计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(10):177−182 [doi: 10.15888/j.cnki.csa.006555]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多面部特征融合的驾驶员疲劳检测算法①

刘炜煌, 钱锦浩, 姚增伟, 焦新涛, 潘家辉

(华南师范大学 软件学院, 佛山 528225)

摘　要: 本文将卷积神经网络 (Convolutional Neural Network, CNN) 应用到视频理解中, 提出一种基于多面部特征融合的驾驶员疲劳检测算法. 本文使用多任务级联卷积网络 (Multi-Task Cascaded Convolutional Networks, MTCNN) 定位驾驶员的嘴部、左眼, 使用 CNN 从驾驶员嘴部、左眼图像中提取静态特征, 结合 CNN 从嘴部、左眼光流图中提取动态特征进行训练分类. 实验结果表明, 该算法比只使用静态图像进行驾驶员疲劳检测效果更好, 准确率达到 87.4%, 而且可以很好地区别在静态图像中很相似的打哈欠和讲话动作.

关键词: 疲劳检测; 多任务级联卷积网络; 光流; 特征融合; 计算机视觉

引用格式: 刘炜煌,钱锦浩,姚增伟,焦新涛,潘家辉.基于多面部特征融合的驾驶员疲劳检测算法.计算机系统应用,2018,27(10):177–182. http://www.c-s-a.org.cn/1003-3254/6555.html

Driver Fatigue Detection Algorithm Based on Multi-Facial Feature Fusion

LIU Wei-Huang, QIAN Jin-Hao, YAO Zeng-Wei, JIAO Xin-Tao, PAN Jia-Hui

(School of Software, South China Normal University, Foshan 528225, China)

Abstract: In this study, Convolution Neural Network (CNN) is applied to video comprehension, and a driver fatigue detection algorithm based on multi-facial feature fusion is proposed. In the study, Multi-Task Cascaded Convolutional Neural Networks (MTCNN) is used to locate the driver's mouth and left eye. CNN is used to extract the static features from the driver's mouth and left-eye image, combined with the dynamic features that CNN extracted from the mouth and left eye optical flow to train for classification. The experimental results show that this algorithm with an accuracy rate of 87.4% is better than only use the static image for driver fatigue detection and it can well distinguish between yawning and speech actions that are similar in static images.

Key words: fatigue detection; Multi-Task Cascaded Convolutional Networks (MTCNN); optical flow; feature fusion; computer vision

1　引言

根据美国国家公路交通安全管理局报告, 有 22% 到 24% 的交通事故是由驾驶员疲劳所引起的, 在驾驶途中驾驶员打瞌睡更会使发生车祸的风险提高 4 到 6 倍. 交通事故频发, 严重威胁到人们的生命财产安全, 因此, 驾驶员疲劳检测的研究有着重要意义.

目前已经有各种技术来测量驾驶员困倦. 这些技术可以大致分为三类: 基于车辆的驾驶模式、基于司机的心理生理特征、基于计算机视觉技术. 驾驶员疲劳检测近年来一直是计算机视觉领域的一个活跃的研究课题. 相较于借助脑电设备采集脑电数据[1,2]进行驾驶员疲劳检测, 它为检测驾驶员状态提供了非侵入性

① 基金项目: 国家自然科学基金青年科学基金 (61503143); 广州市科技计划项目珠江科技新星科技创新人才专项 (201710010038); 广东省自然科学基金博士科研启动项目 (2014A030310244)
Foundation item: Young Scientists Fund of National Natural Science Foundation of China (61503143); Special Project of the Pearl River S&T Nova Program of Guangzhou (201710010038); Ph. D. Research Start-up Project of Natural Science Foundation of Guangdong Provice (2014A030310244)
收稿时间: 2018-02-06; 修改时间: 2018-02-28; 采用时间: 2018-03-13; csa 在线出版时间: 2018-09-28

的机制, 对比检测车辆运行状况、方向盘状况[3,4]的方法, 计算机视觉技术有更好的检测效果. 在现有的基于计算机视觉技术的驾驶员疲劳检测技术中, 有人通过模板匹配、几何特征定眼睛、嘴巴, 计算眨眼率和嘴部动作频率作为判断疲劳驾驶的依据[5], 也有人主要针对眼镜遮挡以及光照变化, 采取级联回归定位特征点, 提出了一种更具鲁棒性的算法[6].

人脸包含了非常重要的信息. 作为驾驶员疲劳指标之一, 脸部动态地表示困倦的特征是打哈欠, 这种行为通常与大脑中缺氧有关. 在这种情况下, 人类的自然反应就是张大嘴巴, 试图呼吸更多的氧气, 这是可以用作疲劳预警的一个面部特征. 另一个面部特征是眨眼率, 眨眼率表示一段时间内眨眼的次数. 在昏昏欲睡的状态下, 一个人的眨眼率会改变, 这个特征可以用来表示疲劳水平.

基于计算机视觉技术的驾驶员疲劳检测是通过安装在仪表板上或镜子下方的低成本摄像头获得驾驶员图像, 包含驾驶员的脸部, 身体的上部, 手部, 座椅的后部或车辆的其他内部部件, 从中获取重要信息如人脸等进行判断.

本文提出了一种使用多任务级联卷积网络 (Multi-Task Cascaded Convolutional Networks, MTCNN) 提取嘴部、左眼区域, 结合嘴部、左眼区域的光流图, 使用卷积神经网络 (Convolutional Neural Network, CNN) 提取特征进行驾驶员疲劳检测的方法, 在 NTHU-DDD 数据集上取得了不俗的效果.

2 人脸检测与关键点定位

真实驾驶视频中的驾驶员嗜睡检测是具有挑战性的, 因为人脸可能受到许多因素的影响, 包括性别、面部姿势、面部表情、光照条件等, 但是车内低成本摄像头只能拍摄低分辨率视频, 因此需要一个高性能的人脸检测器. 即使有了特定的脸部, 定位嘴部、眼部区域也是非常重要的, 这些区域是驾驶员疲劳的面部特征重要区域.

基于人脸正脸的 Haar 特征的 AdaBoost 人脸检测算法[7]在实际复杂多变的环境下效果并不好, 而且无法确定眼部、嘴部区域. MTCNN[8]被称为最快和最精确的人脸检测器之一. 利用级联结构, MTCNN 可以实现联合高速化的人脸检测和对齐. 作为脸部检测和对齐的结果, MTCNN 获得了脸部边界坐标和包含左眼, 右眼, 鼻子, 左唇端和右唇端的位置的五个界标点. 本文使用 MTCNN 进行人脸检测和关键点对齐任务

MTCNN 由 3 个网络结构组成 (P-Net, R-Net, O-Net), 当给定一张照片的时候, 将其缩放到不同尺度形成图像金字塔, 以达到尺度不变. 第一阶段, 浅层的 CNN 快速产生候选窗体; 第二阶段, 通过更复杂的 CNN 筛选候选窗体, 丢弃大量的重叠窗体; 第三阶段, 使用更加强大的 CNN, 实现候选窗体去留, 同时显示五个面部关键点定位.

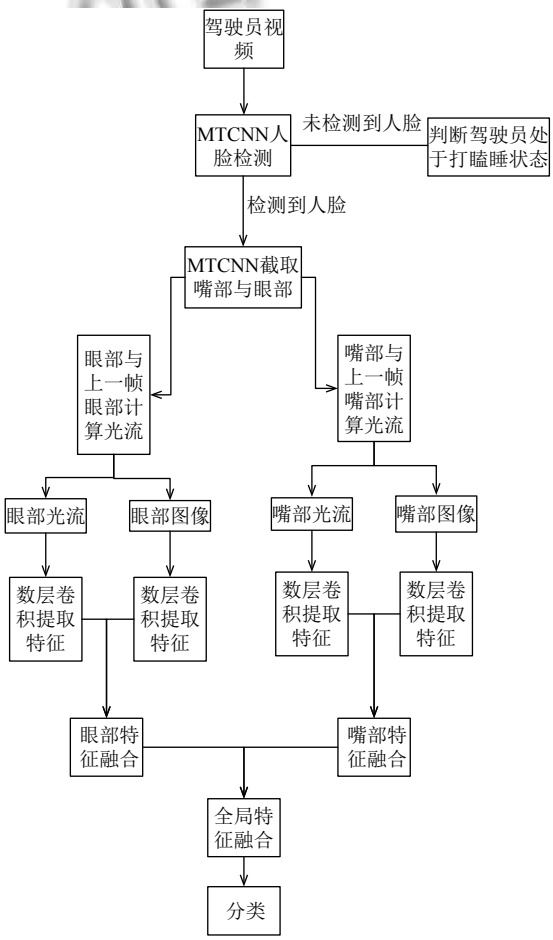


图 1 本文的算法流程图

Proposal Network (P-Net): 该网络结构主要获得了人脸区域的候选窗口和边界框的回归向量. 并用该边界框做回归, 对候选窗口进行校准, 然后通过非极大值抑制来合并高度重叠的候选框.

Refine Network (R-Net): 该网络结构还是通过边界框回归和非极大值抑制来去掉假阳性区域, 但是是由

于该网络结构和 P-Net 网络结构有差异, 多了一个全连接层, 所以会取得更好的抑制假阳性的作用.

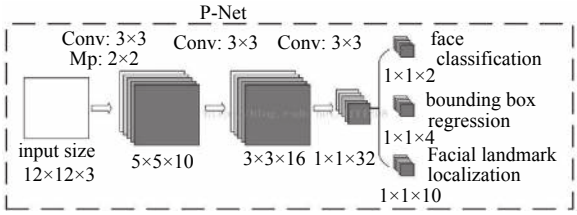


图 2 P-Net

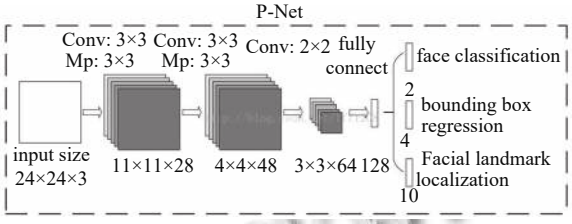


图 3 R-Net

Output Network (O-Net): 该网络结构比 R-Net 网络结构又多了一层卷积层, 所以处理的结果会更加精细. 作用和 R-Net 网络结构作用相似. 但是该网络结构对人脸区域进行了更多的监督, 同时还会得到 5 个坐标, 分别代表左眼、右眼、鼻子、左唇端和右唇端.

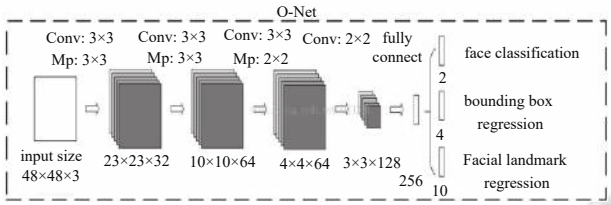


图 4 O-Net

相比于基于局部区域的卷积神经网络 (Region-based Convolutional Neural Network, RCNN) 系列通用检测方法, MTCNN 更加针对人脸检测这一专门的任务, 速度和精度都有足够的提升.

只确定出关键点的坐标是不够的, 需要确定眼部、嘴部区域. 人脸面部器官的分布遵循一定的规律. 根据“三庭五眼”规律, 人脸横向分为三个等分, 额头到眉毛是上庭, 眉毛到鼻头是中庭, 鼻头到下巴是下庭; 人脸纵向分为五个等份, 以一个眼睛长度为一等份, 两眼间距为一等分, 眼睛到太阳穴也是一个等分. 由此可知, 眼睛宽度与嘴巴宽度大致相等.

考虑到嘴部、眼部在打哈欠、眨眼等动作时大小会在一定范围内变化, 本文以眼部坐标为中心、左右唇端距离为长度, 确定一个矩形框, 作为眼部区域; 以左唇端、右唇端中点为中心、左右唇端距离为长度, 确定一个矩形框, 作为嘴部区域.

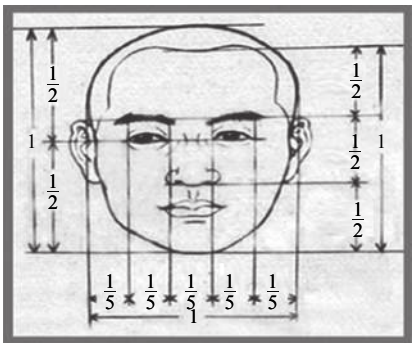


图 5 三庭五眼

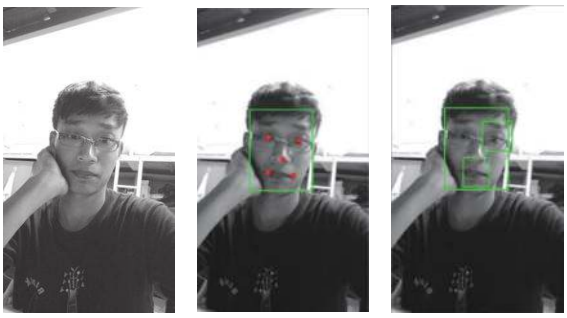


图 6 MTCNN 检测人脸与关键点定位

3 光流计算

作为驾驶员疲劳的指标, 打哈欠、眨眼等并不是一种静态状态, 而是一种动态动作, 因此只有静态的图像是不够的. 光流是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系, 包含了连续帧之间的动态信息. 不同于使用长短时记忆网络 (Long Short Term Memory Network, LSTM)、3D 卷积神经网络 (3D Convolutional Neural Networks, 3DCNN) 这类使用连续帧进行动作识别, 本文使用光流图中包含的动态信息代替连续帧所提供的动态信息. 本文将静态信息与动态信息中的特征融合, 相较于只使用静态图像, 可以更好地进行驾驶员疲劳检测.

真实的三维空间中, 描述物体运动状态的物理概念是运动场. 在计算机视觉的空间中, 计算机所接收到的信号往往是二维图片信息. 由于缺少了一个维度的

信息, 所以其不再适用以运动场描述. 光流场就是用于描述三维空间中的运动物体表现到二维图像中, 所反映出的像素点的运动向量场.

光流是空间运动物体在观察成像平面上的像素运动的瞬时速度, 是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来找到上一帧跟当前帧之间存在的对应关系, 从而计算出相邻帧之间物体的运动信息的一种方法. 假设每一个时刻均有一个向量集合(x,y,t), 表示指定坐标(x,y)在t点的瞬时速度. 设$I(x,y,t)$为t时刻(x,y)点的像素亮度, 在很短的时间Δt内, x和y分别增加$\Delta x,\Delta y$, 可得:

$$I(x+\Delta x,y+\Delta y,t+\Delta t)= I(x,y,t)+\partial I/\partial x\,\Delta x+\partial I/\partial y\,\Delta y+\partial I/\partial t\,\Delta t \tag{1}$$

同时, 考虑到两帧相邻图像的位移足够短, 即:

$$I(x,y,t)=I(x+\Delta x,y+\Delta y,t+\Delta t) \tag{2}$$

因此可得:

$$\partial I/\partial x\,\Delta x+\partial I/\partial y\,\Delta y+\partial I/\partial t\,\Delta t=0 \tag{3}$$

$$\partial I/\partial x\,\Delta x/\Delta t+\partial I/\partial y\,\Delta y/\Delta t+\partial I/\partial t\,\Delta t/\Delta t=0 \tag{4}$$

因:

$$\Delta x/\Delta t=v_x,\Delta y/\Delta t=v_y \tag{5}$$

最终可得出结论:

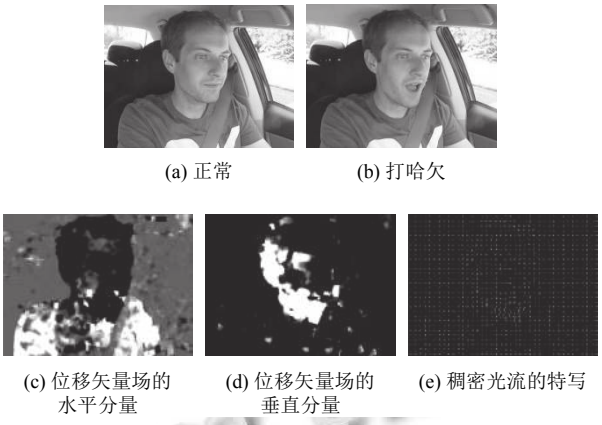
$$\partial I/\partial x\,v_x+\partial I/\partial y\,v_y+\partial I/\partial t=0 \tag{6}$$

这里的v_x, v_y是x和y的速率, 或称为$I(x,y,t)$的光流.

Farneback 算法[9]是一种计算稠密光流的方法. 它首先用二次多项式来逼近两个帧的每个邻域, 这可以用多项式展开变换来有效地完成, 然后通过观察一个精确的多项式如何在平移下进行变换, 从多项式展开系数中导出一个估算光流的方法. 通过这个稠密光流, 可以进行像素级别的图像配准, 所以其配准后的效果也明显优于稀疏光流配准的效果.

在汽车行驶途中, 由于摄像头是固定在车上的, 驾驶员脸部光流是由场景中驾驶员脸部运动所产生的, 本文通过 Farneback 算法计算前后两帧眼部、嘴部的稠密光流反映驾驶员脸部动态变化.

图 7 中, (a)、(b) 是视频中连续的两帧, 视频中驾驶员正在打哈欠; (c) 表示位移矢量场的水平分量; (d) 表示位移矢量场的垂直分量; (e) 是使用 Farneback 光流算法计算出的稠密光流的特写.



(a) 正常 (b) 打哈欠

(c) 位移矢量场的水平分量 (d) 位移矢量场的垂直分量 (e) 稠密光流的特写

图 7 Farneback 光流计算相关图像

4 疲劳检测

CNN 避免了对图像的复杂前期预处理, 可以直接输入原始图像, 以其局部连接和权值共享的特殊结构提取特征, 在语音识别和图像处理方面有独特的优越性.

视频可以分成空间与时间两个部分, 空间部分指独立帧的表面信息, 关于物体、场景等; 而时间部分信息指帧与帧之间的光流, 携带着帧与帧之间的运动信息. 参考文献[10]所提出的网络结构由两个深度网络组成, 分别处理时间与空间的维度. 将视频分帧送入第一个卷积神经网络进行训练来提取静态特征, 同时将从视频中提取出的光流图送进另外一个卷积神经网络来提取动态特征. 最终将两个网络 Softmax 层输出的分数进行一个融合.

由于在自然状态下, 人的左眼和右眼的运动状态是一致的, 因此参考文献[11]提出了一种只将人脸的嘴部区域与左眼区域输入进网络进行驾驶员嗜睡检测的算法. 与输入人脸相比, 这个算法不仅简化了输入, 而且取得了更好的效果.

本文的算法首先对驾驶员进行人脸检测. MTCNN 的三个网络均设置了一个阈值, 这个阈值代表非极大值抑制中的人脸候选窗口的重叠程度. 在设置严格的阈值的情况下, MTCNN 在驾驶员低头, 即嘴部未出现的情况下不会检测到人脸. 如果 MTCNN 未检测到人脸, 则判断驾驶员处于打瞌睡状态. 如果检测到人脸, 则将左眼、嘴部区域进行截取, 送入疲劳检测网络, 结合眼部、嘴部光流图, 判断驾驶员在非打瞌睡状态的时候, 是处于普通、讲话还是打哈欠状态. 不同

于使用 LSTM、3DCNN 从视频上截取的连续帧来进行动作识别, 本文使用 CNN 对原始图像提取静态特征、对光流图提取动态特征, 对一个短时间的动作变化进行分类.

疲劳检测网络包含四个子网, 第一个子网是左眼光流特征提取子网, 第二个子网是左眼特征提取子网, 第三个子网是嘴部光流特征提取子网, 第四个子网是嘴部特征提取子网. 经过嘴部、眼部检测后得到的嘴部、左眼区域和经过计算得到的嘴部光流图、左眼光流图对应输入进四个子网, 经过数层卷积、池化后, 首先将左眼与左眼光流图子网融合, 得到进一步的左眼区域特征, 再将嘴部与嘴部光流图子网融合, 得到进一步的嘴部区域特征, 然后融合后的两个子网再融合输入进全连接层, 得到全局区域的特征, 经过降维, 再输入 Softmax 层进行分类, 得到最终结果. 为了避免过拟合, 在每个卷积层添加了 L2 正则项, 在全连接层前添加了 Dropout 层.

5 实验结果及分析

5.1 NTHU-DDD 数据集

NTHU-DDD 数据集是台湾国立清华大学所开发的数据集, 被用于 2016 年亚洲计算机视觉会议的视频驾驶员瞌睡检测国际研讨会. 数据集在模拟驾驶场景下使用主动红外照明来采集正常驾驶, 打哈欠, 瞌睡、讲话等各种视频数据, 整个数据集 (包括训练、验证、测试数据集) 包含 36 个不同种族的戴/不戴眼镜、在白天/夜晚的照明条件下的数据. 数据集包含了很丰富的各种情景下的正常、瞌睡、讲话、打哈欠的人脸数据, 可以提高本文算法的鲁棒性.

本文将数据集分为两部分. 一部分视频具有打瞌睡动作与正常动作, 用于打瞌睡检测; 一部分视频有正常、讲话、打哈欠动作, 用于疲劳检测网络. 输入进两个网络的图片都先压缩成 50×50 大小.

图 9 中四幅示例图像均来自于实验数据, (a) 是打瞌睡图像, (b) 是正常图像, (c) 是打哈欠图像, (d) 是讲话图像.

5.2 实验结果

表 1 是几种不同的方法进行打瞌睡检测的对比实验结果, 可以看出, 采用文献[10]的方法, 使用全局图像结合光流图进行检测的效果并不好. 不对图像进行预处理, 直接输入全局图像, 输入图像会包含许多无用的信息, 这会带来噪声. 人打瞌睡的特征最重要的表现就是头会低下来, 而实际摄像头拍摄出来的 640×480 图像中驾驶员脸部只占 200×150 左右, 进行压缩后送入 CNN 的图像中携带的人脸信息将会更少, 将很难从中学习到特征. 直接以 MTCNN 是否检测到人脸作为打瞌睡检测的标准效果更好. 从表 2 实验结果可知, 在测试集中, 大部分正常状态都可以被正确检测, 相对来说打瞌睡的检测正确率稍低, 这是因为在测试集中, 有些稍稍低头的状态也被标记为打瞌睡, 这种状况是比较难以识别的.

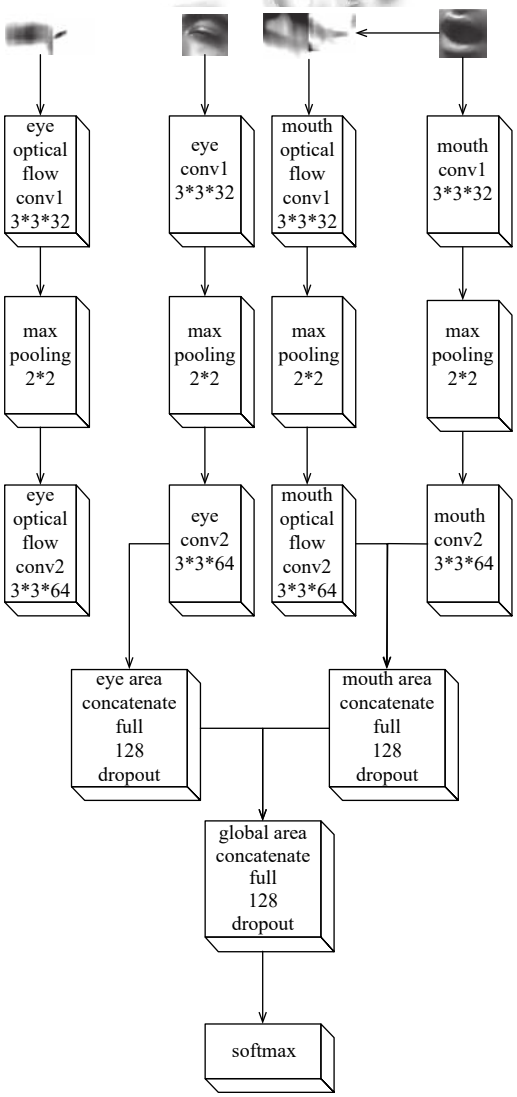


图 8 疲劳检测网络

表 3 是使用两种不同方法进行疲劳检测的对比实验结果. 采用参考文献[11]的方法, 使用嘴部以及左眼作为网络的输入, 没有使用到帧与帧之间的动态信息, 效果不如结合光流图的算法好. 对于一个短时间的动

作变化识别任务, 仅仅使用静态图像是不够准确的, 如打哈欠与讲话时, 人都是张着嘴的, 静态图像无差异, 而光流图则能体现差异. 结合包含动态信息的光流图可以获得更好地效果. 从表 4 的实验结果可知, 讲话和打哈欠这两种在静态图像中很相似的图像, 结合光流图后可以有很好的区别效果.

(a) 瞌睡 (b) 正常

(c) 打哈欠 (d) 讲话

图 9 数据集示例图像

表 1 打瞌睡检测准确率对比实验 (%)

	验证集	测试集
全局图像检测打瞌睡	74.1	68.3
全局图像结合光流图检测打瞌睡[10]	81.6	73.2
MTCNN 检测打瞌睡	93.3	90.7

表 2 MTCNN 测试打瞌睡检测结果

真实类别个数	预测类别个数	
	正常	打瞌睡
正常	1457	43
打瞌睡	236	1264

图 10 稍稍低头的打瞌睡图像

表 3 疲劳检测对比实验 (%)

	验证集准确率	测试集准确率
使用左眼、嘴部检测疲劳[11]	86.1	83.3
使用左眼、嘴部结合光流图检测疲劳	91.2	87.4

6 结束语

本文提出了一种基于多面部特征融合的驾驶员嗜睡检测技术, 不仅避免了对驾驶员身体造成侵入性, 准确率高, 而且将卷积神经网络与光流图结合应用到视频理解中, 取得了不俗的效果. 从实验结果可以看出, 本文的算法具有更高的鲁棒性和准确率, 对于各种不同情景如不同肤色、有无眼镜、不同光照等均适用, 尤其可以很好地区分在静态图像中相似的讲话、打哈欠两种动作.

表 4 使用左眼、嘴部结合光流图测试疲劳检测结果

真实类别个数	预测类别个数		
	正常	讲话	打哈欠
正常	1609	257	134
讲话	49	1801	150
打哈欠	42	124	1834

参考文献

1 Lal SK, Craig A, Boord P, *et al*. Development of an algorithm for an EEG-based driver fatigue countermeasure. Journal of Safety Research, 2003, 34(3): 321–328.

2 郭孜政, 牛琳博, 吴志敏, 等. 基于 EEG 的驾驶疲劳识别算法及其有效性验证. 北京工业大学学报, 2017, 43(6): 929–934.

3 Krajewski J, Sommer D, Trutschel U, *et al*. Steering wheel behavior based estimation of fatigue. Proceedings of the Fifth International Driving Symposium on Human Factors in Driver Assessment Training and Vehicle Design. 2009. 118–124.

4 刘军, 王利明, 聂斐, 等. 基于转向盘转角的疲劳驾驶检测方法研究. 汽车技术, 2016, (4): 42–45.

5 邬敏杰, 穆平安, 张彩艳. 基于眼睛和嘴巴状态的驾驶员疲劳检测算法. 计算机应用与软件, 2013, 30(3): 25–27, 54.

6 耿磊, 袁菲, 肖志涛, 等. 基于面部行为分析的驾驶员疲劳检测方法. 计算机工程, 2018, 44(1): 274–279.

7 Viola P, Jones M. Robust Real-time Face Detection. International Journal of Computer Vision, 2004, 57(2): 137–154.

8 Zhang KP, Zhang ZP, Li ZF, *et al*. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 2016, 23(10): 1499–1503.

9 Farnebäck G. Two-Frame motion estimation based on polynomial expansion. Bigun J, Gustavsson T. Scandinavian Conference on Image Analysis. Berlin: Springer-Verlag, 2003. 363–370.

10 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems. Washington, DC, USA. 2014. 568–576.

11 Reddy B, Kim YH, Yun S, *et al*. Real-time driver drowsiness detection for embedded system using model compression of deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA. 2017. 438–445.