文章编号：0253-374X(2026)01-0150-10
DOI：10. 11908/j. issn. 0253-374x. 24297

基于深度学习的分层多尺度疲劳驾驶检测

张　戟，韩双庆，刘家栋
（同济大学 汽车与能源学院，上海 201804）

摘要：针对驾驶员疲劳检测的问题，建立了深度 HM－LSTM 网络实现基于面部特征的疲劳识别。使用 LOTR 模型实现面部空间特征的提取。考虑到驾驶员疲劳是一个循序渐进的过程，定义 8 种时序疲劳特征，使用堆叠的 HM－LSTM 在短时间序列以及长时间序列上捕捉疲劳度语义特征，从而在早期就准确识别出驾驶员疲劳迹象，通过低成本的视觉方法，进行可靠的驾驶员异常行为监测。使用公开数据集与自制 D1－DDB 数据集实现了模型的训练与系统的集成。

关键词：疲劳检测；深度学习；面部特征提取；卷积神经网络
中图分类号：TP311；TP391　　**文献标志码**：A

Hierarchical Multi-Scale Fatigue Driving Detection Based on Deep Learning

ZHANG Ji, *HAN Shuangqing*, *LIU Jiadong*
(College of Automotive and Energy, Tongji University, Shanghai 201804, China)

Abstract: To address the issue of driver fatigue detection, this paper develops a deep HM-LSTM network to achieve fatigue recognition based on facial features. The LOTR model is used to extract facial spatial features. Considering that driver fatigue is a gradual process, it defines eight temporal fatigue features and uses stacked HM-LSTM to capture fatigue semantic features over both short and long time series, in order to accurately identify signs of driver fatigue at an early stage. This approach provides reliable monitoring of abnormal driver behavior through low-cost visual-based methods. Model training and system integration are conducted by using publicly available datasets and self-constructed D1-DBB datasets.

Keywords: fatigue detection; deep learning; facial feature extraction; convolutional neural network

疲劳驾驶检测是高级驾驶辅助系统(ADAS)的关键功能，旨在预防因疲劳驾驶而造成的事故和伤害。由加拿大 TIRF 组织进行的研究发现，57% 接受调查的驾驶员认为疲劳驾驶是一种极其危险的行为，但有 20% 的人承认曾经疲劳驾驶过[1]。

驾驶员疲倦的表现形式十分多样，例如打哈欠、频繁闭眼睛以及面部肤色变化等等[10]。这些特征均可通过机器视觉的方法获取。然而值得注意的是，闭眼睛等行为不仅与困倦相关，还可能由正常眨眼、交谈以及进食而导致，因此区分困倦与这些正常行为之间的细微区别是本文的重点。

针对人脸关键点定位的研究目前有许多成熟的预测模型。基于坐标回归方法的人脸关键点预测能够实现端到端的模型设计。然而使用多级级联卷积神经网络(Convolutional Neural Network, CNN)来定位面部关键点的设计，增加了计算复杂度，因此增加了推断时间；基于热图的面部关键点定位方法与坐标回归方法相比，可以更好地利用空间信息以提升性能[1-3]。然而这些方法通常依赖于计算密集型的后处理步骤将热图转换为预测的关键点，同样会导致推断时间增加。Feng 等[4]于 2018 年提出了 Wing 模型，使用 Wing 损失函数和姿势基础数据平衡策略的方法，通过设计分段损失函数来增强小和中范围错误的影响，解决了训练集中大角度头部旋转样本不足的问题，实验结果表明，该方法在面部特征定位方面具有显著优势。Wang 等[5]于 2021 年使用 HRNet 进行面部关键点检测，该模型在整个过程中保持高分辨率，从而获得更丰富和更精确的语义信息，具有并行连接高到低分辨率卷积流和在不同分辨率之间重复交换信息的特点。Wang 等[6]于 2020 年提出了自适应 Wing 损失的新型损失函数，灵活适应不同类型的热图，通过其形状调整惩罚损失，使损失函数更关注前景像素的损失而非背景像素的损

收稿日期：2024-09-12
第一作者：张　戟，副教授，工学博士，主要研究方向为智能驾驶、汽车电子控制技术。E-mail：jizhang@tongji. edu. cn
通信作者：韩双庆，硕士生，主要研究方向为智能驾驶。E-mail：2332988@tongji. edu. cn

论文拓展介绍

失，同时引入加权损失地图来处理前景和背景像素之间的不平衡，该模型在面部对准实验中获得了最佳性能。Jin 等[7]于 2021 年提出了 PIPNet 模型，通过采用基于热图回归的检测头，能够在低分辨率特征图上同时进行得分和偏移预测，从而减少计算成本，该模型在保持竞争准确性的同时具有较高的推断速度，是一种性能优越的轻量级面部特征点检测模型。Andres 等[8]于 2022 年提出了 SPIGA 模型，采用多任务学习方法初始化节点位置和由粗到精的描述方案。模型结合了 CNN 和级联图注意力网络回归器，能够学习到面部结构的全局特征，特别擅长处理局部特征点外观变化较大的情况。

本文采用了一种多特征融合的分层多尺度网络用于驾驶员疲劳检测，主要工作包括：对视觉传感器采集的视频进行预处理；构建检测网络；数据收集和设定评价指标；进行不同方法的对比实验。

1　视频数据预处理

视频数据的预处理阶段涵盖多个关键步骤：人脸关键点的精确定位以及从单帧图像中提取眼部和嘴部的特征信息，进而进行时序分析以检测眨眼和张嘴动作。目前一些人脸关键点定位模型已经拥有了很高的推理速度，并且在低照明环境和物体遮挡条件下，都拥有较高的准确率。为了优化开发流程的效率，采用现有的人脸关键点检测模型，并结合自主设计的特征提取方法，以达到疲劳识别目标的目的。

1.1　人脸关键点定位

采用基于直接坐标回归方法的 LOTR（Localization Transformer）模型作为脸部关键点定位模型[9]。LOTR 模型利用转换器模型（Transformer）来提取特征图的空间信息。

如图 1 所示，LOTR 网络主要由三部分组成：视觉主干网络、Transformer 网络和特征点预测头。在主干网络中，采用残差网络 Restnet-50 提取输入图片特征，并且通过 1×1 的卷积核降低特征图维度。输出特征图尺寸为 6×6 像素，通过插值上采样为 192×192 像素特征图以还原更多特征细节。

主干网络如式（1）所示：

$$F=f_{\text{upsample}}(f_{\text{conv1}\times1}(f_{\text{RES}}(\boldsymbol{I}))) \tag{1}$$

式中：F 表示主干网络输出特征，$F\in\mathbb{R}^{W\times H\times 3}$；$f_{\text{upsample}}$ 表示上采样；$f_{\text{conv1}\times1}$ 表示 1×1 卷积；f_{RES} 表示残差网络 Restnet-50；$\boldsymbol{I}$ 表示输入图像，$\boldsymbol{I}\in\mathbb{R}^{W_I\times H_I\times 3}$。

在 Transformer 网络中，F 首先通过 1×1 卷积网络降低维度，随后展开为令牌序列，$X^0\in\mathbb{R}^{WH\times D}$。Transformer 编码器层由 L 层编码层堆叠组成；每个编码层都由 2 个子模块组成：多头注意力（MHSA）和 PFFN[10]。解码器为 L 层解码器层的堆叠，每个解码器层由 3 个子模块组成：MHSA、多头交叉注意力（MHCA）和 PFFN。第 1 个解码器层的第 1 个子层的输入是可训练的面部关键点 Query 序列，$Y^0\in\mathbb{R}^{N\times D}$。第 2 个子层以第 1 个子层和编码器生成的输出 $\boldsymbol{X}^l$ 作为输入，然后计算 2 个序列中令牌之间的关系。特征点预测头采用线性层产生最终的面部关键点坐标 $\hat{\boldsymbol{Z}},\hat{\boldsymbol{Z}}\in\mathbb{R}^{N\times 2}$。

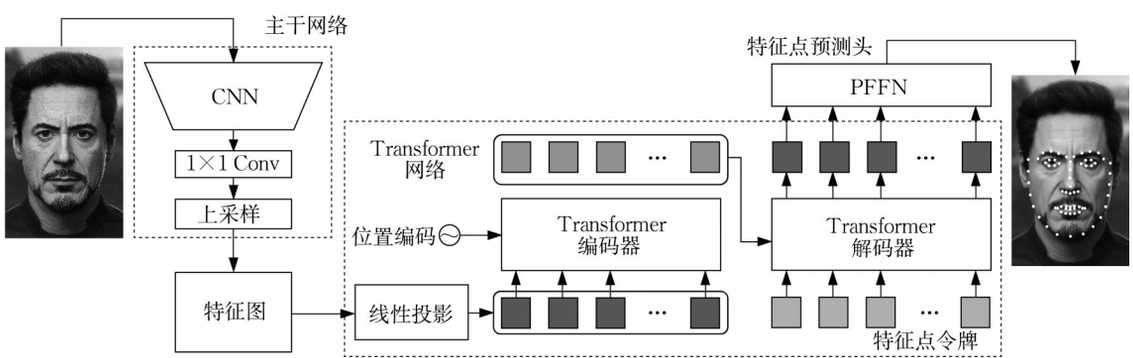


图 1　LOTR 网络结构

Fig. 1　Structure of LOTR network

$$\hat{Z}=f_{\text{Relu}}\left(Y^L w_1^z+b_1^z\right)w_2^z+b_2^z \tag{2}$$

式中：z 为可学习参数；w_1^z 和 b_1^z 分别为特征点预测头中第 1 个线性层的权重和偏置；f_{Relu} 为 Relu 激活函数。

以图 1 的人脸为例，在经过上述主干网络之后得到了若干特征图，如图 2，选取了其中 6 个。

1.2　人脸关键点检测模型验证与对比

为验证模型在低光照情况下的推理准确率以及

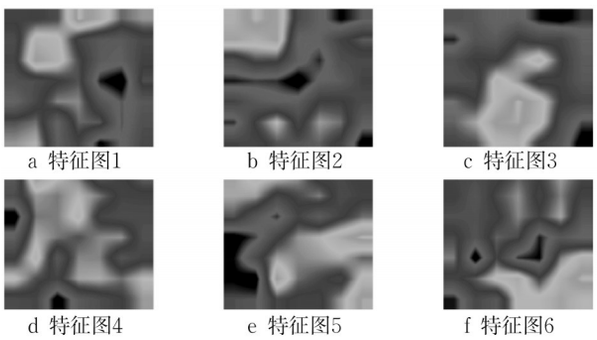


图2 主干网络输出特征

Fig. 2 Characteristics of output of backbone network

模型的推理速度,使用6个典型面部关键点估计模型在WFLW测试集进行验证,并且在同一服务器上测试其推理速度。模型结合了CNN和级联图注意力网络回归器,能够学习到面部结构的全局特征,特别擅长处理局部特征点外观变化较大的情况。

模型性能评估数据集采用野外宽领域人脸标记(WFLW)数据集[11]。数据集中包括遮挡、姿势、化妆、光照、模糊和特殊表情场景的面部标注数据,可用于模型泛化性能和鲁棒性测试。

采用归一化平均误差E_{nme}验证模型的准确性,如式(3)所示[12]:

$$E_{\text{nme}}=\frac{1}{N_{\text{f}}}\sum_{i=1}^{N_{\text{f}}}\frac{\left\|Z_i-\hat{Z}_i\right\|_2}{d} \tag{3}$$

式中:N_{f}为面部关键点的总数;Z_i和$\hat{Z}_i$分别为面部关键点真值和预测值;d为眼睛两角的连线距离,为归一化系数。当E_{nme}值大于阈值(10%)时,认为预测失败。

E_{nme}值越高,代表预测准确度越低。从表1可见,SPIGA模型在多个测试集上都获得了SOTA值,LOTR模型在妆容数据集以及姿态数据集上获得了SOTA值,其性能仅次于SPIGA,接下来研究LOTR和SPIGA在2个部位的表现以及参数量的区别。

表1 6种模型准确度对比

Tab. 1 Comparison of accuracy of six models

单位:%

准则	方法	主干网络	测试集	副测试集					
				姿态	表情	光照	妆容	遮挡	模糊
NME	Wing	ResNet−50	4.99	8.43	5.21	4.88	5.26	6.21	5.81
	HRNet	HRNetV2	4.60	7.94	4.85	4.55	4.24	5.44	5.42
	Awing	Hourglass	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	PIPNet	MobileNetV2	4.48	7.98	4.51	4.35	4.27	5.65	5.19
	SPIGA	Hourglass	4.26	7.44	4.46	4.02	4.09	4.95	4.65
	LOTR	ResNet−50	4.31	7.20	4.52	4.10	4.06	5.12	4.99

由于本文主要关注的面部特征为眼睛轮廓以及嘴部轮廓,所以建立式(4)的评判标准:

$$E_{\text{nme}-\text{p}}=\frac{1}{N_{\text{part}}}\sum_{i\in I_{\text{p}}}\frac{\left\|Z_i-\hat{Z}_i\right\|_2}{d} \tag{4}$$

式中:I_{p}为眼睛与嘴巴所述的索引位置,$I_{\text{p}}\in[I_{\text{e}},I_{\text{m}}]$;$N_{\text{part}}$为眼睛与嘴部的标注点数,$N_{\text{part}}\in[N_{\text{e}},N_{\text{m}}]$。6种候选模型的$E_{\text{nme}-\text{p}}$统计结果如图3所示。

Wing、Awing、SPIGA和LOTR在眼部以及嘴部的预测失误率均低于整体预测失误率,表明在相似模型参数的情况下,基于Resnet-50和Hourglass的特征提取主干网络对于图像拥有更佳的语义理解能力。SPIGA和LOTR在这2个部位的表现分别获得了最佳结果。

在ubuntu20.04系统上分别使用6种模型推理2 500张WFLW图片,并统计其单张图片推理时间。系统配备一块24-core Intel(R) Xeon(R) Gold 6248R @ 3.00GHz处理器和一张Nvidia Quadro RTX 5000显卡,拥有16GB内存。比较结果如表2所示。

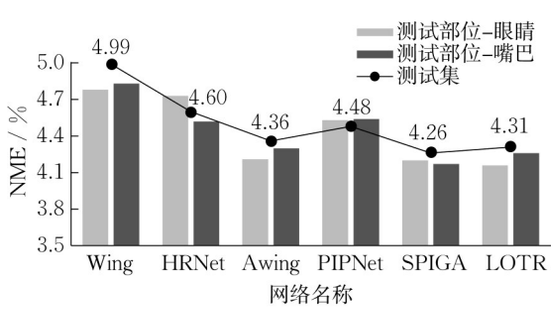


图3 6种模型在眼部与嘴部预测准确度对比

Fig. 3 Comparison of accuracy of six models in eye and mouth prediction

LOTR模型在确保预测准确率的同时,具备较少的参数量和高效的推理速度。综合考虑,采用

表2　6种模型的参数量与推理速度对比

Tab. 2　Comparison of parameter number and inference speed of six models

方法	参数量/10^6	容量/MB	推理速度/ms	
			FPS(CPU)	FPS(GPU)
Wing	22.8	91.0	8.0	31.8
HRNet	9.7	40.7	5.3	40.7
Awing	24.1	91.5	1.8	39.4
PIPNet	26.7	104.1	8.8	60.2
SPIGA	31.2	118.6	6.5	46.3
LOTR	25.2	92.3	7.9	53.2

LOTR模型进行面部关键点识别。

1.3　疲劳特征提取

在日常生活中人们眨眼和在困倦时眨眼具有明显的差异,主要包括4个指标:眨眼频率、眼睛睁开速度、眼睛张开程度以及眨眼持续时间,因此需要针对一段视频数据提取相应特征。同理也需要提取视频中拍摄者的嘴部张开频率、张开速度、张开程度以及张开时间特征。随后单独使用眨眼特征或是使用组合特征来判定受测者是否疲倦。

采用眼睛纵横比(Eye Aspect Ratio, EAR,α)定义拍摄者眼睛睁开的程度,如式(5)所示;采用嘴部张开程度(Mouth Aspect Ratio, MAR,β)定义拍摄者嘴巴的张开程度,如式(6)所示。

$$\alpha=\frac{\|\boldsymbol{E}_3-\boldsymbol{E}_4\|+\|\boldsymbol{E}_5-\boldsymbol{E}_6\|}{2\|\boldsymbol{E}_1-\boldsymbol{E}_2\|} \tag{5}$$

$$\beta=\frac{\|\boldsymbol{M}_3-\boldsymbol{M}_4\|+\|\boldsymbol{M}_5-\boldsymbol{M}_6\|}{2\|\boldsymbol{M}_1-\boldsymbol{M}_2\|} \tag{6}$$

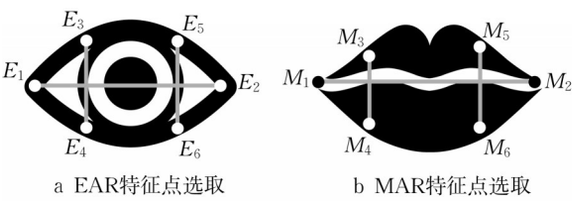
式中:$\boldsymbol{E}_i$为人脸面部关键点的2D坐标,$\boldsymbol{E}_i=[x_{si},y_{si}]$;$\boldsymbol{M}_i$为人脸嘴部关键点的2D坐标,$\boldsymbol{M}_i=[x_{mi},y_{mi}]$。

式(5)、(6)中各点所对应位置见图4a。对视频中每一帧图像提取人脸关键点数据,并计算其EAR数值。设置EAR阈值h_α为0.2,当EAR值小于h_α时,记为负,否则为正。将EAR随帧数的变化绘制成图5所示的时间特征,图中A_{si}、B_{bi}、C_{ei}分别表示眨眼动作起始帧、眨眼动作波谷帧、眨眼动作结束帧。当EAR值从一个波峰下降到波谷,经过阈值,随后从波谷再次上升到波峰,再次经过一个阈值,判定为一次眨眼B_i。图5虚线部位对应一个检测的眨眼动作。

根据每一个眨眼动作计算相应的帧特征值。睁眼持续时间t_i、睁眼程度a_i、睁眼速度v_i、眨眼频率f_i,其关系如式(7)—(11)所示:

$$B_i=[t_i,v_i,f_i,a_i] \tag{7}$$

$$t_i=t_{e,i}-t_{s,i}+1 \tag{8}$$



a EAR特征点选取　　b MAR特征点选取

图4　面部特征点示意

Fig. 4　Schematic diagram of facial feature point

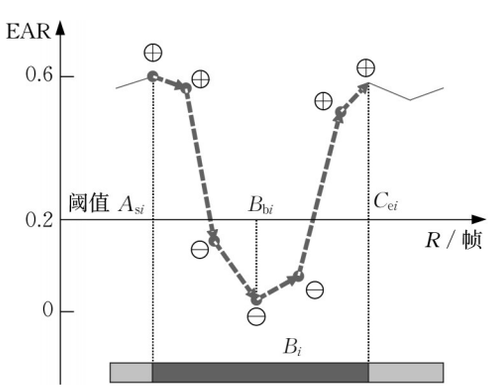


图5　眨眼检测

Fig. 5　Detection of blink

$$f_i=\frac{i}{t_{e,i}} \tag{9}$$

$$\nu_i=\frac{\alpha_i^{s}-\alpha_i^{b}}{t_{e,i}-t_{b,i}} \tag{10}$$

$$a_i=\frac{\alpha_i^{s}+\alpha_i^{e}-2\alpha_i^{b}}{2} \tag{11}$$

式中:α_i^{s}为起始帧对应的EAR值;α_i^{b}为波谷帧对应的EAR值;α_i^{e}为结束帧对应的EAR值。类似地,检测打哈欠张嘴检测的阈值h_β设置为0.5,如图6所示。打哈欠检测所需的各点见图4b,图6中A_{sk}、B_{pk}、C_{ek}分别表示打哈欠动作起始帧、打哈欠动作波峰帧、打哈欠动作结束帧。根据每个打哈欠的动作计算相应的帧特征值:打哈欠持续时间t_k、张嘴程度a_k、张嘴速度v_k、打哈欠频率f_k,如式(12)—(16)所示:

$$Y_k=[t_k,v_k,f_k,a_k] \tag{12}$$

$$t_k=t_{e,k}-t_{s,k}+1 \tag{13}$$

$$f_k=\frac{k}{t_{e,k}} \tag{14}$$

$$\nu_k=\frac{\beta_k^s-\beta_k^p}{t_{e,k}-t_{p,k}} \tag{15}$$

$$a_k=\frac{\beta_k^s+\beta_k^e-2\beta_k^p}{2} \tag{16}$$

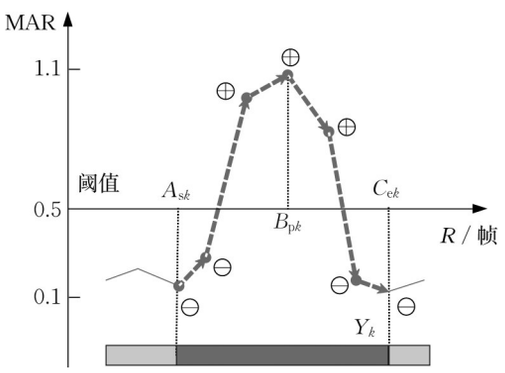


图6 打哈欠检测

Fig. 6 Detection of yawn

2 疲劳驾驶检测网络

采用分层多尺度长短期记忆网络(HM-LSTM)来学习特征,在相同的网络结构基础上,也使用其他模型如LSTM、GRU和Transformer进行对比。

2.1 分层多尺度长短期记忆网络

HM-LSTM的研究源自对DNNs在处理时间序列数据时如何模拟人脑处理信息的层次结构和多时间尺度能力的深入探讨[13]。与LSTM和GRU相比,HM-LSTM引入了对于不同时间尺度上层级信息的适应性处理能力。与RNN模型相比[14-15],HM-LSTM可以通过内部机制动态调整时间层级。对于短时间序列,HM-LSTM网络能够利用其门控机制(如输入门、遗忘门和输出门)来控制信息的流动,从而捕捉到疲劳度的即时变化。这些门控机制允许网络在每个时间步自主决定保留或遗忘哪些信息,这对于捕捉眨眼等短暂行为特别有效。例如,当检测到眨眼行为时,网络可以通过调整细胞状态来反映这一变化,即使这种行为在时间序列中只持续很短时间。对于长时间序列,HM-LSTM网络的层次化结构使其能够捕捉到疲劳度的累积效应。通过在不同层次上处理时间序列数据,HM-LSTM能够识别出疲劳度的长期趋势和周期性模式。这种多层次的处理能力使得HM-LSTM不仅能够捕捉到短期的疲劳信号,还能够理解这些信号随时间的演变,从而更准确地预测疲劳度的变化。

HM-LSTM通过在每个层中引进一个二元边界状态,自适应地调整每层的状态更新频率。这种机制使得HM-LSTM可以在底层捕捉细微的时间变化,在高层捕捉更加抽象和慢速的特征变化。根据边界状态,网络可以决定在t时刻是更新、复制还是刷新当前第l层的状态c_t^l,如式(17)所示:

$$\boldsymbol{c}_t^l=\begin{cases}\boldsymbol{f}_t^l\odot\boldsymbol{c}_{t-1}^l+\boldsymbol{i}_t^l\odot\boldsymbol{g}_t^l, & \text{若}z_{t-1}^l=0\text{及}z_t^{l-1}=1\text{则更新}\\ \boldsymbol{c}_{t-1}^l, & \text{若}z_{t-1}^l=0\text{及}z_t^{l-1}=0\text{则复制}\\ \boldsymbol{i}_t^l\odot\boldsymbol{g}_t^l, & \text{若}z_{t-1}^l=1\text{则重置}\end{cases} \tag{17}$$

随后更新第l层的隐藏层状态h_t^l,如式(18)所示:

$$\boldsymbol{h}_t^l=\begin{cases}\boldsymbol{h}_{t-1}^l, & \text{复制}\\ \boldsymbol{o}_t^l\odot\tanh(\boldsymbol{c}_t^l), & \text{其他}\end{cases} \tag{18}$$

2.2 分层多尺度疲劳驾驶检测网络

一维卷积神经网络在时序信号处理中具有重要应用。在时间序列数据中一维卷积神经网络可以用于提取信号中的关键特征,比如周期性、趋势、脉冲等,从而实现对时序数据的分析和预测。通过权重值不同的卷积核对多个通道的时序数据X($X\in\mathbb{R}^{C_m\times L_m}$)进行卷积处理。输出数据中第$c_y$个通道的第$j$个值$y_{c_y,j}$可以表示为式(19):

$$y_{c_y,j}=\sum_{c_x=0}^{C_n-1}\sum_{k-X}^{0}x_{c_x,j-k}w_k^{c_x,c_y} \tag{19}$$

单层HM-LSTM无法实现其层级式时序信息提取的功能,其最少堆叠层数为2,因此提出如图7所示的基于4层HM-LSTM和一维卷积神经网络的疲劳驾驶检测网络。

该神经网络的输入X为视频中人物的眨眼特征序列、张嘴特征序列或是两者的组合。

$$\boldsymbol{X}=[\,\boldsymbol{B}\,]^{\mathrm{T}} \tag{20}$$

$$\boldsymbol{X}=[\,\boldsymbol{Y}\,]^{\mathrm{T}} \tag{21}$$

$$\boldsymbol{X}=[\,\boldsymbol{B},\boldsymbol{Y}\,]^{\mathrm{T}} \tag{22}$$

$$\boldsymbol{Bk}=[\,\boldsymbol{B}_1^{\mathrm{T}},\boldsymbol{B}_2^{\mathrm{T}},\ldots,\boldsymbol{B}_T^{\mathrm{T}}]^{\mathrm{T}} \tag{23}$$

$$\boldsymbol{Yw}=[\,\boldsymbol{Y}_1^{\mathrm{T}},\boldsymbol{Y}_2^{\mathrm{T}},\ldots,\boldsymbol{Y}_T^{\mathrm{T}}]^{\mathrm{T}} \tag{24}$$

由于个体差异性,针对每个拍摄者,提取其清醒状态下的前10次眨眼与张嘴特征并计算均值与标准差,进行归一化处理。

在完成归一化之后,使用共享权重$\boldsymbol{W}_i^l$和偏置矩阵$\boldsymbol{b}$将输入特征转换为N维空间。定义$\boldsymbol{T}$为输入特征的时序长度。$\boldsymbol{f}_i$表示映射得到的$\boldsymbol{i}$时刻特征。

$$F=f_{\mathrm{Relu}}(\boldsymbol{XW}+\boldsymbol{b}) \tag{25}$$

$$F=[\,\boldsymbol{f}_1^{\mathrm{T}},\boldsymbol{f}_2^{\mathrm{T}},\ldots,\boldsymbol{f}_i^{\mathrm{T}},\ldots,\boldsymbol{f}_T^{\mathrm{T}}\,] \tag{26}$$

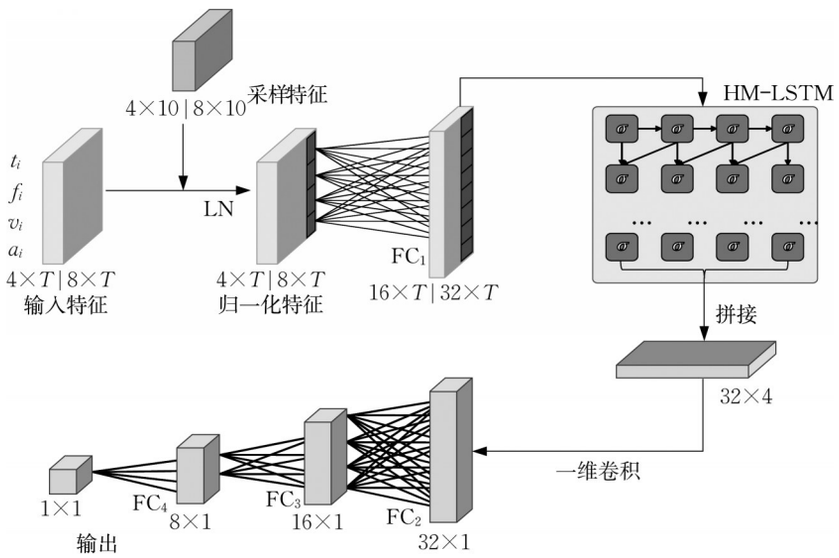


图7　层级式疲劳驾驶检测网络

Fig. 7　Hierarchical network for fatigue driving detection

堆叠HM-LSTM网络在时刻T输出各层的隐藏状态，第l层的隐藏状态表示为$\boldsymbol{h}_l\in\mathbb{R}^{bH}$，$H$为每层的隐藏状态维度。为充分使用各个层级上的多尺度时间特征，将L层HM-LSTM的输出$\boldsymbol{e}_{\text{ll}}$连接成为$H\times L$维张量$\boldsymbol{e}_1=[\boldsymbol{e}_{11}\ \ \boldsymbol{e}_{12}\ \cdots\ \boldsymbol{e}_{1L}]$，并对$\boldsymbol{e}_1$进行1D卷积。

$$\boldsymbol{e}_2=f_{\text{conv1}\times1}D(\boldsymbol{e}_1) \tag{27}$$

随后经过3层全连接层，每层以Relu作为激活函数，输出$\boldsymbol{e}_4$。

在数据集中，针对每个视频，都拥有一个对应的疲劳状态标记。疲劳状态分为3个等级：清醒、轻度疲倦和疲倦。因此，最直观的方式是采用分类的方法输出最终的结果，如式(28)所示：

$$e_5=f_{\text{softmax}}(e_4\boldsymbol{W}_{c4}+\boldsymbol{b}_{c4}) \tag{28}$$

式中：$\boldsymbol{W}_{c4}$、$\boldsymbol{b}_{c4}$为全连接层FC_4中的待学习权重和偏置。

多分类任务对应的损失函数如式(29)所示：

$$l_{\text{loss}}=-\frac{1}{N_{\text{S}}}\sum_{i=0}^{N_{\text{S}}}\sum_{i=1}^{M}y_{ic}\ln(p_{ic}) \tag{29}$$

式中：N_{S}为总样本数；M表示类别，此处为3；y_{ic}为符号函数，如果样本i的真实类别等于c取1，否则取0；P_{ic}为观测样本i属于类别c的预测概率。

另一种方式采用回归的方法如式(30)所示，最终输出e_5。e_5取值范围为[1,10]。在对e_5进行分类时按照以下标准：0～3.3为清醒、3.3～6.6为轻微疲劳、6.6～10为疲劳。

$$e_5=10\times\sigma(e_4\boldsymbol{W}_4+\boldsymbol{b}_4) \tag{30}$$

式中：$\boldsymbol{W}_4$、$\boldsymbol{b}_4$分别表示全连接层FC_4中的待学习权重和偏置矩阵。

当且仅当根据式(31)计算的预测类别与真实类别不符时累加样本损失。回归任务对应的损失函数如式(32)所示：

$$l_{\text{loss}}=\frac{\sum_{i=1}^{N_S}\max(0,\left|e_{5i}-t_i\right|^2-\Delta)}{N} \tag{31}$$

式中：t_i为与样本预测值最接近的类别边界；Δ为距离阈值。

对视频进行整体的疲劳度预测。在某一视频中，假设存在K次眨眼或张嘴的特征。使用长度为T的滑动窗口对特征序列进行分段，每T个连续的特征形成一个特征序列，作为网络的输入。因此每段视频将产生多个疲劳度预测结果，从这些预测结果中选出最频繁预测的疲倦度类别作为此视频的最终分类结果。

3　数据集与评价指标

3.1　UTA-RLDD数据集

采用UTA-RLDD进行模型训练与验证。数据集包括参与者在3种不同状态下的视频：警惕、低警觉和困倦，共计180段。在180个视频中，21个视频中的拍摄者佩戴眼镜，72个视频中拍摄者有相当多的面部毛发。拍摄角度为面部正前方，数据集部分截图如图8所示。拍摄使用的设备为手机或网络摄像头，拍摄帧率不高于30帧。

3.2　评价指标

除了使用损失函数以外，准确率也是重要评价

图8 UTA-RLDD数据集

Fig. 8 UTA-RLDD dataset

指标。定义序列预测准确度S_A和视频预测准确度V_A如式(32)和式(33)所示。S_A表示每条时序长度为T的特征序列被正确预测疲劳度状态的准确率。

$$S_A=\frac{P_{seq}+T_{seq}+C_{seq}}{N_{seq}} \tag{32}$$

式中:P_{seq}、T_{seq}与C_{seq}分别表示被正确预测为疲倦状态、清醒状态和轻度疲倦状态的特征序列数目;N_{seq}为总特征序列数目。V_A表示为每条视频被正确预测疲劳度状态的准确率,定义如式(34)所示:

$$V_A=\frac{P_{vid}+N_{vid}+C_{vid}}{N_{vid}} \tag{33}$$

式中:P_{vid}、N_{vid}和C_{vid}分别表示被正确预测为疲倦状态、清醒状态和轻度疲倦状态的视频数目;N_{vid}为总视频数目。

3.3 模型训练

时序特征编码部分分别选用HM-LSTM模型、LSTM模型、GUR模型与Transformer模型。根据分类或回归的任务,调整输出层FC_4中的神经元个数,使各个模型达到最优。在所有的实验中,均采用相同的时序步长T=30。

4 实验结果与对比

4.1 不同输入特征对比实验

为检验不同特征对疲劳检测准确率的影响,进行如下对比试验。使用眼部特征进行疲倦度推理时,4种模型在验证集上的时序片段准确率(S_A)与损失如图9所示。

在仅使用眼部特征作为模型输入时,GRU获得了最高的S_A准确度。基于LSTM和HM-LSTM的疲劳检测展现出了相似的检测准确度。而基于

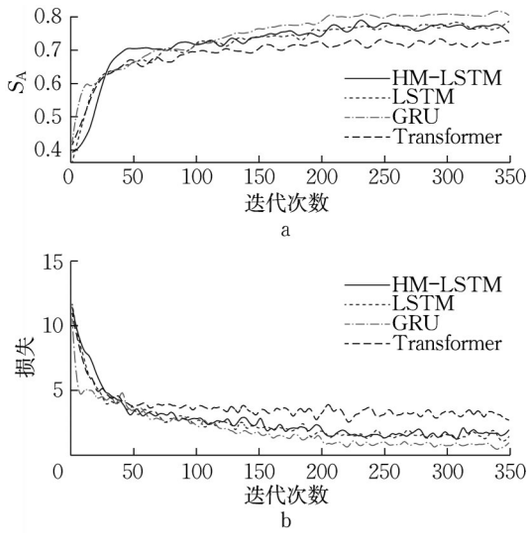


图9 4个不同输入特征模型的回归任务训练结果

Fig. 9 Regression training results of models with four different input features

Transformer的模型检测准确度最低。考虑由于GRU的结构更简单,因此在小数据集的样本上能够获得更好的训练结果。

使用眼部与嘴部的组合特征进行推理,如图10所示。基于HM-LSTM的疲劳网络表现出了最优的检测准确率。相较于单纯的眼部特征,眼部与嘴部的组合特征在语义上提供了另一个维度的疲劳线索。得益于HM-LSTM的层级式设计,层叠的HM-LSTM网络可以从多个特征维度上对语义进行划分,并且得到最准确的预测结果。

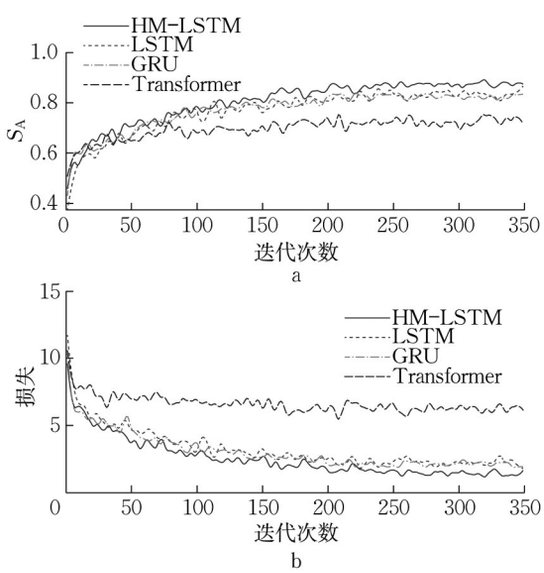


图10 组合特征输入模型回归任务训练结果

Fig. 10 Regression training results of the model with combined feature inputs

4.2　不同任务对比实验

在回归和分类任务中，4 种模型的表现不同，如表 3。

表 3　不同任务下 4 个不同输入特征模型性能对比

Tab. 3　Performance comparison of four different input feature models under different tasks

模型任务	时序编码网络	S_A	V_A
回归	HM−LSTM	0.81	0.78
	LSTM	0.81	0.76
	GRU	0.83	0.84
	Transformer	0.74	0.69
分类	HM−LSTM	0.69	0.70
	LSTM	0.73	0.68
	GRU	0.72	0.74
	Transformer	0.79	0.76

通过混淆矩阵进一步分析 2 种训练策略的优缺点，如图 11 和图 12 所示，分别为在回归任务中取得最优性能的基于 GRU 的疲劳驾驶检测网络的混淆矩阵和在分类任务中取得最优性能的基于 Transformer 的疲劳驾驶检测网络的混淆矩阵。

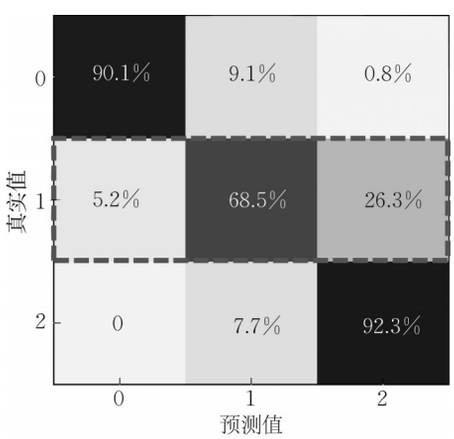


图 11　回归任务最优模型混淆矩阵

Fig. 11　Confusion matrix of optimal regression mode

在图 11 与 12 中，预测值和真实值的标签“0”、“1”、“2”分别对应清醒、轻度疲倦和疲倦 3 个状态。从混淆矩阵中可以看到，2 种模型都能够十分准确地区分拍摄者的清醒状态以及疲劳状态，均接近 90% 的准确率。然而二者在区分轻度疲倦状态时均表现不佳，准确率仅约 70%。清醒状态与疲劳状态拥有十分鲜明的特征表现，容易通过神经网络进行区分，然而轻度疲劳特征并不鲜明，容易造成误分类；其次，拍摄者在视频标注过程中，针对中间的状态存在主观判断，可能造成部分误标注。

进一步分析发现，在回归任务中，26.3% 的轻度疲倦样本被判定为疲倦，5.2% 的轻度疲倦样本被判定为清醒。这是可以接受的，因为 26.3% 的误警率在一定程度上能够起到提早发现驾驶员疲倦迹象的作用，避免事故发生。然而，在分类任务中，有 20.7% 的轻度疲倦样本被判定为清醒，7.0% 的轻度疲倦样本被判定为疲倦。这表明基于分类任务的疲劳检测模型对于疲劳的判定相对宽松，没有起到在早期发现驾驶员疲劳特征的作用。

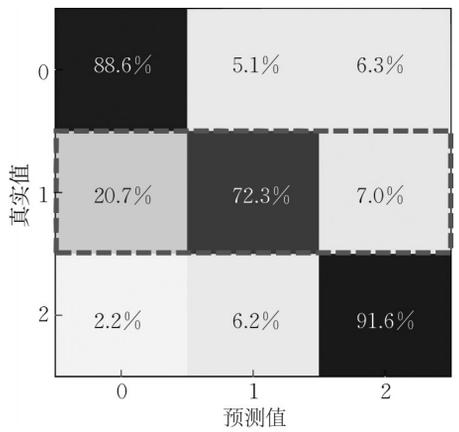


图 12　分类任务最优模型混淆矩阵

Fig. 12　Confusion matrix of optimal classification model

4.3　与其他疲劳检测方法对比

Singh 等[24]使用 CNN 检测人脸特征并且提取眼部特征，随后使用 PERCLOSE 算法，通过在特定时间窗口内检测眨眼次数与眨眼时间来检测疲劳状态。其在 UTA-RLDD 数据集上的检测准确率为 80.6%。

Tamanani 等[25]使用 Haar-cascade 方法和 CNN 对面部区域进行检测，并且使用端到端的方法检测拍摄者的疲劳状态。在 UTA-RLDD 数据集上检测准确率为 93.8%。

这 2 种方法在检测准确率以及模型参数量方面与本文提出的方法对比如 4 所示。

表 4　其他疲劳检测模型对比

Tab. 4　Comparison of other fatigue detection models

方法	准确度/%	参数量/10^6
PERCLOSE	80.6	19.8
Haar−cascade	93.8	56.8
本文	90.3	26.7

本文模型相较最优模型，仍存在 3.5% 的检测准确率差距。由表 4 可见，本文提出的分层多尺度疲劳检测模型在参数量上仅为 Haar-cascade 模型的一半，证明本文设定并提取的特征量与驾驶员疲劳存在很强的相关性，并且在一定程度上丢弃了 Haar-cascade 端到端算法中的冗余信息。

最优疲劳检测模型检测结果可视化如图13和14所示,图13中结果数据已在图中标出。图13展示的是清醒状态的检测片段截图。图14展示的是轻微疲倦状态的检测片段截图,其中有2个片段(图14a和14d)存在检测错误的情况,被归类为疲倦状态。图15展示的是疲劳状态的检测片段截图。

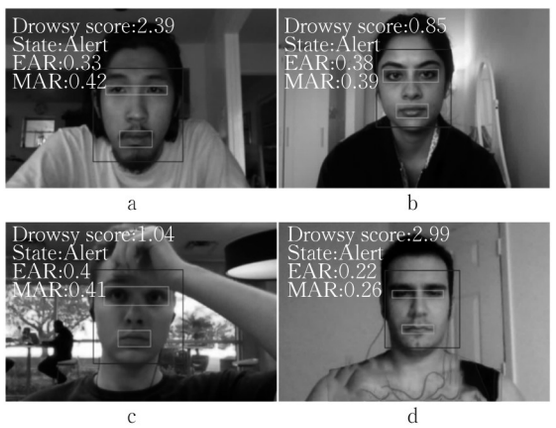


图13 清醒时序片段检测结果

Fig. 13 Detection results of alert temporal segments

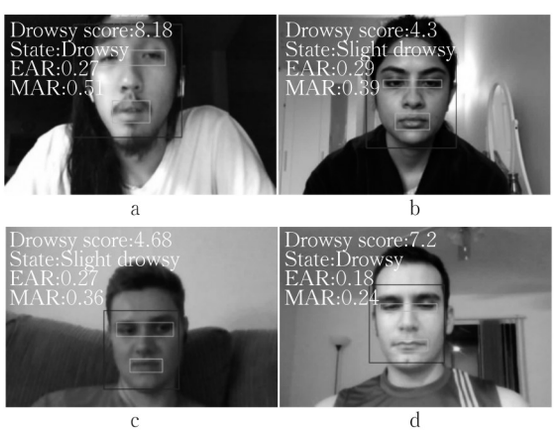


图14 轻度疲倦时序片段检测结果

Fig. 14 Detection results of mild fatigue temporal segments

5 结论

针对驾驶员疲劳检测的问题,设计了一种分层多尺度疲劳驾驶检测网络。从面部的遮挡问题出发,对比了6种典型面部关键点检测模型在不同数据集上的检测准确度等指标,最终选择LOTR网络用于特征提取任务。建立了疲劳驾驶检测网络,并且分别使用GRU、LSTM、HM-LSTM以及Transformer模型进行了时序特征编码,最终对网络预测结果进行分析对比。

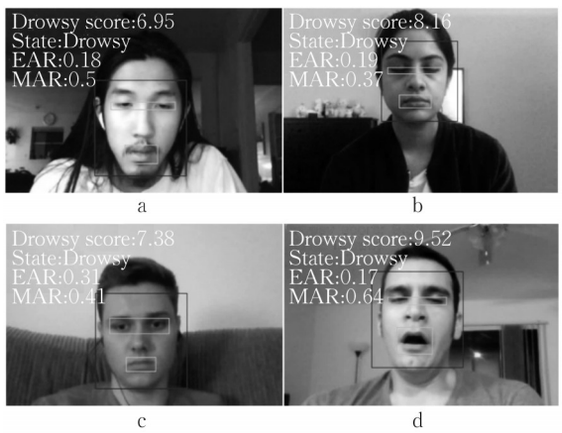


图15 疲倦时序片段检测结果

Fig.15 Detection results of fatigue temporal segments

使用了眼部、嘴部—眼部组合特征2种输入进行疲劳检测,以应对实际驾驶场景下部分特征无法被检测的问题,结果表明使用组合特征的准确率较单一特征提升了7%,同时分析对比了2种不同的损失函数。通过回归任务训练得到的模型在测试集上的检测准确率高于使用分类任务训练得到的模型,通过回归任务训练得到的模型在混淆矩阵上表现出来的检测结果更符合提早检测疲倦状态并且符合进行安全预警的要求。

作者贡献声明:

张 戟:论文整体构思,研究步骤指导,论文修改。

韩双庆:数据处理,神经网络构建和调整,论文撰写。

刘家栋:辅助进行神经网络构建,调参,论文撰写。

参考文献:

[1] KOWALSKI M, NARUNIEC J, TRZCINSKI T. Deep alignment network: a convolutional neural network for robust face alignment[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. Honolulu: IEEE, 2017: 88-97.

[2] XIONG Y, ZHOU Z, DOU Y, *et al*. Gaussian vector: an efficient solution for facial landmark detection [C]// Proceedings of the Asian Conference on Computer Vision (ACCV). Kyoto: Springer, 2020: 759-775.

[3] MAHPOD S, DAS R, MAIORANA E, *et al*. Facial landmarks localization using cascaded neural networks [J]. Computer Vision and Image Understanding, 2021, 205: 103171.

[4] FENG Z H, KITTLER J, AWAIS M, *et al*. Wing loss for robust facial landmark localisation with convolutional neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake

City: IEEE, 2018: 2235-2245..

[5] WANG J, SUN K, CHENG T, *et al*. Deep high-resolution representation learning for visual recognition [J]. 2021, 43 (10): 3349.

[6] WANG X, BO L, FUXIN L. Adaptive wing loss for robust face alignment via heatmap regression[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 6971-6981.

[7] JIN H, LIAO S, SHAO L. Pixel-in-pixel net: towards efficient facial landmark detection in the wild [J]. International Journal of Computer Vision, 2021, 129(12): 3174.

[8] PRADOS-TORREBLANCA A, BUENAPOSADA J M, BAUMELA L. Shape preserving facial landmarks with graph attention networks [J]. arXiv preprint arXiv, 2022: 2210.07233.

[9] QU S, GAO Z, WU X, *et al*. Multi-attention fusion drowsy driving detection model[EB/OL]. (2023-12-28)[2024-10-11]. https://arxiv.org/abs/2312.17052.

[10] KIM D, PARK H, KIM T, *et al*. Real-time driver monitoring system with facial landmark-based eye closure detection and head pose recognition [J]. Scientific reports, 2023, 13 (1): 18264.

[11] SUN Yi, WANG Xiaogang, TANG Xiaoou, *et al*. Deep convolutional network cascade for facial point detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Portland: IEEE, 2013: 3476-3483.

[12] ZHOU E, FAN H, CAO Z, *et al*. Extensive facial landmark localization with coarse-to-fine convolutional network cascade [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops. Sydney: IEEE, 2013: 386-391.

[13] ZHANG J, SHAN S, KAN M, *et al*. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment [C]// Proceedings of the European Conference on Computer Vision (ECCV). Zurich: Springer, 2014: 1-16.

[14] WATCHAREERUETAI U, SOMMANA B, JAIN S, *et al*. Lotr: face landmark localization using localization transformer [J]. IEEE Access, 2022,10: 16530.

[15] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). Long Beach: Curran Associates Inc., 2017: 5998-6008..

[16] WU W, QIAN C, YANG S, *et al*. Look at boundary: a boundary-aware face alignment algorithm[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 2129-2138.

[17] EARP S W, SAMACOITS A, JAIN S, *et al*. Sub-pixel face landmarks using heatmaps and a bag of tricks[J]. arXiv preprint arXiv, 2021:2103.03059.

[18] CHUNG J, AHN S, BENGIO Y. Hierarchical multiscale recurrent neural networks [J]. arXiv preprint arXiv, 2016: 1609.01704.

[19] KOUTNIK J, GREFF K, GOMEZ F, *et al*. A clockwork rnn [C]//Proceedings of the 31st International Conference on Machine Learning (ICML). Beijing: PMLR, 2014: 1863-1871.

[20] HIHI S, BENGIO Y. Hierarchical recurrent neural networks for long-term dependencies[J]. Advances in Neural Information Processing Systems, 1995,8: 493.

[21] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014,27: 3104.

[22] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks [C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver: IEEE, 2013: 6645-6649.

[23] VINYALS O, TOSHEV A, BENGIO S, *et al*. Show and tell: A neural image caption generator[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 3156-3164.

[24] SINGH H K, KUUSIK A. Evaluation of driver status assessment system based on deep learning [D]. Tallinn: Tallinn University of Technology, 2020.

[25] TAMANANI R, MURESAN R, AL-DWEIK A. Estimation of driver vigilance status using real-time facial expression and deep learning[J]. IEEE Sensors Letters, 2021,5(5): 1.