文章编号:1000－2367(2017)03－0096－06 DOI:10.16366/j.cnki.1000－2367.2017.03.014

一种基于双流卷积神经网络跌倒识别方法

袁智,胡辉

(华东交通大学 信息工程学院,南昌 330013)

摘 要:针对跌倒行为的视觉特征难以提取的问题,提出一种由两路卷积神经网络和模型融合部分组成的双流卷积神经网络(Two-Stream CNN)的跌倒识别方法.该方法的一路对视频帧的运动人加框标记后,送三维卷积神经网络(3D-CNN)处理来消除视频背景的干扰;另一路从相邻视频帧获取光流图后,送 VGGNet-16 卷积神经网络处理;最后将 3D-CNN 和 VGGNet-16 的 Softmax 输出识别概率加权融合作为 Two-Stream CNN 输出结果.实验结果表明:标记运动人并经 3D-CNN 处理有效地消除了视频背景的干扰;Two-Stream CNN 跌倒识别率为 96%,比 3D-CNN提高了 4%,比 VGGNet-16 网络提高了 3%.

关键词:跌倒识别;双流卷积神经网络;视频帧;光流图

中图分类号:TP391.4 **文献标志码**:A

中国已经步入老龄化社会,据统计每年大约有 4000 万老人发生跌倒,跌倒后没有即时获得援助会加重老人受伤的程度,严重的情况下甚至可能会导致死亡.因此,及时、准确的判断对老人意外跌倒行为具有重要的研究意义.

随着图像处理技术的快速发展,基于计算机视觉的跌倒识别受到广泛的关注.张金富等[1]通过计算人体关节距离地面的平均高度变化并融合头部和肩部在垂直方向上的速度变化特征进行跌倒识别;瞿畅等[2]利用 Kinect 的深度图像建立人体前景的动态三维包围盒,以三维包围盒的长、宽、高变化作为跌倒识别的特征;彭玉青等[3]结合背景相减法和帧差法提取人体轮廓根据人体高度、宽高比、质心变化表征人体的行为;Vaidehi 等[4]设计一种基于静态人体图像特征的跌倒检测系统,通过提取人体的长宽比和倾斜角度两个特征进行跌倒识别的研究;Rougier 等[5]用椭圆近似表示人体,将椭圆的方向标准差和长短轴比例的标准差作为特征,通过分析运动特征来检测人体形状的变换进行跌倒识别.上述方法都需要人为的对视频做大量的预处理和跌倒特征提取,而特征选择需要丰富的实践经验且直接决定算法的可靠性.2012 年 Alex Krizhevsky 等[6]设计的 AlexNet 卷积神经网络(Convolutional Neural Network,CNN)模型获得 ImageNet 大规模视觉识别挑战赛(Large Scale Visual Recognition Challenge,ILSVRC)冠军,其识别率超过亚军 10%,使得 CNN 成为近几年的研究热点并取得了巨大成功.Karen Simonyan 等[7]采用双流卷积神经网络(Two-Stream Convolutional Neural Network,Two-Stream CNN)进行基于视频的行为识别的研究.Tran 等[8]将 CNN 扩展到三维卷积神经网络(3-Dimensional Convolutional Neural Network,3D-CNN)用以处理视频信息.

本文提出一种基于 Two-Stream CNN 的跌倒识别方法.在文献[9]的基础上,采用对视频帧中的人作加框标记和 3D-CNN 来学习视频的空间维度特征以消除视频背景信息对行为识别的影响;并提取跌倒视频的光流图,用 VGGNet-16[10]学习视频的时间维度特征;将 3D-CNN 和 VGGNet-16 进行线性加权融合后的输出作为 Two-Stream CNN 的跌倒识别结果.实验结果表明,对视频帧中人加框标记后和 3D-CNN 可消除视频背景对行为识别的影响,基于 Two-Stream CNN 的跌倒识别方法有良好的可靠性.

收稿日期:2016-12-25;**修回日期**:2017-03-14.

基金项目:江西省自然科学基金(20142BAB207001)

作者简介(通信作者):胡辉(1970－),男,江西南昌人,华东交通大学教授,博士,主要研究方向为机器视觉,并行算法与并行处理,卫星导航定位,E-mail:gnss523@163.com.

1　卷积神经网络

1.1　二维卷积神经网络

本文将处理静态图片的卷积神经网络称为二维卷积神经网络(CNN).典型的 CNN 结构主要包含卷积层、下采样层、分类层,如图 1 所示[11-12].

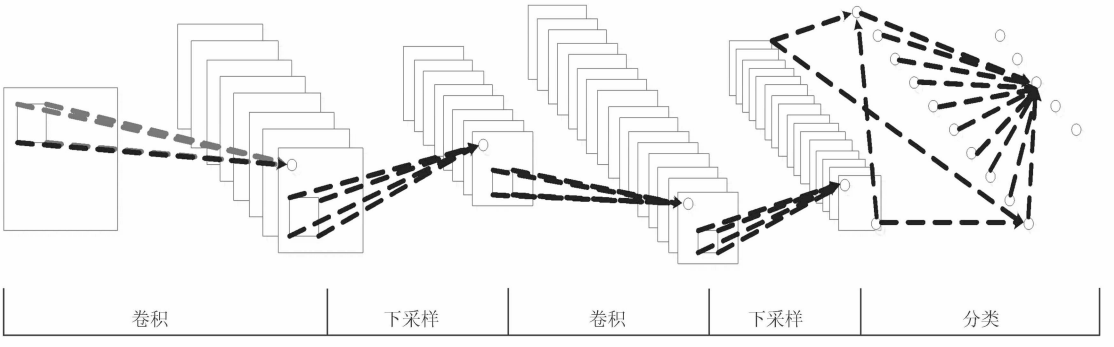


图1　CNN典型结构

CNN 通过卷积计算自动学习图像特征.卷积层接收数据输入层或下采样层的输出数据和多个卷积核进行卷积.卷积的计算过程

$$Y_{i,j}^{l}=f\left(\sum_{i\in m_h}\sum_{j\in n_w}x_{i,j}^{l}G_{i,j}^{l}+b^{l}\right),\tag{1}$$

其中,$Y_{i,j}^{l}$ 表示第 l 层输出的第(i,j) 点的值,$x_{i,j}^{l}$ 表示第 l 层第(i,j) 点的输入值,G 表示卷积核,b^{l} 表示偏置项,m_h,n_w 表示第 l 层中局部感受野的窗口大小.

CNN 的结构中,经过一次卷积后输出多个特征图.对特征图进行下采样使网络对图像旋转、平移、尺度变换具有鲁棒性,且减小了网络训练计算量.常用的下采样主要有均值采样、最大值采样.均值采样对采样窗口内的特征值求平均后作为采样结果,均值采样计算公式

$$Y_{i,j}=\frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}X_{i\times H+h,j\times W+w},\tag{2}$$

其中,$Y_{i,j}$ 表示下采样的输出值,$X_{i\times H+h,j\times W+w}$ 表示输入值,H,W 表示采样窗口的长,宽.

最大值下采样方法是取采样窗口内的最大值作为采样结果,计算公式为

$$Y_{i,j}=\max_{0\le h\le H-1,0\le w\le W-1}(x_{i\times H+h,j\times W+w}).\tag{3}$$

CNN 采用 Softmax 作为分类方法.Softmax 是 Logistic 回归模型在多分类问题上的推广.在 Softmax 回归中,类别标签 y 可以取 $k(k\ge 2)$ 个不同的值.对于训练集$\{(x^1,y^1),\cdots,(x^n,y^n)\}$,有 $y^i\in\{0,1,\cdots,k-1\}$(类别标签从 0 开始取值).

对于给定的测试集输入 x^i,Softmax 输出 x^i 属于每一个类别的概率值 $p(y^i=j\mid x^i)$,Softmax 回归的计算公式

$$h_\theta(x^i)=\begin{bmatrix}p(y^i=0\mid x^i,\theta)\\p(y^i=1\mid x^i,\theta)\\\vdots\\p(y^i=(k-1)\mid x^i,\theta)\end{bmatrix},\tag{4}$$

其中,x^i 表示测试集输入,y^i 表示类别标签,θ 为卷积神经网络的参数.

1.2　三维卷积神经网络

3D-CNN 在 CNN 的基础上增加了时间维度,用以处理视频信息.二者的主要区别在于 3D-CNN 的卷积层和下采样层需要考虑时间维度信息.

3D-CNN 的卷积层输入为连续的视频帧.三维卷积计算公式

$$Y_{i,j,k}^{l}=f(\sum_{i\in m_h}\sum_{j\in n_w}\sum_{k\in q_t}x_{i,j,k}^{l-1}G_{i,j,k}^{l}+b^{l}),\tag{5}$$

其中，$Y_{i,j,k}^{l}$ 表示第 l 层输出中的第(i,j,k) 点的输出值，$x_{i,j,k}^{l}$ 表示第 l 层中的第(i,j,k) 点的输入值.

3D 均值下采样计算公式为

$$y_{i,j,k}=\frac{1}{HWT}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\sum_{t=0}^{T-1}x_{i\times H+h,j\times W+w,k\times T+t},\tag{6}$$

其中，T 表示采样窗口帧长.

三维最大值下采样的计算公式为：

$$y_{i,j,k}=\max_{0\le h\le(H-1),0\le w\le(W-1),0\le t\le(T-1)}(x_{i\times H+h,j\times W+w,k\times T+t}).\tag{7}$$

2 基于双流卷积神经网络的跌倒识别

Two-Stream CNN 由两路卷积神经网络和模型融合部分组成. 首先，用背景相减法检测视频中的运动目标(人)，对检测到的人加框标记后作为 3D-CNN 模型的训练数据；其次，用光流法提取视频的运动信息，将光流图作为 VGGNet-16 模型的训练数据；最后将 3D-CNN 模型和 VGGNet-16 模型的 Softmax 输出加权融合作为 Two-Stream CNN 模型的输出结果. Two-Stream CNN 基本结构如图 2 所示.

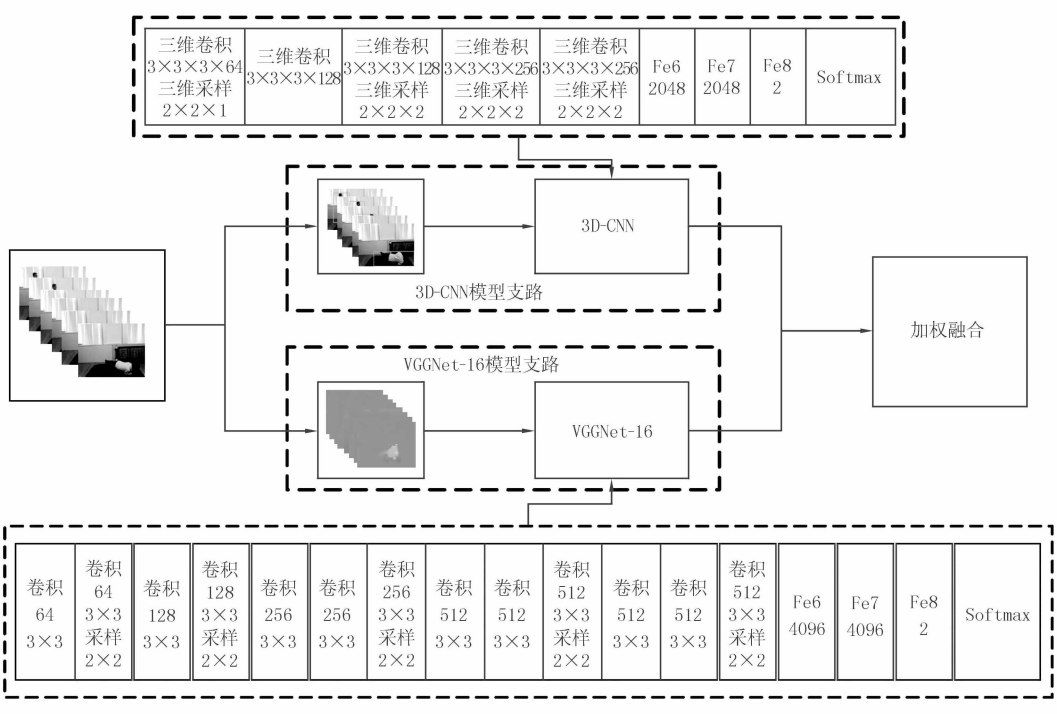


图2 Two-Stream CNN基本结构

2.1 VGGNet-16 支流

光流法是一种被广泛应用于运动目标检测方法，利用光流法计算视频中运动目标像素点在每一帧上的变化得到视频光流图，保留了视频中运动信息，将光流图作为卷积神经网络模型输入，可以直接学习行为的运动特征.

VGGNet-16 的网络结构如图 2 中所示，根据本文数据集包含的行为类别数目将 VGGNet-16 模型的 Fc8 层分类参数设置为 2.

VGGNet-16 的输入数据为视频光流图，对每一个视频的光流图每 10 帧叠加后作为一个输入样本.

2.2 3D-CNN 支流

3D-CNN 的结构参考 VGGNet-16 卷积神经网络模型结构设计，包含 5 个三维卷积层，4 个三维下采样层，3 个全连接层，采用 Softmax 作分类.

3D-CNN 中所有卷积核大小均为 3×3×3,卷积核的个数第 1 层为 64,第 2 和第 3 层为 128,第 4 和第 5 层为 256. 第 1 层卷积后,为了使更多的时间维度特征信息参与卷积计算,下采样窗口选择 2×2×1(1 为时间维度参数),剩下的下采样层窗口均为 2×2×2,所有的下采样均采用最大值下采样. 全连接层 Fc6 和 Fc7 维度为 2048,Fc8 根据数据集类别设置为 2. Softmax 层输出最终的分类结果.

通过前期研究发现,视频背景信息对模型的识别结果有极大的影响,为了解决此问题本文采用先检测视频中的运动目标(人),对人加框标记后,再进行跌倒识别.

3D-CNN 的输入为对运动目标(人)加框后的视频帧,将视频每 8 帧作为一个输入样本,视频分辨率为 112×112.

2.3　Two-Stream CNN

视频帧和光流图包含了视频中的物体信息和运动信息,是基于视频的行为识别两个重要的元素. 本文参考文献[9,13]将基于 3D-CNN 支流 Softmax 输出和 VGGNet-16 支流 Softmax 输出进行线性加权融合,3D-CNN支流权值取 1/3,VGGNet-16 支流权值取 2/3. Two-Stream CNN 输出结果计算公式

$$R = \frac{1}{3}R_1 + \frac{2}{3}R_2, \tag{8}$$

其中,R 为 Two-Stream CNN 的输出结果,R_1 为 3D-CNN 支流 Softmax 输出,R_2 为 VGGNet-16 支流 Softmax 输出.

3　实验和结果分析

3.1　数据预处理和模型评价方法

本文采用了 Le2i[14]、SDU[15] 等两种视频数据集,其中,Le2i 数据集包含 Home,Coffee Room,Office,Lecture Room 等 4 种背景,共 250 个视频,其中跌倒 192 个,行走、做家务、坐等日常行为 58 个;视频分辨率为 320×240,帧率为 25. SDU 数据集包含 200 个视频,由 20 个人在实验室内模拟“跌倒”、“行走”、“下蹲”、“弯腰”、“站立”、“躺地”每种行为 10 次,视频分辨率为 640×480,帧率为 30.

上述数据集中视频大小、帧率不一致,本文首先将所有视频都统一到相同的大小:320×240,帧率:25. 其次,上述数据集单个行为的视频都很少,难以满足卷积神经网络训练数据需求,所以本文对数据集做了数据增强处理. 通过对视频进行水平翻转、对比度、亮度、加噪处理,使数据集扩大了 8 倍.

本文采用 THUMOS Challenge 2014[16] 行为识别竞赛的模型评价方法. 将测试集中的每一个视频进行测试,若模型 Softmax 输出概率值 TOP-1 的类别标签和输入视频的真实标签相同,认为识别正确,统计模型正确识别的视频比例作为模型的识别率. 模型识别率计算公式 $P = \frac{n}{N}$,其中,P 为最终识别率,N 为测试的视频个数,n 为模型 Softmax 输出概率值 TOP-1 的类别标签和输入视频的真实标签相同的视频个数.

3.2　模型训练和结果分析

选择 Le2i 中 Home 和 Coffee Room 背景中“跌倒”、Office 和 Lecture Room 背景中“行走”行为视频与 SDU 中的“跌倒”和“行走”行为视频作为本文的数据集,其中“行走”作为“跌倒”的负样本. 将选出的视频作预处理后取其中视频的 60%作为训练集;将剩余的视频作为测试集. 训练集和测试集视频数如表 1 所示.

将训练集未加框标记视频帧分别直接输入 VGGNet-16 和 3D-CNN. VGGNet-16 初始学习率设为 0.001,每经过 10 000 次迭代学习率减小 90%,总共迭代 30 000 次;3D-CNN 初始学习率设为 0.01,每经过 10 000 次迭代学习率减小 90%,总共迭代 40 000 次. 用测试集分别测试 VGGNet-16 和 3D-CNN 模型,结果如表 2 所示.

从表 2 中可看出,VGGNet-16 对于 Le2i Office 和 Lesi Lecture Room 场景中跌倒视频识别率很低;3D-CNN对于 Le2i Office 和 Lesi Lecture Room 场景中跌倒视频可以正确识别但识别率较低. 这说明背景对模型识别有很大影响,3D-CNN 可消除视频背景的干扰.

对训练集视频中的人加框标记后输入 3D-CNN. 图 3 为跌倒视频连续 2 帧和对应加框标记后视频帧. VGGNet-16 初始学习率设为 0.001,每经过 10 000 次迭代学习率减小 90%,总共迭代 30 000 次;3D-CNN

初始学习率设为 0.01,每经过 10 000 次迭代学习率减小 90%,总共迭代 40 000 次.用测试集分别测试 VGGNet-16 和 3D-CNN 模型,结果如表 3 所示.对视频中的人加框标记后,3D-CNN 模型的识别率提高了 9%.这说明对视频中的人加框标记后进一步消除了视频背景的干扰.

表 1 训练集、测试集视频数

背景	训练集视频数		测试集视频数	
	跌倒	行走	跌倒	行走
SDU	960	960	320	320
Le2i Home	384	0	96	96
Le2i Coffee Room	448	0	112	112
Le2i Office	0	410	102	102
Le2i Lecture Room	0	358	90	90
总数	1792	1728	720	720

表 2 VGGNet-16 和 3D-CNN 输入视频帧测试结果 (%)

模型	不同背景"跌倒"行为识别率					平均识别率
	SDU	Le2i Home	Le2i Coffee Room	Le2i Office	Le2i Lecture Room	
VGGNet-16	70	78	80	7	4	48
3D-CNN	85	86	83	78	85	83

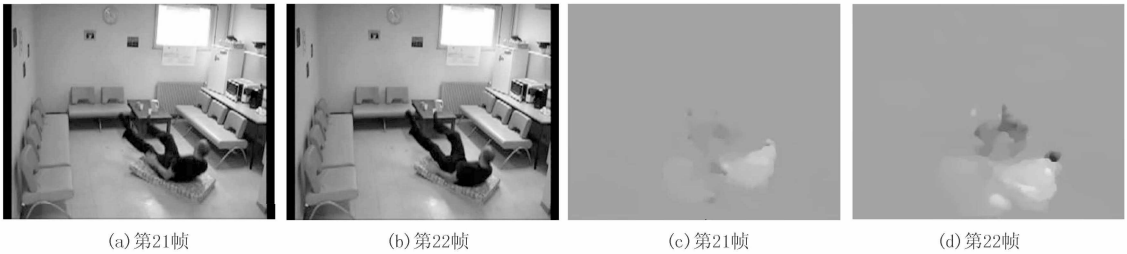
(a)第21帧 (b)第22帧 (c)第21帧 (d)第22帧

图3 加框标记的视频帧

表 3 3D-CNN 加框视频帧测试结果 (%)

模型	不同背景"跌倒"行为识别率					平均识别率
	SDU	Le2i Home	Le2i Coffe Room	Le2i Office	Le2i Lecture Room	
3D-CNN	94	93	90	88	95	92

利用光流法提取训练集视频的光流图输入 VGGNet-16.图 4 为跌倒视频连续两帧和对应光流图.VGGNet-16 初始学习率设为 0.003,每经过 10 000 次迭代学习率减小 90%,总共迭代 30 000 次.提取测试集光流图测试 VGGNet-16 模型,并根据公式(8)得到 Two-Stream CNN 的测试结果.



(a)第21帧 (b)第22帧 (c)第21帧 (d)第22帧

图4 跌倒视频的光流图

输入光流图的 VGGNet-16 模型识别率分别为 93%,模型融合后 Two-Stream CNN 识别率为 96%,比

3D-CNN 模型提高了 4%和比 VGGNet-16(输入光流图)模型提高了 3%.

4 总　结

本文提出了一种基于双流卷积神经网络跌倒识别方法(Two-Stream CNN). 该方法对视频中运动目标加框标记后，输入 3D-CNN；并提取视频的光流图后输入 VGGNet-16；最后将 3D-CNN 和 VGGNet-16 进行加权融合. Two-Stream CNN 充分利用视频的空间维度和时间维度信息，且不需要人为提取跌倒的视觉特征. 实验结果表明，对运动目标加框标记和 3D-CNN 方法有效地消除了视频背景的干扰；Two-Stream CNN 的跌倒识别率为 96%，比 3D-CNN 和 VGGNet-16 分别提高 4%和 3%.

参　考　文　献

[1] 张金富. 基于 Kinect 的跌倒检测报警监护系统[D]. 黑龙江大学，2016.

[2] 瞿畅，孙杰，王君泽，等. 基于 Kinect 体感传感器的老年人跌倒自动检测[J]. 传感技术学报，2016，29(3)：378-383.

[3] 彭玉青，高晴晴，刘楠楠，等. 基于多特征融合的跌倒行为识别与研究[J]. 数据采集与处理，2016，31(5)：890-902.

[4] 白勇，孙晓雯，秦昉，等. 基于 SVD 特征降维和支持向量机的跌倒检测算法[J]. 计算机应用与软件，2016，34(34)：247-251.

[5] Vaidehi V，Ganapathy K，Mohan K，et al. Video based automatic fall detection in indoor environment [C]//International Conference on Recent Trends in Information Technology. Piscataway：IEEE，2011：1016-1020.

[6] Rougier C，Meunier J，Rousseau J，et al. Robust Video Surveillance for Fall Detection Based on IIuman Shape Deformation[J]. IEEE Trans. Circuits and Systems for Video Technology，2011，21：611-622.

[7] Krizhevsky A，Sutskever I，IIinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems，2012，25(2)：1097-1105.

[8] Annane D，Chevrolet J C，Chevret S，et al. Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Advances in Neural Information Processing Systems，2014，1(4)：568-576.

[9] Tran D，Bourdev L，Fergus R，et al. Learning spatiotemporal features with 3d convolutional networks[C]//International Conference on Computer Vision. Piscataway：IEEE，2015：4489-4497.

[10] Simonyan K，Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Com Sci，2015：1409-1419.

[11] 李伟，彭玉峰. 基于形态学小波理论和 SVM 神经网络的人脸识别[J]. 河南师范大学学报(自然科学版)，2012，40(5)：61-64.

[12] 刘霞，焦建锋. 具有时滞的递归神经网络模型的分支分析[J]. 河南师范大学学报(自然科学版)，2016，43(1)：1-7.

[13] Ye II，Wu Z，Zhao R W，et al. Evaluating Two-Stream CNN for Video Classification[C]//Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. New York：ACM，2015：435-442.

[14] Charfi I，Miteran J，Dubois J，et al. Definition and Performance Evaluation of a Robust SVM Based Fall Detection Solution[C]//International Conference on Signal Image Technology and Internet Based Systems. Washington DC：IEEE Computer Society，2012：218-224.

[15] 薛冰霞. 基于多模特征融合的人体跌倒检测算法研究[D]. 济南：山东大学，2015.

[16] Jiang Y G，Liu J，Zamir A R，et al. TIIUMOS challenge：Action recognition with a large number of classes[EB/OL]. [2016-12-05]. http://crcv. ucf. edu/ICCV13-Action-Workshop.

A Fall Detection Method Based on Two-Stream Convolutional Neural Network

Yuan Zhi，Hu Hui

(School of Information Engineering，East China Jiaotong University，Nanchang 330013，China)

Abstract：It is difficult to extract suitable visual feature for fall detection. To solve this problem，a fall detection method based on two-stream convolutional neural network (Two-Stream CNN) method was proposed. The 3-Dimensional Convolutional Neural Network (3D-CNN) stream input the marked video frame to eliminate the interference of video background. The VGG-Net-16 convolutional neural network stream input the optical flow frame. Finally the Softmax of 3D-CNN and VGGNet-16 were fused as the Two-Stream CNN output. Experimental results show that，the marked video frame and 3D-CNN method can effectively eliminate the interference of the video background. The recognition rate of Two-Stream CNN is 96%，which is increased by 4% compared with 3D-CNN，3% compared with VGGNet-16 network.

Keywords：fall detection；two-stream convolutional neural network；video frame；optical flow frame

[责任编辑　杨浦]