计算机科学
COMPUTER SCIENCE

基于多尺度特征融合的驾驶员注意力分散检测方法

张宇欣，陈益强

引用本文

张宇欣，陈益强. 基于多尺度特征融合的驾驶员注意力分散检测方法[J]. 计算机科学, 2022, 49(11): 170-178.

ZHANG Yu-xin, CHEN Yi-qiang. Driver Distraction Detection Based on Multi-scale Feature Fusion Network[J]. Computer Science, 2022, 49(11): 170-178.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于无监督集群级的科技论文异质图节点表示学习方法

Scientific Paper Heterogeneous Graph Node Representation Learning Method Based onUnsupervised Clustering Level

计算机科学, 2022, 49(9): 64-69. https://doi.org/10.11896/jsjkx.220500196

多源异构环境下的车联网大数据混合属性特征检测方法

Mixed Attribute Feature Detection Method of Internet of Vehicles Big Datain Multi-source Heterogeneous Environment

计算机科学, 2022, 49(8): 108-112. https://doi.org/10.11896/jsjkx.220300273

基于重参数化多尺度融合网络的高效极暗光原始图像降噪

Re-parameterized Multi-scale Fusion Network for Efficient Extreme Low-light Raw Denoising

计算机科学, 2022, 49(8): 120-126. https://doi.org/10.11896/jsjkx.220200179

蒙汉神经机器翻译研究综述

Survey of Mongolian-Chinese Neural Machine Translation

计算机科学, 2022, 49(1): 31-40. https://doi.org/10.11896/jsjkx.210900006

基于多源位置数据的居民出行频繁模式挖掘

Frequent Pattern Mining of Residents' Travel Based on Multi-source Location Data

计算机科学, 2021, 48(7): 155-163. https://doi.org/10.11896/jsjkx.200800072

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000040

基于多尺度特征融合的驾驶员注意力分散检测方法

张宇欣[1,2]　陈益强[2]
1 全球能源互联网发展合作组织　北京 100031
2 中国科学院计算技术研究所　北京 100094
(yuxin-zhang@geidco.org)

摘　要　近年来,道路交通事故的发生逐年增加。驾驶员注意力不集中是造成交通事故的主要原因之一。该项工作利用多源数据来检测驾驶员是否注意力分散。由于每个数据源能为其余数据源提供一定的信息,即多源数据之间的关联性较强,因此对不同来源的数据进行同等处理或对多源特征进行简单的连接整合会导致特征耦合度高,不能保证挖掘任务的有效性。另外,注意力分散驾驶可能受到许多因素的影响,当已知类别的集合中不存在驾驶员注意力分散的类型时,常见的有监督方法可能会导致分类错误。对此,提出了一种基于多尺度特征融合的驾驶员注意力分散检测方法(Multi-Scale Feature Fusion Network,MS-FFN)。首先,通过多个嵌入式子网络从多源数据中学习低维表示。然后,提出一种多尺度特征融合方法,从时空关联性的角度聚合这些特征表示,降低多源特征之间的耦合度。最后,设计基于卷积长短期记忆的编解码模型进行无监督检测。在训练阶段,模型仅对正常驾驶实例进行训练,确定正常数据的一类分类边界。在检测阶段,计算模型重构误差并将其作为每一个测试数据的评分,从而做出细粒度的检测决策。该方法在公开的驾驶员行为数据集上取得了很好的实验结果,优于现有方法。

关键词:驾驶员注意力分散;无监督学习;多源;多尺度融合;编解码器
中图法分类号　TP183;TP391

Driver Distraction Detection Based on Multi-scale Feature Fusion Network

ZHANG Yu-xin[1,2] and CHEN Yi-qiang[2]
1 Global Energy Interconnection Development and Cooperation Organization,Beijing 100031,China
2 Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100094,China

Abstract　The occurrence of road traffic accidents has increased year by year. Driver inattention during driving is one of the major causes of traffic accidents. In this paper,we utilize multi-source data to detect driver distraction. However,the correlations derived from multi-source data will generate feature of high-dimensional entanglement. Existing methods perform similar processing for data of different sources or simply stick to concatenate multi-source features,which are not easy to catch the key feature of high-dimensional entanglement. And distracted driving can be affected by many factors. Supervised methods might cause misclassification when the type of driver distraction does not exist in the set of the known categories. Therefore,we propose a multi-dcale feature fusion network approach to tackle these challenges. Basically,it first learns low-dimensional representations from multi-source data through multiple embedding subnetworks,and then proposes a multi-scale feature Fusion method to aggregate these representations from the perspective of spatial-temporal correlation,thereby reducing the entanglement of feature. Finally,we utilize a ConvLSTM encoder-decoder model to detect driver distraction. Experimental results on a public loaded drive dataset show that the proposed method outperforms the existing methods.

Keywords　Driver distraction,Unsupervised learning,Multi-source,Multi-scale fusion,Encoder-decoder

1　引言

驾驶车辆是一项复杂的任务,驾驶员注意力分散增加了驾驶过程中发生撞车的风险。驾驶注意力分散指驾驶员在驾驶车辆时进行另一项活动而使注意力从驾驶中移开。美国国家公路交通安全管理局(NHTSA)[1]的数据显示,2019 年因注意力分散导致车祸的有 3142 人,2012－2019 年间在涉及注意力分散驾驶的机动车事故中至少丧生 26000 人。

注意力分散驾驶分为 3 种主要类型:视觉、认知和人为型注意力分散。驾驶员注视前方道路以外的事物时,会受到视觉干扰。例如,驾驶员驾驶时观察车外广告牌会在视觉上分散注意力;GPS 和数字娱乐系统之类的汽车电子设备也会

到稿日期:2021-10-08　返修日期:2022-03-15
基金项目:国家重点研发计划(2020YFC2007104)
This work was supported by the National Key R&D Program of China(2020YFC2007104).
通信作者:陈益强(yqchen@ict.ac.cn)

分散驾驶员的注意力。认知型注意力分散指驾驶员的注意力不集中在驾驶上，比如与乘客交谈，陷入沉思或收听广播。音频可以使驾驶员的注意力从驾驶以及整个周围环境中转移。人为型注意力分散指驾驶员出于某种原因将手从方向盘上移开，这种分散注意力的方式包括开车时进食、饮水、吸烟、从钱包或公文包中取东西。当驾驶员发送或阅读短信时，他们的视线会离开路面约 5 s，当车辆以 55 mph 的速度行驶时，车辆 5 s 行驶的距离足以达到足球场的长度（约 105 m）。可以看出，注意力分散的影响因素很多，即未知、新型注意力分散类别普遍存在，使用一般的有监督分类方法难以涵盖所有异常驾驶行为，因此在驾驶员注意力检测场景中研究无监督检测方法具有重大意义。该类方法能够解决注意力分散数据难以标定的问题，即使存在未知、新型的异常驾驶行为也能够被检测到。

检测驾驶员是否注意力分散的方式有很多，大多数研究者使用摄像头来检测驾驶过程中驾驶员注意力分散的情况。例如，通过头部姿势、嘴部动作和眼睛注视方向等信息[2]来估计驾驶员当前时刻的注意力所在区域，也可以将汽车行驶信息[3]作为重要的信息源。由于驾驶员的行为直接影响车辆的行驶数据，因此可以使用控制器局域网（Controller Area Network，CAN bus）进行分析，包括车辆速度、方向盘角度和制动值等车辆行驶信息。此外，有研究者使用麦克风来检测驾驶员的注意力分散程度和疲劳程度[4]，也可通过脑电图、心电图和其他类似的生理传感器来估计驾驶员的生理和情绪状态[5]，从而判断驾驶员当前注意力分散情况。

近年来，随着多传感器采集技术的发展，人们发现单源数据的分类问题只关注于对一种特定数据的分析和处理，相较于单一通道，来自多个源头的数据更接近大数据背景下信息流真实的形态，具有全面性和复杂性。目前，不少研究者利用多源信息，将不同传感器之间的数据进行相互补充，以提升检测系统的性能。即便如此，这类研究仍然面临着挑战。第一个挑战是如何处理多个同类或异类传感器。为了整合视频、音频和生理信号等传感器，文献[6]考虑从多源数据中提取特征，但是由于特征提取和检测算法是分开训练的，因此该类方法在训练过程中很容易陷入局部最优。第二个挑战是如何融合多源数据。有研究者提出了早期融合（特征级融合）[7-8]和后期融合（决策级融合）[9]。早期融合包括简单地在输入级别上连接多源数据的特征，而后期融合则执行决策投票。然而，他们没有考虑多源数据存在关联性而引起的特征耦合度高的问题，这些融合方法很难提取和学习到传感器之间与传感器内部的有效特征。第三个挑战是如何准确找到清晰的决策边界。如上所述，注意力分散驾驶可能受到许多因素的影响，当已知类别的集合中不存在驾驶员注意力分散的类型时，常见的有监督方法[8]可能会导致分类错误。

为了解决这些问题，本文提出了基于多尺度特征融合的驾驶员注意力分散检测方法 MSFFN，主要贡献有 4 个方面。

（1）利用不同的架构如 Bi-LSTM 和 MobileNet 对多种传感器产生的数据分别进行特征提取。其目的是建立端到端模型，共同优化特征提取和检测模型，避免模型陷入局部极小值。

（2）提出一种多尺度特征融合方法来聚合多源数据特征。这种方法可以对传感器数据内部的时序相关性及不同传感器数据之间的空间相关性进行有效融合，降低多源特征之间的耦合度。

（3）引入基于 ConvLSTM 的无监督编解码模型。该方法的优势在于训练过程不受未知或新型注意力分散实例的影响；且由于在人工监督下准确标记注意力分散实例的成本很高，该方法在训练过程中只需要正常驾驶实例。

（4）在模拟驾驶的公开数据集上进行的实验表明，所提出的 MSFFN 具有优于最新技术的性能。为了进一步验证所提模型的效果，进行了消融分析、类别识别率分析、可视化分析以及参数敏感性分析，结果表明 MSFFN 在该数据集上具有较好的性能。

2 相关工作

2.1 面向单源数据的驾驶员注意力分散检测

用于检测驾驶员注意力分散的方法可分为 4 个主要类别：基于视觉信号、基于音频信号、基于车辆行驶信息和基于可穿戴传感器的方法。表 1 列出了面向单源数据的检测方法的优缺点。

表 1 面向单源数据的驾驶员注意力分散检测方法的优缺点

Table 1 Pros and cons of driver distraction detection method based on mono-source data

单源信号	优点	缺点
视觉信号	精度高	易缺失信息，如面部遮挡或光线改变
声音信号	易检测声音活动	难以检测视觉型注意力分散
车辆行驶信息	易采集	用户个体差异大
可穿戴传感器	计算复杂度低	多噪声

2.1.1 视觉信号检测方法

利用面部表情、身体姿势或路况等视觉信息来进行驾驶员注意力分散检测，研究工作主要集中在面部信息的检测，例如头部、眼睛或嘴巴的动作，其中打哈欠、点头、视线估计和眨眼等行为是最常用的特征，在文献[2,10]中得到了广泛使用。但是，面部的某些部分被遮盖或照明的变化对模型的鲁棒性有很大影响，因此人们考虑添加其他设备来弥补缺失的信息。一些研究利用 Kinect 摄像机来检测驾驶员的身体姿势，该设备能够检测到驾驶员的手是否在方向盘上[11]；也有研究者提出了使用后置摄像头[12]检测道路上的运动物体，通过推断运动物体是否位于驾驶员视线区域范围内来量化驾驶员的视觉注意力情况。

2.1.2 声音信号检测方法

利用声学特征来进行驾驶员注意力分散检测，能够检测到环境音和语音活动，例如电话交谈、人与人的对话以及周围环境的声音。文献[4]使用麦克风来记录和分析驾驶员的语音信息，该语音识别系统能够检测驾驶员的疲劳程度。尽管基于声音信号的检测方法有一定的准确性，但是该类方法很难检测到所有类型的驾驶注意力分散，尤其是视觉注意力分散和人为型注意力分散。

2.1.3 车辆行驶信息检测方法

利用车辆行驶信息来进行驾驶员注意力分散检测，这些

信息来自控制器局域网(CAN-Bus),包括车速、方向盘转角位置、油门踏板使用情况和刹车使用情况[3,13]。CAN-Bus是车辆必备的装置,因此车辆行驶记录数据很容易获得,但由于个体差异性大,仅依赖该信息无法准确判断驾驶员注意力分散的情况。

2.1.4 可穿戴传感器检测方法

文献[14]利用生理传感器如脑电信号、肌电信号、心率信号、皮肤电信号等来评估驾驶员的注意力分散情况。此外,文献[15]使用运动传感器如加速度计、陀螺仪来检测驾驶员的注意力分散行为。然而,相对于其他设备的信号来说,这些可穿戴传感器产生的信号含噪声较多,因此单独使用该类信号需要解决噪声干扰的问题。

2.2 面向多源数据的驾驶员注意力分散检测

在驾驶员注意力分散检测中,大多数方法执行两阶段框架:首先针对不同传感器分别提取特征,然后使用多类分类器检测驾驶员是否注意力集中。Du等[8]从3种信号源中提取特征:面部表情、语音信号和汽车行驶信息;然后学习了一个多项式融合网络用于检测驾驶员注意力。Dehzangi等[6]融合了运动传感器信号(加速度计和陀螺仪)、心电图信号、皮肤电信号和控制器局域网的特征,将它们送入集成分类器中识别驾驶员的3种不同状态。Lechner等[16]设计了一个轻量级框架,用于集成和融合多种传感器信号,以检测驾驶员的注意力分散情况。

由于在实际应用场景中很难人为标注所有类型的驾驶员行为,尤其是发生注意力分散的情况,因此使用无监督学习方法进行驾驶员注意力分散检测的趋势正在增长。其基本思想是仅提供正常驾驶的样本来训练模型,学习正常模式的边界,基于样本在该模型中产生的偏差来检测驾驶员是否注意力分散。Tanprasert等[17]结合了半监督异常检测和神经网络,将提取到的加速计和GPS传感器的特征作为输入,评估当前驾驶员的身份认证。Zhang等[18]使用无监督的深度异常检测方法,对视频、音频、运动传感器、生理传感器等多种信号进行端到端的模型训练,以此实现驾驶员的注意力分散检测。

3 MSFFN

如图1所示,MSFFN方法由3个主要部分组成:1)多源数据表示学习将多源时序数据(如生理传感器信号、可见光视觉信号以及红外热成像信号等)作为输入,使用嵌入式子网络得到不同的特征输出;2)将多种特征输入多尺度特征融合层中进行多源特征的统一表示;3)无监督检测模块将多尺度矩阵作为输入,通过基于ConvLSTM的编解码模型得到重构误差。在训练阶段,对提出的方法进行了端到端的训练。在检测阶段,将重构误差作为每个测试数据的分数,从而做出细粒度的检测决策。如果该分数大于某个阈值,则测试数据被分类为"注意力分散驾驶",否则为"正常驾驶"。

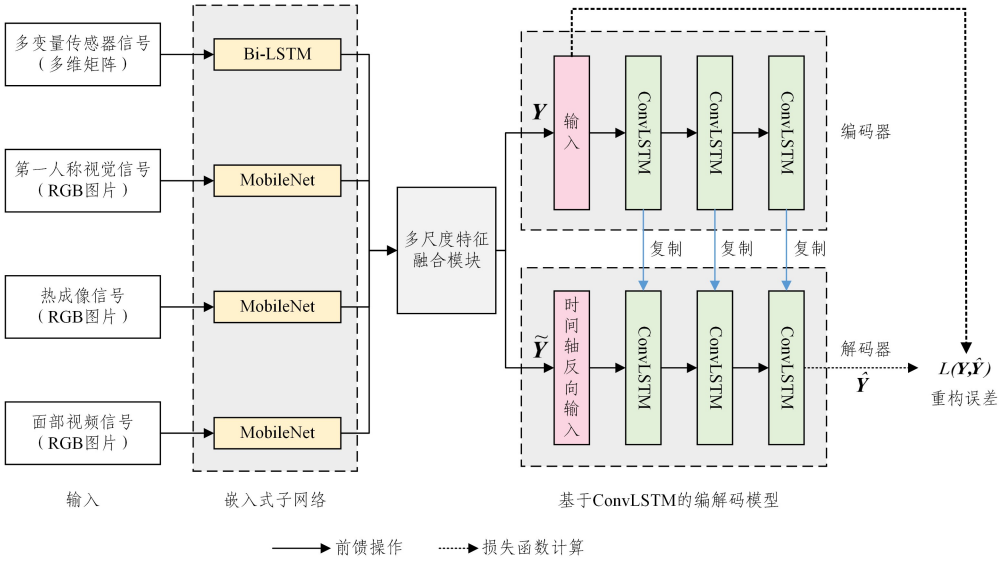

图1 MSFFN架构图

Fig. 1 Architecture diagram of MSFFN

3.1 问题定义

给定一个具有$M(M>1)$个传感器的多源时序数据集,由于每个传感器信号的采样率有可能不同,因此这里将所有传感器信号统一成相同的长度T。本文使用了生理传感器信号、车辆行驶信息、可见光视觉信号以及红外热成像信号,考虑到不同传感器的结构特性不同,将数据分别定义为$\mathcal{X}=\{(\boldsymbol{X}_s,\boldsymbol{X}_{v1},\boldsymbol{X}_{v2},\boldsymbol{X}_{v3})\mid \boldsymbol{X}_s\in\mathbb{R}^{T\times N_1\times M_1},\boldsymbol{X}_{v1}\in\mathbb{R}^{T\times N_2\times M_2\times C},\boldsymbol{X}_{v2}\in\mathbb{R}^{T\times N_3\times M_3\times C},\boldsymbol{X}_{v3}\in\mathbb{R}^{T\times N_4\times M_4\times C}\}$,包括多变量传感器信号$\boldsymbol{X}_s$(生理传感器信号和车辆行驶信息)、3种视频信号$\boldsymbol{X}_{v1},\boldsymbol{X}_{v2},\boldsymbol{X}_{v3}$,其中$N_1$表示多变量传感器信号的时间窗口大小,$M_1$表示传感器的数量,$N_2$,$N_3$,$N_4$和$M_2$,$M_3$,$M_4$分别表示图像的高度和宽度,$C$是通道数(如RGB)。

3.2 数据预处理

本文方法使用了驾驶员行为监测的公开数据集,将数据集中的生理传感器信号、车辆行驶信息、可见光视觉信号以及红外热成像信号等多源时序数据作为模型的输入。预处理过程中将原始数据转换为通用格式,以下小节详细地介绍针对数据集中不同传感器的数据预处理。

3.2.1 多变量传感器信号

将生理传感器(呼吸率、心率、皮肤电信号)和车辆行驶信息(行驶速度、行驶加速度、制动值、方向盘转向角以及车道位置等)统称为多变量传感器信号。我们通过上采样或下采样的方式将所有信号转换为一致的采样频率。上采样采用临近插值法，选择最近点的值，从而产生分段恒定的插值。下采样是插值的逆过程，它以整数或非整数因子增加采样间隔，从而降低采样频率。最终将信号采样率统一为 20 Hz。

3.2.2 可见光视觉信号

将可见光视觉信号分为面部视频信号和第一人称视觉信号两种形式。人脸是视觉信息最重要的来源，面部表情、头部位置和头部旋转可用于评估驾驶员的行为，该面部视频信号以 27 Hz 频率采样，为了降低计算复杂度，我们根据经验将该视觉信号从 27 Hz 下采样到 1 Hz。第一人称视觉信号是从驾驶员视角记录其可见道路范围，在该数据集中还配有眼动仪，能够将驾驶员凝视的方向以绿点跟踪叠加在第一人称视角的视频中，该视频对驾驶员的注意力分散检测起到重要作用。同理，将该信号从原始的 10 Hz 下采样到 1 Hz。

3.2.3 红外热成像信号

红外热成像技术利用热像和温度数据，与可见光形成互补，用于驾驶员行为监测时能够进一步提高识别准确率和安全性。同理，将该信号从原始的 25Hz 下采样到 1Hz。

3.3 多源数据表示学习模块

面向多源数据的检测方法通常是分别进行特征提取和检测，这使得两种基线方法的联合性能很容易陷入局部极小值。因此，本文提出嵌入式子网络作为特征提取模块，完成端到端训练。

3.3.1 基于多变量传感器信号的嵌入式子网络

我们采用滑动窗口策略将时序信号分段为短信号，将包含所有传感器的二维矩阵 $X_s \in \mathbb{R}^{N_1 \times M_1}$ 用作双向长短期记忆网络(Bi-directional Long Short-Term Memory，Bi-LSTM)的输入。Bi-LSTM[19]被广泛用于处理传感器信号，方法是计算由前向和后向输出串联而形成的编码矢量。编码后的向量表示为 $\boldsymbol{Z}_s \in \mathbb{R}^d$，$d$ 表示特征的维度。

3.3.2 基于可见光视觉信号的嵌入式子网络

我们实现 MobileNet[20]，以获取关键的视觉特征。MobileNet 是一种轻量级的深度卷积神经网络，能够显著降低网络的复杂度和模型大小，适用于移动设备或计算能力低的设备。令 $\boldsymbol{X}_{v1} \in \mathbb{R}^{N_2 \times M_2 \times C}$ 表示面部图像输入数据，$\boldsymbol{X}_{v2} \in \mathbb{R}^{N_3 \times M_3 \times C}$ 表示第一人称视觉输入数据，编码后的向量表示为 $\boldsymbol{Z}_{v1} \in \mathbb{R}^d$ 和 $\boldsymbol{Z}_{v2} \in \mathbb{R}^d$。

3.3.3 基于红外热成像信号的嵌入式子网络

由于红外热成像信号也属于视觉信号的一种，我们仍然选择了 MobileNet 以获取感兴趣区域的关键特征。令 $\boldsymbol{X}_{v3} \in \mathbb{R}^{N_4 \times M_4 \times C}$ 表示热成像输入数据，编码后的向量表示为 $\boldsymbol{Z}_{v3} \in \mathbb{R}^d$。

3.4 多尺度特征融合模块

为了将传感器数据内部的时序相关性及不同传感器数据之间的空间相关性进行有效融合，本文构建了基于多尺度融合的表征学习模型，使得多源数据得到统一表示。很多针对多源数据的研究[21-22]都使用了特征级联或特征协调作为多源数据融合的方法。本文工作中介绍了一种多尺度特征融合的方法，利用端到端的策略同时学习传感器内部和传感器之间的特征。该方法能够降低特征耦合度，使模型得到更有效的特征。首先将 4 个嵌入式子网络的输出分成 h 个时间步，$\boldsymbol{Z}_s = \{\boldsymbol{Z}_s^1, \cdots, \boldsymbol{Z}_s^h\} \in \mathbb{R}^{h \times d}$，$\boldsymbol{Z}_{v1} = \{\boldsymbol{Z}_{v1}^1, \cdots, \boldsymbol{Z}_{v1}^h\} \in \mathbb{R}^{h \times d}$，$\boldsymbol{Z}_{v2} = \{\boldsymbol{Z}_{v2}^1, \cdots, \boldsymbol{Z}_{v2}^h\} \in \mathbb{R}^{h \times d}$，$\boldsymbol{Z}_{v3} = \{\boldsymbol{Z}_{v3}^1, \cdots, \boldsymbol{Z}_{v3}^h\} \in \mathbb{R}^{h \times d}$，其特征融合表示被计算为：

$$\begin{aligned}\boldsymbol{Z}^h &= (\boldsymbol{Z}_s^h * \boldsymbol{Z}_s^h, \boldsymbol{Z}_{v1}^h * \boldsymbol{Z}_{v1}^h, \boldsymbol{Z}_{v2}^h * \boldsymbol{Z}_{v2}^h, \boldsymbol{Z}_{v3}^h * \boldsymbol{Z}_{v3}^h, \boldsymbol{Z}_s^h * \boldsymbol{Z}_{v1}^h, \boldsymbol{Z}_s^h \\ &\quad * \boldsymbol{Z}_{v2}^h, \boldsymbol{Z}_s^h * \boldsymbol{Z}_{v3}^h, \boldsymbol{Z}_{v1}^h * \boldsymbol{Z}_{v2}^h, \boldsymbol{Z}_{v1}^h * \boldsymbol{Z}_{v3}^h, \boldsymbol{Z}_{v2}^h * \boldsymbol{Z}_{v3}^h) \\ &= (\boldsymbol{Z}_{ss}^h, \boldsymbol{Z}_{v1v1}^h, \boldsymbol{Z}_{v2v2}^h, \boldsymbol{Z}_{v3v3}^h, \boldsymbol{Z}_{sv1}^h, \boldsymbol{Z}_{sv2}^h, \boldsymbol{Z}_{sv3}^h, \boldsymbol{Z}_{v1v2}^h, \boldsymbol{Z}_{v1v3}^h, \\ &\quad \boldsymbol{Z}_{v2v3}^h) \end{aligned} \tag{1}$$

其中，$\boldsymbol{Z}^h \in \mathbb{R}^{d \times m}$ 是第 h 时间步不同传感器之间所有可能的组合形式，以矩阵表示；m 表示可能组合的数量；$*$ 表示向量之间的元素积。为了表示时序段中的时间信息，我们构造了一个有着 h 时间步的多尺度矩阵 $\boldsymbol{Z} \in \mathbb{R}^{d \times m \times s}$：

$$\boldsymbol{Z} = \left(\boldsymbol{Z}^h, \frac{\sum_{k=\delta}^{h} \boldsymbol{Z}^k}{h-\delta}, \cdots, \frac{\sum_{k=1}^{h} \boldsymbol{Z}^k}{h-1}\right) \tag{2}$$

其中，δ 是缩放因子，s 代表需要的不同尺度的矩阵数量。例如，当 $s=1$ 时，仅有矩阵 $\boldsymbol{Z}=\boldsymbol{Z}^h$ 被计算。当 $s=2$ 时，矩阵 $\boldsymbol{Z}=\left(\boldsymbol{Z}^h, \frac{\sum_{k=1}^{h} \boldsymbol{Z}^k}{h-1}\right)$ 被计算。当 $s=3$ 时，矩阵 $\boldsymbol{Z}=\left(\boldsymbol{Z}^h, \frac{\sum_{k=h/2}^{h} \boldsymbol{Z}^k}{h/2}, \frac{\sum_{k=1}^{h} \boldsymbol{Z}^k}{h-1}\right)$ 被计算。多尺度矩阵 $\boldsymbol{Z}$ 能够捕获来自不同源头的数据之间(或多个传感器之间)的相关性，同时捕获同源数据内部(相同或相似传感器内部)的时间信息。

3.5 无监督检测模块

由于在实际采集环境中很难精确地标记多传感器信号并表示所有类型的注意力分散行为，在本节中，我们使用基于 ConvLSTM 的编解码模型进行驾驶员注意力分散检测。该模块使用的是无监督模型，其优势在于不需要标签信息对模型参数进行调整。

编解码模型通常用于机器翻译或时间序列预测。为了使该模型能应用于驾驶注意力分散检测方法中，本文使用了自动编码器，这是一种特定类型的编解码模型，其输入与输出相同，具有一个中间隐藏层，用于存储输入的潜在表示，可以通过最小化重构误差来训练网络，该误差用于测量原始输入与重构之间的差异。给定 $\boldsymbol{Y} = \{\boldsymbol{Z}^1, \boldsymbol{Z}^2, \cdots, \boldsymbol{Z}^t\} \in \mathbb{R}^{t \times d \times m \times s}$ 作为编解码模型的输入，其中 t 表示时间维度，$\boldsymbol{Z}$ 表示多尺度矩阵的输出，所提出的网络结构使用多个堆叠的 ConvLSTM 层[23]。该网络有两个主要元素：编码器和解码器。

编码器：旨在学习低维表示形式来表征数据的重要特征。编码器按时间顺序接收输入特征 $\boldsymbol{Y}$，受序列到序列模型(Sequence-to-Sequence)[24]的启发，每个 ConvLSTM 层都将前一层的输出转换为 2 个状态向量(细胞状态和隐藏状态)。保留每个 ConvLSTM 层的最后一个时间步的状态，并丢弃编码器的输出，其状态向量将在下一步中用作解码器的“条件”。

解码器：解码器的输出表示为 $\hat{\boldsymbol{Y}}$，这是原始输入 $\boldsymbol{Y}$ 的

重构；解码器的输入 $\widetilde{\boldsymbol{Y}}$ 与原始输入 $\boldsymbol{Y}$ 相同，但时间维度相反，即 $\widetilde{\boldsymbol{Y}}=\{\boldsymbol{Z}^t,\boldsymbol{Z}^{t-1},\cdots,\boldsymbol{Z}^1\}$。每个解码器的 ConvLSTM 层使用编码器中相应层提供的编码表示进行初始化，如图 1 所示，解码器将编码器的状态向量复制为初始状态。

3.6 模型训练与检测

通过最小化重构误差 $\mathcal{L}(\boldsymbol{Y},\hat{\boldsymbol{Y}})$ 训练模型，使用的误差是均方误差(Mean-Square Error，MSE)，用于测量重构输入 $\hat{\boldsymbol{Y}}$ 与原始输入 $\boldsymbol{Y}$ 的近似程度，如式(3)所示。

$$\mathcal{L}(\boldsymbol{Y},\hat{\boldsymbol{Y}})=\frac{1}{L}\sum_{i=1}^{L}(\boldsymbol{Y}_i-\hat{\boldsymbol{Y}}_i)^2 \tag{3}$$

其中，$(\boldsymbol{Y}_i-\hat{\boldsymbol{Y}}_i)$ 被定义为残差，L 代表样本个数。

自动编码器通常假设样本的类别不同，其压缩得到的特征也有一定差异，即训练数据仅包含正常样本，则将注意力分散样本输入模型时，重构误差会更高。因此，根据检测阶段的重构误差将样本分类为“正常驾驶”或“注意力分散”。

给定正常驾驶实例作为训练集 $\mathcal{Y}=\{\boldsymbol{X}^1,\boldsymbol{X}^2,\cdots,\boldsymbol{X}^L\mid X^i=(\boldsymbol{X}_s^i,\boldsymbol{X}_{v2}^i,\boldsymbol{X}_{v1}^i,\boldsymbol{X}_{v3}^i)\}$，其中 $\boldsymbol{X}_s^i\in\mathbb{R}^{T\times N_1\times M_1}$，$\boldsymbol{X}_{v1}^i\in\mathbb{R}^{T\times N_2\times M_2\times C}$，$\boldsymbol{X}_{v2}^i\in\mathbb{R}^{T\times N_3\times M_3\times C}$，$\boldsymbol{X}_{v3}^i\in\mathbb{R}^{T\times N_4\times M_4\times C}$，$i\in L$，$L$ 代表样本数量，T 代表信号的时间长度。决策标准为：

$$D=\frac{1}{L}\sum_{i=1}^{L}[Err(\boldsymbol{X}^i)+Z\cdot\sqrt{(Err(\boldsymbol{X}^i)-\mu)^2}] \tag{4}$$

其中，$Err(\boldsymbol{X}^i)$ 表示 $\boldsymbol{X}^i$ 的误差值，μ 表示所有 $Err(\boldsymbol{X}^i)$ 的均值。根据参数敏感性分析结果，当 $Z=1.645$ 时，模型精度达到最高。通过计算标准正态分布的百分位数来确定阈值，标准公式为 $X=\mu+Z\cdot\theta$，μ 是均值，θ 是变量 X 的标准偏差，当 $Z=1.645$ 时，等同于计算正态分布的第 95 百分位数。模型训练过程见算法 1。

在检测阶段，如果 $Err(\boldsymbol{X}^i)$ 大于 D，则该测试样本 $\boldsymbol{X}^i$ 被标定为注意力分散，反之为正常驾驶。

算法 1 MSFFN 的训练过程

输入：正常数据集 $\mathcal{Y}=\{(\mathbf{X}_s,\mathbf{X}_{v1},\mathbf{X}_{v2},\mathbf{X}_{v3})\mid\mathbf{X}_s\in\mathbb{R}^{T\times N_1\times M_1},\mathbf{X}_{v1}\in\mathbb{R}^{T\times N_2\times M_2\times C},\mathbf{X}_{v2}\in\mathbb{R}^{T\times N_3\times M_3\times C},\mathbf{X}_{v3}\in\mathbb{R}^{T\times N_4\times M_4\times C}\}$，包括多变量传感器信号 $\mathbf{X}_s$，3 种视频信号 $\mathbf{X}_{v1}$，$\mathbf{X}_{v2}$ 和 $\mathbf{X}_{v3}$，时间步长 h 和其他超参数 s，t，d 和 m

输出：决策阈值 D 和模型参数 θ

1. 将每个样本沿时间轴分割成 $\mathbf{X}_s\in\mathbb{R}^{t\times h\times N_1\times M_1}$，$\mathbf{X}_{v1}\in\mathbb{R}^{t\times h\times N_2\times M_2\times C}$，$\mathbf{X}_{v1}\in\mathbb{R}^{t\times h\times N_3\times M_3\times C}$，$\mathbf{X}_{v3}\in\mathbb{R}^{t\times h\times N_4\times M_4\times C}$，其中 $T=t\times h$；
2. 随机初始化模型参数 θ；
3. while 没有达到停止准则 do
4. 计算每一个样本中的低维特征 $\mathbf{Z}=\{(\mathbf{X}_s,\mathbf{X}_{v1},\mathbf{X}_{v2},\mathbf{X}_{v3})\mid\mathbf{Z}_s\in\mathbb{R}^{t\times h\times d},\mathbf{Z}_{v1}\in\mathbb{R}^{t\times h\times d},\mathbf{Z}_{v2}\in\mathbb{R}^{t\times h\times d},\mathbf{Z}_{v3}\in\mathbb{R}^{t\times h\times d}\}$；
5. 生成多尺度特征 $\mathbf{Y}\in\mathbb{R}^{t\times d\times m\times s}$；
6. 计算 $\hat{\mathbf{Y}}$ 和 $\mathbf{Y}$ 之间的重构误差；
7. 更新参数 θ；
8. end while
9. 使用训练集(仅包含正常数据)计算 D；
10. return 优化后的 θ 和 D.

4 实验验证

本文通过一个驾驶模拟的公开数据集来验证 MSFFN 方法的有效性。

4.1 数据集

Taamneh 等[25]于 2017 年发布了一个与驾驶行为相关的多源时序数据集，它是在驾驶模拟器上进行的受控实验。该实验涉及了 68 名志愿者，他们在 4 种场景下进行驾驶模拟，场景分为无干扰、认知干扰、情绪干扰和感知干扰。图 2 给出了驾驶模拟器的设置以及驾驶过程中记录的多源信号。

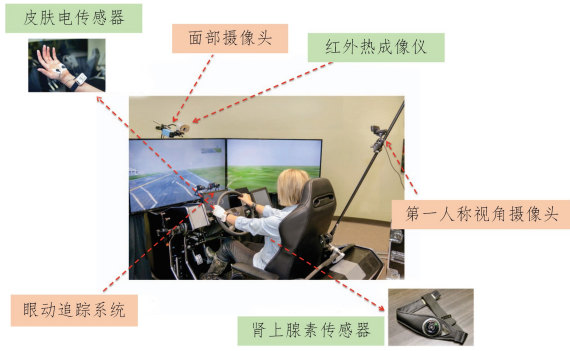

图 2 模拟驾驶采集系统

Fig. 2 Collection system of driving simulation

(1)红外热成像仪：使用 Tau 640 长波红外(LWIR)相机，热敏度小于 50 毫开尔文，空间分辨率为 640×480 像素，以 25fps 的帧频收集热成像数据。热成像信号提取了鼻周出汗区域信息，被称为鼻周皮肤电信号。

(2)面部摄像头：使用 FireWire CCD 单色变焦相机，空间分辨率为 640×480 像素，以 27fps 的帧频收集面部视频数据。该摄像头放置于热成像仪旁边，便于空间位置的精确配准。摄像头距测试者约 1.2 m，保证面部被全部覆盖。

(3)第一人称视角摄像头：使用 HD Pro 网络摄像头 C920，空间分辨率为 960×540 像素，以 10 fps 的帧率收集驾驶画面数据。摄像头位于测试者后方，对准驾驶模拟器的中央屏幕，记录测试者视角的驾驶画面。此外，还使用了 faceLAB 进行眼动追踪，该系统安装在仪表盘上方，用红外光照亮测试者的面部。faceLAB 的软件可显示测试者的凝视点，并投射到模拟器屏幕上形成一个移动的绿点，因此在第一人称视角摄像头录制的过程中也能采集到眼动信息。

(4)皮肤电传感器：使用 Shimmer3 GSR 传感器，采集测试者手心的皮肤电信息。该传感器的测量范围为 104 700 千欧，采集过程中通过蓝牙将数据无线传输到主机。

(5)肾上腺素传感器：使用 Zephyr BioHarness 3.0 传感器，检测测试者的心率和呼吸率。该传感器戴在测试者衣服内的胸带上，心率检测范围 25240 次/分，呼吸率检测范围 4-70 次/分。

(6)车辆行驶记录：通过修改驾驶模拟器的程序，保存了不断变化的驾驶参数，包括行驶速度、行驶加速度、制动值、方向盘转向角以及车道位置。

实验设计将分为以下 6 种场景，其中 PD，RD 和 LD 为正常驾驶，CD，ED 和 MD 为注意力分散驾驶。

(1)练习驾驶(PD)：测试者以规定的速度在四车道高速公路的直线段上驾驶，熟悉驾驶模拟器，速度每两千米改变

一次（80 km/h→50 km/h→100 km/h）。

（2）放松驾驶（RD）：测试者以 70 km/h 的速度在四车道高速公路上直行，反方向的车道上每千米约 3 辆车，测试者行驶 5.2 km 后须改变车道直到行驶 1.2 km 后才可以返回原车道。

（3）无负荷驾驶（LD）：驾驶员正常驾驶，车速为 70 km/h，测试者可以选择变道行驶。

（4）认知负荷驾驶（CD）：测试者在驾驶过程中被问数学问题和逻辑分析问题，测试者需要口头回答，并尽可能准确地回答问题。

（5）情绪负荷驾驶（ED）：测试者在驾驶过程中被问两组问题，一组是具有较少针对性的问题，一组是具有较多针对性的问题，测试者需要尽可能准确地回答问题。

（6）感知负荷驾驶（MD）：测试者在驾驶过程中收到多条短信，要求测试者通过手机进行交流。

4.2 实现细节

由于数据集中包含缺失数据，因此最终选择了 19 名测试者的数据，女士 12 名，男士 7 名，包含年轻人 13 位，老人 6 位。经过预处理后，得到的正常驾驶数据 128 600 条，注意力分散驾驶数据 150 640 条，将数据每 20 s 分割成一个样本，滑动窗口为 5 s，最终得到的样本维度为 $\boldsymbol{X}_s \in \mathbb{R}^{20\times20\times11}$，$\boldsymbol{X}_{v1} \in \mathbb{R}^{20\times128\times96\times3}$，$\boldsymbol{X}_{v2} \in \mathbb{R}^{20\times128\times96\times3}$，$\boldsymbol{X}_{v3} \in \mathbb{R}^{20\times128\times72\times3}$。模型使用基于 Tensorflow 的 Keras[26] 开发，在 Ubuntu 64 位上将 Intel © Xeon© CPU E5-2637 v4 3.50 GHz 和 2 个 NVIDIA GTX 1080Ti 图形卡用作实验环境。

4.3 实验结果

实验的评估指标为精确率（Pre）、召回率（Rec）、F1 分数（F1）和准确率（Acc），mPre 表示不同类别的平均精确率，mRec 和 mF1 同理。如表 2 所列，本文所提方法的平均准确率达到 96.20%；且在实际的应用场景中，该方法的计算时间在可接受的范围内。

表 2 MSFFN 实验结果

Table 2 Experimental results of MSFFN

类别	*Pre*/%	*Rec*/%	*F1*/%	*Acc*/%	测试时间/ms
正常	95.62	94.98	95.30	96.20	2.3
注意力分散	96.59	97.03	96.81		2.4

4.4 有效性验证

4.4.1 消融实验

通过消融实验来观察 MSFFN 的效果。为保证相同的参数和网络结构，具有单一信号源的 MSFFN 取消了多尺度特征融合，并保留了一个嵌入式子网络。其中多变量传感器信号指生理传感器信号和车辆行驶信息，由于我们将其合并为多维矩阵，因此在这里并没有将它们分成两种信息源进行分析。以此类推，具有两种信号源的 MSFFN 保留了两个嵌入式子网络，具有 3 种信号源的 MSFFN 保留了 3 个嵌入式子网络。从表 3 可以看出，与单一、两种和 3 种信号源相比，具有 4 种信号源的 MSFFN 达到了最佳性能。我们还可以观察到，视觉信号如热成像信号、第一人称视觉信号、面部视频信号在获得较高 F1 分数上起着至关重要的作用，多变量传感器信号可以作为辅助信息，从不同程度进一步提升模型性能。

表 3 多源数据下 MSFFN 的实验评估

Table 3 Experimental evaluation of MSFFN from multi-source data

方法	*mPre*/%	*mRec*/%	*mF1*/%	*Acc*/%	测试时间/ms
S	60.33	60.09	60.21	61.99	0.7
V1	66.55	66.41	66.48	63.85	0.8
V2	85.98	87.06	86.52	86.51	1.0
V3	77.63	78.16	77.90	75.94	0.9
S+V1	80.02	81.11	80.56	79.82	0.9
S+V2	82.65	83.63	83.14	81.68	1.0
S+V3	78.40	78.96	78.68	79.14	0.9
V1+V2	89.07	89.67	89.37	89.62	1.5
V1+V3	89.95	90.36	90.16	90.44	1.5
V2+V3	84.20	85.24	84.72	84.73	1.4
S+V1+V2	90.18	90.92	90.55	90.72	1.7
S+V2+V3	85.27	86.49	85.87	85.64	1.4
S+V1+V3	92.34	92.12	92.52	92.52	1.5
V1+V2+V3	92.92	93.46	93.19	93.36	2.1
S+V1+V2+V3	**96.10**	**96.01**	**96.05**	**96.20**	2.4

注：S 为多变量传感器信号，V1 为面部视频信号，V2 为热成像信号，V3 为第一人称视觉信号

4.4.2 可视化分析

为了验证所提模型得到的多尺度融合特征是否有效，我们将训练好的模型在测试数据集上进行验证，得到的特征利用 t-SNE 降维方法[27] 降到二维空间中。如图 3 所示，红色代表正常驾驶实例，蓝色代表注意力分散驾驶实例，其中图3(a)代表嵌入式子网络提取的特征图，图 3(b)代表多尺度特征融合后的特征，图 3(c)代表经过 ConvLSTM 编码器得到的特征。从图中可以观察到，在嵌入式子网络对 4 种信号进行特征提取后，得到的特征仍存在关联性强、耦合度高的问题，且正常和注意力分散数据难以区分；进行多尺度特征融合后，得到的特征明显具有可分性，结合 ConvLSTM 编解码器，正常数据与注意力分散数据之间的边界更加清晰。由此可见，MSFFN 能够降低多源特征之间的耦合度，找到最优特征空间。

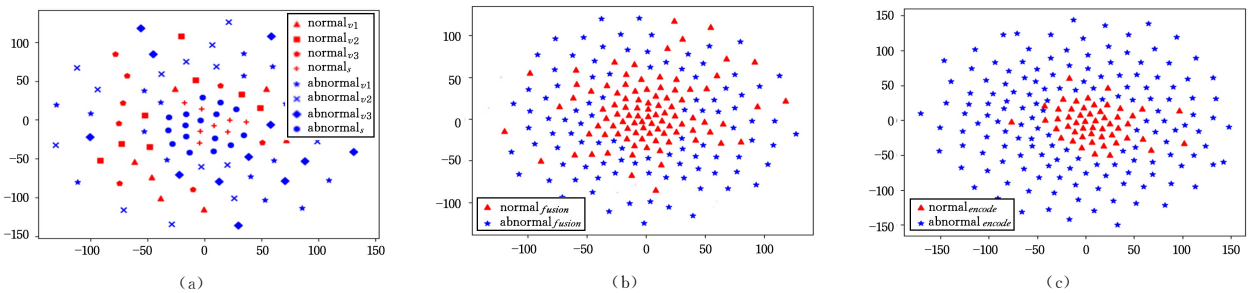

图 3 MSFFN 可视化分析（电子版为彩图）

Fig. 3 Visualization analysis of MSFFN

4.4.3 类别识别率分析

该数据集的注意力分散类别有3类，为了验证MSFFN能否准确划分不同类别的数据，本小节进行了注意力分散驾驶的识别率分析，如图4所示。可以观察到，感知负荷驾驶检测准确率高于认知和情绪负荷驾驶，原因是认知和情绪的干扰是通过问问题让测试者进行思考，但测试者的视线仍大概率直视前方，对驾驶的干扰较小，而感知的干扰是属于外部干扰，需要进行手机短信的交流，因此被检测出是异常驾驶行为的概率较高。

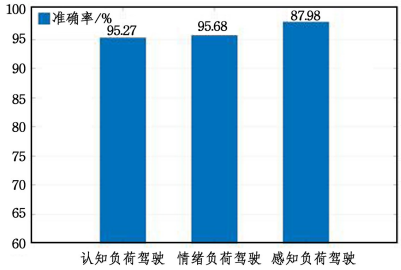

图4 不同类别的注意力分散识别率的比较

Fig. 4 Comparison of driver distraction classification

4.4.4 参数敏感性分析

本节评估了MSFFN的参数敏感性，通过调整超参数MSFFN获得了最佳性能。如表4所列，我们应用控制变量法[28]来评估参数的敏感性，设置了超参数：T是单个样本的时间长度，d表示嵌入式子网络的输出维度，s表示尺度大小。实验结果表明，当$T=20$，$d=100$，$s=3$时，模型可获得最佳性能。

表4 MSFFN参数敏感性分析

Table 4 Parameter sensitivity analysis of MSFFN

MSFFN	*mPre*/%	*mRec*/%	*mF*1/%	*Acc*/%	测试时间/ms
固定 $d=100, T=20$					
$s=1$	79.14	59.56	67.97	67.01	2.4
$s=2$	77.18	59.25	67.04	66.65	2.4
$s=3$	**96.10**	**96.01**	**96.05**	**96.20**	2.4
固定 $d=100, T=20$					
$T=12$	94.46	94.10	94.28	94.49	2.2
$T=20$	**96.10**	**96.01**	**96.05**	**96.20**	2.4
固定 $d=100, T=20$					
$d=50$	92.86	92.73	92.79	93.06	2.3
$d=100$	**96.10**	**96.01**	**96.05**	**96.20**	2.4
$d=200$	95.62	95.91	95.77	95.90	2.4

4.5 对比实验

本节将MSFFN分成3个部分，分别与相应方法进行对比。

4.5.1 多源数据表示学习对比实验

为了验证MSFFN中嵌入式子网络的有效性，我们针对每一种信号源都选择了不同的嵌入式子网络进行比较，对比了现有的表征学习方法，如CNN[29]，LSTM[30]，VGG19[31]以及GoogleNet[32]模型。实验结果如表5(a)所列，在测试时间接近的情况下，使用Bi-LSTM对多变量传感器信号进行特征提取的方法优于现有表征学习方法。从表5(b)－表5(d)中观察到，使用MobileNet对视觉信号进行特征提取的方法花费的时间最少却可获得最佳性能。

表5 MSFFN与其他表征学习方法对比实验

Table 5 Comparison between MSFFN and other representation learning methods

(a)多变量传感器信号表征学习

方法	*mPre*/%	*mRec*/%	*mF*1/%	*Acc*/%	测试时间/ms
CNN	51.26	51.29	51.27	51.09	0.7
LSTM	59.01	59.01	59.01	57.62	0.6
Bi-LSTM	**60.33**	**60.09**	**60.21**	**61.99**	0.7

(b)面部视频信号表征学习

方法	*mPre*/%	*mRec*/%	*mF*1/%	*Acc*/%	测试时间/ms
VGG	51.72	51.55	51.63	48.08	2.0
GoogleNet	52.52	0.80	51.64	42.93	1.2
MobileNet	**66.55**	**66.41**	**66.48**	**63.85**	0.8

(c)热成像信号表征学习

方法	*mPre*/%	*mRec*/%	*mF*1/%	*Acc*/%	测试时间/ms
VGG	57.13	56.99	57.06	54.54	1.7
GoogleNet	58.06	54.23	56.08	47.46	1.2
MobileNet	**85.98**	**87.06**	**86.52**	**86.51**	1.0

(d)第一人称视觉信号表征学习

方法	*mPre*/%	*mRec*/%	*mF*1/%	*Acc*/%	测试时间/ms
VGG	48.93	48.96	48.94	50.66	1.5
GoogleNet	53.49	50.21	51.80	41.05	1.0
MobileNet	**77.63**	**78.16**	**77.90**	**75.94**	0.9

4.5.2 多尺度特征融合对比实验

本节将本文提出的多尺度融合方法与现有融合方法进行了比较，如常用的早期融合方法(Concat)[33]和张量融合方法(Outer)[34]。用这两种融合方法替换了MSFFN的多尺度特征融合模块。实验结果如表6所列。可以看出，我们的多尺度融合方法取得了最佳结果，优于其他融合方法，计算时间与拼接融合方法的时间近似。对于张量融合方法，其性能不佳是由于在多源数据之间使用外积，张量融合的输出是高维的，从而导致模型的拟合不足，因此在我们选择的多源数据集中该方法不能得到很好的训练结果。

表6 MSFFN与其他融合方法的对比

Table 6 Comparison between MSFFN and other fusion methods

方法	*mPre*/%	*mRec*/%	*mF*1/%	*Acc*/%	测试时间/ms
Concat	86.79	87.49	87.14	87.39	2.3
Outer	70.91	70.62	70.76	66.68	2.8
MSFFN	**96.10**	**96.01**	**96.05**	**96.20**	2.4

4.5.3 无监督检测对比实验

为了验证我们的编解码模型的有效性，将其与现有的编解码模型进行对比，如全连接自动编码器(FC-AE)[35]、LSTM自动编码器(LSTM-AE)[36]和卷积自动编码器(CAE)[37-38]。在不改变MSFFN其他模块的情况下，使用这些方法替换MSFFN中基于ConvLSTM的编解码模型。如表7所列，MSFFN的准确率达到了96.20%，超过了其他方法至少10%，这反映出ConvLSTM在处理时空相关性方面优于FC，LSTM和CNN等模型结构。

表 7 MSFFN 与其他编解码模型的对比

Table 7 Comparison between MSFFN and other encoder-decoder methods

方法	*mPre*/%	*mRec*/%	*mF1*/%	*Acc*/%	测试时间/ms
FC-AE	45.66	46.73	46.19	51.76	0.2
LSTM-AE	85.60	87.51	86.55	85.79	1.9
CAE	79.23	79.92	79.57	79.91	1.6
MSFFN	**96.10**	**96.01**	**96.05**	**96.20**	2.4

4.5.4 参数量对比分析

对上述所有的对比方法进行模型参数量的统计分析，结果如表 8 所列。其中，单位 M 代表百万，MB 代表兆字节。可以看出，MSFFN 方法能够在同比参数量的情况下取得最优的结果，这使得其在实际应用中更具灵活性。

表 8 参数量对比分析

Table 8 Comparison of model parameters

方法	*Acc*/%	参数量/M	模型大小/MB
VGG_{V1}	48.08	21.4	81.6
VGG_{V3}	50.66	21.1	80.6
$GoogleNet_{V1}$	42.93	12.5	47.9
$GoogleNet_{V3}$	41.05	8.6	32.7
$MobileNet_{V1}$	63.85	5.4	20.5
$MobileNet_{V3}$	75.94	5.4	20.5
CNN_S	51.09	0.7	2.6
$LSTM_S$	57.62	0.6	2.3
$BiLSTM_S$	61.99	0.6	2.4
Concat	87.39	14.1	53.7
Outer	66.68	181.4	692.1
FC-AE	51.76	13.6	51.9
LSTM-AE	85.79	14.2	54.3
CAE	79.91	13.5	51.7
MSFFN	**96.20**	**14.9**	**57.2**

结束语 本文研究介绍了一种基于多尺度特征融合的驾驶员注意力分散检测方法。它由 3 个主要模块组成：多源数据表示学习、多尺度特征融合和无监督检测模块。首先，构建 4 个嵌入式子网络，从不同的数据源中提取低维特征。为了降低多源特征的耦合度，提出了一种多尺度特征融合方法，从时空关联性的角度聚合这些多源特征表示。最后，使用基于 ConvLSTM 的编解码模型实现无监督检测。由 3 个模块组成的 MSFFN 进行了端到端的训练。为了验证 MSFFN 的有效性，我们在公开的驾驶员行为数据集上进行了实验验证，研究结果表明，MSFFN 的性能优于其他基准模型，在平均检测精度上领先了现有方法至少 10%。

参 考 文 献

[1] ASCONE D, TONJA LINDSEY T, VARGHESE C. An examination of driver distraction as recorded in NHTSA databases [R]. United States. National Highway Traffic Safety Administration, 2009.

[2] ERAQI H M, ABOUELNAGA Y, SAAD M H et al. Driver Distraction Identification with an Ensemble of Convolutional Neural Networks [J]. Journal of Advanced Transportation, 2019, 2019(PT. 1): 1-12.

[3] HANSEN J, BUSSO C, ZHENG Y, et al. Driver Modeling for Detection and Assessment of Driver Distraction: Examples from the UTDrive Test Bed [J]. IEEE Signal Processing Magazine, 2017, 34(4): 130-142.

[4] DHUPATI L S, KAR S, RAJAGURE A, et al. A novel drowsiness detection scheme based on speech analysis with validation using simultaneous EEG recordings [C] // 2010 IEEE International Conference on Automation Science and Engineering. IEEE, 2010: 917-921.

[5] MURPHY-CHUTORIAN E, TRIVEDI M M. Head Pose Estimation and Augmented Reality Tracking: An Integrated System and Evaluation for Monitoring Driver Awareness [J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(2): 300-311.

[6] DEHZANGI O, SAHU V, TAHERISADR M, et al. Multi-modal system to detect on-the-road driver distraction [C] // 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 2191-2196.

[7] HU R, SINGH A. Transformer is all you need: Multimodal multitask learning with a unified transformer [J]. arXiv: 2102.10772, 2021.

[8] DU Y, RAMAN C, BLACK A W, et al. Multimodal polynomial fusion for detecting driver distraction [J]. arXiv: 1810.10565, 2018.

[9] WANG H, MEGHAWAT A, MORENCY L P, et al. Select-additive learning: Improving generalization in multimodal sentiment analysis [C] // 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017: 949-954.

[10] KUTILA M, JOKELA M, MARKKULA G, et al. Driver distraction detection with a camera vision system [C] // 2007 IEEE International Conference on Image Processing. IEEE, 2007: VI-201-VI-204.

[11] CRAYE C, KARRAY F. Driver distraction detection and recognition using RGB-D sensor [J]. arXiv: 1502.00250, 2015.

[12] XIAO D, FENG C. Detection of drivers visual attention using smartphone [C] // 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2016: 630-635.

[13] YANG J, CHANG T N, HOU E. Driver distraction detection for vehicular monitoring [C] // IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society. IEEE, 2010: 108-113.

[14] WALI M K, MURUGAPPAN M, AHMMAD B. Wavelet Packet Transform Based Driver Distraction Level Classification Using EEG [J]. Mathematical Problems in Engineering, 2013, 2013(pt. 13): 841-860.

[15] SATHYANARAYANA A, NAGESWAREN S, GHASEMZADEH H, et al. Body sensor networks for driver distraction identification [C] // IEEE International Conference on Vehicular Electronics and Safety. IEEE, 2008.

[16] LECHNER G, FELLMANN M, FESTL A, et al. A lightweight framework for multi-device integration and multi-sensor fusion to explore driver distraction [C] // International Conference on Advanced Information Systems Engineering. Cham: Springer, 2019: 80-95.

[17] TANPRASERT T, SAIPRASERT C, THAJCHAYAPONG S. Combining unsupervised anomaly detection and neural networks

for driver identification[J]. Journal of Advanced Transportation, 2017, UNSP 6057830.
[18] ZHANG Y, CHEN Y, GAO C. Deep unsupervised multi-modal fusion network for detecting driver distraction[J]. Neurocomputing, 2021, 421: 26-38.
[19] ALJARRAHI A A, ALI A H. Human activity recognition using PCA and BiLSTM recurrent neural networks[C]// 2019 2nd International Conference on Engineering Technology and its Applications(IICETA). IEEE, 2019: 156-160.
[20] HOWARD A G, ZHU M, CEHN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv: 1704. 04861, 2017.
[21] VUKOTIC V, RAYMOND C, GRAVIER G. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications[C]// Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. 2016: 343-346.
[22] WANG W, OOI B C, YANG X, et al. Effective multi-modal retrieval based on stacked auto-encoders[J]. Proceedings of the VLDB Endowment, 2014, 7(8): 649-660.
[23] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting [C] // Advances in Neural Information Processing Systems. 2015: 802-810.
[24] ZHANG H, LI J, JI Y, et al. A character-level sequence-to-sequence method for subtitle learning[C]// 2016 IEEE 14th International Conference on Industrial Informatics(INDIN). IEEE, 2016: 780-783.
[25] TAAMNEH S, TSIAMYRTZIS P, DCOSTA M, et al. A multimodal dataset for various forms of distracted driving [J]. Scientific Data, 2017, 4(1): 1-21.
[26] KETKAR N. Introduction to keras [M]// Deep Learning with Python. Springer, 2017: 97-111.
[27] VAN DER MAATEN L, HINTON G. Visualizing data using tsne. [J]. Journal of Machine Learning Research, 2008, 9(2605): 2579-2605.
[28] RAMANAN K. Control Techniques for Complex Networks[J]. Journal of the American Statistical Association, 2009, 104(487): 1274-1275.
[29] YANG J B, NGUYEN M N, SAN P P, et al. Deep convolutional neural networks on multichannel time series for human activity recognition [C]// Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.
[30] LIM W, JANG D, LEE T. Speech emotion recognition using convolutional and recurrent neural networks [C]// 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA). IEEE, 2016: 1-4.
[31] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409. 1556, 2014.
[32] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
[33] NGIAM J, KHOSLA A, KIM M, et al. Multimodal deep learning [C]// ICML. 2011.
[34] ZADEH A, CHEN M, PORIA S, et al. Tensor fusion network for multimodal sentiment analysis[J]. arXiv: 1707. 07250, 2017.
[35] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.
[36] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using lstms [C]// International Conference on Machine Learning. 2015: 843-852.
[37] DUMAN T B, BAYRAM B, İNCE G. Acoustic anomaly detection using convolutional autoencoders in industrial processes [C]// 14th International Conference on Soft Computing Models in Industrial and Environmental Applications(SOCO 2019). Cham: Springer International Publishing, 2020: 432-442.
[38] PARK S, KIM M, LEE S. Anomaly detection for http using convolutional autoencoders [J]. IEEE Access, 2018, 6: 70884-70901.

ZHANG Yu-xin, born in 1993, Ph.D, is a member of China Computer Federation. Her main research interests include anomaly detection, deep learning and activity recognition.

CHEN Yi-qiang, born in 1973, Ph. D, professor, is a senior member of IEEE and fellow of China Computer Federation. His main research interests include human computer interaction, pervasive computing and wearable computing.

(责任编辑:柯颖)