTU-RGB+D dataset without using the IB objective

叠,动作类别5、7分布区域也有重叠。解码器无法对重叠的隐藏样本进行区分,最终导致分类错误。在其余动作类别中也有相似的主成分分析结果。

使用NTU-RGB+D60对经过预处理后的2D姿态序列进行训练。根据不同的验证集划分方式与损失函数设定分别对比动作识别网络在验证集上的

准确率,结果见表3。

表3 不同损失函数对预测准确率的影响

Tab.3 Effect of different loss functions on prediction accuracy

分类方法	损失函数组合	准确率/%	
		Top-1	Top-5
X-Sub	$L_1+L_2+L_3$	90.86	98.13
	$L_1+L_2$	90.59	98.09
	$L_1+L_3$	90.39	98.12
	$L_1$	90.28	98.23
X-View	$L_1+L_2+L_3$	93.59	98.98
	$L_1+L_2$	93.61	98.87
	$L_1+L_3$	93.50	99.02
	$L_1$	93.37	98.64

由表3可见,在X-Sub和X-View分类方法下,采用不同的损失函数组合会导致准确率的变化不同,基于信息瓶颈目标引入的损失函数 $L_2$ 与 $L_3$ 均导致模型预测准确率上升。在X-Sub分类方法中,使用 $L_1$ 、 $L_2$ 和 $L_3$ 损失函数的组合可以获得较高的准确率(90.86%),而在X-View分类方法中,损失函数的选择对准确率的影响相对较小。考虑到X-Sub分类方法存在跨个体的数据差异,因此样本空间分布差别更大。这说明,当数据分布差异较大时,引入信息瓶颈目标能够使模型更好地学习到数据之间的特征,从而提升模型的泛化性能。

4.3 驾驶员动作识别结果

使用KIT Drive&Act数据集与D1-DDB驾驶行为数据集(混合数据集)一起训练动作类别网络。以5:1的比例,使用X-Sub方法划分训练集与数据集。模型在验证集上的验证结果见表4,在21种驾驶动作上的预测结果混淆矩阵见图8。

表4 驾驶行为识别训练结果

Tab.4 Driving behavior recognition training results

分类方法	姿态维度	是否引入信息瓶颈目标	准确率/%	
			Top-1	Top-5
X-Sub	2D	是	87.60	97.55
		否	85.27	96.78

引入信息瓶颈目标后,Top-1准确率提升了2.33%,达到87.60%。该提升幅度远高于NTU数据集,表明在数据量较小且噪声较大的驾驶场景中,信息瓶颈对特征筛选的效果更为显著。

值得注意的是,驾驶场景的整体识别率低于NTU数据集。这主要是由于车内环境导致严重的下肢遮挡(如图9所示,仅14个关键点可用,图中frame表示帧),且驾驶动作(如喝水、进食)的空间位移幅度较小,因此特征区分度降低。

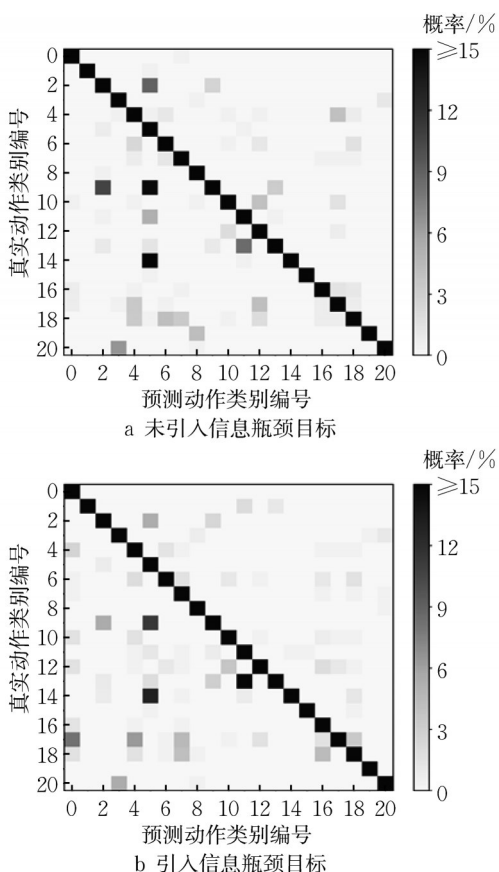


图8 驾驶动作识别验证结果混淆矩阵  
Fig.8 Confusion matrix of driving action recognition validation results

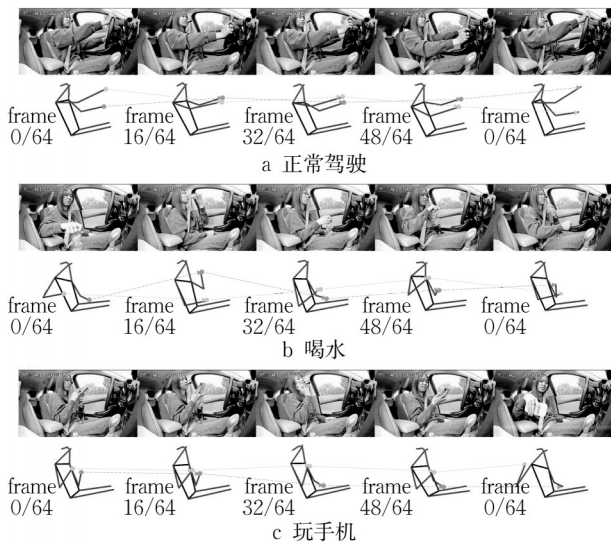


图9 驾驶动作空间分布特征  
Fig.9 Spatial distribution characteristics of driving actions

对如图9所示的2D时序驾驶姿态进行信息压缩并采样,获得对应隐藏样本。对隐藏样本进行PCA降维,如图10、11所示。结果表明,即使在特征

重叠严重的情况下,信息瓶颈目标也能有效分离大部分动作类别,提升了对细微动作差异的洞察力。

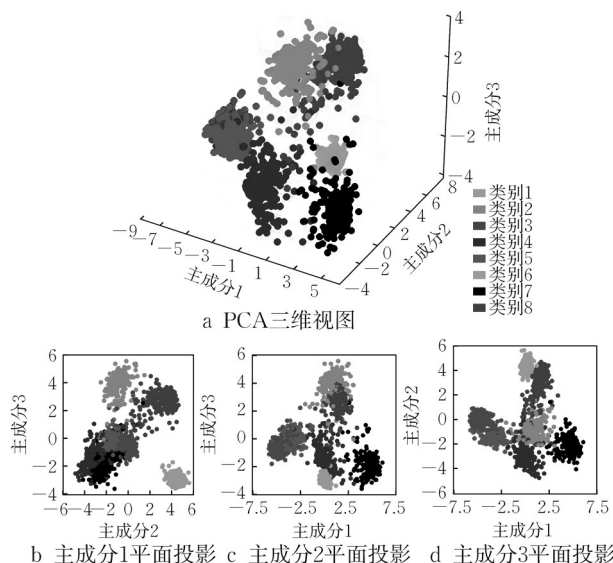


图10 混合数据集使用信息瓶颈目标时PCA降维结果  
Fig.10 PCA dimensionality reduction results on mixed dataset using the IB objective

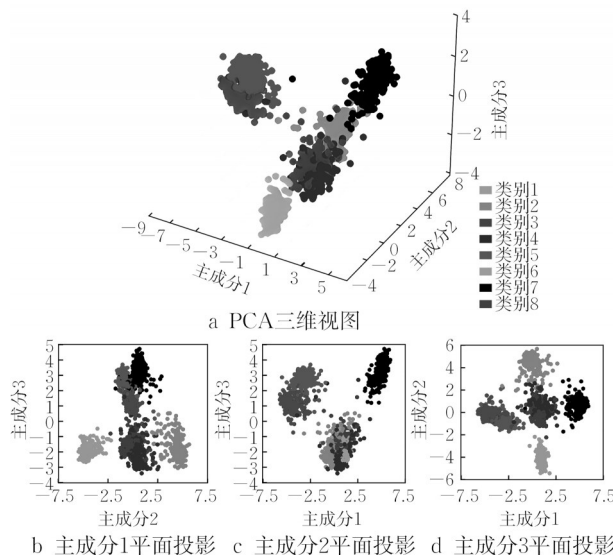


图11 混合数据集不使用信息瓶颈目标时PCA降维结果  
Fig.11 PCA dimensionality reduction results on mixed dataset without using the IB objective

### 5 结语

将信息瓶颈理论与图卷积网络相结合,实现了基于时序2D姿态估计的动作识别任务。相较于3D姿态估计,2D姿态估计丢失了一个维度信息,从而降低动作识别准确率。通过使用信息瓶颈目标,有效提取输入特征中与网络输出高度相关的信息,并

将其压缩至隐藏空间,以 NTU-RGB+D 数据集为例,输入特征维度为  $T \times N \times C = 1\ 920$ ,通过 CTR-GCN 压缩至 256 维隐藏空间,压缩比约为 13.3%,最终提高动作识别精度。基于信息瓶颈目标及其变分上界,重新定义了 CTR-GCN 的损失函数。信息瓶颈目标的引入使得模型在 NTU-RGB+D X-Sub 验证集上的动作识别准确率从 90.28% 提升至 90.86%。通过主成分分析发现,信息瓶颈目标的引入能够有效分离不同动作类别隐藏样本在隐藏空间的分布区域。使用 D1-DDB 驾驶动作数据集与 KIT Drive&Act 数据集一起进行模型训练,本文提出的动作识别网络在 21 种驾驶行为上的识别准确率达到 87.60%。

#### 作者贡献声明:

张 戟:研究思路及学术指导。

白亚坤:理论分析,模型搭建,论文撰写。

韩双庆:协助仿真和实验,论文撰写。

刘家栋:协理论分析,模型设计指导。

#### 参考文献:

- [1] SHULMAN S. Distracted driving event; traffic fatality data release [EB/OL]. (2024-04-01) [2024-04-22]. <https://www.nhtsa.gov/speeches-presentations/distracted-driving-event-put-phone-away-or-pay-campaign>.
- [2] WHO. 道路交通伤害[EB/OL]. (2023-12-13) [2024-04-22]. <https://www.who.int/zh/news-room/fact-sheets/detail/road-traffic-injuries>.
- [3] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018: 7444-7452.
- [4] LI M, CHEN S, CHEN X, *et al.* Actional-structural graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3595-3603.
- [5] SHI L, ZHANG Y, CHENG J, *et al.* Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 12026-12035.
- [6] SONG Y, ZHANG Z, SHAN C, *et al.* Constructing stronger and faster baselines for skeleton-based action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022,45(2): 1474.
- [7] CHEN Y, ZHANG Z, YUAN C, *et al.* Channel-wise topology refinement graph convolution for skeleton-based action recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 13359-13368.
- [8] CHI H, HA M H, CHI S, *et al.* InfoGCN: representation learning for human skeleton-based action recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 20186-20196.
- [9] SHI C, WANG Y, JIA F, *et al.* Fisher vector for scene character recognition: a comprehensive evaluation [J]. Pattern Recognition, 2017,72: 1.
- [10] 任国印,吕晓琪,李宇豪.基于2D转3D骨架的多特征融合实时动作识别[J].激光与光电子学进展,2021,58(24): 241. REN Guoyin, LÜ Xiaoqi, LI Yuhao. Real-time action recognition based on multi-feature fusion of 2D-to-3D skeleton [J]. Laser & Optoelectronics Progress, 2021,58(24): 241.
- [11] SHAN W, LIU Z, ZHANG X, *et al.* Diffusion-based 3D human pose estimation with multi-hypothesis aggregation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 14761-14771.
- [12] DOERSCH C. Tutorial on variational autoencoders[EB/OL]. (2016-06-19) [2024-04-22]. <https://arxiv.org/abs/1606.05908>.
- [13] SHAHROUDY A, LIU J, NG T, *et al.* NTU RGB+D: a large scale dataset for 3D human activity analysis [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1010-1019.
- [14] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [EB/OL]. (2016-09-09) [2024-04-22]. <https://arxiv.org/abs/1609.02907>.
- [15] ALEMI A A, FISCHER I, DILLON J V, *et al.* Deep variational information bottleneck [EB/OL]. (2016-12-01) [2024-04-22]. <https://arxiv.org/abs/1612.00410>.
- [16] FISCHER I. The conditional entropy bottleneck [J]. Entropy, 2020, 22(9): 999.
- [17] REZENDE D, MOHAMED S. Variational inference with normalizing flows [C]//Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR, 2015: 1530-1538.
- [18] ABDI H, WILLIAMS L J. Principal component analysis [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433.