DOI:10.19651/j.cnki.emt.2208888

轻量级 CNN 实时跌倒预测及嵌入式系统实现*

杜群贵 钟 威
(华南理工大学机械与汽车工程系 广州 510640)

摘 要: 为了实现实时而准确的跌倒预测,同时将深度学习模型移植到于可穿戴端设备中运行,提出了一种轻量级卷积神经网络模型。借鉴深度可分离网络的轻量级模型思想,设计了网络结构,并优化通道数和卷积核尺寸,在保证准确率基本不变的情况下大大减小了模型计算复杂度。为将算法部署于可穿戴跌倒保护设备,提出了模型在嵌入式端的实时运行框架,并将算法编写为 C 程序,移植到了 STM32 单片机中。此模型在 Sisfall 数据集中获得了 97.5%的准确率,204.3 ms 的裕量时间。移植的模型仅有 11.65 KB 大小,在 STM32 单片机中的算法延时仅为 8.24 ms。实验结果表明,该模型具有较高的预测精度和很好的实时性,为跌倒预测算法和跌倒保护装置的开发提供了进一步的参考。

关键词: 跌倒检测;深度可分离网络;嵌入式;可穿戴设备
中图分类号: TP391.4 **文献标识码:** A **国家标准学科分类代码:** 510.4030

Lightweight CNN real-time fall prediction and embedded system implementation

Du Qungui Zhong Wei
(School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510640, China)

Abstract: In order to achieve real-time and accurate fall prediction, and transplant the deep learning model to run on wearable devices, a lightweight convolutional neural network model is proposed. Drawing on the lightweight model idea of DSC network, the network structure is designed, and the number of channels and the size of convolution kernel are optimized, which greatly reduces the computational complexity of the model while keeping the accuracy rate basically unchanged. In order to deploy the algorithm in the wearable fall protection device, a real-time running framework of the model on the embedded side is proposed, and the algorithm is written as a C program and transplanted to the STM32 microcontroller. This model achieves 97.5% accuracy with 204.3 ms lead time on the Sisfall dataset. The transplanted model is only 11.65 KB in size, and the algorithm delay in the STM32 microcontroller is only 8.24 ms. The experimental results show that the model has high prediction accuracy and good real-time performance, which provides a further reference for the development of fall prediction algorithms and fall protection devices.

Keywords: fall detection; depth-wise separable network; embedded; wearable device

0 引 言

跌倒是一个值得关注的人身安全问题,跌倒不仅会导致扭伤、骨折等伤害,还有可能会危及生命,这些风险在身体素质偏低的老年群体中更为突出。本文中跌倒保护器是一种穿戴式气囊保护装置,其利用传感器感知人体运动,并结合跌倒预测算法来触发气囊快速充气,保护人体关键部位,以大大减轻跌倒冲击[1]。

由于跌倒失衡到触地的间隔仅有 300~400 ms,跌倒预测算法必须在此期间内挖掘足够的信息,并快速做出决策,因此要求算法具备高实时性。一些学者使用了基于阈值的跌倒预测算法,当某些特征指标超过阈值时即判断人将要跌倒。如 Ahn 等[2]基于三轴加速度合成的三角形面积阈值构建了算法;Jung 等[3]利用合成加速度、合成角速度和姿态角特征阈值构建了算法。这些固定阈值算法可以轻易地移植到嵌入式设备上,并保证实时预测,但很难达到高准确度。机器学习算法在分类问题中具有良好的表现,杨智超等[4]从频域提取特征,建立了 SVM 跌倒检测模型;Kim 等[5]分别建立了 KNN(K-nearest neighbor)、RF(random forest)、SVM(support vector machine)等 6 种机

收稿日期:2022-01-19
* 基金项目:广东省自然科学基金(2021A1515012258)项目资助

器学习模型预测跌倒,比较了不同算法的性能。然而传统机器学习方法依赖于启发式提取手工特征,如果提取不合适的特征会丢失重要信息,影响机器学习上限。深度学习方法在模式识别问题中具有独特的优势,它能自动学习特征,无需过多人工干预。Wang 等[6]提出了一种基于多源卷积网络的跌倒预测算法,在私有数据集上报道了 99.3% 的准确率,但其网络庞大,计算复杂,难以在嵌入式设备上实时预测。Yu 等[7]构建了 ConvLSTM 算法,在 KFall 数据集上达到了 99.32% 的敏感度和 99.01% 的特异性,但 Yu 同样指出算法仅在 PC 上实现了预测,有待在嵌入式设备上进行实际验证。如何利用深度学习的优势,并在嵌入式设备上实时运行深度学习决策,实现即准确又快速的跌倒预测是当前面临的一项挑战。

基于上述背景,本文在兼顾准确率与实时性的基础上设计了一种用于跌倒预测的轻量级卷积神经网络(convolutional neural networks, CNN)模型,使用深度可分离网络来替代传统卷积网络,降低特征提取过程的复杂度,减少模型参数并提高计算速度。为将算法部署在嵌入式设备,设计一种两级运行框架,并将算法移植到 STM32 单片机验证其有效性和实用性。

1　算法模型

1.1　深度可分离卷积

深度可分离卷积(depthwise separable convolution, DSC)是一种面向轻量化模型的卷积策略,其主要思想是将输入分多个通道单独卷积(depth-wise convolution),再使用点卷积(point-wise convolution)融合各个通道的特征,这种方式降低了卷积维度,可以有效减少卷积核参数和不必要的计算。DSC 卷积网络结构如图 1 所示。

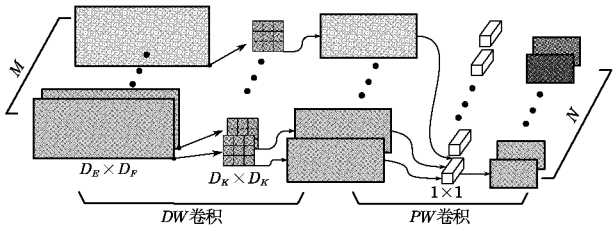


图 1　深度可分离卷积网络结构

图 1 中输入特征图大小为 $D_E \times D_F$,输入通道数为 M,卷积核大小为 D_K,输出通道数为 N。若使用普通卷积网络输出同样的通道数,其计算成本为:

$$D_E \times D_F \times M \times N \times D_K \times D_K \tag{1}$$

而使用深度可分离卷积网络时,其计算成本为:

$$D_E \times D_F \times M \times D_K \times D_K + M \times N \times D_E \times D_F \tag{2}$$

将式(2)比式(1):

$$\frac{D_E \times D_F \times M \times D_K \times D_K + M \times N \times D_E \times D_F}{D_E \times D_F \times M \times N \times D_K \times D_K} = \frac{1}{N} + \frac{1}{D_K^2} \tag{3}$$

随着 D_K 和 N 的增大,DSC 相对传统 CNN 的计算成本也急剧下降。在著名的轻量级网络 MobileNet 中,使用 DSC 网络可以使计算量下降到原来的 1/8 左右,而其精度相比于传统卷积网络只是略微下降[8]。

1.2　网络结构

使用可穿戴惯性传感器可以测量人体的加速度,角速度信号。涂亚庆等[9]的跌倒检测算法使用了加速度信号;吕艳等[10]使用了加速度及角速度信号;而任晶秋等[11]发现使用姿态角可以提升跌倒检测的性能。故本文不仅使用了加速度和角速度作为输入,还使用 Mahony 互补滤波算法提取姿态角参数,提出了如图 2 所示的网络结构来进行预测。左侧支路包含了三轴加速度信号,三轴角速度信号及三轴角度信号,经过深度可分离网络高效提取特征。其中 DW 层输入通道数为 9,卷积步长为 2,卷积核大小为 K;PW 层卷积步长为 1,输出通道数为 N。

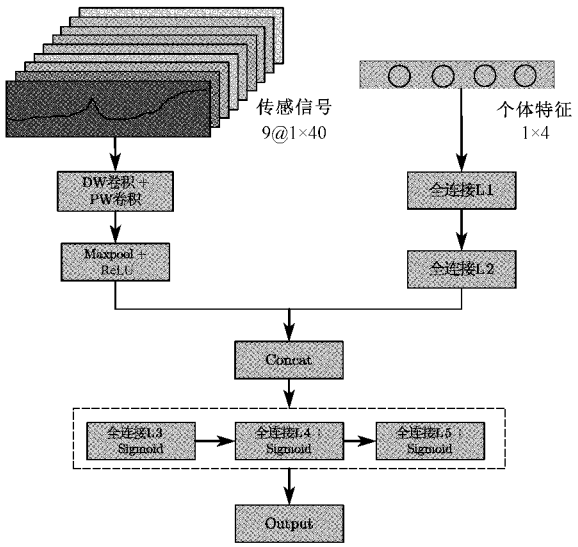


图 2　轻量级 CNN 网络结构

此外,考虑到具有不同生理特征的个体行动模式也不同,故将个体生理特征(身高、体重、年龄、性别)作为另一输入支路,并且经过两层全连接层后与卷积支路的特征做拼接。这将个体生理特征与传感器信号输出关联了起来,挖掘个体特征与传感器输出之间的隐含联系。拼接的特征向量经过三层全连接层后输出分类结果,其中全连接层采用了 Sigmoid 激活函数。

使用卷积神经网络处理连续时间序列分类问题时需要利用滑动时间窗,而时间窗的大小影响着算法性能。依据 Aziz 等[12]的研究结果及跌倒过程真实历程,本文采用了 200 ms 时间窗,它能较好地保持算法的鲁棒性。

模型中采用的损失函数为二进制交叉熵(BCE)损失函数。损失函数如式(1)所示,其中 S 为样本数,第 i 个样本的标签为 y_i,正类概率为 p_i。为了保证收敛速度及精度,本文使用了固定步长衰减学习率,如式(2)所示,其中 α_0 代表初始学习率,*epoch* 代表训练迭代次数,*stepsize* 代表步

长，γ 为衰减率。

$$L=\frac{1}{S}\sum_i L_i=-\frac{1}{S}\sum_i[y_i\ln(p_i)+(1-y_i)\ln(1-p_i)] \tag{4}$$

$$\alpha=\alpha_0\gamma^{epoch//stepsize} \tag{5}$$

1.3　嵌入式端模型运行框架

跌倒是低概率事件，人大部分时间处于静止或者低动态运动，而这些动作几乎没有跌倒的风险。当算法部署在可穿戴设备上时，应当避免低风险动作触发复杂的神经网络计算，以减少功耗。本文为轻量级 CNN 设计了在嵌入式设备中运行的两级模型框架，其流程如图 3 所示。

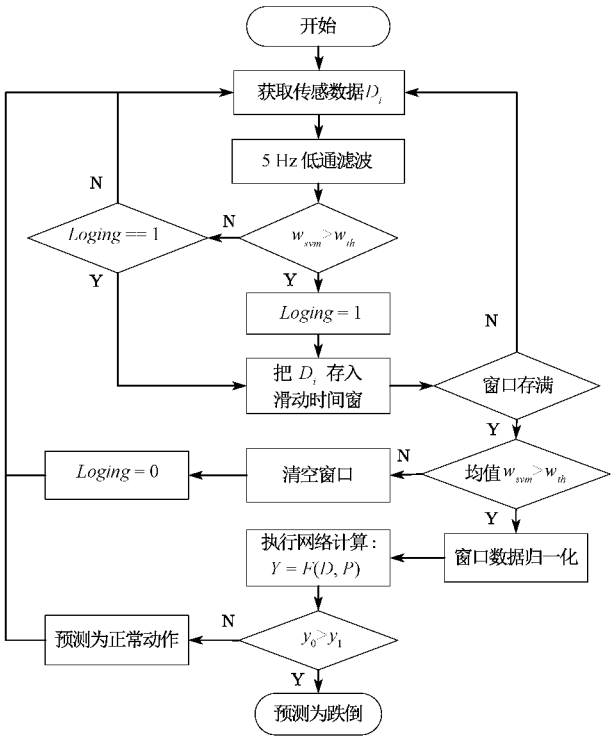


图 3　两级模型框架

图 3 中 $\boldsymbol{D}_i$ 代表传感器当前获取的信号向量，包含三轴角速度信号及三轴加速度信号。w_{svm} 代表当前帧的三轴合成角速度，w_{th} 为合成角速度阈值，$\boldsymbol{Y}$ 为 CNN 输出的向量，$\boldsymbol{Y}=(y_0,y_1)$。该框架包含两级分类，即阈值分类器和 CNN 分类器。在第 1 阶段，计算合成角速度，若角速度低于阈值且不处于存储状态，则说明人处于轻微运动状态，并不存在跌倒的风险。此时程序将直接获取下一帧的传感数据。当角速度超过阈值时，则说明人存在跌倒风险，接下来 200 ms 的传感数据会被无条件存入时间窗。时间窗被存满后，若时间窗内的平均角速度大于阈值，则执行 CNN 模型，否则清空时间窗。这一步过滤了一些低动态的运动，避免了不必要的计算。若计算结果 $y_0>y_1$，则说明算法预测了一个跌倒动作，微控制器则改变特定引脚的电平，从而触发气囊系统来保护人体。通过这种方法，可以让高精度的 CNN 模型只运行很少的时间，大大减少微控制器的功耗。

2　实验设计

2.1　数据集描述

数据集是建立算法的基础，Casilari 等[13]评估了该领域主要的 12 个数据集，而其中 Sisfall 数据集[14]采集了更丰富的动作并具有更广的样本分布，因此被本文所采用。该数据集采自 38 名实验对象，年龄分布在 19～75 岁之间。Sisfall 数据集涵盖 19 种正常行为动作和 15 种跌倒动作，累计 1 798 个跌倒活动和 2 707 个正常行为活动。惯性传感器绑在实验者腰部，测量其三轴加速度和三轴角速度信号，采样频率为 200 Hz。

2.2　模型训练

Sisfall 数据集中的样本按照 7∶1.5∶1.5 的比例划分为训练集、验证集和测试集。模型训练过程中将 L4 全连接层设为 dropout 层，使训练阶段该层每个神经元有一定概率被抑制输出，提高泛化能力。基于 Python 和 Pytorch 深度学习框架建立了网络模型，其超参数如表 1 所示。

表 1　模型超参数

参数	设定值
批大小	20
迭代次数 *epoch*	150
初始学习率 α_0	0.005
衰减步长 *stepsize*	12
γ	0.5
dropout 概率	0.3

2.3　算法评估指标

跌倒预测算法的评估指标通常有准确度(accuracy，ACC)，敏感度(sensitivity，SEN)，特异性(specificity，SPE)，裕量时间(lead time，LT)及算法延时(delay)。

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$SEN=\frac{TP}{TP+FN} \tag{7}$$

$$SPE=\frac{TN}{TN+FP} \tag{8}$$

$$LT=T_i-T_d \tag{9}$$

其中，TP 代表发生跌倒且成功预测的样本数；FN 代表发生跌倒但漏检的样本数；TN 代表正常行为动作被正确识别的样本数；FP 代表正常行为动作引发虚警的样本数。T_i 是撞地时刻，T_d 是预测到跌倒的时刻。裕量时间代表了算法预测到跌倒后到实际跌倒的时间差，它预留给嵌入式算法及机械系统响应延时。算法延时是处理器执行一次模型决策所需要的时间，由于本算法面向嵌入式设备，故研究中使用单片机芯片作为处理器来测量算法延时。

2.4　嵌入式系统实现

跌倒保护器是一种智能穿戴设备，也是跌倒预测算法

所需要部署到的地方。基于深度学习的算法也许可以获得很高的准确率,但其模型通常较为复杂,在计算资源有限的嵌入式设备中可能无法实时运行。因此需要搭建用于跌倒预测的嵌入式平台,并且将算法移植到硬件上来验证其实时性。

图 4 展示了模型移植到 STM32 的过程,具体流程为:1)在 PC 端使用 Python 语言和 Pytorch 框架训练模型;2)导出训练得到的各层权值矩阵 $\boldsymbol{W}$ 和偏置向量 $\boldsymbol{b}$;3)根据卷积神经网络模型编写决策函数的 C 语言程序,包括卷积函数、激活函数、池化函数、全连接函数等;4)将编写的函数和硬件驱动程序打包,通过 Keil 和 ST-Link 工具下载到嵌入式设备中,并采用 Level-3 优化 C 程序,缩减模型大小。

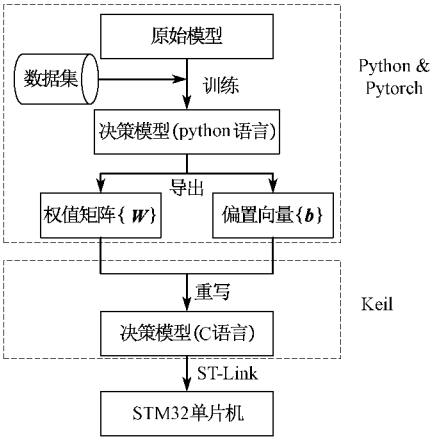


图 4　模型移植流程

嵌入式端采用了 STM32F103RCT6 芯片,该 STM32 芯片是常见的单片机芯片,具备 72 MHz 时钟频率,128 KB ROM 和 48 KB RAM。惯性传感器采用了 MPU6050 模块,它能测量三轴加速度、三轴角速度,并自带 DMP(digital motion processor)解算三轴姿态角,减轻了芯片解算角度的负担。蓝牙采用了 HC-02 模块,用于和 PC 端通讯。嵌入式端硬件结构如图 5 所示。

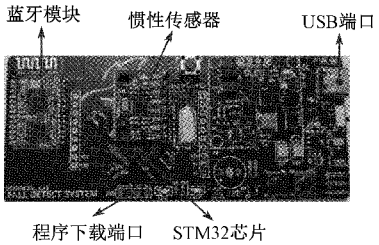


图 5　嵌入式端结构

3　实验结果与分析

3.1　模型结构优化

DSC 网络主要的超参数有输出通道数 N 和卷积核大小 K。首先确定深度可分离网络的通道数 N,设计了 4 种不同通道数的实验作对比。通道数分别为 3、4、5、6,卷积核大小一致为 3,步长一致设为 2,4 种模型的训练结果如图 6 所示。4 条曲线都很快收敛,而通道数为 5 和 6 的模型表现一致,均获得了最高的准确率。卷积层的输出通道数越多,产生的计算量也成正比增加,因此本文选择的 DSC 通道数为 5,既保证了算法准确率,又控制了模型大小。

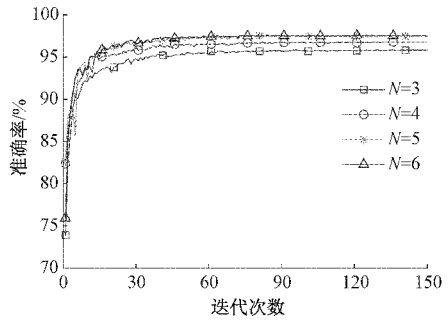


图 6　通道数与验证集准确率关系

再固定输出通道数为 5,比较 DW 卷积层中不同卷积核大小对性能的影响。选取的卷积核大小分别为 3、5、7、9,padding 设为 0,卷积步长为 2,实验结果如图 7 所示。可见使用尺寸 $K=3$ 卷积核的模型获得了最高准确率,显著高于其余 3 种模型,说明较小的尺寸能够更好地处理信号。同时,采用小的卷积核也可以减少特征提取过程的计算量,能在一定程度上降低算法延时。

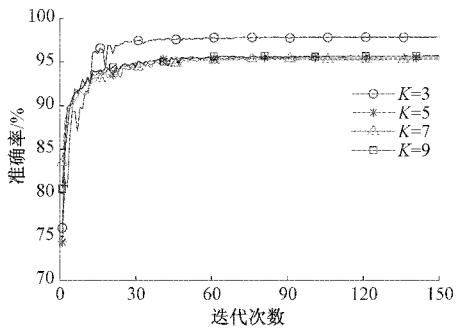


图 7　卷积核尺寸与验证集准确率对比

3.2　两级模型参数调节

图 3 所示的两级模型主要参数为三轴合成角速度阈值 w_{th},显然该阈值越大,DSC 网络模型更不容易被轻易调用,节省计算开支,但可能漏报;而越小则更不容易遗漏预测跌倒,但耗费计算量。图 8 给出了 w_{th} 和准确率及计算时间(遍历 Sisfall 数据集所需要的时间)的关系。当阈值为 0 时,该两级模型退化为单一 DSC 网络模型,其准确率也达到最高值 97.5%。随着阈值的增加,计算时间急剧下降,而准确率曲线在一定区间内保持不变,然后加速下降。依据此关系,选取的 w_{th} 为 13.2 deg/s,保证了准确率最大化的同时降低了 DSC 模型被频繁调用的几率。

3.3　嵌入式端性能评估

按照图 4 所示流程将 DSC 卷积网络模型部署到 STM32 单片机,通过 Keil 编译后得到该模型的资源占用

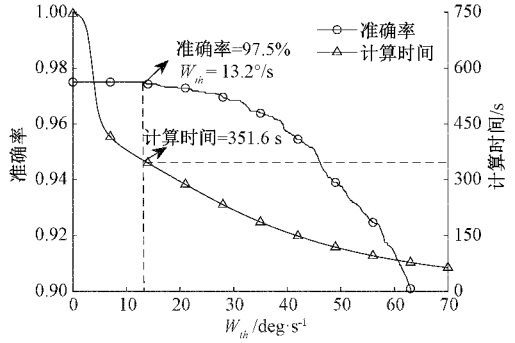


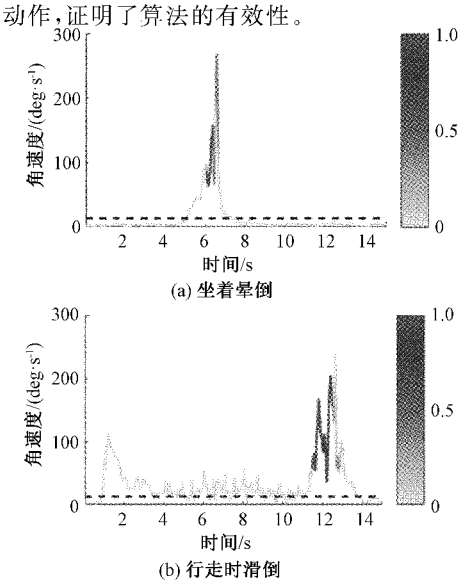
图 8　角速度阈值对准确率和计算时间的影响

情况，并通过在线调试测量了算法延时，结果如表 2 所示。模型仅仅占用了 10.78 KB 的 ROM 和 11.65 KB 的 RAM，远低于硬件总资源。这说明该模型足够轻量化，STM32 芯片完全能胜任运行该模型。

表 2　STM32 平台的资源占用

指标	参数值
ROM	10.87 KB/128 KB
RAM	11.65 KB/48 KB
Delay	8.24 ms

使用图 3 中的两级模型和 6 个测试活动验证算法的真实表现。将模型的输出转为概率值，并用颜色深度表示跌倒概率，测试结果如图 9 所示。由于跌倒时刻也往往是合成角速度峰值出现的时刻，故图 9 中使用了三轴合成角速度来观测动作剧烈程度，并用黑色虚线标明 w_{th}（即激活 DSC 模型的阈值）。图 9 中每一个跌倒动作((a)、(b)、(c))中曲线颜色都在峰值前由浅色快速变为深色，意味着算法成功在跌倒撞击前预测到了跌倒动作；而正常行为动作中虽然也出现了很高的角速度峰值，但算法并没有将其误判为跌倒动作，证明了算法的有效性。



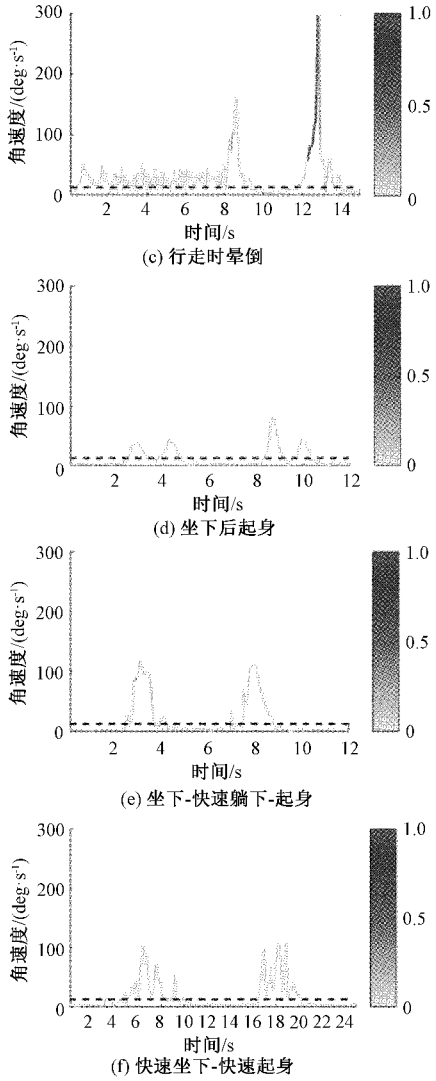


图 9　算法的实际验证

3.4　相关算法对比

表 3 分别给出了几种不同类型算法在 Sisfall 数据集中的表现。其中 LR(logistic regression)算法是典型的机器学习算法，采用了 Kim 构建的特征[6]，并用 ILFS(infinite latent feature selection)方法[15]选择特征。ILFS 是一种能考虑所有可能的特征子集而对特征重要性排序的方法，本文基于 ILFS 选择了最优的 30 个特征。可以看出，基于阈值的算法具有很低的算法延时和较高的裕量时间，但准确率最高仅有 92.4%。值得指出的是，裕量时间并不是越长越好[12]。表 3 中可以看出 Ahn 的算法裕量时间是最长的，而其特异性却是最低的。这可能是因为该阈值算法没有抓住跌倒失衡的关键特征，在人失衡之前产生了误报。

算法特异性是值得关注的，特异性不高意味着算法易对一些正常行为动作产生虚警，导致误触发气囊装置，降低用户体验，而且耗费材料和成本。LR 算法的特异性为 92.28%，而提出的轻量级 CNN 模型达到了 95.42%的特

表 3 相关算法对比

算法	算法表现				
	ACC/%	SEN/%	SPE/%	LT/ms	Delay/ms
阈值[2]	90.3	100	83.9	401.0	0.03
阈值[3]	92.4	96.1	90.5	280.0	0.025
LR	94.59	96.92	92.28	327.1	5.83
本文	97.5	99.57	95.42	204.3	8.24

异性,一方面可认为人工提取的特征并没有卷积网络自动提取的合适,一方面可归结为 LR 算法欠拟合。必须指出的是,人跌倒是低概率事件,而正常行为动作无时不刻在发生,虽然该轻量级 CNN 的特异性高于其他三者,但仍然具有一定的提升空间。

4 结 论

本文使用深度可分离网络设计了一种轻量级卷积网络用于跌倒预测,融合了加速度、角速度、姿态角信号及个体生理特征,构建了网络模型。该模型在 Sisfall 数据集中获得了 97.5%的准确率,204.3 ms 裕量时间,相比于其他算法获得了更高的准确率。轻量化是该模型的显著特点,模型中使用了更小的卷积核和合理的通道数实现了更高的准确率。为将网络模型部署于 STM32 单片机设计了能够实时运行的两级框架,在保证准确率的同时大大降低了嵌入式端计算负担。并且在 STM32 单片机平台的验证表明该算法延时为 8.24 ms,满足了可穿戴设备实时性要求,也为跌倒保护器的开发提供了进一步参考。

参考文献

[1] ZHONG Z, CHEN F, ZHAI Q, et al. A real-time pre-impact fall detection and protection system[C]. 2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), 2018: 1039-1044.

[2] AHN S, KIM J, KOO B, et al. Evaluation of inertial sensor-based pre-impact fall detection algorithms using public dataset[J]. Sensors(Basel), 2019, 19(4). DOI: 10.3390/s19040774.

[3] JUNG H, KOO B, KIM J, et al. Enhanced algorithm for the detection of preimpact fall for wearable airbags[J]. Sensors(Basel), 2020, 20(5). DOI: 10.3390/s20051277.

[4] 杨智超, 李国辉, 李佳韵, 等. 基于分散熵和支持向量机的运动状态识别[J]. 国外电子测量技术, 2019, 38(7): 28-31.

[5] KIM T H, CHOI A, HEO H M, et al. Machine learning-based pre-impact fall detection model to discriminate various types of fall[J]. Journal of Biomechanical Engineering, 2019, DOI: 10.1115/1.4043449.

[6] WANG L, PENG M, ZHOU Q. Pre-impact fall detection based on multi-source CNN ensemble[J]. IEEE Sensors Journal, 2020, 20(10): 5442-5451.

[7] YU X, JANG J, XIONG S. A large-scale open motion dataset(KFall) and benchmark algorithms for detecting pre-impact fall of the elderly using wearable inertial sensors[J]. Front Aging Neurosci, 2021, DOI: 10.3389/fnagi.2021.692865.

[8] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. ArXiv Preprint, 2017, ArXiv: 1704.04861.

[9] 涂亚庆, 陈鹏, 陈宝欣, 等. 基于离散特征的跌倒检测智能方法及应用[J]. 仪器仪表学报, 2017, 38(3): 629-634.

[10] 吕艳, 张萌, 姜昊昊, 等. 采用卷积神经网络的老年人跌倒检测系统设计[J]. 浙江大学学报(工学版), 2019, 53(6): 1130-1138.

[11] 任晶秋, 蒋杨, 张光华. AHRS 的老人跌倒检测算法[J]. 电子测量与仪器学报, 2020, 34(12): 190-196.

[12] AZIZ O, RUSSELL C M, PARK E J, et al. The effect of window size and lead time on pre-impact fall detection accuracy using support vector machine analysis of waist mounted inertial sensor data[C]. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2014: 30-33.

[13] CASILARI E, SANTOYO-RAMÓN J A, CANO-GARCÍA J M. Analysis of public datasets for wearable fall detection systems[J]. Sensors, 2017, 17(7): 1513.

[14] SUCERQUIA A, LÓPEZ J D, VARGAS-BONILLA J F. SisFall: A fall and movement dataset[J]. Sensors, 2017, 17(1): 198.

[15] ROFFO G, MELZI S, CASTELLANI U, et al. Infinite latent feature selection: A probabilistic latent graph-based ranking approach [C]. 2017 IEEE International Conference on Computer Vision(ICCV), 2017: 1407-1415.

作者简介

杜群贵,工学博士,教授,主要研究方向为机器动力学,模式识别。

E-mail: ctqgdu@scut.edu.cn

钟威,硕士研究生,主要研究方向为传感器数据挖掘,物联网。

E-mail: zwspored@foxmail.com