基于人体关键点的分心驾驶行为识别

夏瀚笙,沈　峘,胡　委
(南京航空航天大学 能源与动力学院,江苏 南京 210016)

摘　要:驾驶员分心驾驶是造成交通事故的主要原因之一,利用车载设备识别驾驶员是否存在分心行为是当下亟须解决的问题。识别驾驶员是否存在分心行为的关键,在于正确理解驾驶员的姿态。对此,文中提出一种使用驾驶员的人体关键点位置信息来帮助卷积神经网络识别驾驶员是否分心驾驶的方法。通过加入人体关键点的位置信息,可以有效地使得卷积神经网络关注于驾驶员的姿态,减少背景信息的干扰。使用 Alpha Pose 系统获取驾驶员上半身 9 个关键点的坐标,利用高斯公式分别以每个关键点为中心生成热力图。热力图包含关键点位置的响应,离关键点越近的位置,响应值越大。在 VGG16 和 ResNet50 的基础上,探讨 8 种结构,分别将 9 张热力图和不同的特征图融合,作为下一个卷积的输入。实验结果表明,该方法在 State Farm 数据集上达到了 94.934% 的准确率,优于其他方法。

关键词:分心驾驶;人体关键点;卷积神经网络;热力图;深度学习
中图分类号:TP391.4;TP183　　**文献标识码**:A　　**文章编号**:1673-629X(2019)07-0001-05
doi:10.3969/j.issn.1673-629X.2019.07.001

Detecting Distraction of Drivers Using Human Pose Keypoints

XIA Han-sheng, SHEN Huan, HU Wei
(School of Energy and Power Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: Detecting distraction of drivers is one of the main causes of traffic accidents. Using in-vehicle equipment to identify whether the driver has distracted behavior is an urgent problem to be solved. The key to identify whether the driver has distracted behavior is to correctly understand the driver's posture. For this, we propose a method to help the convolutional neural network identify whether the driver is distracted by driving by human keypoints. By adding the position information of human keypoints, the convolutional neural network can effectively focus on the driver's attitude and reduce the interference of background information. The Alpha Pose system is used to obtain the coordinates of 9 keypoints of the driver's upper body, and Gauss formula is used to generate the heat map with each keypoint as the center. The heat map contains the response of the keypoints. The closer to the keypoints, the higher the response value. On the basis of VGG16 and ResNet50, 8 structures are discussed, and 9 heat maps and different characteristic graphs are respectively fused as the input of the next convolution. The experiment shows that the proposed method has an accuracy rate of 94.934% in the State Farm Dataset, which is better than other methods.

Key words: driver distraction; human pose keypoints; convolutional neural network; heat maps; deep learning

1　概　述

近年来,随着国内汽车保有量的增长,交通事故发生的频率也在逐年增加。其中,驾驶员在驾驶过程中注意力不集中是导致交通事故发生的主要原因之一。驾驶员注意力不集中主要有两个方面,分别是疲劳驾驶和分心驾驶。疲劳驾驶是驾驶员在感到疲倦的情况下仍然驾驶,驾驶员没有足够的精神状态。与疲劳驾驶不同,驾驶员在分心驾驶时,仍然具有良好的精神状态,但是忙于其他事情,例如打电话,发短信,喝水等。如果驾驶员的驾驶状态可以由车载设备检测到,并及时地提醒驾驶员注意安全,则可以很好地避免事故的发生。

多年来,对驾驶员异常行为识别的研究一直是个热门的方向。早期的研究[1-5]主要集中在识别驾驶员

收稿日期:2018-08-22　　**修回日期**:2018-12-26　　**网络出版时间**:2019-03-21
基金项目:航空科学基金(20120952022)
作者简介:夏瀚笙(1993-),男,硕士,研究方向为计算机视觉、姿态估计和人脸识别;沈　峘,博士,硕导,通信作者,研究方向为数字图像处理、计算机视觉与检测识别技术。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190321.0904.018.html

是否疲劳驾驶。采用的方式是要求驾驶员佩戴传感器以获得驾驶员的生理信息,例如血压,心率和脑电波等。这些方法成本较高并且准确率低。更重要的是,侵入式的研究方法,会对驾驶员的驾驶行为造成干扰。随着计算机视觉技术的发展,越来越多的研究人员围绕计算机视觉的方法展开研究。非侵入式的,不会影响驾驶员的驾驶体验,同时成本大大降低。可利用安装在车辆仪表盘上的相机来收集驾驶员的图像,然后通过图像的分析,判断驾驶员状态是否出现异常。

文献[6]通过结合 AdaBoost 与核相关滤波算法进行人脸检测及跟踪,然后采用级联回归方法定位特征点,提取眼睛和嘴巴部分,再利用卷积神经网对眼睛和嘴巴状态进行识别,从而进行疲劳驾驶识别。文献[7]基于肤色模型对驾驶人人脸进行检测,然后利用基于 PCA 的人脸局部特征来识别驾驶员特征,判断驾驶员是否有疲劳驾驶。

相比于疲劳驾驶,分心驾驶则更为常见,近年来,有关分心驾驶行为的研究逐渐增多。文献[8]使用 Faster R-CNN 网络[9]来检测驾驶员的双手位置,并判断驾驶员手中是否拿着手机。文献[10]建立以径向基为核函数的驾驶人分心状态判别 SVM 模型,采用遗传算法(GA)优化 SVM 模型惩罚参数 C 和核函数参数 g。文献[11]提出了基于反向双目的驾驶状态检测方法。根据 Hough 算法进行车道线检测和识别,计算车辆偏航率;同时采用多点透视算法对驾驶员头部姿态进行估计;再建立基于高斯隶属度函数模糊判断规则,根据车辆偏航率与驾驶员头部姿态对驾驶员驾驶状态进行识别。

在著名的机器学习竞赛网站 Kaggle 上,State Farm 公司举办了一个关于分心驾驶识别的竞赛。参加比赛的队伍中,toshi-k 队伍利用检测器检测出驾驶员的身体轮廓,然后按照检测出来的身体轮廓将图片裁剪,再利用深度卷积网络进行识别。Soteria 公司提出的 Soteria 系统将 VGG16 网络[12]的全连接层改为全局均值池化(global average pooling,GAP)以减少网络参数,同时防止因训练数据不足而导致过拟合。

图 1 为 State Farm 数据集的驾驶员行为示例。可以观察到,当驾驶员手握方向盘,同时正视前方时,他很大可能正专注于驾驶,如图 1(a)所示;当驾驶员低着头,并且稍微抬起胳膊,他很有可能是在看手机、发消息,如图 1(b)所示;当驾驶员抬起胳膊,靠近耳朵,他很有可能是在打电话,如图 1(c)所示;当驾驶员抬起手靠近嘴巴时,他很有可能是在喝水或饮料,如图 1(d)所示。

因此,正确地识别驾驶员是否存在分心行为,关键在于卷积神经网络理解驾驶员的姿态行为。为此,文中提出一种使用驾驶员的人体关键点位置信息来帮助卷积神经网络识别驾驶员分心驾驶的方法。首先介绍 Alpha Pose 系统,并利用该系统获取驾驶员的人体关键点坐标位置。然后根据驾驶员上半身的 9 个关键点坐标,利用高斯公式生成 9 张关键点的热力图。最后基于 VGG16 网络和 ResNet50 网络[13],探讨 8 种将热力图和卷积层的输出特征融合的方式。

(a)安全驾驶　(b)右手发信息

(c)右手打电话　(d)喝水或饮料

图 1　驾驶员行为示例

2　Alpha Pose 系统

2.1　系统简介

Alpha Pose 系统是由上海交通大学提出的一个开源的多人姿态估计系统,如图 2 所示。

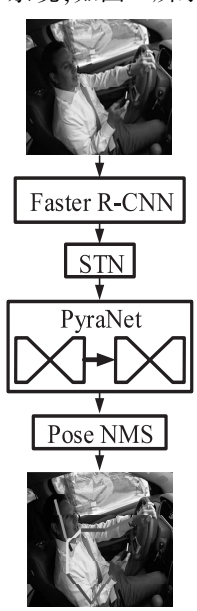


图 2　Alpha Pose 系统

Alpha Pose 系统采用的是自上而下的形式,由四部分组成,即 Faster RCNN 网络、空间变换网络(spatial transformer network,STN)[14]、PyraNet[15]以及姿态极大值抑制模块。Faster RCNN 用于检测人的位置,输出的是人的位置框坐标。空间卷积网络用来调整 Faster RCNN 得到人的位置框。PyraNet 是 Stacked

Hourglass 网络[16]的改进版,用来获取人体关键点的坐标。姿态极大值模块用来消除 Faster RCNN 检测出来的多余人体框位置,获取最终的人体关键点坐标。

通过 Alpha Pose 系统,可以精确地得到输入图片中人物的人体关键点的坐标位置。分心驾驶行为主要依据的是驾驶员上半身的姿态,因此,文中只使用 Alpha Pose 系统获得驾驶员上半身的 9 个关键点位置,分别是:头顶、颈部、胸膛、左肩、右肩、左肘、右肘、左腕、右腕。

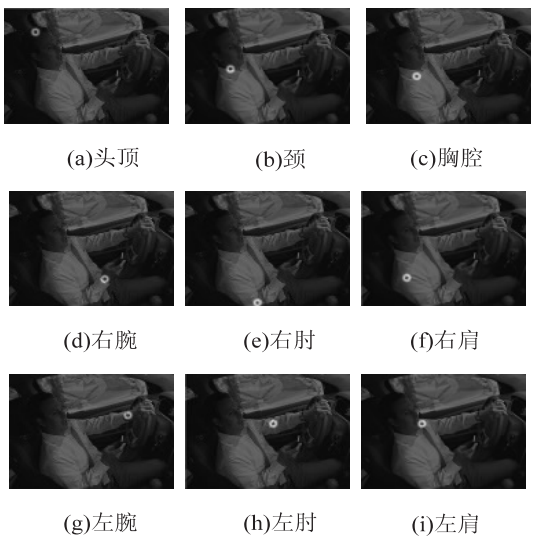
2.2　热力图

为了将关键点的位置信息融入到卷积网络,需要根据 9 个关键点的坐标位置,由高斯公式生成 9 张热力图。高斯公式如下所示:

$$\text{Response} = \exp(-((x-i)^2 + (y-j)^2)/2\sigma^2) \tag{1}$$

其中,x,y 是热力图中每个像素点的坐标;i,j 是对应的关键点的坐标;σ 是关键点响应的范围。热力图中所有像素点的值在 [0, 1] 之间,距离关键点的位置越近,响应值越大。

根据高斯公式生成的 9 张热力图,如图 3 所示。



(a)头顶　(b)颈　(c)胸腔

(d)右腕　(e)右肘　(f)右肩

(g)左腕　(h)左肘　(i)左肩

图 3　热力图

3　基于人体关键点的分心驾驶行为识别

3.1　网络结构设计

在深度卷积网络中,随着卷积层数的增加,卷积层学习到的特征也逐渐从低阶发展到高阶。例如,在常见的分类卷积网络中,网络的第一个卷积层学习到的可能是边、角、曲线等低阶特征,而第二个卷积层学习的则是第一个卷积层输出的低阶特征的组合,如半圆、矩形等。因此,将姿态信息(即热力图)和卷积网络的不同层的特征图进行融合,产生的效果也不一样。文中以 VGG16 网络和 ResNet50 网络为基础网络,尝试 8 种结构将热力图融合到不同的卷积输出层中,以获得最好的实验效果。VGG16 网络结构和 ResNet50 网络结构如图 4 所示,实验中需要将最后一个全连接层的输出数改为 10。

输入:224×224(RGB)	
Conv1.	conv3-64　conv3-64
Pool1(Max)	
Conv2.	conv3-128　conv3-128
Pool2(Max)	
Conv3.	conv3-256　conv3-256　conv3-256
Pool3(Max)	
Conv4.	conv3-512　conv3-512　conv3-512
Pool4(Max)	
Conv5.	conv3-512　conv3-512　conv3-512
Pool5(Max)	
FC	4 096
FC	4 096
FC	1 000

(a)VGG16 网络

输入:224×224(RGB)	
Conv1.	conv7-64
Pool1(Max)	
Conv2.	[conv1-64 cov3-64 conv1-256] ×3
Conv3.	[conv1-128 conv3-128 conv1-512] ×4
Conv4.	[conv1-256 conv3-256 conv1-1 024] ×6
Conv5.	[conv1-512 conv3-512 conv1-2 048] ×3
Pool2(Average)	
FC	1 000
Softmax	

(b)ResNet50 网络

图 4　VGG16 网络模型和 ResNet50 网络模型

图 4 中,Conv1. 表示网络第一阶段,Conv2. 表示网络第二阶段,以此类推。Conv3-64 表示卷积操作,其中卷积核大小为 3 × 3,输出通道数为 64,以此类推。Pool 表示池化操作,其中 Max 是最大值池化,Average 是均值池化。FC 表示全连接层。

基于 VGG16 的 4 种结构分别是:

结构 a:9 张热力图和 3 通道的 RGB 图像串接,通道数变为 12,作为 Conv1. 的输入。

结构 b:9 张热力图经过 1 × 1 的卷积,输出尺寸为 112 × 112、通道数为 64 的特征图,和相同尺寸的 Pool1 的输出特征图相加,作为 Conv2. 的输入。

结构 c:9 张热力图经过 1 × 1 的卷积,输出尺寸为 56 × 56、通道数为 128 的特征图,和相同尺寸的 Pool2 的输出特征图相加,作为 Conv3. 的输入。

结构 d:9 张热力图经过 1 × 1 的卷积,输出尺寸为 28 × 28、通道数为 256 的特征图,和相同尺寸的 Pool3 的输出特征图相加,作为 Conv4. 的输入。

基于 ResNet50 的 4 种结构分别是:

结构 e:9 张热力图和 3 通道的 RGB 图像串接,通道数变为 12,作为 Conv1. 的输入。

结构 f:9 张热力图经过 1 × 1 的卷积,输出尺寸为 112 × 112、通道数为 64 的特征图,和相同尺寸的 Conv1. 的输出特征图相加,作为 Pool1 的输入。

结构 g:9 张热力图经过 1 × 1 的卷积,输出尺寸为 56 × 56、通道数为 64 的特征图,和相同尺寸的 Pool1 的输出特征图相加,作为 Conv2. 的输入。

结构 h:9 张热力图经过 1 × 1 的卷积,输出尺寸为 56 × 56、通道数为 256 的特征图,和相同尺寸的 Conv2. 的输出特征图相加,作为 Conv3. 的输入。

3.2　损失函数

损失函数采用的是 Softmax 损失,如式 2 所示:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(W_{y_i}^{\mathrm{T}}x_i + b_{y_i})}{\sum_{j=1}^{n}\exp(W_j^{\mathrm{T}}x_i + b_j)} \tag{2}$$

其中,N 是批量的大小,n 是样本类别数;x_i 表示第 i 个样本的特征向量,y_i 为其对应的标签;W_j 是类别 j 类别对应的权值,b_j 是类别 j 对应的偏置,W_{y_i} 是类别 y_i 对应的权值,b_{y_i} 是类别 y_i 对应的偏置。

4　实　验

4.1　State Farm 数据集

State Farm 是 State Farm 公司在 Kaggle 上发布的一个竞赛数据集。它包含 81 个驾驶员,共 102 150 张图。所有的图片尺寸都是 640 × 480 像素。驾驶员的行为分为 10 个类别,分别是:安全驾驶、左手发信息、右手发信息、左手打电话、右手打电话、调收音机、喝水或饮料、向后拿东西、化妆或者抓耳挠腮、和乘客说话。

该数据集的图片都是从视频上截取的视频帧,同一个驾驶员所对应的图片高度相关。因此为了验证实验的准确性,训练和测试数据需要按照驾驶员来分,同一个驾驶员对应的所有照片只能是在训练集和测试集中选其一。

文中选择 26 个驾驶员对应的照片用做测试集,约占总的图片数的 22%,剩下的 55 个驾驶员对应的照片作为训练集,约占总的图片数的 78%。

4.2　数据预处理及训练参数

文中实验均在 Caffe 框架上进行。训练过程中,所有图片首先被缩放到 224 × 224 尺寸。然后做数据增广,包括随机旋转(最大旋转角 30 度),随机水平翻转。

训练参数上,每一批量的训练样本数为 128,动量为 0.9,权重衰减为 0.000 5,使用在 ImageNet 上训练好的 VGG16 模型进行微调,总共训练 15 个 epoch,初始学习率为 0.01,分别在 5 个 epoch 和 10 个 epoch 的时候下降一次,下降因子为 0.1。

4.3　实验结果

8 种网络结构的实验结果如表 1 所示。

表 1　4 种网络结构的实验结果对比

网络结构	准确率/%
原始 VGG16	83.357
结构 a	80.461
结构 b	81.132
结构 c	87.142
结构 d	84.264
原始 ResNet50	86.236
结构 e	83.234
结构 f	84.102
结构 g	94.934
结构 h	90.239

将 Toshi-k 的方法和 Soteria 系统的方法在同样的数据集上进行训练和测试,结果如表 2 所示。

表 2　不同方法在相同数据集上的实验结果

网络结构	准确率/%
原始 VGG16	83.357
原始 ResNet50	86.236
Soteria 系统	83.158
Toshi-k	91.973
结构 c	87.142
结构 g	94.934

在 VGG16 网络结构中,结构 a 和结构 b 相比于原始的 VGG16 网络不但没有提升,反而有所下降。结构 c 的提升最明显,结构 d 略有提升。原因在于人体关键点的信息属于高阶特征,在 VGG16 网络结构中,如果直接将热力图信息和原图或者 Pool1 层输出的特征图进行融合,反而会对后面网络的特征学习造成干扰。而到 Pool3 时,由于输出的通道数有 256 个,远大于热力图的 9 个通道数,同时特征图的变小,导致姿态信息

对网络的帮助效果很小。

同样,在 ResNet50 网络结构中,结构 e 和结构 f 相比于原始的 ResNet50,性能下降。结构 g 性能提升明显,结构 h 性能略有提升。因为直接将热力图信息和原图或者 Conv1. 阶段输出的特征图进行融合,会对后面网络的特征学习造成干扰。而到 Conv2. 阶段,虽然输出的特征图尺寸没变,但是通道数有 256 个,减弱了姿态信息对网络的效果提升。

可以看出,文中提出的结构 g 性能要优于 Soteria 系统和 Toshi-k 方法。

5　结束语

由于驾驶员是否出现分心驾驶行为和驾驶员的姿态密切相关,因此文中提出通过在 VGG16 网络中添加驾驶员的姿态信息,来帮助识别驾驶员的分心行为。为了验证方法的有效性,在 State Farm 数据集上进行了实验验证,尽管实验结果比较理想,但是文中的工作仍有一些不足之处。首先,训练的样本量较少,没有充分利用深度网络的学习能力,特别是有全连接层的网络,在训练的时候需要谨慎调参,防止过拟合;其次,数据集拍摄的角度是在驾驶员的右侧,左边有时遮挡很严重,影响了识别的效果。因此,接下来的工作可以围绕增加数据量以及数据的复杂程度,尝试 ResNet101 等更为深层的网络来提高识别的效果。

参考文献:

[1] CRAYE C,RASHWAN A,KAMEL M S,et al. A multimodal driver fatigue and distraction assessment system[J]. International Journalof Intelligent Transportation Systems Research,2016,14(3):173-194.

[2] SAHAYADHAS A,SUNDARAJ K,MURUGAPPAN M,et al. A physiological measures-based method for detecting inattention in drivers using machine learning approach[J]. Biocybernetics & Biomedical Engineering,2015,35(3):198-205.

[3] 彭军强,吴平东,殷　罡. 疲劳驾驶的脑电特性探索[J]. 北京理工大学学报,2007,27(7):585-589.

[4] 李君羡,潘晓东. 基于脑电分析的连续驾驶疲劳高发时间判断[J]. 交通科学与工程,2012,28(4):72-79.

[5] 薛　雷. 考虑驾驶员生物电信号的疲劳驾驶检测方法研究[D]. 长春:吉林大学,2015.

[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[J]. 计算机工程,2018,44(1):274-279.

[7] 才　博. 基于人脸识别驾驶员疲劳检测系统设计与开发[D]. 大连:大连理工大学,2016.

[8] LE T H N,ZHENG Y,ZHU C,et al. Multiple scale faster-RCNN approach to driver's cell-phone usage and hands on steering wheel detection[C]//IEEE conference on computer vision and pattern recognition workshops. Las Vegas,NV,USA:IEEE,2016:46-53.

[9] REN Shaoqing,HE Kaiming,GIRSHICK R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,39(6):1137-1149.

[10] 张　辉,钱大琳,邵春福,等. 驾驶人分心状态判别支持向量机模型优化算法[J]. 交通运输系统工程与信息,2018,18(1):127-132.

[11] 王　冠,李振龙. 基于反向双目识别的驾驶员分心检测[J]. 科学技术与工程,2018,18(17):82-88.

[12] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//International conference on learning representations. [s. l.]:[s. n.],2015:44-54.

[13] HE Kaiming,ZHANG Xiangyu,REN Shaoqing,et al. Deep residual learning for image recognition[C]//Computer vision and pattern recognition workshops. [s. l.]:IEEE,2016:770-778.

[14] JADERBERG M,SIMONYAN K,ZISSERMAN A. Spatial transformer networks[C]//Advances in neural information processing systems. [s. l.]:[s. n.],2015:2017-2025.

[15] YANG Wei,LI Shuang,OUYANG Wanli,et al. Learning feature pyramids for human pose estimation[C]//IEEE international conference on computer vision. [s. l.]:IEEE,2017:1290-1299.

[16] NEWELL A,YANG Kaiyu,DENG Jia. Stacked hourglass networks for human pose estimation[C]//European conference on computer vision. [s. l.]:Springer,2016:483-499.