第一节 总则

介绍

老年人和患者的安全和救援成为计算机视觉和现代技术的迫切目标。老年人跌倒后往往无法站起来。因此，老年人更容易因跌倒而意外死亡[1]。目前，许多自动跌倒检测系统用于监测家庭和医院中的老人和病人。

跌倒检测系统可分为：环境设备、可穿戴设备和基于视觉的设备（摄像头）。基于可穿戴设备的检测包括使用加速度计和陀螺仪等设备[2]。虽然这些设备易于使用，价格低廉，并且可以帮助监控各地的人员，但此类设备存在用户忘记、容易被盗或忘记给空电池充电等缺点。这些缺点使此类设备成为老年人不方便的选择。

基于环境的方法依赖于使用外部传感器收集老年人周围活动目标区域的瞬时数据。压力和振动传感器是环境传感器的例子，它们被放置在地板上以指示人员的瞬时位置。虽然这种方法便宜且非侵入式，但此类传感器由于对活动区域内的所有物体都很敏感，导致检测率低，因此会引起高误报率。

由于摄像头的普及和计算机视觉的进步，基于跌倒检测的视觉方法为照顾老年人提供了一种有前途的解决方案。与以前的系统相比，基于视觉的系统最适合定位和监控多个人，以及同时检测多个动作。此外，长者不必携带专门的设备。摄像头易于安装，可提供有关人员的大量信息，例如位置、动作和运动 [3]。在这种方法中，移动物体被检测，分类为人，并在视频的每一帧中进行跟踪。使用计算机视觉技术，根据为被检测者提取和分析的有用特征（例如提取的轮廓）进行跌倒识别。最近，深度学习在计算机视觉的各种问题中作为一种强大的技术出现。利用深度学习检测老年人跌倒，提高了跌倒检测的准确性，减少了误报。

本文组织如下：第 2 节介绍了跌倒检测系统的相关工作。在第 3 节中，方法论和建议的方法。在第 4 节中实验设置和结果。在第 5 节 结论和讨论

第二节.

相关工作

摄像头无处不在，无需携带任何设备，这使得基于视觉的方法最适合自动检测跌倒。大多数视觉跌倒检测系统的步骤包括视频采集、视频分析和报警通信。视觉跌倒检测系统的性能主要取决于对视频帧的分析来检测跌倒。视频分析步骤最困难的挑战是由于背景移动、照明变化等多种原因导致检测结果不佳导致人体形状变形。其他挑战，如遮挡、相机位置以及不同动作（如坐姿和跌倒）中人体形状的相似性，也可能导致跌倒检测方法的性能不佳。

M. Khan 等[5]提出了运动梯度和人体形状变化特征的混合，以检测人体跌倒。大动作用于描述人类跌倒的特征。跟踪人体运动用于确定全局运动方向，从而决定人体运动的方向。跌落与其他活动不同，具体取决于针

对特定阈值测试的纵横比。在蹲下动作的情况下，由于跌倒和蹲下动作之间的纵横比相似，可能会产生误报。

Yu 等[6]提出了另一种基于人体轮廓分析的人体姿态识别跌倒检测方法。为了淡化对人形的错误分析问题，采用了背景减法技术。然后，使用投影直方图和拟合椭圆作为全局和局部特征来描述不同姿势;最后，利用图支持向量机（DAGSVM）对下降进行分类。这种方法面临两个挑战。第一个挑战是遮挡问题，secobd 挑战是场景中多个运动物体的存在问题。Shukla 和 Tiwari[7]提出了一种基于两个特征的人体跌倒检测方法。这两个特征是人体中心到地面的高度和人体运动的历史。该方法通过应用背景减法来检测物体，从而提供移动物体的完整轮廓。然后，去除噪声以正确跟踪对象。通过确定物体和 MHI 的质心来进行分类。根据阈值角度分类的跌落：如果轮廓的中心点的值小于特定阈值;那么运动是跌倒，否则就不是跌倒。

Zerrouki 等[8]提出了一种基于视觉监测中人体轮廓形状变化的跌倒检测方法。该系统采用 HMM 和 SVM 来识别视频中的跌倒和非跌倒，这取决于从 RGB 相机中提取的信息，这会导致某些跌倒的错误分类，尤其是在低质量图像的情况下。使用传统的 RGB 相机也很难提取人体的轮廓，尤其是在黑暗或遮挡的情况下。

Doulamis 和 N. Doulamis[9]介绍了一种自适应深度学习方法来检测人类跌倒。首先，基于监督学习方法，将介绍与背景隔离开来。之后，该网络在人体模型中进行训练，该模型将人体描述为人脸下方的一个地方，这是使用二维功能实现的，该功能导致相对于其在该区域中的位置属于人体的像素。最后，我们依靠一种决策机制，该机制允许调整网络权重以适应当前的视觉条件，而决策机制确保当前环境是否与获得的知识“相似”，从而对人类堕落进行分类。由于运动中缺乏活动或存在各种运动而发生误报，因此忽略了人的运动。

Marcos 和 G. Azkune[10]提出了一种跌倒检测方法，它根据 CNN（卷积神经网络）对一系列帧中跌倒的存在进行分类。使用一组光流图像，使用在 ImageNet 上预训练的 VGG-16 CNN 模型对特征进行提取和分类。虽然这种方法提供了很高的准确性和效率，但由于缺乏足够的训练网络，它遭受了事件错误分类的困扰。

为了应对前面提到的挑战，本文提出的框架结合了手工制作的特征和卷积神经网络 RetinaNet 和 Mobilenet。RetinaNet 在用于人类检测的卷积神经网络中具有最高的准确性。此外，Retinanet 还应对了传统背景减法技术的挑战。因此，所提出的框架依赖于 RetinaNet 来准确检测人类在一系列帧中的位置。然后，将边界框（围绕检测到的人）纵横比和头部位置检测为检测到的人形特征的表示。此外，还获得了将运动物体的运动聚合在单个图像中的运动历史图像（MHI）。这些特征用于形成用于训练移动网络的特征图，该特征图用于将人类行为分类为跌倒或不跌倒。MobileNet 使用两个全局超参数，在准确性和延迟之间进行有效权衡。所提出的框架利用了卷积神经网络 RetinaNet 和 Mobilenet 的效率和准确性。为了评估所提框架的有效性，利用 UR 和 FDD 数据集对所提框架进行了评估，实验结果证明了所提框架的效率与现有方法相比达到了 98%的准确率。

第三节.

建议的方法

本文提出的跌倒检测框架依赖于 CNN 网络来检测视频中移动的人体，并将人体运动分类为“跌倒”或“不跌倒”。所提框架的三个阶段如图 1 所示。拟议框架的第一阶段是检测和跟踪视频帧中移动的人类。RetinaNet 是用于检测移动人类的选定模型。RetinaNet 显示出高水平的准确度和可接受的检测速度[11]。该框架的第二阶段是提取人体运动和人体形状的特征。这些特征包括人体周围边界框的宽高比、显示物体运动方向的运动历史图像、头部跟踪以及指示物体相对于地面位置的方向。这些特征在帧序列中被跟踪并输入到第三阶段，以对人体运动“跌倒”或“不跌倒”进行分类。所提方法的第三阶段采用 MobileNets[12]的修改版本，根据第二阶段提取的特征将人体运动分类为“跌倒”或“不跌倒”。MobileNet 将提取的特征提供给第一个卷积层，并将结果传递给后续卷积层，这些卷积层又通过应用深度卷积对输入进行分类，直到到达最后一层，最后一层根据最后一层的输出做出决定。

A. 人体检测

以前的大多数跌倒检测方法都是基于背景减法来检测物体，然后将这些物体分类为人类或非人类。这些基于背景减法的方法面临着光照变化、动态背景、伪装等挑战。这些挑战可能导致人体形状变形，从而影响跌倒检测框架的性能。与以往的方法不同，所提出的方法使用深度卷积神经网络“RetinaNet”作为人体检测网络。RetinaNet 在用于人体检测的其他卷积神经网络中具有最高的准确性，并应对了传统背景减法技术的挑战[13]。

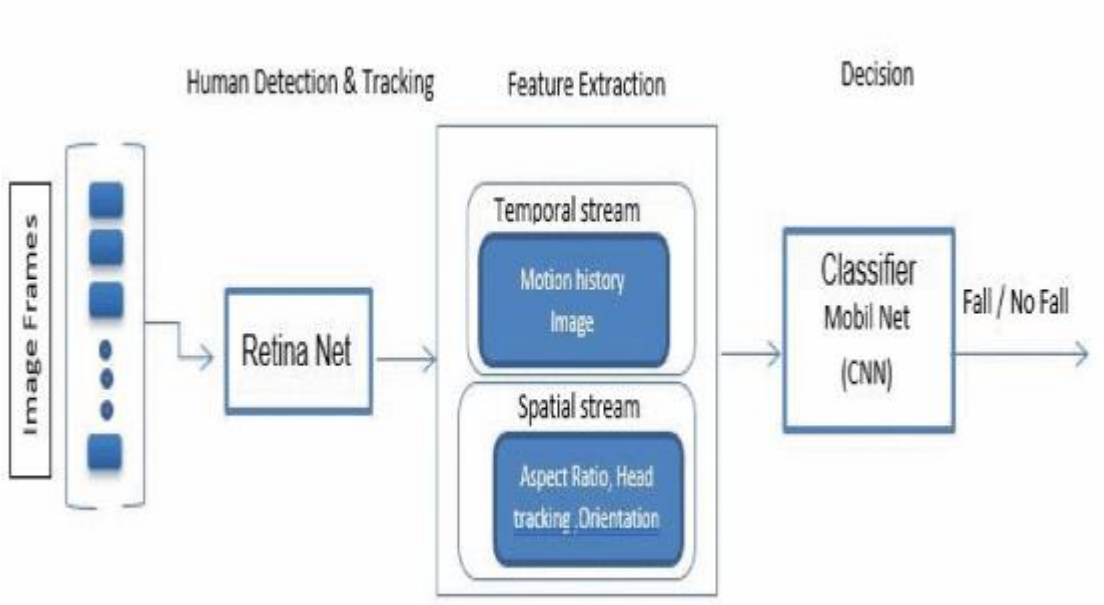


图 1.

拟议框架阶段的概述。

显示全部

RetinaNet 的分层结构包括两个子网和一个骨干网络。RetinaNet 主干网提取特征，然后两个子网层根据主干网的输出对对象进行分类和回归。图 2 显示了 RetinaNet 的总体架构。第一个子网根据主干网的输出对对象进行分类。第二个子网使用主干输出来构造边界框回归。输入帧的特征图由特征金字塔网络（FBN）提取，FBN 是 RetinaNet 的骨干。类似于 RPN 中的锚框用于标识每个位置的对象边界。这些锚框经过修改，可进行具有适配阈值的多类检测。这些锚点分别包括金字塔 P3 到 P7 级别上 32 到 512 的区域，纵横比为 1：2、1：1 和 2：1。因此，每个级别和现有级别有 9 个锚点，根据输入图像到网络，覆盖从 32 到 813 像素的比例范围。如图 3 所示，每个锚点设置为一个向量，每个向量设置为 K 个对象类之一和一个用于盒回归的 4 向量。然后，回归子网根据阈值为 0.5（经验设置）的交集并集 （IoU） 将每个锚点分配给附近的真值对象框。此后，分类子网预测每个 A 锚点和 K 对象类在每个位置的存在或不存在对象。如果锚点未分配给地面实况，则在训练期间将忽略此锚点。该子网的输出是对象在场景中的空间位置，如图 4 所示。

B. 特征提取

虽然所提方法依靠 CNN 进行人体检测和分类以提高精度，但所提方法通过手动提取检测对象的特征，而不是自动提取特征，从而缩短了计算时间。跌倒检测系统的主要问题之一是选择要分类为跌倒或不跌倒的特征。从初始数据空间移动到特征空间，使跌倒检测更易于识别。该方法基于检测到的物体形状和运动特征来决定坠落。为了保留丰富的对象信息，选择了边界框的宽高比（apect ratio）、对象运动历史图像和头部跟踪要素。

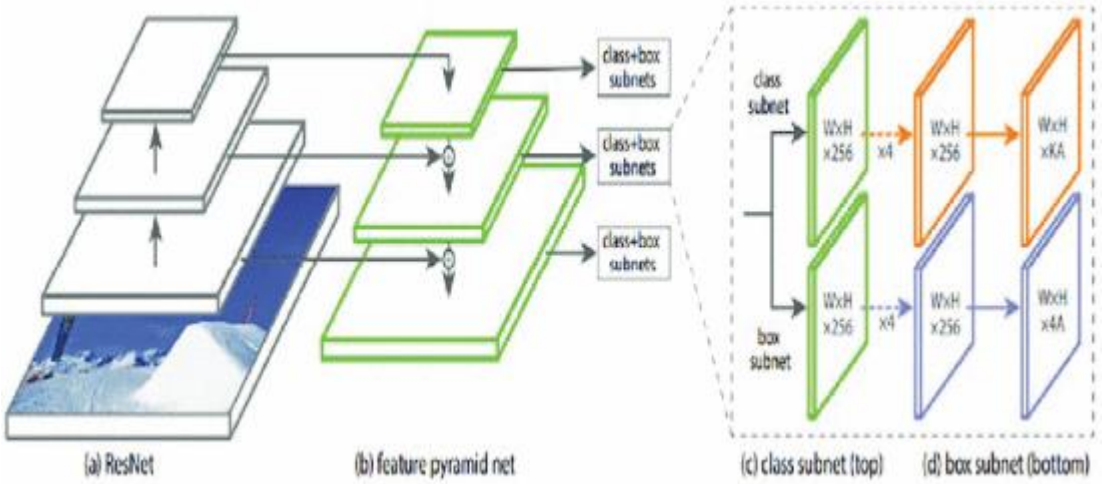


图 2.

RetinaNet 模型架构。

显示全部

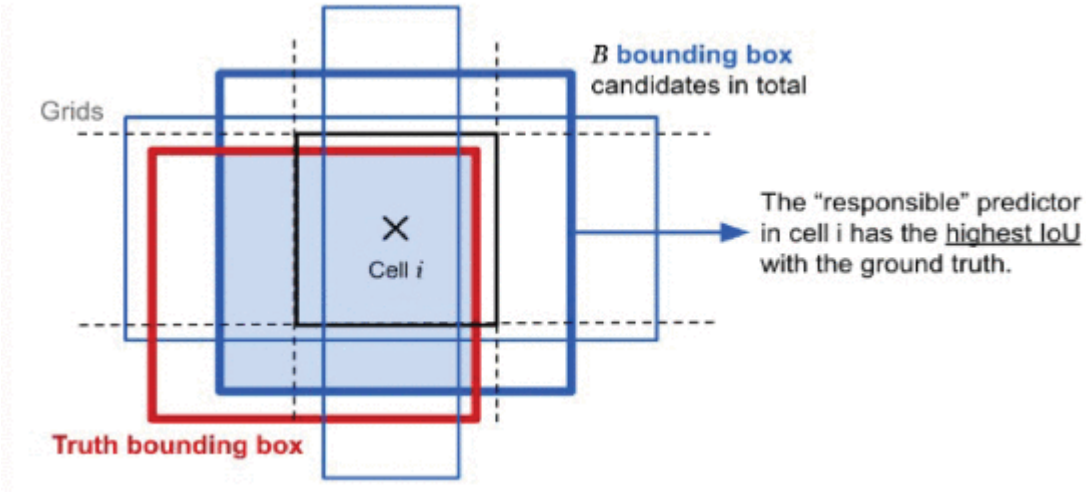


图 3.

边界框生成。

显示全部

1） 纵横比

在目标检测中，边界框用于描述目标位置[14]。边界框可以定义为一个矩形正方形，由左上角的 x 和 y 轴坐标以及右下角的 x 和 y 轴坐标定义。纵横比表示框的宽度和高度之间的比率。按理说，跌倒时的宽高比大，运动或蹲下时小。

2） 运动历史图像 （MHI）

运动历史图像（MHI）是由 Bobick 和 Davis 引入的，作为一种动作表示方法，它将运动物体的运动聚合在单个图像中[15]。MHI 是一种模板图像，用于识别和理解人体动作以及运动方向。人体运动的聚合以及运动的方向被馈送到下一阶段的分类器。该方法主要依靠运动聚合和运动方向作为决定人体坠落的强指标。三菱重工的定义由式说明：

$$H_\tau(x，y，t)=\{\tau,\text{麦克斯}(0，H_\tau(x，y，t-1）， \text{如果 D（x，y，t）=1 否则}(1)$$

查看源代码

图 4.

IOU 阈值下的检测结果 =0.5。

显示全部

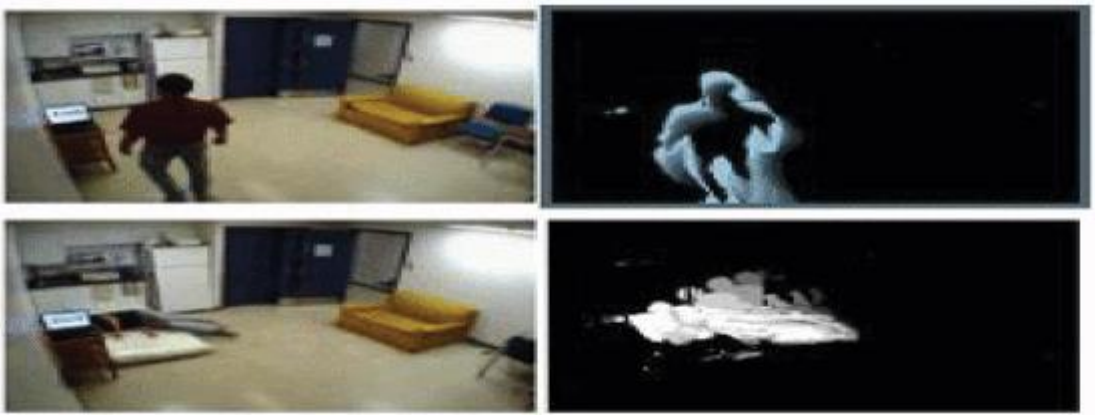


图 5.

两种不同动作的 MHI 图像示例。

显示全部

其中 D（x，y，t） 是指二进制图像在时间 t 处的像素 （x，y） 的值，带有“1”的 D 表示运动区域。采用帧差分法得到 D（x，y，t）。$H_\tau(x、y、t)$是一个标量值图像，其中上次运动发生的区域比前一帧中的运动更亮。图 5 显示了两个不同动作的两个运动历史图像的示例。

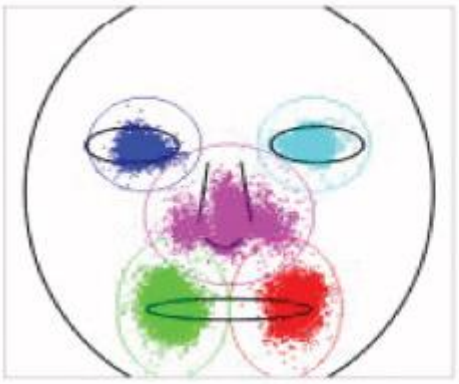
3） 头部追踪

所提出的方法依赖于检测和跟踪移动人体的头部，因为人体坠落过程中的头部运动是显而易见的。头部由椭圆近似。所提方法通过应用人脸检测系统检测图像中的人脸来检测人头。利用相同 Haar-like 的增强级联来检测帧中的人脸[16]。如图 6 所示，一组两个特征组用于检测正面视图和剖面图中的人脸。在正面视图中检

测人脸的特征是：左右眼、鼻子、左右嘴角。人脸的整个可见部分的特征用作在剖面图中检测人脸的特征。通过在输入帧上应用滑动窗口来搜索正面或轮廓特征，为检测到的人脸定位一个边界框。一旦检测到人脸，就会使用神经网络来估计头部姿势的三个旋转角度：滚动、俯仰和偏航，如图 7 所示。使用具有一个隐藏层的多层前馈网络，每个单元通过激活函数传递输入的加权信号来计算其输出。该网络的输入是坐标 （x，y），它决定了面部特征的位置以及面部特征对 i 和 j($d_x=x_{我}-x_j$, $d_y=y_{我}-y_j$)这两个值都在 [−0.5， +0.5] 范围内归一化。头部探测器盖的角度范围分别为 [-25， +25]、[-40， +40] 和 [-90， +90]，分别用于俯仰、滚动和偏航。

C. 分类

本文提出的跌倒检测框架依赖于一个轻量级的深度卷积神经网络 MobileNet。MobileNet 是一种快速、准确、小尺寸的深度卷积神经网络，可用于分类和检测。MobileNet 网络的修改版本用于在前一阶段对提取的特征进行分类，并给出最终决定是否下降。使用公共数据集 FDD [17]对网络进行重新训练。使用 RetinaNet 对数据集的训练视频中的人体进行检测，然后按照前文所述提取边界框、MHI 和头部检测器的特征纵横比。对于训练视频的每一帧，都会计算 MHI。框架和 MHI 都用于馈送 MobileNet 的第一层。输入 RGB 帧包括视觉特征，MHI 表示视频中人类的运动特征。输入帧和 MHI 的大小调整为 [224]、[224]、[3] 维的向量。MobileNet 模型的主要思想基于深度可分离卷积，其中每个块由一个 3x3 深度卷积组成，该卷积过滤输入帧，然后是 1x1 逐点卷积。MobileNet 有 28 层，按深度和逐点卷积作为单独的层进行计数，然后是池化层和全连接层。每一层之后是批量归一化和 ReLU 激活。输入帧和 MHI 通过第一个卷积层，它是一个完全卷积层，然后应用 3 x 3 深度卷积，然后应用 1 x 1 逐点卷积。此后，将这些过滤后的值组合在一起以创建特征图。第一层的输出用于馈送下一层，并应用第一层的相同先前操作来创建特征图，但尺寸比前一层小，深度更深。该过程一直持续到到达池化层，池化层将特征图缩放到更小的尺寸和更深的深度，以达到预定义的大小（例如 7×7），以避免过度拟合。对于分类，密集层的输入是最终池化层的输出，该池化层被展平，然后馈送到密集层中。最后，在输出层中使用 SoftMax，以便将输出值转换为范围 [0]、[1]，并给出动作下降或不下降的最终决定。图 8 说明了 MobileNet 的体系结构。



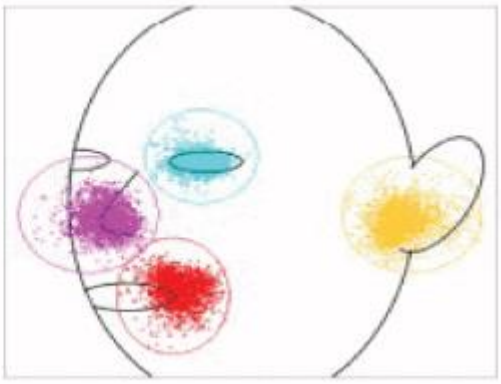


图 6.

正面和左侧剖面图的面部特征检测器示例。

显示全部

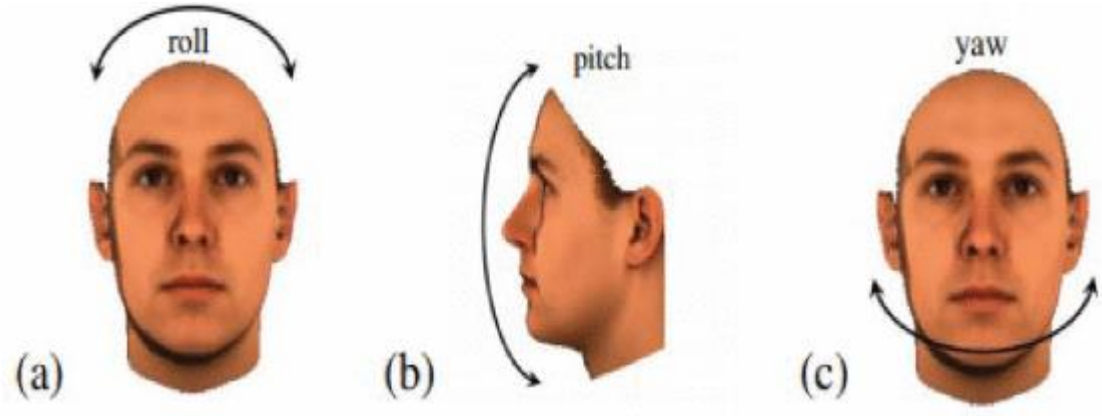


图 7.

三个头部角度 （a） 横滚角、（b） 俯仰角和 （c） 偏航角。

显示全部

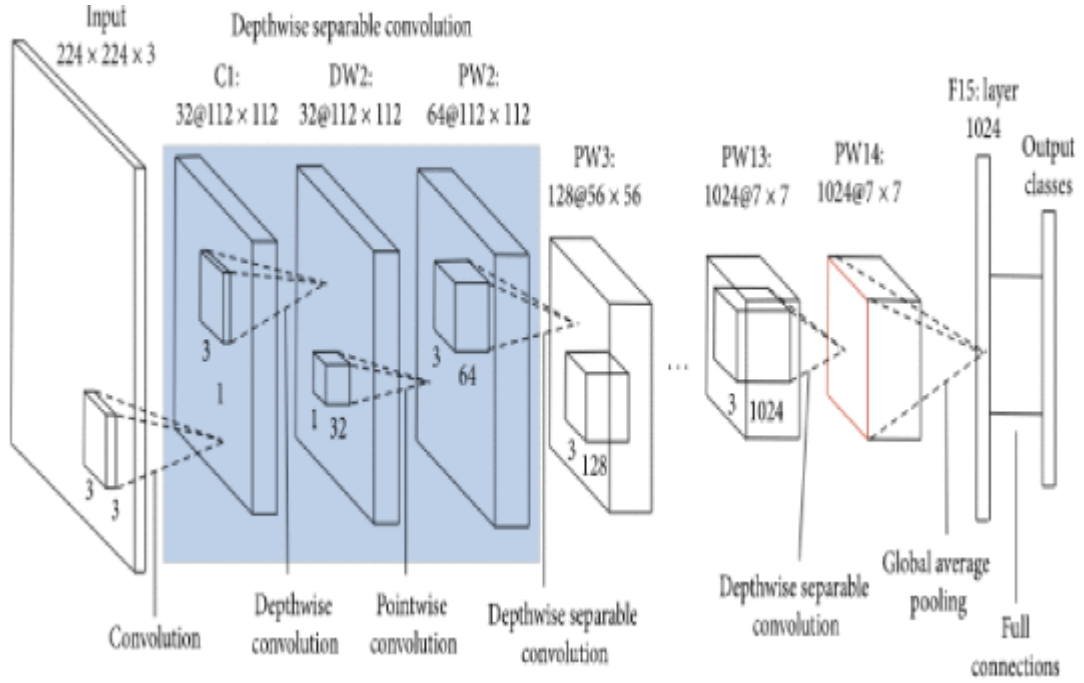


图 8.

MobileNet 的体系结构。

显示全部

第四节.

实验结果

所提出的框架的训练和测试是使用 python 版本 3.0、TensorFlow 和 Keras 实现的。Framework 在 Google Colab（colabrarory 虚拟机）上在线训练，在云上提供免费的 GPU 和内存资源。所提框架的测试和评估值来自训练模型的第 200 纪元。使用两个数据集对所提出的框架进行了测试和评估：UR 跌倒检测（URFD） 数据集和跌倒检测数据集 FDD。UR 跌倒数据集 （URFD） 包括 30 个跌倒视频和 40 个日常活动视频，称为无跌倒。FDD 数据集包括 191 个视频，帧率为 25 帧/秒，分辨率为 320x240 像素[18]。视频帧有三个通道（即 R、G 和 B），并调整为 224x224x3。FDD 数据集记录在不同的位置，允许定义多个评估协议（“家庭”、“咖啡室”、“办公室”和“演讲室”）。之所以选择 FDD 数据集来训练所提出的框架，是因为它保留了每一帧上被检测者的坐标和坠落动作的坐标。因此，训练是在一小部分数据集视频（“Coffee_room_01”、“Coffee_room_02”、“Home_01”、“Home_02”）上进行的。数据集视频序列包含照明变化、遮挡和纹理背景等挑战。视频中只有一个演员。

MobileNet 训练了 200 个 epoch，在学习过程中针对训练和验证数据评估每个 epoch 的损失函数。所提出的方法应用在 30 个 epoch 处提前停止，因为损失没有改善。这是一种理想的培训策略，可以在不需要时避免全面培训。对于损失函数，值 1 用于 w1 类权重，因为不需要增加“无下降”类别。图 9 显示了每个时期从训练和验证中获得的损失。如图所示，当 epoch 增加时，训练值和验证值都接近 0，输出值中的训练值和验证值收敛，这被认为是所提框架的精度指示。所提框架的测试和评估值来自训练模型的 epoch 200，训练 FDD 数据集的 34 epoch 大约需要 72 分钟。该实验使用表（1）所示的以下超参数进行。

Parameter	value
EPOCH	200
LEARNING_RATE	0.01
MOMENTUM	0.95
WEIGHT_DECAY	0.001
BATCH_SIZE	32

表一.

Mobilnet 训练的参数设置

显示全部

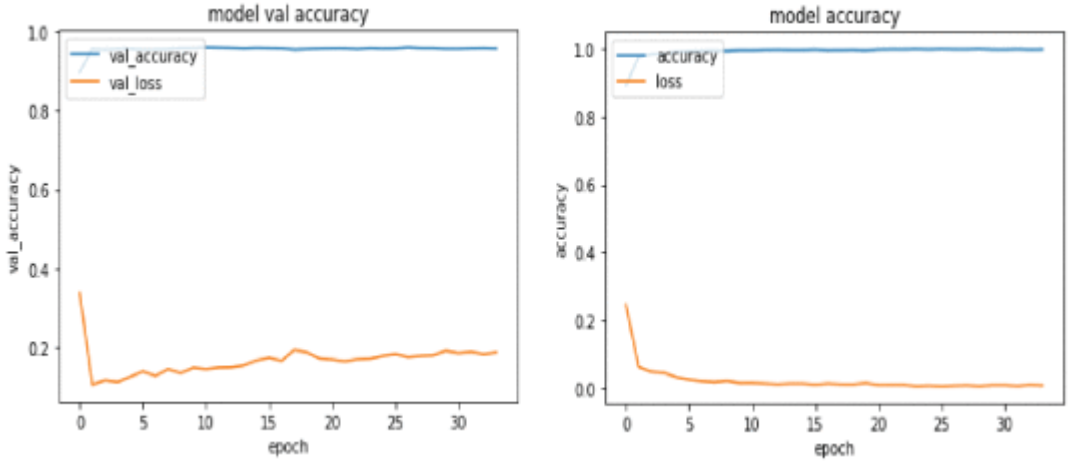


图 9.

每个 epoch 的训练和验证损失。

显示全部

A. 定性评价

为了评估所提方法的有效性，将所提框架的实验结果与几种跌倒检测方法进行了定性和定量的对比。将所提出的框架与其他使用手工制作特征的方法进行了比较，如 Yu al.方法[19]、Charfi 等[20]和 Rougier 等[21]。此外，将所提出的框架与基于 CNN 的跌倒检测方法（如 Adrian 等[22]和 Fakhruddin 等[23]进行了比较。图 10 显示了前面描述的框架和方法在不同帧和场景中的视觉结果比较。图 10 的第一行是从视频场景中生成的输入帧，第二行是用于评估过程的真值，下一行是不同可比方法的结果，最后一行显示了所提出的框架的结果。如图 10 所示，Yu 等[19]、Charfi 等[20]和 Rougier 等[21]基于手工特征提取的方法的结果似乎不稳定，在场景中人与其他物体之间的遮挡问题和高误分类的情况下，结果很差。Adrian 等[22]的方法基于光流图像，该方法涉及对连续帧进行预处理的大量计算，并且在照明变化的情况下无法检测到下降。Fakhruddin 等人的方法需要大量数据来检测下降，这消耗了大量的网络带宽。所提出的框架提供了准确的人体跌倒决策，并且在不同情况下（例如遮挡和照明变化）在视觉上优于几种可比较的方法。

B. 定量评价

为了评估所提出的框架相对于可比较的跌倒检测方法的效率，使用了准确性、灵敏度、特异性和精密度指标。为了计算这些指标，参数真阳性 （TP） 表示检测为跌倒的跌倒事件数、假阳性 （FP） 表示检测为跌倒事件（也称为误报）的非跌倒事件数、真阴性 （TN） 表示检测为非跌倒事件的非跌倒事件数，以及表示检测为非跌倒事件的跌倒事件数的假阴性 （FN）。准确度、灵敏度、特异性和精密度指标指标的计算方法如下：

$$A\,c\,c\,ura\,cy=(TP+TN)/(TP+FP+FP+FN)(2)$$

查看源代码

$$\text{Sens t i vity/reca l l}=(TP)/(TP+FN) \quad (3)$$

查看源代码

$$\text{SPECIf 我 city}=TN/(TN+FP) \quad (4)$$

查看源代码

$$\text{Perc i sion}=TP/(TP+FP) \quad (5)$$

查看源代码

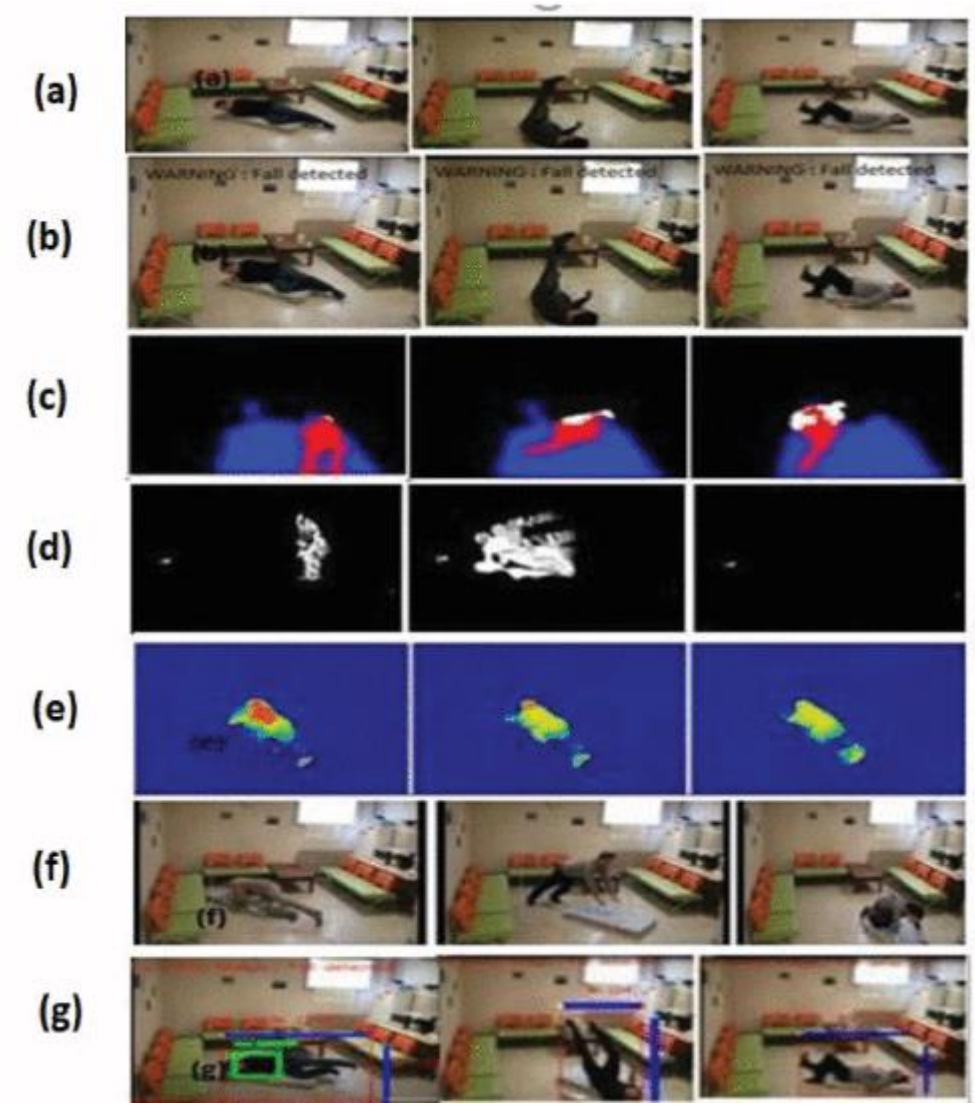


图 10.

视觉结果比较 （a） 输入帧，（b） 地面实况帧，（c） Yu 等人 [19] 的检测输出，（d） Rougier 等人 [21] 的检测输出，（e） Adrian 等人 [22] 的检测输出，（f） Charfi 等人 [20] 的检测输出，以及 （g） 所提出方法的检测输出。

显示全部

图 11 显示了使用基于 FDD 数据集的框架得到的混淆矩阵。虽然在一些测试视频中，由于类似跌倒的活动，出现了“跌倒”和“不跌倒”之间的混淆，但大多数样本都正确分类，所提出的框架达到了 98.8%的整体准确率。

为了保证所提框架基于 UR 和 FDD 数据集的鲁棒性，与最先进的跌倒检测方法进行了多次比较。所提出的框架优于 Yu 等[19]、Charfi 等[20]和 Rougier 等[21]方法的结果，因为所提出的框架使用 RetinNet 来检测人类，避免在静止情况下出现任何错误分类，并避免背景和检测到的人类之间的重叠。此外，为能够很好地检测到的人类提取手工制作的特征（由于 RetinNet 的效率）来训练 MobileNet 网络有助于实现更高的准确性。所提出的框架采用 MobileNet 作为分类器，与[22]中提出的方法中使用的 VGG 相比，它小了 30 倍，速度快了约 10 倍。MobileNet 中的深度卷积证明了高精度、低延迟和低功耗，模型性能为 0.04 fps。所提出的框架在测试数据上实现了 98.5% 的准确率，其中框架在“Coffee_room_01”、“Coffee_room_02”、“Home_01”、“Home_02”上进行了训练，作为使用数据的一部分。与[23]中提出的工作相比，所提出的框架准确率最高，达到了 92.3%。此外，所提出的框架在分析躺姿的灵敏度方面达到了 98%，这在跌倒检测中是非常理想的，因为跌倒后的姿势被认为是躺着的。同样，所提出的框架无论人体位置、照明变化或面向任何相机角度的侧面，都显示出高精度。图 12 显示了所提出的框架与其他最先进的跌倒检测方法之间的比较。

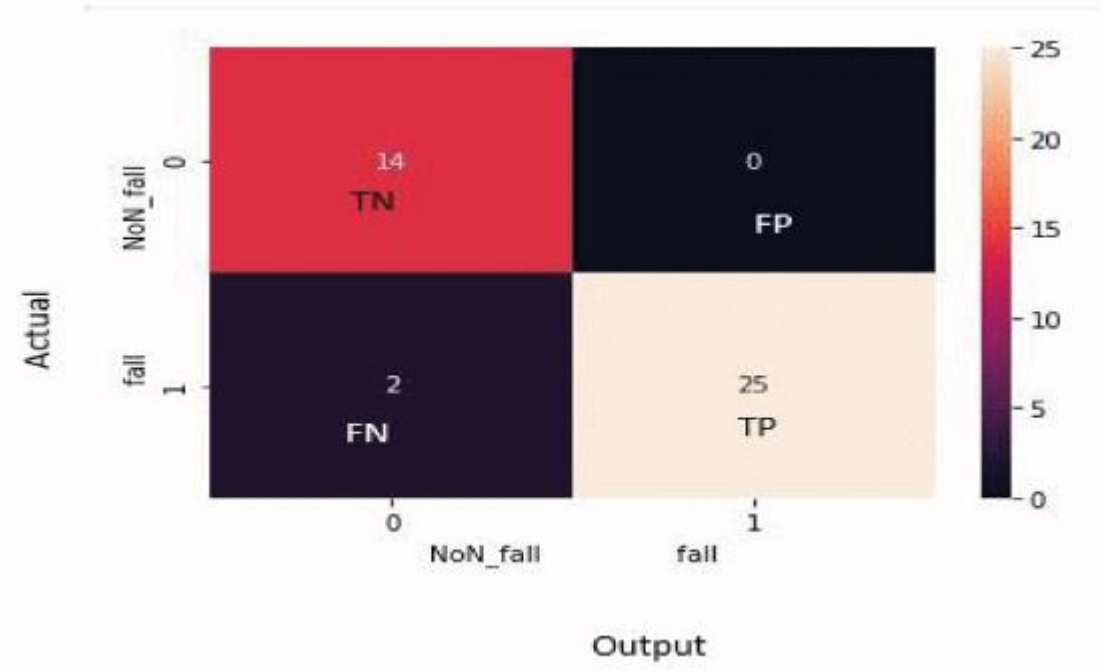


图 11.

FDD 数据集的混淆矩阵。

显示全部

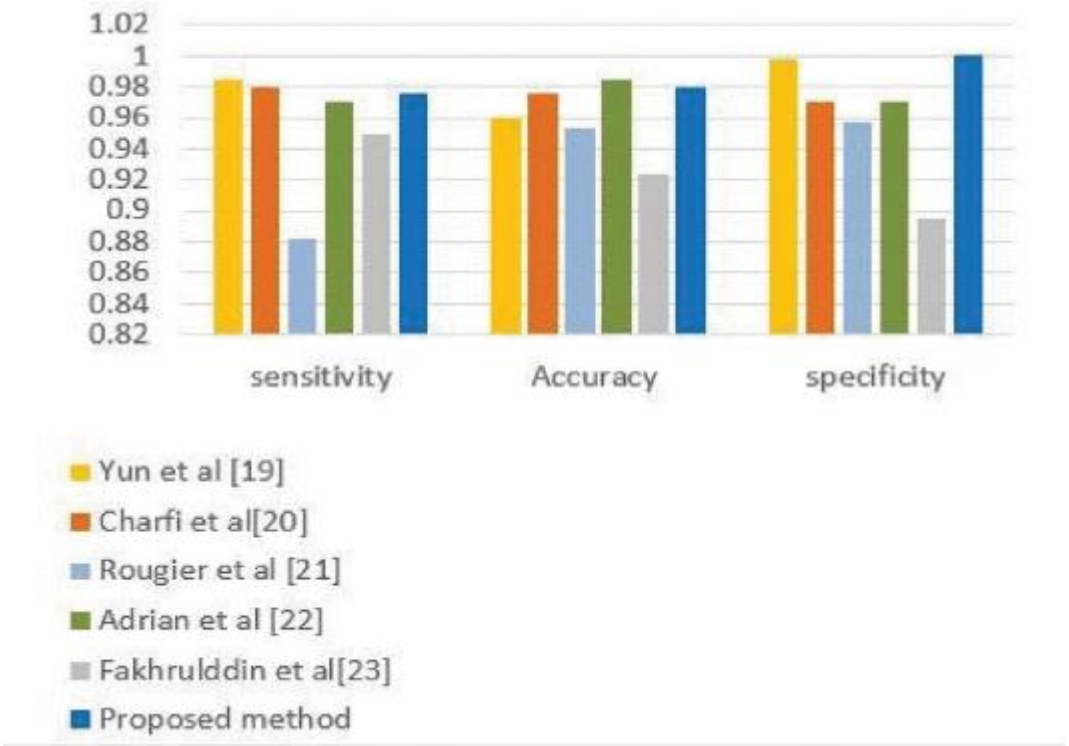


图 12.

准确性、灵敏度和特异性 所提出的框架与最先进的跌倒检测方法之间的比较。

显示全部

第五节

结论

本文提出了一种基于卷积神经网络和手工特征相结合的人体跌倒检测框架。该方法基于对视频中人体外观、运动和形状变化的分析。所提出的跌倒检测分三个阶段进行：人工检测、特征提取和动作分类以跌倒或不跌倒。RetinaNet 是一种单阶段目标检测 CNN，用于准确确定视频中人的位置。特征提取是所提框架的第二阶段，其中提取了表示检测到的人体运动和身体形状（MHI、纵横比和头部检测）的特征。这些特征构成了一个特征图，用于提供 MobileNet 模型。所提方法的第三阶段使用 MobileNets 的修改版本，根据第二阶段提取的特征将人体运动分类为“跌倒”或“不跌倒”。实验结果表明，所提框架在不同场景和情况下检测人体跌倒的效率高达 98%。在未来的工作中，使用卷积神经网络进行特征提取旨在代替手工制作的特征，同时分析计算时间，以手工制作特征和 CNN 特征提取来分析所提出的框架。