面向助老行为识别的三维卷积神经网络设计

李秀智[1,2], 张 冉[1,2], 贾松敏[1]
(1. 北京工业大学信息学部, 北京 100124; 2. 数字社区教育部工程研究中心, 北京 100124)

摘 要: 针对室内老人跌倒问题,提出一种室内人体跌倒行为识别方法. 首先,提出基于卷积核分解与分组卷积的轻量化3D网络;之后融合浅层2D子网络与轻量化3D子网络,并采用随机滑动组合采样策略改进3D卷积行为识别网络. 为进一步提高网络泛化性能,对视频帧进行视觉显著性检测,通过加强背景纹理与人物行为之间关联性提高真实场景识别准确度. 实验结果表明:该网络参数量为6.9×10^6,时间复杂度降低至8.04×10^9;实现算法在室内跌倒行为识别任务上达到81.5%的准确度.

关键词: 行为识别; 跌倒检测; 3D卷积神经网络; 视觉显著性; 卷积核分解; 分组卷积
中图分类号: U 461; TP 308 **文献标志码**: A **文章编号**: 0254 - 0037(2021)06 - 0589 - 09
doi: 10.11936/bjutxb2020040005

Design of 3D Convolutional Neural Network for Action Recognition for Helping the Aged

LI Xiuzhi[1,2], ZHANG Ran[1,2], JIA Songmin[1]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124, China)

Abstract: To solve the problem of action recognition in indoor environment, a method for human falling recognition in indoor environment was proposed. First, a lightweight 3D network, which uses grouping convolution and factorization to lighten the network structure for action classification, was proposed. Then 2D subnetworks and lightweight 3D sub-networks were fused to improve behavior recognition network based on the 3D convolution. Finally, visual saliency detection was performed on video frames to improve the accuracy of real scene recognition by enhancing the correlation between background texture and human behavior. Results show that the network's parameter is reduced to 6.9×10^6 and the floating point of operations is reduced to 6.9×10^9. The algorithm achieves 81.5% accuracy in the task of indoor fall behavior recognition.

Key words: action recognition; fall detection; 3D convolution neural network; visual saliency; factorization of convolution kernal; group convolution

随着我国社会发展和人口老龄化程度不断加快,空巢老人数量呈现明显上升趋势. 当独居老人发生摔倒等意外情况时,如何在第一时间实施医疗救助? 一种方案是在家中布置大量传感器[1]或身体携带相应传感设备. 相比之下,另一种基于视觉的人体行为识别技术,仅使用视频图像流输入就可以分析其中人体行为动作,再将识别出的跌倒等危险动作报警信号通过通信模块发送给亲人、护工等,可减少繁冗的传感器使用和携带传感设备[2-4]的不便. 亲人或护工可通过报警信息与视频监控及时查

收稿日期: 2020-04-10
基金项目: 北京市教育委员会科技计划资助项目(JZ041001201701)
作者简介: 李秀智(1979—), 男, 副教授,主要从事智能机器人导航、机器视觉方面的研究, E-mail: xiuzhi.lee@163.com

看老人健康状态,对受到意外伤害的老人及时救助、提高老人生活质量、减轻老人生活自理压力都具有重要的现实意义.

基于视觉的人体行为识别技术中最重要的就是行为识别网络. 人体行为识别通常以视频流为数据源,综合考察一个时间序列上的图像信息,继而实现一个完整动作的识别. 在深度学习应用于该领域前,Wang 等[5]提出的改进稠密轨迹(improved dense trajectories,iDT)算法是人体行为识别中的经典算法. iDT 算法的前身为稠密轨迹(dense trajectories,DT)算法,其基本思路为利用光流场获取视频序列中的一些轨迹,之后从轨迹中提取出 4 种特征,最后对特征进行编码,再基于编码结果训练支持向量机(support vector machine, SVM)来完成分类任务. iDT 算法的改进之处在于,它利用前后 2 帧视频之间的光流以及关键点进行匹配,从而减弱相机运动带来的影响.

基于深度学习的行为识别方法中,Simonyan 等[6]首次提出双流卷积神经网络,该网络分为相同结构的空间网络和时序网络. 其基本原理为:首先,对视频序列中每 2 帧计算密集光流,得到密集光流时序序列;然后,对视频图像和密集光流分别训练神经网络;最后,将结果融合得到最终动作. Feichtenhofer 等[7]在双流网络基础上,利用卷积神经网络进行时序网络和空间网络的融合,进一步提高分类效果. Wang 等[8]提出的时序分割网络(temporal segment networks,TSN)同样是基于双流网络,但不同于双流网络采用单帧或单堆帧,TSN 使用整个视频中稀疏采样获得一系列短片段,每个片段都将给出其本身对于行为类别的初步分类,最后融合这些片段结果得到最终分类. 另一类基于深度学习的行为识别主流方法为 3D 卷积神经网络. Ji 等[9]认为对于基于视频分析的问题 2D 卷积神经网络不能很好捕获时序上的信息,因此,提出 3D 卷积神经网络. Tran 等[10]在此思想上提出 C3D 网络,采用 8 次卷积操作和 4 次池化操作,最终经过 2 次全连接层和 Softmax 层后得到最终分类结果. Carreira 等[11]基于 Inception-V1 模型,将 2D 卷积扩展到 3D 卷积,提出了 I3D 模型,但该模型参数量巨大,对硬件要求较高.

根据上述问题,本文提出一种实时的室内人体跌倒行为识别方法,基本实现了在室内环境下跌倒及某些日常行为动作行为识别,实验结果证明视觉显著性检测对于室内跌倒行为识别有积极作用.

1 实时室内跌倒行为识别构建框架

本文所述实时室内跌倒行为识别框架如图 1 所示. 视觉显著性算法可以根据图像将显著性部分与背景纹理部分分割,加强背景纹理与人物行为之间的关联性. 基于3D 卷积神经网络的室内跌倒动作识别网络,通过稀疏采样对视频流中人体行为进行分类识别. 其中,2D 子网络对视频图像提取低层特征,3D 子网络对 2D 子网络的输出进行组合,进一步提取高层特征,最后由输出层输出行为分类结果.

2 基于卷积核分解与分组卷积的 3D 网络

2.1 基于卷积核分解与分组卷积的 3D 卷积模块

3D 卷积核是视频行为识别中重要的角色. 3D

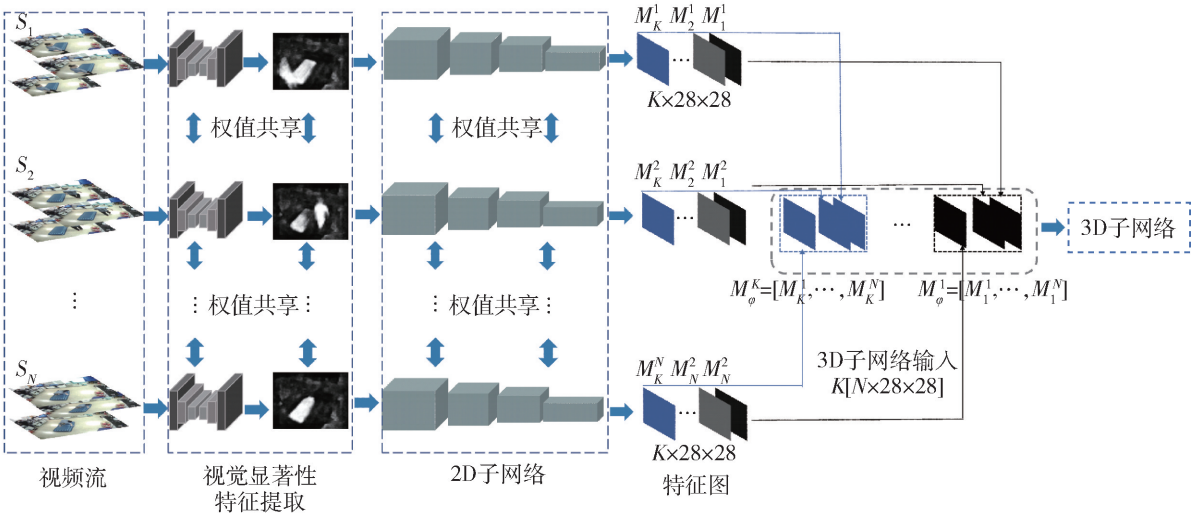
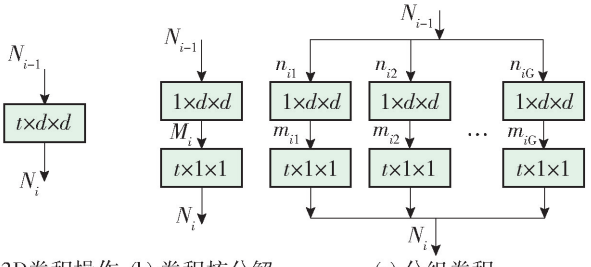


图 1 室内跌倒行为识别构建框架

Fig. 1 Indoor falling recognition framework

卷积核相较于 2D 卷积核多了时序维度上的卷积，这使 3D 卷积具备提取时序维度特征的能力，可以使得 3D 卷积神经网络更好地捕捉视频流的运动信息，有利于视频中人体的行为识别.



(a) 3D卷积操作　(b) 卷积核分解　　(c) 分组卷积

图 2　3D 卷积核分解与分组卷积

Fig. 2　3D grouping convolution and factorization

3D 卷积操作如图 2 中(a)所示. 传统 3D 卷积核将空间信息与时序信息一起卷积不易进行优化，于是本文采用卷积核分解[12]将 $t\times d\times d$ 卷积核分解为 $1\times d\times d$ 与 $t\times 1\times 1$，如图 2 中(b)所示. t 为 3D 卷积核中时空维度卷积参数，d 为 3D 卷积核中空间卷积参数. 时空分解后的 2 个卷积核分别对应处理视频图像序列的空间信息与时序信息. 通过时空分解 3D 卷积核，分离了空间信息与时序信息，增加了网络的非线性表达能力，易于网络优化. 同时，为了保留分解后的卷积核与原 3D 卷积核的表达能力，通过超参数 M_i 调节时空信息间的子空间数，使分解后的卷积核参数与原 3D 卷积核参数保持一致. 图 2(a)中原 3D 卷积核参数量等于图 2(b)中时空分解后的 2 层卷积核参数，即

$$N_{i-1}td^2N_i = N_{i-1}d^2M_i + M_itN_i \tag{1}$$

式中：N_{i-1} 为输入通道；N_i 为输出通道.

超参数 M_i 为

$$M_i = \left[\frac{td^2N_{i-1}N_i}{d^2N_{i-1}+tN_i}\right] \tag{2}$$

分组卷积能降低网络的时间复杂度，大幅降低训练参数量且不易过拟合. 如图 2 中(c)所示，将输入通道 N_{i-1} 分为 G 个组，每组分别进行卷积操作. 图 2(b)中时空分解后 3D 卷积核参数量 N 的计算式为

$$N = N_{i-1}d^2M_i + M_itN_i \tag{3}$$

图 2(c)中分组时空分解后 3D 卷积核参数量为未分组前的 $1/G$，计算公式为

$$\left(\frac{N_{i-1}}{G}d^2\frac{M_i}{G}+\frac{M_i}{G}t\frac{N_i}{G}\right)G = \frac{N_{i-1}d^2M_i+M_itN_i}{G} \tag{4}$$

本文采用残差模块作为基础结构，融合卷积核分解与分组卷积，提出基于分组卷积与卷积核分解的 3D 卷积模块，如图 3 所示.

基于分组卷积与卷积核分解的 3D 卷积模块分为 Conv a 与 Conv b. Conv a 模块功能为通过 $1\times1\times1$ 卷积改变通道数量，实现升维，Conv b 模块不改变通道维度. 由于使用分组卷积，通道间信息交换减少，所以采用 Multiplexer 模块弥补通道间的信息交换. 该模块为一个 2 层 $1\times1\times1$ 的卷积，第 1 个 $1\times1\times1$ 的卷积会将通道数量降低到 $1/k$，第 2 个 $1\times1\times1$ 的卷积再升维至输出通道数，因此，该模块的时间复杂度是一层 $1\times1\times1$ 卷积的 $2/k$，具体计算公式为

$$N_{i-1}lhw\frac{N_{i-1}}{k}+\frac{N_{i-1}}{k}lhwN_{i-1} = \frac{2}{k}N_{i-1}lhwN_{i-1} \tag{5}$$

式中 l、h、w 分别为特征图的时间维度长度和空间维

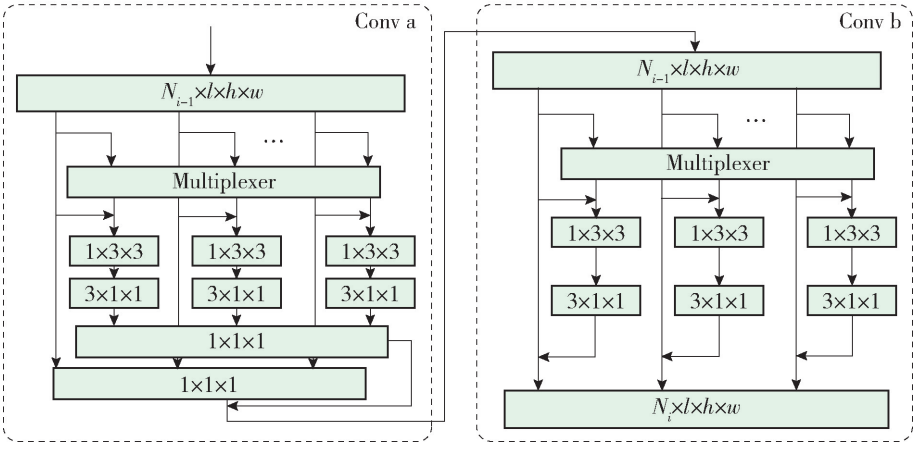


图 3　基于分组卷积与卷积核分解的 3D 卷积模块

Fig. 3　3D convolution unit based on grouping convolution and factorization

度的高与宽.

2.2 基于分组卷积与卷积核分解的 3D 卷积神经网络结构

本文将 2.1 节设计的 3D 卷积模块扩展为基于分组卷积与卷积核分解的 3D 卷积神经网络. 本文网络结构设计参考了 ResNet-34 的网络结构. 因为将 2D 子网络的输出作为 3D 子网络的输入,所以选取了中高层网络结构,舍弃了前几层低维卷积层,即从 Conv 3a 开始;同时,修改了一些通道数量,卷积层后均有批量归一化(batch normalization,BN)层及线性修正单元(rectified linear unit,Relu). 具体 3D 网路结构如图 4 所示.

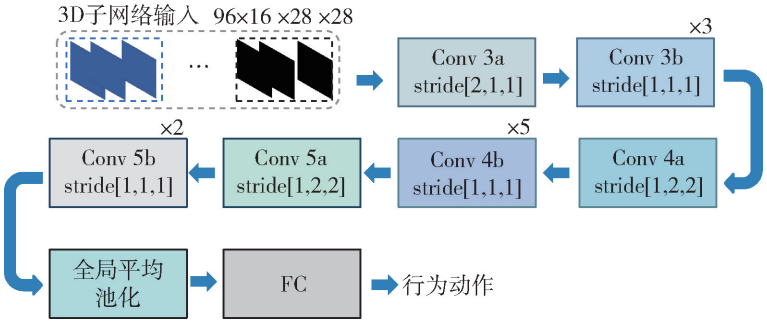


图 4 基于分组卷积与卷积核分解的 3D 卷积神经网络结构

Fig. 4 3D Convolution neural network architecture based on grouping convolution and factorization

3D 子网络输入为 2D 子网络输出的特征图组合,大小为 96 × 16 × 28 × 28;Conv 3a 将通道数扩展到 192,并通过时序维度卷积步长设置为 2,将特征图输出大小压缩为 192 × 8 × 28 × 28;之后通过 3 个 Conv b 模块,进入 Conv 4a 模块,将通道数扩展到 354,并通过空间维度步长设置为 2,将特征图输出大小压缩为 354 × 8 × 14 × 14;之后通过 5 个 Conv 4b 模块,进入 Conv 5a 模块,将通道数扩展为 768,同样通过空间维度步长设置为 2,将特征图输出大小压缩为 768 × 8 × 7 × 7;最后连接平均池化层和全连接层输出最终预测动作结果. 分组卷积设置的组数 G 为 16,即将各模块输出通道数平均分为 16 组. 网络细节见表 1.

表 1 基于分组卷积与卷积核分解的 3D 卷积神经网络结构

Table 1 3D convolution neural network architecture based on grouping convolution and factorization

卷积层	数量	通道数	输出	步长
Conv 3a	1	192	16 × 28 × 28	[2,1,1]
Conv 3b	3	192	8 × 28 × 28	[1,1,1]
Conv 4a	1	354	8 × 14 × 14	[1,2,2]
Conv 4b	5	354	8 × 14 × 14	[1,1,1]
Conv 5a	1	768	8 × 7 × 7	[1,2,2]
Conv 5b	2	768	8 × 7 × 7	[1,1,1]
全局平均池化			1 × 1 × 1	

3 基于 3D 卷积的行为识别网络

3.1 视频采样

不同于单张图片、视频序列中的连续图像,因为背景相同,所以具有连续性与冗余性. 因此,在使用视频序列进行训练时,为了避免图形处理器(graphics processing unit,GPU)资源浪费,提高模型训练效率,将训练集视频序列平均分为 N 段,再将每段中随机提取的一帧作为模型训练输入——这样的稀疏随机采样策略在减少冗余信息的同时还可以在训练中引入更复杂的多样性,从而提高模型的泛化能力.

在实际应用环境中,由于视频序列是源源不断的,可以采用随机滑动组合采样算法识别当前行为动作.

如图 5 所示,网络模型推断时,设置可以容纳 N 帧的滑动组合组作为行为识别网络的输入. 为了保持视频序列的时间上下文信息,同时设置历史记忆组与新视频流组. 历史记忆组与新视频流组以 5∶5 的比例组成滑动组合组. 当视频流开始时,滑动组合组稠密采集 N 帧输入网路,同时通过稀疏采样收集到 $N/2$ 帧存入新视频流组,将新视频流组中的 $N/2$ 帧替换掉原滑动组合组中的一半. 此时生成的 N 帧为新滑动组合组,而原滑动组合组此时称为历史记忆组. 每次预测时,从 2 组视频中各采样一半来更新滑动组合组,并将其作为网络模型的输入,预测出当前的行为动作结果. 将当前的预测结果和平均预测结果进行平均后得到最终的输出.

3.2 2D 与 3D 网络融合

虽然 3D 卷积神经网络完全可以胜任行为识别任务,但尽管使用网络模型压缩技术令 3D 网络时

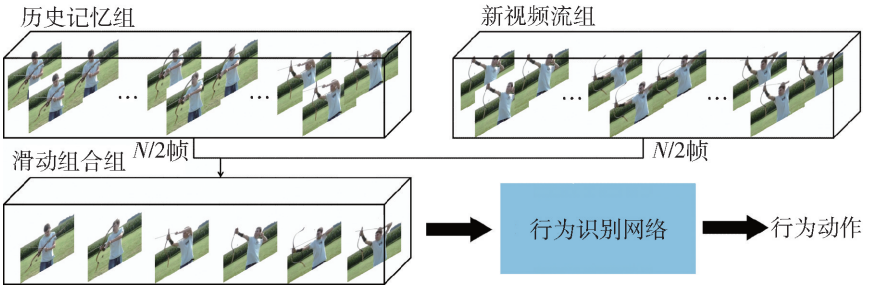


图 5　随机滑动组合采样算法

Fig. 5　Random sliding combined sampling

间复杂度与参数量大幅减少，3D 网络与 2D 网络相比依然不是同一个量级的网络. 因此，采用 2D 加 3D 的网络结构更加轻量且 2D 网络与 3D 网络结合并不会降低网络行为识别精度[13]，同时在视频的高层语义抽象层捕获视频的时间动态信息要优于在视频的底层像素级捕获视频的时间动态信息. 因此，本文采用底层 2D 网络加顶层 3D 网络结构在行为识别准确性与网络结构方面是最优选择.

本文采用的 2D 子网络结构如图 6 所示，网络为 GoogleNetV2 中前半部分(从输入层到 inception-3c 层). 每个卷积层后都有 BN 层和 Relu 激活层.

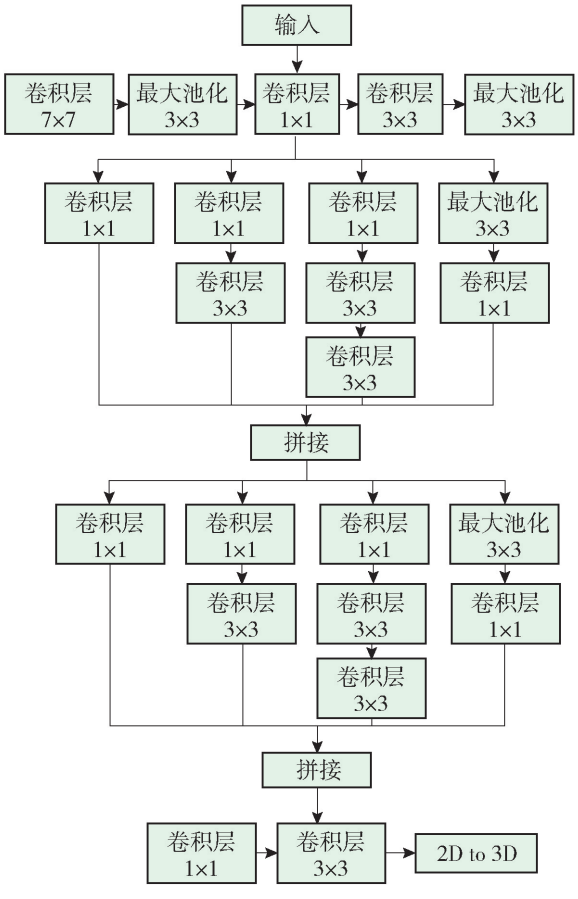


图 6　2D 子网络结构

Fig. 6　2D subnetwork architecture

在 2D 转化为 3D 时，本文将 16 张连续视频帧经过 2D 网络生成的 96 × 28 × 28 大小的特征图堆叠在一起，形成大小为 16 × 96 × 28 × 28 的特征图组，生成 96 × 16 × 28 × 28 的 3D 网络输入，即 96 个通道的时间维度为 16，空间维度为 28 × 28 的 3D 特征图，如图 1 中 2D 子网络至 3D 子网络结构.

3.3　基于 3D 卷积的行为识别网络训练

行为识别任务中 UCF101 数据集[14]有大量来自网络的视频片段，每个视频包含一个完整动作. 本文选取 UCF101 数据集作为预训练数据集，之后使用训练好的预训练模型在室内跌倒数据集上微调，实现识别室内人体行为的任务. 针对室内环境，Auvinet 等[15]建立了多摄像头室内跌倒数据集，该数据集由天花板上处于各个不同位置的摄像机所采集到的图像序列组成，如图 7 所示.

(a) 视角1

(b) 视角2

(c) 视角3

图 7　室内跌倒数据集

Fig. 7　Indoor fall dataset

数据集中,以单人演绎 24 个场景为基础,将人体动作分为 8 个类别,分别为行走或站立、摔倒、躺倒在地、蹲伏、向上移动、向下移动、坐以及躺在沙发上. 之后按照 5∶1的比例将数据集分为训练集与测试集,分别按照格式编辑训练集与测试集的训练文件. 训练集共 113 751 帧,测试集共 47 068 帧.

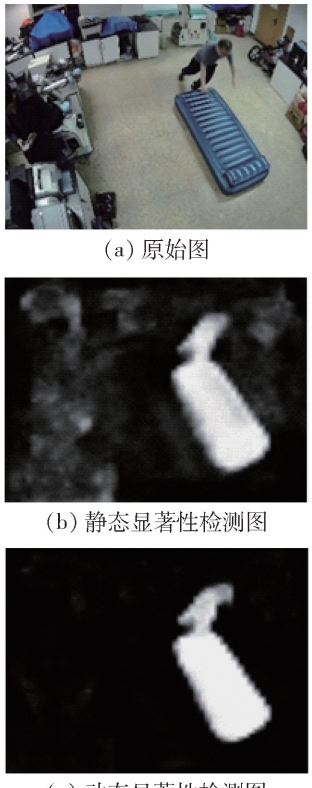
本文在配有 Intel i7-6700 CPU @ 3. 40 GHz、16 GB 内存的 GTX1070 GPU 和操作系统 Ubuntu16. 04 LTS 的电脑上进行训练和测试,深度学习框架选择 Caffe,开发语言是 Python. 首先将 UCF101 数据集进行剪切并编辑真值文件,网络输入首先减去像素均值并剪裁为 224 ×224,之后使用标准随机梯度下降的方法. 学习率设置为 0. 001,动量为 0. 9,权重衰减为 0. 000 50 进行训练优化时,每经过 5 000 次迭代,将学习率降低 0. 1,经过 2 万次迭代得到预训练模型. 之后在预训练模型上使用同样优化参数,每经过 3 000 次迭代,将学习率降低 0. 1. 经过 1 万次迭代得到最终室内人体行为识别模型.

4 基于视觉显著性检测的特征帧分析

在室内人体行为识别网络的真实场景应用中,由于每个应用环境都不一样,为确保网络模型的泛化性能,本文使用视觉显著性检测作为行为识别的前端处理,提取出对行为识别准确度有益的显著性特征. 因此,选取 Wang 等[16]提出的基于全卷积神经网络的视频显著性检测模型. 该模型通过全卷积神经网络分别构造出 2 个模块:静态显著性网络与动态显著性网络. 静态显著性网络通过图片的训练产生显著性检测结果,而动态显著性网络会根据时序特征,判断出动态的显著性检测结果. 视觉显著性检测效果如图 8 所示.

可以看出,显著性检测图相较于原始 RGB 视频帧,摒弃掉很多图像信息,仅展示图像中某些显著性纹理,并且由灰度图形式输出. 同时,静态显著性检测图与动态显著性检测图相比,拥有更多环境纹理信息,而动态显著性检测图更关注于动态物体和显著性较大的物体. 因此,可以认为静态显著性检测图是包含背景纹理信息与人物行为信息且不包含颜色信息的特征图;动态显著性检测图是不包含背景纹理仅包含人物行为信息且不包含颜色信息的特征图.

同时,为证明视频帧中颜色信息对行为识别是否有益,设计出 2 种基于视觉显著性的视频帧处理算法与原 RGB 视频帧对比. 第 1 种算法为图像融合:将原视频帧分别与静态视觉显著性检测图和动态视觉显著性检测图线性融合. 由于视觉显著性检测具有提取显著性物体的性质,此种做法将原 RGB 视频帧中显著性物体与背景纹理对比度加强但保留图像中 RGB 信息. 第 2 种算法为图像加权:将显著性检测图像作为原 RGB 视频帧的权值,选择只将显著性物体还原颜色特征. 具体 2 种算法效果如图 9 所示.



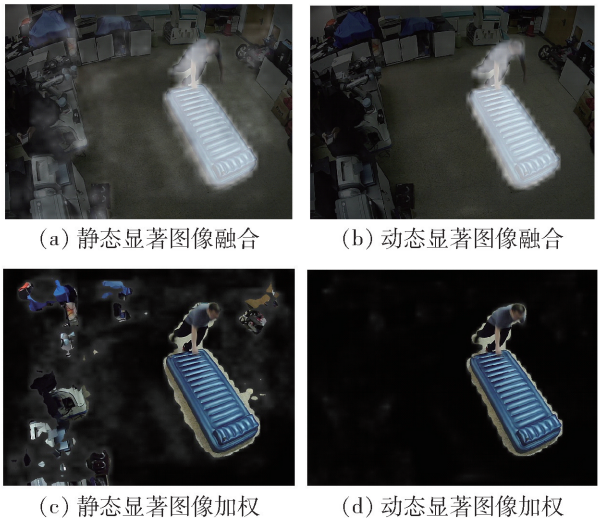
(a) 原始图

(b) 静态显著性检测图

(c) 动态显著性检测图

图 8 显著性检测图与原始视频帧

Fig. 8 Saliency image and original image



(a) 静态显著图像融合 (b) 动态显著图像融合

(c) 静态显著图像加权 (d) 动态显著图像加权

图 9 基于视觉显著性的图像加权与图像融合

Fig. 9 Merge image & weighted image base on saliency image

5　实验结果与分析

5.1　行为识别网络时间复杂度与参数量计算

网络时间复杂度和参数量决定了网络训练/预测时间以及网络的模型大小. 如果复杂度过高，则会导致模型训练和预测耗费大量时间，无法做到快速预测，很难适应实时性强的应用场景. 同时，若模型量级巨大，则在终端部署时，会浪费大量空间，导致成本增加. 为验证基于 3D 卷积的行为识别网络的性能，本文计算了网络的时间复杂度与参数量并与当前领域内效果优秀的网络进行对比. 参数量与时间复杂度计算公式分别为

$$\text{Params} = K_d K_h K_w C_{in} C_{out} \tag{7}$$

$$\text{FLOPs} = K_d K_h K_w DWHC_{in} C_{out} \tag{8}$$

式中：$K_h K_w$ 为卷积核大小；C_{in} 为输入通道数；C_{out} 为输出通道数；D、W 和 H 分别为输出特征图的深度、宽和高，与当前主流网络对比结果如表 2 所示.

表 2　时间复杂度与参数量对比结果

Table 2　Comprison results of FLOPs and parameter

网络名称	Params	FLOPs
Two-Stream[6]	1.2×10^7	
C3D[10]		3.9×10^{10}
Res3D[17]		1.9×10^{10}
I3D-RGB[11]	1.2×10^7	1.1×10^{11}
S3D[13]	8.8×10^6	6.6×10^{10}
R(2+1)D-RGB[12]	6.4×10^7	1.5×10^{11}
本文	6.9×10^6	8.0×10^9

本文提出的基于 3D 卷积的行为识别网络参数量低于 Params 与 FLOPs 最大的 R(2+1)D-RGB 网格 1 个量级且 FLOPs 低于 2 个量级. 可以看出，本文提出的基于 3D 卷积的行为识别网络 Params 与 FLOPs 均为最低，说明在网络模型中可以更快速地预测出结果，并且由于参数量少，模型更加精致，易于部署.

5.2　实时室内跌倒行为识别实验

为验证基于 3D 卷积的行为识别网络的行为识别准确率，本文采用测试集回灌与实际场景测试 2 种验证方式. 基于 3D 卷积的行为识别网络回灌测试集采用跌倒数据集中的测试集；实际场景为实验室录制的模拟视频. UCF101[14] 是一个人类动作视频数据，是从 YouTube 上剪辑的 101 类真实世界中的不同种类人类动作视频. 表 3 为以基于 3D 卷积的行为识别网络在 UFC101 数据集和跌倒测试集的准确率，分别与当前主流行为识别网络进行对比.

表 3　基于 3D 卷积的行为识别网络准确率(测试集)

Table 3　Precision of action recognition networks based on 3D convolution network architecture(text set)

%

网络名称	UCF101[19]	跌倒数据集[13]
Two-Stream[6]	88.0	65.3
C3D[10]	82.3	53.3
Res3D[17]	85.8	57.8
I3D-RGB[11]	95.6	81.6
S3D[13]	96.8	84.3
R(2+1)D-RGB[12]	96.8	83.9
本文	91.3	81.5

根据表 3 实验结果，在 UCF101 数据集与跌倒数据集中本文提出的网络准确率分别为 91.3% 和 81.5%，略低于当前主流 3D 网络，高于双流网络及 C3D、Res3D 网络. 尽管准确率略低几个百分点，但是本文的 Params 及 FLOPs 远小于以上网络，在实际部署应用中更加有利.

实验室模拟视频为 10 段不同摔倒动作的模拟视频，摔倒动作分为 2 次向前摔倒(FF1&FF2)、向后摔倒(BF)、失去平衡侧摔(LF)以及坐时摔倒(SF)；每组动作有 2 个摄像头视角录制. 图 10 展示了 SF 的一组行为识别结果，第 1 排左起分别为行走、向下移动和摔倒；第 2 排左起分别为躺在地上、向上移动和蹲伏.

图 10　实时室内跌倒行为识别实验结果

Fig. 10　Experimental results of real-time indoor falling recognition

图 10 表明基于 3D 卷积的行为识别网络能较好地识别出相应的动作，经过统计得出实际场景模拟测试集的准确率如表 4 所示.

从表 4 中可以看出，实际场景测试结果不如测试集回灌效果，这是由于背景环境以及光线等引起，说明实际场景中还需针对不同环境做出优化，才能

增加模型的行为识别准确度.

表 4 基于 3D 卷积的行为识别网络准确率(实际场景)

Table 4 Precision of action recognition networks based on 3D convolution network architecture(real) %

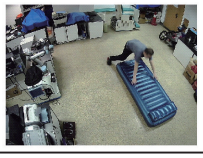
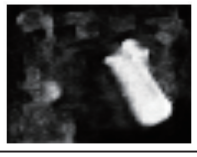
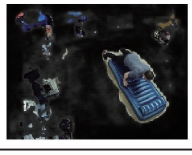
实际场景	准确率
向前摔倒	68.3
向后摔倒	52.5
失去平衡侧摔	70.5
坐时摔倒	66.3

5.3 视觉显著性检测特征帧实验

基于 5.2 节可以看出在实际场景中面临的问题. 为了验证视觉显著性检测特征帧对基于 3D 卷积的行为识别网络的影响,本实验根据 2 种基于视觉显著性的视频帧处理算法,设置不同融合比例以寻找最适合实际场景的算法及比例. 显著性图像融合图中将显著性检测图与原 RGB 图按照不同比例进行融合;显著性图像加权图中将按灰度阈值进行加权融合. 静态显著性图像比例融合及阈值加权实验结果如表 5 所示. 由此可以看出,图像融合图与图像加权图以 2 种方式补充了颜色与背景纹理信息.

表 5 静态显著性图像比例融合与阈值加权图

Table 5 Static saliency merge image in proportion & weighted image

原始帧	静态视觉显著性检测帧	比例融合 9:1	比例融合 5:5	阈值加权 50	阈值加权 200
					

5.4 实时室内跌倒行为识别实验

为了验证视觉显著性视频帧处理的可靠性,本文将 4.3 节处理的视觉显著性视频帧作为网络输入. 以跌倒动作为例,显著性图像阈值加权实验结果与显著性图像比例融合实验结果如图 11 所示.

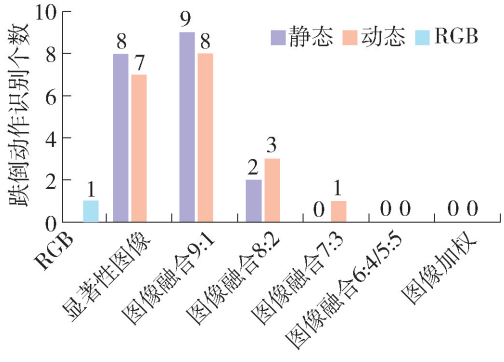


图 11 视觉显著性检测特征帧行为识别实验结果

Fig. 11 Action recognition experimental results of feature frames in visual saliency detection

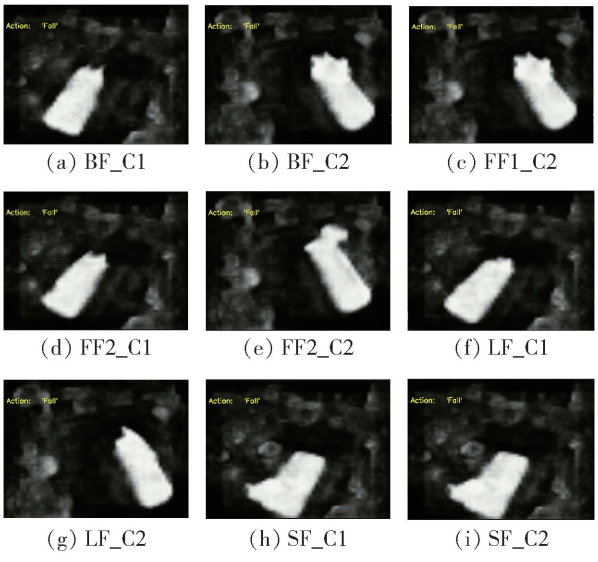


图 12 视觉显著性检测 9:1融合行为识别实验结果

Fig. 12 Action recognition experimental results of visual saliency merge image detection in proportion 9:1

由结果可知:RGB 视频帧输入仅识别出一次跌倒动作;视觉显著性融合算法提升了原算法的跌倒识别准确率,其中视觉显著性检测帧与原 RGB 帧按照 9:1比例融合效果最好,识别出 9 次跌倒动作,如图 12 所示,图中左上角黄色为识别出的跌倒动作“Fall”;视觉显著性加权算法针对行为识别任务没有积极作用,针对跌倒动作均无有效识别.

6 结论

1) 提出了一个基于 3D 卷积神经网络的室内实时行为识别网络. 设计并实现了基于分组卷积与卷积核分解的 3D 网络,精简了网络结构,使网络参数量及时间复杂度显著降低,Params 减少到 6.9×10^{6},FLOPs 减少到 8.04×10^{9}.

2）根据基于分组卷积与卷积核分解的3D 网络进一步融合 2D 网络，提出基于 3D 神经网络的行为识别网络并且为加快模型推断速度提出随机滑动组合采样算法，网络性能针对跌倒测试集可达到 81.5% 的准确率.

3）根据视觉显著性检测模型提出视觉显著性检测融合与视觉显著性检测加权 2 种算法，其中视觉显著性检测融合算法按照 9∶1比例效果最好. 根据 2 种算法结果可以推断出本文行为识别网络对颜色信息不敏感，但背景纹理信息对于行为识别有积极作用.

参考文献：

[1] 何颖，黄艳，王腾，等. 社区独居老人智能监控系统的手环设计[J]. 数字技术应用，2019，37(10)：163-164，166.
HE Y，HUANG Y，WANG T，et al. Bracelet design of intelligent monitoring system for the elderly living alone in the community [J]. Digital Technology & Application，2019，37(10)：163-164，166. (in Chinese)

[2] 裴利然，姜萍萍，颜国正. 基于支持向量机的跌倒检测算法研究[J]. 光学精密工程，2017，25(1)：182-187.
PEI L R，JIANG P P，YAN G Z. Research on fall detection algorithm based on support vector machine [J]. Optics and Precision Engineering，2017，25(1)：182-187. (in Chinese)

[3] 米晓萍，李雪梅. 基于物联网智能的独居老人自动监控方法研究[J]. 计算机仿真，2014，31(2)：378-381.
MI X P，LI X M. Old people who live alone automatic monitoring method based on IoT intelligence research [J]. Computer Simulation，2014，31(2)：378-381. (in Chinese)

[4] 张国梁，贾松敏，张祥银，等. 采用自适应变异粒子群优化 SVM 的行为识别[J]. 光学精密工程，2017，25(6)：1669-1678.
ZHANG G L，JIA S M，ZHANG X Y，et al. Adaptive mutation particle swarm optimization for SVM behavior recognition [J]. Optics and Precision Engineering，2017，25(6)：1669-1678. (in Chinese)

[5] WANG H，CORDELIA S. Action recognition with improved trajectories [C] // IEEE International Conference on Computer Vision. Piscataway：IEEE，2013：3551-3558.

[6] SIMONYAN K，ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems，2014，27：568-576

[7] FEICHTENHOFER C，PINZ A，ZISSERMAN A. Convolutional two-stream network fusion for video action recognition [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2016：1933-1941.

[8] WANG L，XIONG Y，WANG Z，et al. Temporal segment networks：towards good practices for deep action recognition [C] // European Conference on Computer Vision. Berlin：Springer，2016：20-36.

[9] JI S，XU W，YANG M，et al. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35(1)：221-231.

[10] TRAN D，BOURDEV L，FERGUS R，et al. Learning spatiotemporal features with 3d convolutional networks [C] // 2015 IEEE International Conference on Computer Vision. Piscataway：IEEE，2015：4489-4497.

[11] CARREIRA J，ZISSERMAN A，QUO V. Action recognition? a new model and the kinetics dataset [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2017：4724-4733.

[12] TRAN D，WANG H，TORRESANI L，et al. A closer look at spatiotemporal convolutions for action recognition [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway：IEEE，2018：6450-6459.

[13] XIE S，SUN C，HUANG J，et al. Rethinking spatiotemporal feature learning：speed-accuracy trade-offs in video classification [C] // European Conference on Computer Vision. Berlin：Springer，2018：305-321.

[14] SOOMRO K，ZAMIR A R，SHAH M. UCF101：a dataset of 101 human actions classes from videos in the wild [J/OL]. [2012-01-03]. https：// arxiv. org/abs/1212.0402.

[15] AUVINET E，ROUGIER C，MEUNIER J，et al. Multiple cameras fall dataset [R]. Montreal：DIRO Université de Montréal，2010.

[16] WANG W，SHEN J，SHAO L. Video salient object detection via fully convolutional networks [J]. IEEE Transactions on Image Processing，2017，27(1)：38-49.

[17] TRAN D，RAY J，SHOU Z，et al. Convnet architecture search for spatiotemporal feature learning [J/OL]. [2017-08-16]. https：// arxiv. org/abs/1708.05038.

(责任编辑　梁　洁)