控制与决策
Control and Decision

基于MobileNetV3与ST-SRU的危险驾驶姿态识别

赵俊男, 佘青山, 穆高原, 吴秋轩, 席旭刚

引用本文:
赵俊男,佘青山,穆高原,吴秋轩,席旭刚. 基于MobileNetV3与ST-SRU的危险驾驶姿态识别[J]. 控制与决策, 2022, 37(5): 1320-1328.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.1144

基于MobileNetV3与ST-SRU的危险驾驶姿态识别

赵俊男，佘青山†，穆高原，吴秋轩，席旭刚

(杭州电子科技大学 自动化学院，杭州 310018)

摘 要: 针对危险驾驶行为引起的交通安全事故频发的现状,提出一种基于MobileNetV3和ST-SRU的危险驾驶姿态识别系统. 首先,修改MobileNetV3的网络结构使其适用于人体姿态估计任务,输出关节点的热力图和偏移量图,用来估计J个关节点的二维坐标位置;其次,定义ST-SRU骨架动作识别算法,利用动作的骨架序列数据对动作进行分类. 实验结果表明: MobileNetV3姿态估计算法在自建的AI Challenger上肢姿态数据集上测得PCP值(percentage correct parts)达到95.6 %,测试1 000次用时仅为5.03 s;利用自建的危险驾驶行为数据集将训练好的姿态估计和动作识别模型移植到嵌入式平台,实现了实时的危险驾驶姿态识别系统.

关键词: MobileNetV3；人体姿态估计；骨架动作识别；ST-SRU；危险驾驶姿态识别

中图分类号: TP391.4；TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1144 **开放科学(资源服务)标识码(OSID):**

引用格式: 赵俊男,佘青山,穆高原,等. 基于MobileNetV3与ST-SRU的危险驾驶姿态识别[J]. 控制与决策, 2022, 37(5): 1320-1328.

Dangerous driving pose recognition based on MobileNetV3 and ST-SRU

ZHAO Jun-nan, SHE Qing-shan†, MU Gao-yuan, WU Qiu-xuan, XI Xu-gang

(College of Automation，Hangzhou Dianzi University，Hangzhou 310018，China)

Abstract: In the face of frequent traffic accidents caused by dangerous driving behaviors, this paper proposes a dangerous driving pose recognition system based on MobileNetV3 and ST-SRU. Firstly, the network structure of MobileNetV3 is modified to be used for human pose estimation, and the heatmaps and offsets of joint points are output to estimate the 2D coordinate positions of J joint points. Then, the ST-SRU skeleton action recognition algorithm is defined, and the actions are classified by using skeleton sequence data. The experimental results show that the PCP (percentage correct parts) of MobileNetV3 pose estimation algorithm is 95.6 % on the self-built AI Challenger upper limb attitude dataset, and the time of 1 000 tests is only 5.03 seconds. By using the self-built dangerous driving behavior dataset, the trained pose estimation and action recognition model is transplanted to the embedded platform, and the real-time dangerous driving pose recognition system is realized.

Keywords: MobileNetV3；human pose estimation；skeleton action recognition；ST-SRU；dangerous driving pose recognition

0 引 言

据世界卫生组织统计,因为车祸导致的死亡数每年大概会达到125万之多,相当于现在的每小时约有146人丧命于车祸,而其中90.3 %的车祸是由于人为因素导致的[1]. 例如,驾驶者会在开车过程中进行发短信、看消息、进食等危险的姿势行为. 为了减少该现象的发生,危险驾驶姿态识别系统已成为该问题的研究热点.

引发交通事故的因素多种多样,其中人的因素占主导地位[2]. 文献[3]对482名实验者进行了为期两个月的相关数据收集,调查结论显示,发短消息、查看消息和拿手机是3种最常见的危险驾驶行为;文献[4]利用调查问卷和借助驾驶模拟实验平台等方法对驾驶行为进行综合分析,得到典型的危险驾驶行为:看车内电子设备、进食、与他人说话、玩手机、抽香烟[5].

随着道路监控和测速雷达等监管设备的普及,出现了很多基于路测设备的危险驾驶行为识别应用方案. 比如基于图像分析、基于光学字符识别等人工智能技术可以实现对驾驶人的打电话、吃东西等危险驾驶行为进行监测. 但是,基于道路监控设备的危险驾驶行为识别方案存在以下3个问题:

1) 普及率低. 路测监控设备的部署一般在车流

收稿日期: 2020-08-17；**录用日期:** 2021-03-03.

基金项目: 国家自然科学基金项目(61871427)；浙江省重点研发计划项目(2019C04018).

†通讯作者. E-mail: qsshe@hdu.edu.cn.

量密集路段,然而司机经过这些路段很容易加倍小心. 反而是在部署了监控设备的低车流量路段,司机会存在侥幸心理,危险驾驶行为的比例更高.

2) 交互性差. 路测设备有识别危险驾驶行为的能力,但无法及时将危险信息与驾驶员进行交互,无法及时预警和规避危险,很大程度只能作为发生危险后的视频证据,因此仍然存在很大的危险性.

3) 维护成本高. 路测设备往往是露天的,常年经历高温、霜冻、雨雪天气等,必须时常对设备进行检修维护,才能维持监测设备的有效性,这是一项不小的人力、财力、物力的支出.

目前,危险驾驶行为识别的研究主要是基于车载设备. 从交通视角来看,危险驾驶行为除了打电话、喝水等,还包括驾驶人行为特性的跟驰、超车等. 因此,本文将具体研究驾驶行为中的驾驶人姿态识别方面的工作. 采集驾驶员数据的设备主要有单目摄像头和Kinect深度相机. 基于单目摄像头的设备往往只能对头部的姿态情况进行识别,不会对其他身体部位的姿态进行关注. 而利用Kinect相机可以直接得到人体关节点的3D运动信息,但该设备的成本高、尺寸大,与移动设备及应用的便携和低成本特点不符合.

近几年,众多研究人员将深度学习与轻量级移动设备结合,提出了不少非常有实践意义的算法. 2017年,谷歌提出了轻量级卷积神经网络MobileNets[6]算法,是用于移动和嵌入式视觉应用的卷积神经网络模型;2019年,该团队提出了性能更加先进的轻量级卷积神经网络MobileNetV3模型[7];2018年,文献[8]提出了一种相较于长短期记忆网络(LSTM)运行速率更快、计算复杂度更低的简单循环单元(simple recurrent units, SRU),在部分领域工作中可以更快速地完成LSTM能完成的任务,比如语义识别、预测任务等.

本文旨在结合MobileNetV3和SRU的运行速度快、效率高的特点,研究基于MobileNetV3的人体关节点特征提取,获取驾驶员上肢的关节点的坐标信息[5];设计处理骨架序列的ST-SRU模型,根据骨节点2D坐标数据,对危险驾驶姿态进行分类,最终将设计的算法通过深度学习训练获取相应的模型,并移植至移动硬件设备,设计一种嵌入式轻量级危险驾驶姿态识别系统[5]. 本文的创新点可以分为以下3个方面:

1) 大多数深度学习的CNN模型的参数量大、模型层数多,以此为基本网络结构进行人体姿态估计训练所获得的模型几乎不能在移动硬件设备上移植运行. 故本文提出以MobileNetV3为基础来改进的姿态估计方法,大幅减少参数量,加快模型的推理速度.

2) 动作识别算法部分以SRU单元为基础,加入时空概念提出的ST-SRU模型,在加快推理速度的同时保持识别的高效性.

3) 现有行为识别硬件设备因价格高和便携性差等问题,难以在市场上大面积推广使用,故本文设计一种使用单目摄像头的嵌入式轻量级危险驾驶姿态识别系统[5],以低价的嵌入式平台为载体,实现实时的危险驾驶姿态识别,具有很好的实用性和普及性.

1　基于MobileNetV3和ST-SRU的姿态识别

图1即为本文提出的危险驾驶姿态识别系统. 图1中:$x_{j,t}$为骨架序列中第t帧的第j个关节的2D坐标[5],$t \in \{1,2,\ldots,T\}, j \in \{1,2,\ldots,8\}$.

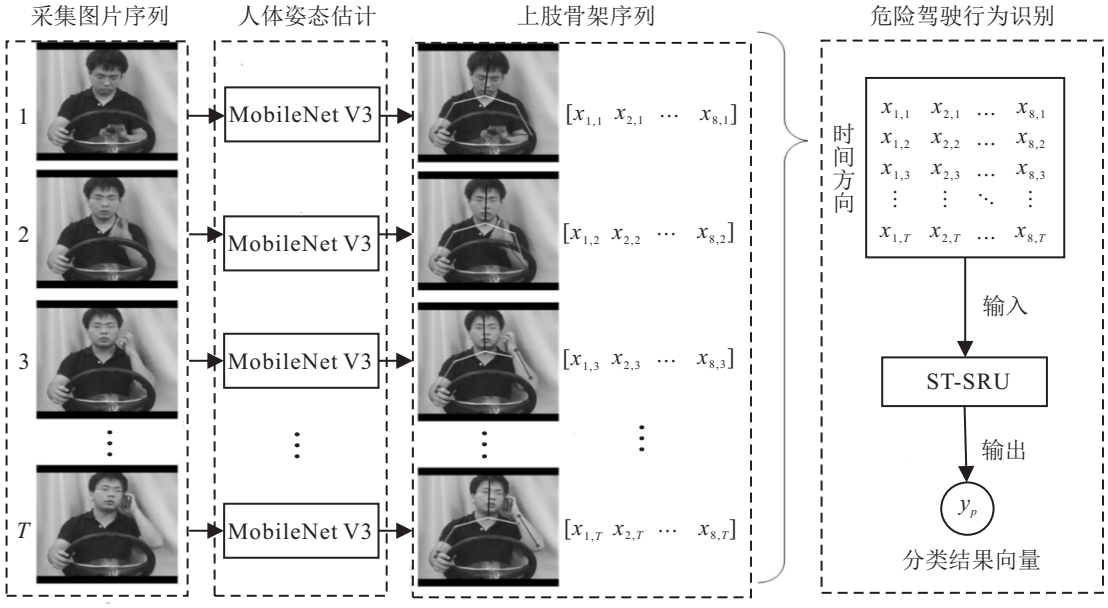

图1　危险驾驶姿态识别算法流程

1.1 基于MobileNetV3的姿态估计

近年来,在人体姿态估计领域表现较为优秀的算法有堆叠沙漏网络(stacked hourglass network)[9]和卷积姿态机(convolutional pose machines, CPM)[10],这两种都是基于热力图回归的人体姿态估计方法. 本文也将采用热力图回归的方法估计关节点的位置,通过对MobileNetV3的网络结构进行修改,建立一种用于人体姿态估计的CNN模型[5].

1.1.1 姿态估计模型输出

姿态估计模型的输入是规格大小为224×224的RGB图片. 本文假设估计的关节点数量为J,输出部分分成热力图和偏移量图. 其中:输出热力图的尺寸为224×224,一共J张;输出的偏移量图的尺寸为224×224,共$2J$张,包括X、Y方向的偏移量图各有J张. 这是参考了文献[11]提出的方法.

获得模型输出数据后,将通过热力图和偏移量图估计每个关节点的坐标. 第j个关节点的估计坐标可以通过下列的公式计算得到:

$$mx_j, my_j = \arg\max(h_j(x, y)), \tag{1}$$

$$px_j = mx_j + ox_j(mx_j, my_j), \tag{2}$$

$$py_j = my_j + oy_j(mx_j, my_j). \tag{3}$$

其中:h_j为第j张热力图;(mx_j, my_j)为第j张热力图像素值最大的坐标,说明该坐标最接近第j个关节点. ox_j和oy_j分别为第j张X和Y方向上的偏移量图. (mx_j, my_j)只能作为大概的估计坐标值,式(2)和(3)将估计的偏移量与估计的坐标值相加,获得了效果更好的关节点的估计坐标值. 为了最终获得较为精确的关节点2D估计坐标数据,将获得的热力图与偏移量图的数据进行融合. 图2即为完整的人体姿态估计算法的步骤流程.

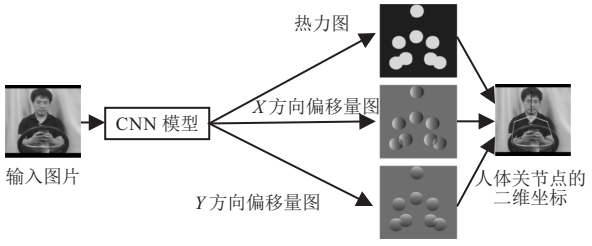

图 2 人体姿态估计算法流程

1.1.2 训练样本的标签

现有的人体姿态估计数据集提供的是以每个人体关节点的2D坐标为数据集的标签. 1.1.1节已经给出了人体姿态估计算法的流程,算法要完成训练必须有热力图及偏移量图两种类型的标签. 人体姿态数据集一般都会提供人体的J个关节点的坐标信息,根据这些坐标信息生成J张热力图标签和$2J$张偏移量图标签,尺寸大小也为224×224. 通过以下公式得到数据集的新标签.

1) 热力图标签

$$\mathrm{lh}_j(x, y) = \begin{cases} 1, \ \sqrt{(x - \mathrm{lx}_j)^2 + (y - \mathrm{ly}_j)^2} \leqslant R; \\ 0, \ \text{otherwise}. \end{cases} \tag{4}$$

其中:$\mathrm{lh}_j(x, y)$为第j个热力图标签在位置(x, y)上的值;热力图的标签是二值置信图,点(x, y)在以$(\mathrm{lx}_j, \mathrm{ly}_j)$为圆心且$R$为半径的圆内则像素值为1,不在范围内则为0. 用于训练的样本提供有8个关节的2D坐标信息,共生成了8张热力图标签.

2) 偏移量图标签

$$\mathrm{lox}_j(x, y) = \begin{cases} \mathrm{lx}_j - x, \ \sqrt{(x - \mathrm{lx}_j)^2 + (y - \mathrm{ly}_j)^2} \leqslant R; \\ 0, \ \text{otherwise}. \end{cases} \tag{5}$$

$$\mathrm{loy}_j(x, y) = \begin{cases} \mathrm{ly}_j - y, \ \sqrt{(x - \mathrm{lx}_j)^2 + (y - \mathrm{ly}_j)^2} \leqslant R; \\ 0, \ \text{otherwise}. \end{cases} \tag{6}$$

其中:$\mathrm{lox}_j(x, y)$为第j个X方向偏移量图标签在坐标(x, y)的值,$\mathrm{loy}_j(x, y)$为第j个Y方向偏移量图标签在坐标(x, y)的值. 根据算法训练的需求,要在X、Y两个方向上给各个骨节点生成偏移量图,样本提供的是8个骨节点的2D坐标信息,则需要生成16张偏移量图标签.

1.1.3 损失函数

本文假设姿态估计模型生成的热力图h, X、Y方向上偏移量图o_x和o_y,其相应的训练样本的标签为$\bar{h}$、$\bar{o}_x$和$\bar{o}_y$,并且它们的维度都是$224 \times 224 \times J$. 关于热力图预测可视为二分类问题,本文引入的热力图损失函数用交叉熵损失函数定义,表现形式如下:

$$\mathrm{hs} = \sigma(h), \tag{7}$$

$$F_h = -\bar{h} \times \log(\mathrm{hs} + d) - (1 - \bar{h}) \log(1 - \mathrm{hs} + d), \tag{8}$$

$$L_h(\theta) = \sum_{x=1}^{224} \sum_{y=1}^{224} \sum_{j=1}^{J} F_h(x, y, j). \tag{9}$$

其中:σ为sigmoid激活函数,通过sigmoid激活函数将热力图h的每个点的值映射到区间$[0, 1]$;以防传递给log函数的参数是零,将d设置为一个极小值;F_h为hs与$\bar{h}$的每个像素点间的交叉熵值,它的维度

是 $224 \times 224 \times J$; $L_h(\theta)$ 即为最终定义的热力图损失函数.

在 X、Y 方向的偏移量图预测可视为回归问题,其损失函数使用最小平方误差定义,即

$$F_x = \text{square}(\bar{o}_x - o_x \times \bar{h}), \tag{10}$$

$$F_y = \text{square}(\bar{o}_y - o_y \times \bar{h}), \tag{11}$$

$$F_o = (F_x + F_y)/2, \tag{12}$$

$$L_o(\theta) = \sum_{x=1}^{224}\sum_{y=1}^{224}\sum_{j=1}^{J} F_o(x,y,j). \tag{13}$$

其中: square( )表示获取矩阵的每个元素的平方;根据1.1.2节中热力图标签的介绍可知,$\bar{h}$ 在关节点附近则值是1,否则为0,这样 $o_x \times \bar{h}$、$o_y \times \bar{h}$ 就忽略了 o_x、o_y 在 $\bar{h}$ 相应的位置等于0上的值; F_x 为 $o_x \times \bar{h}$ 和 $\bar{o}_x$ 相应位置元素的差值取平方; F_y 为 $o_y \times \bar{h}$ 和 $\bar{o}_y$ 相应位置元素的差值取平方; $L_o(\theta)$ 即为最终定义的偏移量图损失函数,为 F_o 所有位置元素之和,相当于 $\bar{o}_x$、$\bar{o}_y$ 的最小平方误差的平均值.

这两部分的损失函数以一定比例融合,形成本文所需的损失函数

$$L(\theta) = \lambda_h L_h(\theta) + \lambda_o L_o(\theta). \tag{14}$$

其中: λ_h 和 λ_o 是比例系数,为了平衡 $L_h(\theta)$ 和 $L_o(\theta)$ 的比例,这里将 λ_h:λ_o 设为3:2.

1.1.4　MobileNetV3模型

2017年,谷歌提出了MobileNets[6]模型,采用深度可分离卷积的操作,比标准的卷积拥有更少的计算量. 轻量级卷积神经网络MobileNetV2包含线性瓶颈结构(linear bottlenecks)和倒置残差结构(inverted residuals)[12], MobileNetV3模型[7]则在网络结构中使用了压缩奖惩(squeeze and excitation, SE)结构和一种新的激活函数Hard-swish[7],由此改进了模型性能. MobileNetV3的提出者只使用其进行了基本的图像分类的实验,验证了其优越的性能. 而本文是要以MobileNetV3为基础网络结构进行人体姿态估计任务,因此需要先对MobileNetV3模型的部分网络结构进行修改,再应用于人体姿态估计[5].

step 1: 原倒数第3层为一个 7×7 的池化层,滤波器数为576,将其替换成 1×1 的普通卷积层,滤波器数为 $3J$. 修改这层的目的是调整特征通道数,使得输出的特征通道的维度是 $7^2 \times 3J$.

step 2: 原倒数第2层为 1×1 的普通卷积层,现修改成转置卷积层,同样是为了特征通道的调整. 将通过转置卷积层后的特征维度设为 $56^2 \times 3J$,因为224正好是56的4倍.

step 3: 原最后一层是普通卷积层,现修改成双线性插值层. 第1节中设置了模型输出的维度是 $224^2 \times 3J$,现通过该层将特征维度调整成 $224^2 \times 3J$,分别包括了维度是 $224^2 \times J$ 的热力图、X 和 Y 方向的偏移量图,通过式(1)～(3)的估计得到人体关节点的2D坐标信息.

本文修改的最后3层网络结构如表1所示,人体姿态估计的网络如图3所示,修改后的MobileNetV3将用于人体的2D姿态估计任务.

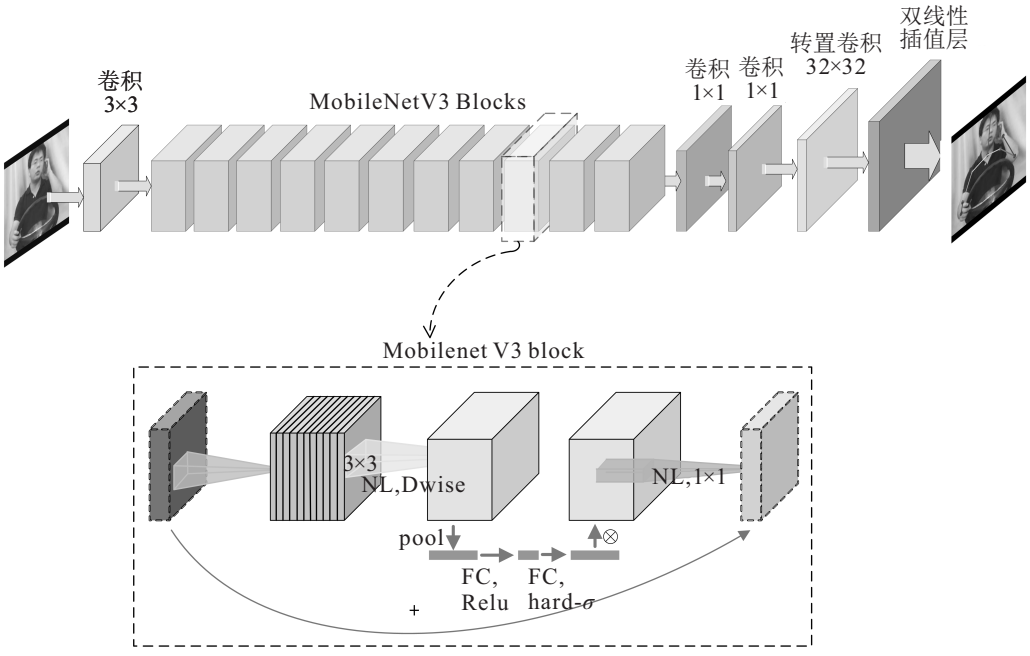

图 3　基于MobieNetV3的人体姿态估计CNN结构

表1 修改的最后3层网络层及参数设置

输入特征维度	操作及滤波器大小	滤波器数	步长	SE	NL
$7^2 \times 576$	标准卷积1×1	$3J$	1	–	HS
$7^2 \times 3J$	转置卷积32×32				
$56^2 \times 3J$	双线性插值层	=	8	=	HS

1.2 基于ST-SRU的行为识别方法

通过人体姿态估计模型理论上可以得到驾驶员的骨骼点的2D坐标信息,要对所获取的骨节点数据进行建模分析,才能对相应的动作类别进行识别. 本文采用时空SRU模型,对骨架序列进行分类. ST-SRU表达式如下:

$$\tilde{I}_{j,t} = WI_{j,t}, \tag{15}$$

$$f^S_{j,t} = \text{sigmoid}(W^S_f I_{j,t} + b^S_f), \tag{16}$$

$$f^T_{j,t} = \text{sigmoid}(W^T_f I_{j,t} + b^T_f), \tag{17}$$

$$r_{j,t} = \text{sigmoid}(W_r I_{j,t} + b_r), \tag{18}$$

$$c_{j,t} = f^T_{j,t} \odot c_{j,t-1} + f^S_{j,t} \odot c_{j-1,t} + (1 - f^T_{j,t}) \odot (1 - f^S_{j,t}) \odot \tilde{I}_{j,t}, \tag{19}$$

$$h_{j,t} = r_{j,t} \odot \tanh(c_{j,t}) + (1 - r_{j,t}) \odot \tilde{I}_{j,t}. \tag{20}$$

其中: W、W^S_f、W^T_f和W_r均为大小是(d,d)的可学习矩阵, d为内部状态大小; $\odot$为哈达玛积. ST-SRU有两个遗忘门,对应来自两个域的上下文信息: $f^S_{j,t}$为空间域, $f^T_{j,t}$为时间域. $h_{j,t}$为输出门,可以通过复位门$r_{j,t}$进行调整.

本文采用了3层叠加的ST-SRU模型, ST-SRU网络框架如图4所示. 每个单元代表在时空步(j,t)的内部状态为$c_{j,t}$. 每个单元接收先前关节和相同关节的先前帧的内部状态[13]. 在逐步处理完骨架序列后,将在最后一步的输出状态$c_{j,t}$输入到softmax分类器,得到危险驾驶姿态分类结果.

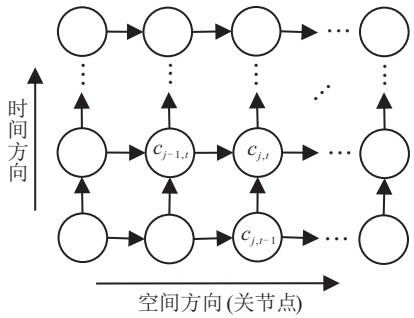

图4 ST-SRU网络框架说明

2 实验结果及分析

2.1 姿态估计实验

本文实验的设备是8 G内存、NVIDIA GTX1050-Ti的计算机,使用的深度学习框架为tensor flow 1.8.0. 用于比较的4种算法包括:

1) MobileNetV3: 为了满足人体姿态估计任务已将算法进行了修改.

2) 基于深度神经网络的人体姿态估计算法DeepPose[14]: 将姿态估计表述为关节点直接回归的问题. 同时提出使用级联方式获取更精确的坐标位置.

3) CPM[12]: CPM包含一系列卷积网络,为每个部分的位置重复生成2D置信图. 这里实验的CPM网络采用的是3阶的.

4) stacked hourglass[11]: 由多个Hourglass模块集成来捕捉图像在不同尺度等级上的特征.

2.1.1 实验数据集

第1个数据集是LSP(leeds sports pose dataset)[15]数据集. 它是一个运动姿势数据集,包含约2 000个姿势注释,每个样本提供了14个关节点的2D坐标,主要来自于体育活动,人物的姿态复杂,具有一定的挑战性.

第2个数据集是从AI challenger的姿态估计数据集中筛选并整理的上肢姿态数据集. 筛选的方法是根据数据集提供的关节点2D坐标数据进行计算,挑选出人体上肢姿态从正面至半侧面的样本图,最后根据人体上肢位置将筛选的图片进行剪裁、缩放等操作. 筛选完后共有64 304个样本,每个样本包含了8个人体骨节点的2D坐标注释. 这些骨节点有头部、颈、左右手腕、左右肩膀、左右手肘[15].

2.1.2 实验设置

LSP共包含2 000张体育运动的图像,其中1 000张用作训练,另1 000张用作测试. 训练采用的梯度优化算法为RMSprop (root mean square propagation)[16]算法和mini-batch梯度下降,设全局学习率为1e-3,衰减系数为0.9, mini-batch为30.

对于上肢姿态数据集,筛选出的样本一共64 304张,为满足训练需要,将其划分成3部分: 训练集、验证集和测试集,划分比例为8:1:1. 训练采用的梯度优化算法为RMSprop算法和mini-batch梯度下降,设全局学习率为1e-3,衰减系数为0.9, mini-batch为36.

在对人体姿态估计准确性评价中广泛使用PCP

(percentage correct parts)[17]标准进行评定. PCP标准中,人体的某个部位端点在真实部位长度的特定范围内,且人体的关节是部位端点,则认为正确地检测出这个部位[18]. 本节实验也使用该标准作为实验结果的评价指标.

2.1.3　LSP数据集上姿态估计

本文遵循文献[17]计算出不同方法在LSP数据集上的PCP值,如表2所示. 上臂、前臂、大腿和小腿各有左右两个PCP值[5].

表2　4种方法在LSP数据集上的部件PCP值　%

方法	头部	躯干	上臂	前臂	大腿	小腿	均值
MobileNetV3	86.1	92.9	62.6 62.3	47.2 47.7	80.3 80.7	72.6 74.4	70.7
DeepPose	81.3	90.9	56.4 56.6	38.2 38.7	77.6 77.4	71.3 71.5	66.0
CPM	92.2	95.1	85.1 85.6	77.5 78.7	86.2 86.1	82.8 82.3	85.2
stacked hourglass	89.8	95.7	83.3 83.1	75.7 75.9	85.9 84.5	79.5 80.3	83.4

根据表2的结果分析可知,本文所提方法的平均PCP值达到了70.7%,比DeepPose (66.0%)的PCP值高了4.7%,但与其他两种方法对比可以发现,其值比stacked hourglass[9] (83.4%)少了12.7%,比CPM (85.2%)低了14.5%. 经过分析发现,产生这种结果的原因是LSP数据集提供的关节点数量是14个,实验是单人全身姿态估计,而本文所提算法是轻量级的CNN模型,其特征提取能力较弱,对关节点数量多且姿态复杂样本的估计任务表现出了弱势.

除了上述实验,还需要通过参数量和运行效率对所提算法进行实验和分析. 本实验统计了4种方法在LSP数据集上执行姿态估计任务的参数量和同一样本测试1 000次的耗时,为保证实验结果的可信性,重复实验5次后取均值.

根据表3的结果可知,本文所提方法的参数量为2 863 360,分别是stacked hourglass、CPM和DeepPose的29.5%、17.3%和13.2%. DeepPose作为一种基准对比,测试耗时是MobileNetV3的33倍,估计效果也不如本文提出的方法. stacked hourglass (27.63 s)和CPM (178.31 s)的耗时统计分别是本文所提方法的3.72倍和24.0倍. 这两种方法虽然获得了更好的估计效果,但是姿态估计任务增加的耗时降低了更多的性能.

表3　4种方法的参数量和测试1 000次的时间

方法	参数量	耗时/s
MobileNetV3	2 863 360	7.42
DeepPose	21 695 900	232.54
CPM	16 585 365	178.31
stacked hourglass	9 716 480	27.63

2.1.4　上肢姿态数据集上姿态估计

该数据集的每张图片经过处理后,只保留了人体上肢的8个关节点的2D坐标,这8个关节点包括头部、颈部、左右腕、左右肘、左右肩[18]. 将这8个上肢关节点划分为7个部分,以满足PCP的评价标准:左上臂、右上臂、左小臂、右小臂、左肩膀、右肩膀、头部. 表4中记录了4种算法在该数据集上的实验结果.

表4　4种方法在上肢姿态数据集上的PCP值　%

方法	头部	左肩膀	左上臂	左前臂	右肩膀	右上臂	右前臂	均值
MobileNetV3	99.3	97.1	96.5	91.1	98.5	96.5	90.2	95.6
DeepPose	96.3	95.0	94.4	90.0	95.5	94.2	88.7	93.4
CPM	99.1	99.8	98.8	96.2	99.2	98.0	96.7	98.3
stacked hourglass	99.5	98.7	97.4	93.8	98.8	97.6	93.9	97.1

根据表4的结果分析可知,在上肢姿态数据集上本文所提方法的平均PCP值达到了95.6%,与stacked hourglass (97.1%)相比少了1.5%,比CPM (98.3%)低了2.7%,比DeepPose (93.4%)的平均PCP值高2.2%. 表2与表4对比分析能看出,执行人体姿态估计的样本关节点数量从14个减少到8个时,4种方法的姿态估计任务的平均PCP值都有所提升,本文所提方法与stacked hourglass、CPM方法的差距进一步减小.

下面将通过参数量和运行效率对所提算法进行分析.

如表5所示,与表3对比MobileNetV3的参数量减少了43%,降低至1 636 480,这是由模型自身结构导致的,需要估计的关节点数从14个减少到8个,本文所提方法的参数量也成倍降低. 从耗时方面看,

stacked hourglass、CPM 和DeepPose三种方法的耗时统计分别是本文方法的5.03倍、33.25倍和43.14倍,这几乎与LSP数据集上的实验结论相同.

表5 4种方法的参数量和测试1 000次的时间

方法	参数量	耗时/s
MobileNetV3	1 636 480	5.35
DeepPose	21 646 730	230.81
CPM	16 572 286	177.88
stacked hourglass	9 708 800	26.93

2.2 危险驾驶姿态识别实验

2.2.1 危险驾驶行为数据集

关于动作数据集设计和采集方面,比较流行的方式是采用Microsoft Kinect,它能直接得到人体关节点的3D位置坐标数据. 研究过程中发现没有适合于本系统开发的危险驾驶行为数据集,故本文参考UT-Kinect[19]动作数据集的设计,自行采集一个危险驾驶行为识别的数据集,以方便本文的嵌入式轻量级危险驾驶姿态识别系统的开发.

在数据集设计中,选择了驾驶过程中出现的典型的7种危险动作:左手使用手机、右手使用手机、双手使用手机、左手打电话、右手打电话、左手喝水以及右手喝水[5],还有规范驾驶姿态,一共采集的动作类别有8类,每一类的动作实例如图5所示.

(a) 正常驾驶 (b) 左手使用手机 (c) 右手使用手机 (d) 双手使用手机

(e) 左手接电话 (f) 右手接电话 (g) 左手喝水 (h) 右手喝水

图5 8种驾驶行为

根据UT-Kinect[19]动作数据集的设计,本文的数据采集实验共有10名被试者,9名男性,1名女性,采集设备是PC机、720P分辨率的USB摄像头和拍摄支架. 摄像头放置在桌面上与PC机连接,与被试者的距离为120 cm、与竖直方向上的夹角为30°,采集过程中会根据被试者的身高情况适当微调摄像头的角度. 首先,每位被试者会将每类动作做2次,本文采集到的是15帧/s的动作图像序列和人体姿态估计模型输出的上肢8个关节点的2D坐标数据. 这里的人体姿态估计模型即为本文提出的基于MobileNetV3的二维姿态估计算法训练得到的模型. 接着,将采集到的动作图片序列和骨架序列进行每类动作的起始和结束的位置标注,并根据动作类别进行排序编号. 最后对数据进行整理和保存. 危险驾驶姿态数据集一共有8个动作类别,每个动作有20个样本,共采集整理得到160个样本. 对于训练ST-SRU模型而言,动作类别的样本是一个动作骨架序列,维度均为8×2,包含8个上肢关节点的2D坐标数据,样本标签是动作的类别.

考虑到ST-SRU的模型训练任务,需要从采集好的危险驾驶姿态数据集中的骨架序列中截取固定长度的序列. 参考了文献[13]的做法,从骨架序列样本中按顺序随机采样,最后获取T帧的子序列. 同时,这样的操作可以从原始骨架序列中多次采样,扩充样本数量,防止网络在训练过程中发生过拟合现象[19].

本实验中的参数有采样子序列的长度T和ST-SRU隐藏层的节点数d,T在集合$\{5,10,15,20,25\}$中搜索,d在区间$[32,128]$内搜索且步长为8,它们都是经过5-fold cross-validation获得的. 最终获取的最优参数为$T=10$和$d=48$.

实验选取Adam算法进行优化,全局学习率为2e-3,mini-batch为32. 虽已通过采样子序列的方法进行了样本扩充,但是样本数量仍然比较小. 因此为了尽量避免出现过拟合现象,在模型训练中采用变分丢弃法(variational dropout)[20],Dropout为0.5. 在实验过程中为了减少过拟合也采用早停法提前终止训练.

2.2.2 简单实现危险驾驶姿态系统

本小节将实现第1节提出的危险驾驶姿态识别系统流程,将对应算法训练得到的模型移植到树莓派[21]上,完成整个系统的研发.

硬件平台采用了Raspeberry Pi 3B+开发板,运行频率为1.4 GHz,内存为1 GB. 将720P高清USB摄像头与树莓派连接,进行数据采集.

本节设计的嵌入式危险驾驶姿态识别系统的算法流程如下:

step 1: 利用USB摄像头采集到$1\,280\times 720$的动作图片序列,根据人物在图片中的位置,用裁剪、缩放、中值滤波等图像预处理操作,获得大小为$224\times$

224的图片,作为模型训练的输入;

step 2: 将step 1得到的图片进行标准化,以达到模型的输入要求;

step 3: 将step 2标准化后的图片输入本文提出的基于MobileNetV3改进的2D人体姿态估计模型,得到上肢8个关节点的2D坐标数据,维度为8×2的张量;

step 4: T帧图片序列生成相同长度T的骨架序列,维度是$T \times 8 \times 2$的张量. 先将S进行零-均值(Z-score)标准化[22],再将标准化后的骨架序列输入到ST-SRU模型,最后得到危险驾驶的姿态类别.

在2.2.1节设计并制作了危险驾驶姿态数据集,下面将利用该数据集对本文算法进行可行性分析. 通过5次5-fold cross-validation后取平均值,最终的危险驾驶姿态识别的结果如图6所示.

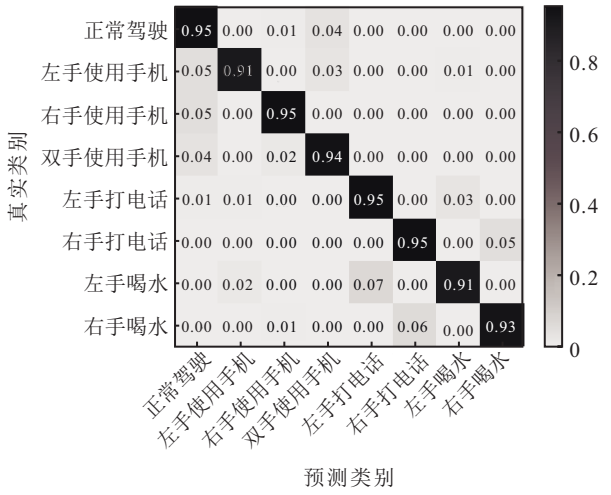

图6　危险驾驶姿态识别的混淆矩阵

从图6中可以看出,ST-SRU对8种姿态类别有不错的识别率,该方法的可行性得到了验证.

图7是系统最终的运行效果,由于该开发板的性能有限,并未开发可视化的交互界面,只是通过一个终端窗口对识别结果输出. 在这款开发板上的系统识别速率为0.3帧/s.

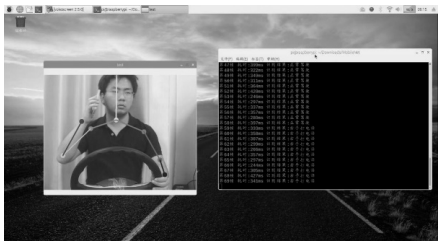

图7　危险驾驶姿态识别系统实现

3　结　论

本文提出了一种基于MobileNetV3和ST-SRU的危险驾驶姿态识别系统,在MobileNetV3网络中通过修改模型结构使其适用于人体姿态估计任务. 同时提出ST-SRU动作识别算法,能够将每帧骨架序列中的每个关节点在时间和空间依赖关系上进行建模,加快模型推理速度,提高动作识别的效果. 实验表明,所提出的算法在公开数据集和自建的危险驾驶姿态数据集上都能有很好地应用,设计的危险驾驶姿态识别系统实现了精度与速度间较好的折衷,达到了本文的研究目的. 本文提出的二维人体姿态估计算法只能从RGB图片中得到关节点的二维坐标,而二维信息对动作的表达能力远不如三维坐标信息,这将是本文算法改进的一个方向. 此外,本文采集的危险驾驶行为数据集仅仅基于模拟的开车场景,与真实场景下的实测数据仍然存在一定差异,被试者的数量也不够多,这些因素都会对模型训练和评估产生影响. 因此,下一步工作将研究基于轻量级CNN的3D姿态估计算法,并在实测场景数据采集上进一步完善.

参考文献(References)

[1] World Health Organization (WHO). Road traffic injuries overview[EB/OL]. (2020-09-19)[2021-03-21]. http://www.who.int/health-topics/road-safety.

[2] Niu Z L, Lin M, Chen Q, et al. Correlation analysis between risky driving behaviors and Characteristics of commercial vehicle drivers[M]. Advances in Intelligent Systems and computing. Cham: Springer International Publising, 2016: 677-685.

[3] Lansdown T C. Individual differences and propensity to engage with in-vehicle distractions — A self-report survey[J]. Transportation Research, Part F: Traffic Psychology and Behaviour, 2012,15(1): 1-8.

[4] 肖遥. 分心行为对交通安全和交通效率的影响分析与建模[D]. 北京: 清华大学, 2016.
(Xiao Y. Analyzing and modeling of the influence of driving distraction on traffic safety and traffic efficiency[D]. Beijing: Tsinghua University, 2016.)

[5] 穆高原. 基于深度学习的危险驾驶行为识别研究[D]. 杭州: 杭州电子科技大学, 2020.
(Mu G Y. Study on dangerous driving behavior recognitoin based on deep learning[D]. Hangzhou: Hangzhou Dianzi University, 2020.)

[6] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[J/OL]. 2017, arXiv: 1704.04861.

[7] Howard A, Sandler M, Chen B, et al. Searching for MobileNetV3[C]. International Conference on Computer Vision. Seoul, 2019: 1314-1324.

[8] Lei T, Zhang Y, Wang S I, et al. Simple recurrent units for highly parallelizable recurrence[C]. Proceedings of the 2018 Conference on Empirical Methods in Natural

Language Processing. Stroudsburg: Association for Conputational Linguistics, 2018.

[9] Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation[M]. Computer Vision. Cham: Springer International Publishing, 2016: 483-499.

[10] Wei S H, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016: 4724-4732.

[11] Papandreou G, Zhu T, Kanazawa N, et al. Towards accurate multi-person pose estimation in the wild[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, 2017: 3711-3719.

[12] Sandler M, Howard A, Zhu M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 4510-4520.

[13] Liu J, Shahroudy A, Xu D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[M]. Computer Vision. Cham: Springer International Publishing, 2016: 816-833.

[14] Tosheyv A, Szegedy C. DeepPose: Human pose estimation via deep neural networks[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, 2014: 1653-1660.

[15] 石跃祥, 许湘麒. 基于改进 DenseNet网络的人体姿态估计 [J]. 控制与决策, 2021, 36(5): 1206-1212.
(Shi Y X, Xu X Q. Improved DenseNet network for human pose estimation[J]. Control and Decision, 2021, 36(5): 1206-1212.)

[16] Kurbiel T, Khaleghian S. Training of deep neural networks based on distance measures using RMSProp[J/OL]. 2017, arXiv: 1708.01911.

[17] Eichner M, Ferrari V, Zurich S. Better appearance models for pictorial structures[C]. Proceedings of the British Machine Vision Conference (BMVC). London: British Machine Vision Association, 2009: 6-17.

[18] 冯健颖. 基于卷积神经网络的人体姿态估计研究 [D]. 哈尔滨: 哈尔滨工业大学, 2018.
(Feng J Y. Research of human pose estimation based on convolutional neural network[D]. Harbin: Harbin Institute of Technology, 2018.)

[19] Xia L, Chen C C, Aggarwal J K. View invariant human action recognition using histograms of 3d joints[C]. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2012: 20-27.

[20] Gal Y, Ghahramani Z. A theoretically grounded application of dropout in recurrent neural networks[C]. Proceedings of Advances in Neural Information Processing Systems 29 (NIPS 2016). Barcelona, 2016: 1019-1027.

[21] Sajjad M, Nasir M, Muhammad K, et al. Raspberry Pi assisted face recognition framework for enhanced law-enforcement services in smart cities[J]. Future Generation Computer Systems, 2020, 108: 995-1007.

[22] Xing Y, Lv C, Wang H J, et al. Driver activity recognition for intelligent vehicles: A deep learning approach[J]. IEEE Transactions on Vehicular Technology, 2019, 68(6): 5379-5390.

作者简介

赵俊男 (1996–), 男, 硕士生, 从事人体姿态估计、动作识别的研究, E-mail: 663261972@qq.com;

佘青山 (1980–), 男, 教授, 博士, 从事动作识别、主动康复机器人、脑机交互、机器学习等研究, E-mail: qsshe@hdu.edu.cn;

穆高原 (1995–), 男, 硕士生, 从事深度学习、动作识别的研究, E-mail: 1171791288@qq.com;

吴秋轩 (1978–), 男, 副教授, 博士, 从事自重构机器人运动控制、仿生软体机器人运动控制、光伏发电系统等研究, E-mail: wuqx@hdu.edu.cn;

席旭刚 (1975–), 男, 副教授, 博士, 从事脑认知与人工智能、生物医学信号处理等研究, E-mail: xixugang@hdu.edu.cn.

(责任编辑：闫 妍)