ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 1,Jan. 2023,pp. 132－140
DOI:10. 16337/j. 1004-9037. 2023. 01. 011

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于CNN-LSTM双流融合网络的危险行为识别

高治军[1]，顾巧瑜[1]，陈　平[2]，韩忠华[1]

(1. 沈阳建筑大学信息与控制工程学院,沈阳 110168;2. 中北大学信息与通信工程学院,太原 030051)

摘　要：针对目前人体危险行为识别过程中由于时空特征挖掘不充分导致精度不够的问题,对传统双流卷积模型进行改进,提出了一种基于CNN-LSTM的双流卷积危险行为识别模型。该模型将CNN网络与LSTM网络并联，其中CNN网络作为空间流,将人体骨架空间运动姿态分为静态与动态特征进行分别提取,两者融合作为空间流的输出;在时间流中采用改进的可滑动长短时记忆网络,以增加人体骨架时序特征的提取能力;最后将两个分支进行时空融合,利用Softmax对危险动作做出分类识别。在公开的NTU-RGB＋D数据集和Kinetics数据集上的实验结果表明,改进后模型的平均跨角度(Cross view,CV)精度达到92.5%,平均跨视角(Cross subject,CS)精度为87.9%。所提方法优于改进前及其他方法,可以有效地对人体危险动作做出识别,同时对于模糊动作也有较好的区分效果。

关键词：人体骨架;危险行为识别;卷积神经网络;长短时记忆网络;融合实验

中图分类号：TP391　　**文献标志码**:A

Dangerous Behavior Recognition Based on CNN-LSTM Dual-Stream Fusion Network

GAO Zhijun[1], GU Qiaoyu[1], CHEN Ping[2], HAN Zhonghua[1]

(1. School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China;2. School of Information and Communication Engineering, North University of China, Taiyuan 030051, China)

Abstract: To solve the problem of insufficient spatial and temporal feature in the process of dangerous behavior recognition, this paper improves the traditional dual-stream convolution model and proposes a new dual-steam convolution dangerous behavior recognition model based on CNN-LSTM. In this model, CNN network and LSTM network are connected in parallel. CNN network is used as the spatial flow. The spatial motion attitude information of human skeleton is divided into static and dynamic. These features are fused as the output of the spatial flow. In order to increase the ability of extracting temporal features of human skeleton, an improved temporal sliding LSTM network is used in the time stream. Finally, the two branches are fused in time and space, and the dangerous actions are classified and identified by Softmax. Experimental results on NTU RGB D and Kinetics datasets show that the average cross view (CV) accuracy of the improved model is 92.5% and the average cross subject (CS) accuracy is 87.9%. The proposed method is superior to that before improvement and other methods. It can effectively recognize dangerous human actions and has good discrimination effect for fuzzy actions.

Key words: human skeleton; dangerous behavior recognition; convolutional neural network (CNN); long short-term memory (LSTM); fusion experiment

基金项目:国家重点研发计划(2018YFF0300304-04)。
收稿日期:2021-11-12;**修订日期**:2022-10-10

引　言

人体危险行为识别技术是人工智能的重要应用之一,对于预防犯罪及处理突发事故非常重要[1]。其具体的过程是利用计算机对摄像机采集到的视频进行目标检测与分析,随后给出具体的人体行为分类,其中特征提取方式是决定识别精度的关键,也是需要研究的重点。在实际检测人体过程中难免会受到遮挡物遮挡、运动视角变换和背景复杂等因素干扰,从而导致人体运动信息提取完整性差。因为人体骨架来源于人体本身,对于人体的运动表征更加明显,所以为解决上述问题,本文利用人体的骨架关节信息作为特征提取。随着深度学习的不断发展和硬件设备的不断完善,利用Openpose,Alphapose等人体姿态识别技术[2]可以直接获得人体骨架的光流姿态等运动信息,可使得相关工作更加高效。目前,基于骨架的人体行为识别技术已经受到了国内外学者的广泛研究。Ding等[3]针对卷积神经网络(Convolutional neural network, CNN)在处理动态骨架数据时性能不足的问题,提出一种基于骨架的方形网格结构,用于将动态骨架转换为三维网格结构数据,以增强CNN网络对数据的深度特征捕捉;Jia等[4]设计了一种双流时间卷积网络(Two-stream temporal convolutional networks, TS-TCNs),充分利用骨架序列的帧间向量和帧内向量,提高一般时间卷积网络对长时间依赖动作的处理能力;田志强等[5]利用图卷积神经网络,利用时序散度模型描述骨骼点的运动状态,放大了不同人体动作的类间方差,增加骨骼点运动的部分细节信息;Wen等[6]设计了一种分层级的网络结构,这种结构具有基于图卷积网络(Graph convolutional network, GCN)来编码分层空间结构,并使用可变的时间密集块来利用骨骼序列不同范围上的局部时间信息;Su等[7]利用多循环网络融合的行为检测模型增加对骨架序列的时序特征提取能力;Peng等[8]对时空图卷积网络(Spatial temporal GCN, ST-GCN)的整体框架进行改进,并将输入图序列缩减到欧氏空间,实现多尺度时间滤波器的引入,可有效捕获人体骨架的动态信息。除上述深度学习方法外,Wang等[9]提取单个关节的运动特征和多个关节的关系特征作为人体运动识别的综合特征,从运动学和空间几何学角度发掘了人体运动时的关节特征。

虽然现有的人体骨架行为检测方法已具有较好的应用,但忽视了对人体运动的时空综合信息提取。人体行为识别过程属于类间识别,除不同种类的动作之间存在差异外,同种动作还会因为动作幅度的不同存在一定的差异。虽然时空双流卷积模型较为典型,但是原始的双流法模型较简单,对于复杂、模糊的人体动作识别显得较乏力。针对以上问题,本文提出一种改进的双流卷积危险行为识别模型。首先描述了双流模型的架构,其次介绍了两个分支的具体设计思路,然后进行了仿真实验验证,最后对本文工作进行了总结。

1　基于CNN-TS LSTM的双流融合网络

本文提出的危险行为识别检测模型如图1所示。首先将包含人体动作的视频分割成为视频帧,再利用Openpose进行人体目标锁定以及关节点提取。接着将带有骨骼点序列信息的视频帧输入到双流网络中。在空间流中通过CNN对人体骨架运动姿态的空间信息进行特征提取;时间流中使用滑动长短时记忆网络(Temporal sliding LSTM,TS-LSTM)提取人体运动过程中的骨架时序信息,最后进行时空特征融合并用Softmax进行人体动作分类。

1.1　CNN网络分支

作为深度学习中经典的神经网络,卷积神经网络专门用于处理数据网格结构,如图片数据就可认为是像素组成的二维网格数据。卷积操作正是通过对输入数据一系列特征进行自动提取,其计算过程通过卷

图1　基于CNN-LSTM危险行为识别网络结构图

Fig.1　Network structure diagram of dangerous behavior recognition based on CNN-LSTM

积核在输入中按Stride给定的步长依次移动，每移动一步就是进行了一次卷积运算，运算式为

$$\boldsymbol{y}_{mn}=\sum_{j=0}^{J-1}\sum_{i=0}^{I-1}\boldsymbol{x}_{m+i,n+j}*\boldsymbol{w}_{ij}+\boldsymbol{b} \tag{1}$$

式中：$\boldsymbol{x}$为二维输入向量，$\boldsymbol{w}$为大小为$I*J$的卷积核，$\boldsymbol{b}$为偏置，$\boldsymbol{y}$为卷积的输出。本文在空间流中选用VGG-16模型，该模型采用了16层的深度网络，包含13个卷积层和3个全连接层，卷积核为3×3，步长为1。

该网络层的输入为带有人体骨骼信息的视频帧数据，首先对包含人体骨骼点的图片进行人体运动空间姿态静态特征的提取，然后通过两个视频帧在Δt时差下相同关节点的位置变换得到隐藏动态特征。该过程的网络示意图如图2所示。

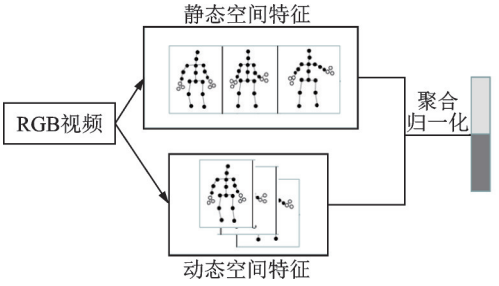

图2 基于CNN的行为识别

Fig.2 CNN based behavior recognition

具体而言，空间流在提取骨架信息过程中把单一动作定义为p，动作序列的总数为T，设单独时刻为$t(t\in T)$，其动作特征为f_t^p。现将该网络层在特征提取过程中得到的信息进行整合；将每帧动作的序列值进行聚合归一，即对该动作关节向量$\boldsymbol{i}$变化过程中得到的最大值和最小值进行整合，其过程分别为

$$h_i=\min f_t^p(\boldsymbol{i}) \tag{2}$$

$$H_i=\max f_t^p(\boldsymbol{i}) \tag{3}$$

所有静态特征的集合表示为

$$V_{\text{stat}}^p=[h_1,h_2,\cdots,h_m;\ H_1,H_2,\cdots,H_m] \tag{4}$$

式中V_{stat}^p表示视频序列的静态特征。

人体骨架空间姿态运动中的动态特征由相邻视频帧中关节点数据的序列差表示。设相邻间隔之间的差值为Δf_t^p，则$\Delta f_t^p=f_{t+\Delta t}^p-f_t^p$，在$m$张视频帧中可以整合到$m-1$个相应的数据值，将其作为隐藏特征对静态空间姿态数据进行加强。该过程可表示为

$$V_{\text{dyn}}^p=[\Delta h_1,\Delta h_2,\cdots,\Delta h_{m-1};\ \Delta H_1,\Delta H_2,\cdots,\Delta H_{m-1}] \tag{5}$$

空间层的输出由上述的动态特征数据和静态特征数据共同决定，其具体处理方法为：从动静态最大值融合、动静态最小值融合、动态最大值与静态最小值融合、动态最小值与静态最大值融合、动静态特征均值5种方式中选择精度最高的融合方式。

1.2 时间流TS-LSTM分支

在实际检测中仅仅依靠空间流卷积提取到的人体骨骼姿态信息还不足以表示该动作的全部特征，故在区分相同动作和相似动作上的效果还不够，因此还要利用时间层网络进一步提取人体骨架运动过程中的时序信息并与空间流的特征进行融合互补。

LSTM网络是一种改进的循环神经网络，它经常用于训练包含时间序列的动态模型。传统的LSTM网络用于人体行为识别时只是模拟骨架关节的整体时间动态，而不考虑每个关节的详细时间动态，这样就会忽略很多重要的细节。为解决这个问题，本文在时间流中对滑动长短时记忆网络进行改进，设计了长中短期3个周期模块用来加强关节局部信息，通过设置滑动窗口来逐一对骨架时序信息进行提取并整合动作的属性，其滑动运算过程的模型图如图3所示。

改进后的时间流TS-LSTM模块如图4所示。该模型整体由中长短期3个模块组成，并在其基础上增加了Concat层、Sumpool层、Linear层以及Dropout层，用来增强特征提取的能力。其中Concat层的

作用是将TS-LSTM层输出数据的维度进行整合;Sumpool层和Meanpool层用于过滤无关参数防止网络在训练中过度拟合;Linear层为全连接层,其作用是对数据进行平滑处理,把不同神经元提取的局部特征整合成完整的人体骨架运动信息;Dropout层用来防止上一层的参数过度拟合,并停止无效部分神经元的工作。本文网络3个周期的参数根据由各自模块中TS-LSTM单元中的通道数量与1的占比决定,因为第1个模块中只有一个通道,故周期数为1,第2个模块中每个通道的周期为0.5,第3个模块中每个通道的周期为0.33。

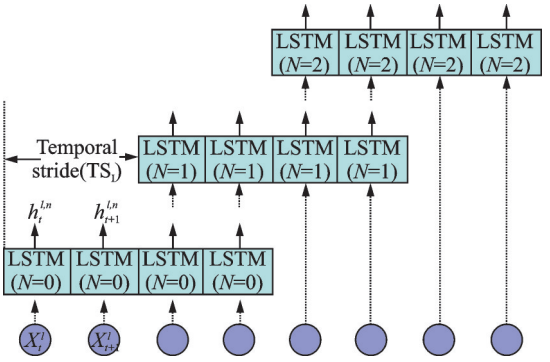

图3 TS-LSTM网络模型结构图

Fig.3 Structure diagram of TS-LSTM network model

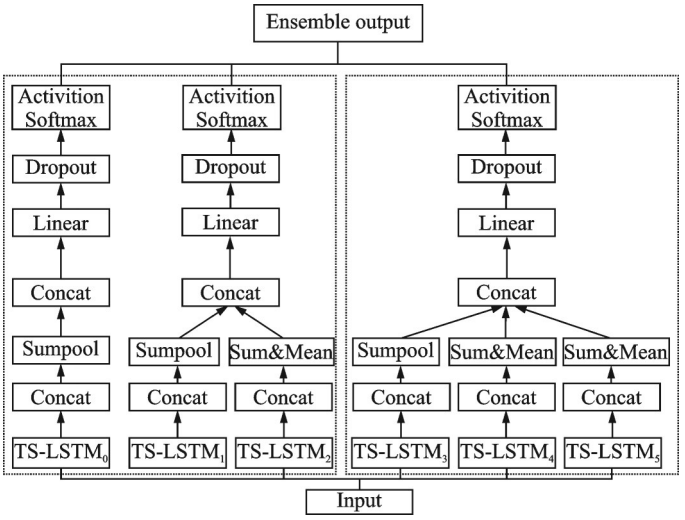

图4 TS-LSTM层危险行为识别网络结构图

Fig.4 Network structure diagram of TS-LSTM layer dangerous behavior recognition

现说明其具体运算过程,其中LSTM的个数设为N_l,它的滑动窗口设置为W_l,时间步长设置为TS_l,将不同的LSTM单元组成不同的滑动窗口,通过调整滑动窗口的尺寸以及时间步长实现更高效的特征提取。假设OX_t^l为t时刻下第l个单元层的输出,那么此LSTM单元的输出和3个门可以分别表示为

$$\boldsymbol{i}_t^{l,n}=\sigma(\boldsymbol{W}_{ix}^{l,n}\boldsymbol{OX}_t^{l,n}+\boldsymbol{W}_{ih}^{l,n}\boldsymbol{h}_{t-1}^{l,n}+\boldsymbol{b}_i^{l,n}) \tag{6}$$

$$\boldsymbol{f}_t^{l,n}=\sigma(\boldsymbol{W}_{fx}^{l,n}\boldsymbol{OX}_t^{l,n}+\boldsymbol{W}_{fh}^{l,n}\boldsymbol{h}_{t-1}^{l,n}+\boldsymbol{b}_f^{l,n}) \tag{7}$$

$$\boldsymbol{c}_t^{l,n}=\boldsymbol{f}_t^{l,n}\boldsymbol{c}_{t-1}^{l,n}+\boldsymbol{i}_t^{l,n}\tanh(\boldsymbol{W}_{cx}^{l,n}\boldsymbol{OX}_t^{l,n}+\boldsymbol{W}_{ch}^{l,n}\boldsymbol{h}_{t-1}^{l,n}+\boldsymbol{b}_c^{l,n}) \tag{8}$$

$$\boldsymbol{o}_t^{l,n}=\sigma(\boldsymbol{W}_{ox}^{l,n}\boldsymbol{OX}_t^{l,n}+\boldsymbol{W}_{oh}^{l,n}\boldsymbol{h}_{t-1}^{l,n}+\boldsymbol{b}_o^{l,n}) \tag{9}$$

式中:$\boldsymbol{i}_t^{l,n}$为输入门,$\boldsymbol{f}_t^{l,n}$为遗忘门,$\boldsymbol{o}_t^{l,n}$为输出门,$\boldsymbol{c}_t^{l,n}$为激活单元,$\boldsymbol{h}_t^{l,n}$为该单元的输出量,$\boldsymbol{W}_{mn}^{l,n}$为$l$层TS-LSTM中第$n$个单元到第$m$个单元的连接权重矩阵,此时每个LSTM单元的更新式为

$$\boldsymbol{h}_t^{l,n}=\boldsymbol{o}_t^{l,n}\tanh(\boldsymbol{c}_i^{l,n}) \tag{10}$$

接下来设第m个时间序列输入为$\boldsymbol{X}_t^l$,将它表示为$\boldsymbol{X}_t^{l,m}$并代入到式(6~9)中,同理将$\boldsymbol{h}_t^{l,n}$表示为$\boldsymbol{h}_t^{l,n,m}$,并代入到式(10)中,可得到

$$q_{\mathrm{S}}^{l,m}=\sum_{t=0}^{w_l-1}\mathrm{concat}\left(\left[h_{n*TS_{l+t}}^{l,n,m}\right]_{n=0}^{n=N_l-1},0\right);\ \left[(\bullet)\right]_{n=0}^{n=N_l-1}=\left[(\bullet)_0(\bullet)_1\cdots(\bullet)_{N_l-1}\right] \tag{11}$$

$$q_{\mathrm{M}}^{l,m}=\frac{q_{\mathrm{S}}^{l,m}}{W_l} \tag{12}$$

式(11)和(12)分别表示l层第n个单元在m时序下的MeanPool值和SumPool值。接着将其串联可得到

$$\begin{cases} r_{\mathrm{S}}^{m}=\left[q_{\mathrm{S}}^{0,m}\right]^{\mathrm{T}} \\ r_{\mathrm{M}}^{m}=\left[\mathrm{concat}\left(\left[q_{\mathrm{S}}^{1,m},q_{\mathrm{S}}^{2,m}\right],1\right)\right]^{\mathrm{T}} \\ r_{\mathrm{L}}^{m}=\left[\mathrm{concat}\left(\left[q_{\mathrm{S}}^{3,m},q_{\mathrm{S}}^{4,m},q_{\mathrm{S}}^{5,m}\right],1\right)\right]^{\mathrm{T}} \end{cases} \tag{13}$$

TS-LSTM模块的线性激活算法如式(14～16)。

$$a_{\mathrm{S}}^{m}=w_{\mathrm{S}}\cdot r_{\mathrm{S}}^{m}+b_{\mathrm{S}} \tag{14}$$

$$a_{\mathrm{M}}^{m}=w_{\mathrm{M}}\cdot r_{\mathrm{M}}^{m}+b_{\mathrm{M}} \tag{15}$$

$$a_{\mathrm{L}}^{m}=w_{\mathrm{L}}\cdot r_{\mathrm{L}}^{m}+b_{\mathrm{L}} \tag{16}$$

式中:w和b分别代表线性层的权重和偏差。将激活后设a_{S}^{m},a_{M}^{m}和a_{L}^{m}的第k种动作值分别设为$a_{\mathrm{S}}^{m,k}$、$a_{\mathrm{M}}^{m,k}$和$a_{\mathrm{L}}^{m,k}$,最后将其归一化得到

$$Pr(c|a_{\mathrm{S}}^{m})=\frac{\exp(a_{\mathrm{S}}^{m,c})}{\sum_{k=0}^{N_c-1}\exp(a_{\mathrm{S}}^{m,k})} \tag{17}$$

$$Pr(c|a_{\mathrm{M}}^{m})=\frac{\exp(a_{\mathrm{M}}^{m,c})}{\sum_{k=0}^{N_c-1}\exp(a_{\mathrm{M}}^{m,k})} \tag{18}$$

$$Pr(c|a_{\mathrm{L}}^{m})=\frac{\exp(a_{\mathrm{L}}^{m,c})}{\sum_{k=0}^{N_c-1}\exp(a_{\mathrm{L}}^{m,k})} \tag{19}$$

式中:N_c和c分别表示动作种类数和对应类的索引。利用交叉熵函数的样本的最大似然估计值,其过程如式(20)所示。

$$e=-\sum_{m=0}^{N_{\mathrm{M}}-1}\sum_{c=0}^{N_c-1}y_c^m\ln\{Pr\} \tag{20}$$

式中:$N_{\mathrm{M}}-1$和y_c^m分别表示样本训练总数和实际动作标签。为突出同种动作由于运动幅度不同而产生的类间差距,利用最小目标函数对其进行再次训练,最后测试时取式(17～19)三个激活函数结果的平均值。

1.3 融合策略

在得到最终输出前需要对双流网络进行时空融合,其中损失函数选用交叉熵函数,得到当前预测样本属于所有类别的概率,具体的计算方式为

$$\hat{y}_c=\lambda\frac{\sum_n\hat{y}_{ct}}{n}+(1-\lambda)\frac{\sum_n\hat{y}_{st}}{n} \tag{21}$$

式中:$\hat{y}_{ct}$和$\hat{y}_{cs}$分别为空间流和时间流的对于当前输入的预测概率向量,λ为(0－1)的常数。为寻找较为合适的λ值,将该问题描述成一个简单的最优解取值问题,使用粒子群算法通过不同的λ取值,得到最适用于本问题的解。

2 实验结果及分析

2.1 数据集

本文使用的危险行为数据集来自于公开的NTU-RGB+D数据集和Kinetics数据集。NTU-RGB+D数据集是目前应用最广泛的3D人体行为数据集之一,该数据集是由3台Kinect相机采集到的视频集,包含60个不同种类的56 880个动作样本视频,其中包含单人动作和双人动作。该数据集的评价指标为跨视角(Cross subject,CS)精度和跨角度(Cross view,CV)精度。

Kinetics数据集同样也是人体行为识别领域应用比较广泛的数据集,它包含的动作种类较多,视频数量基数大。为了与NTU-RGB+D数据集尽量一致,对超过两人的动作视频,选择骨骼数据最为完整健壮的两个作为输入;同时对一些置信度低的视频片段进行删除,不计入最后结果。Kinectics数据集的评价指标为Top-1精度和Top-5精度。

因为Kinetics数据集包含动作种类较多,所以本文在其中筛选出与NTU-RGB+D数据集中有交集的15种动作,约16 000个剪辑视频片段,其中训练集占2/3,测试集占1/3。这15种危险动作分为单人动作和双人动作,其中定义为危险的单人动作有踢东西、背痛、脖子痛、呕吐、跌倒、头疼、恶心、扔东西、胸疼和蹒跚;定义为危险的双人动作有推搡、指向某人、拍打、脚踢和摸钱包。在训练前对每个视频片段依据时长进行随机取帧,其范围为200～250。

2.2 实验过程及结果

为确定融合方式以及相关参数,一共进行4组实验。

实验1 CNN层动静态融合方式的确定

实验1设定的融合策略为5种,分别是动静态特征最大值融合、动静态特征最小值融合、动态特征最大值与静态特征最小值融合、动态特征最小值与静态特征最大值,以及动静态特征均值融合。从这5种方式中选择精度最高的作为空间流的输出。

为证明本文方法的普适性,每组实验将分别进行两次。将VGG16模型最后一个F层分类参数设置为45;初始学习率为0.001,切割的视频尺寸规范到224×224,每个动作视频取帧范围为200～250。将CS/CV和Top-1/Top-5共同作为评价指标,所得精度分别如表1和2所示。

表1 动静态特征融合策略的CS/CV精度

Table 1 CS/CV accuracy of dynamic and static feature fusion strategy

融合方式	CS精度/%	CV精度/%
动静态特征最大值	79.3	80.8
动静态特征最小值	81.1	83.4
动态特征最大值+静态特征最小值	83.5	86.2
动态特征最小值+静态特征最大值	82.8	86.7
动态特征均值+静态特征均值	83.7	87.6

表2 动静态特征融合策略的Top-1和Top-5精度

Table 2 Top-1 and Top-5 accuracy of dynamic and static feature fusion strategy

融合方式	Top-1精度/%	Top-5精度/%
动静态特征最大值	33.6	51.1
动静态特征最小值	34.4	53.5
动态特征最大值+静态特征最小值	35.7	54.2
动态特征最小值+静态特征最大值	35.9	54.6
动态特征均值+静态特征均值	36.5	55.3

通过表1和2的结果可以确定,空间流中的动态特征和静态特征的融合方式为均值融合,其中CS精度为83.7%,CV精度为87.6%;Top-1精度为36.5%,Top-5精度为55.3%。

实验2 消融实验

为验证TS-LSTM网络相比于传统LSTM网络的优势以及中网络中增加Concat层、Sumpool层、Linear层以及Dropout层的意义,进行了一组消融实验,并分别在两个数据集中进行测试,其精度分别如表3和4所示。

表3 消融实验中的CS/CV精度

Table 3 CS/CV accuracy in ablation experiment

融合方式	CS精度/%	CV精度/%
LSTM	78.5	80.3
TS-LSTM	80.4	82.7
TS-LSTM+Concat	82.3	84.9
TS-LSTM+ Concat + Sumpool	83.5	86.2
TS-LSTM+ Concat + Sumpool+Linear	84.1	87.4
TS-LSTM+ Concat + Sumpool+Linear+Dropout	85.3	88.1

表4 消融实验中的Top-1和Top-5精度

Table 4 Top-1 and Top-5 accuracy in ablation experiment

融合方式	Top-1精度/%	Top-5精度/%
LSTM	33.9	52.0
TS-LSTM	34.7	54.5
TS-LSTM+Concat	35.6	55.4
TS-LSTM+ Concat +Sumpool	36.1	55.9
TS-LSTM+ Concat +Sumpool+Linear	36.4	56.5
TS-LSTM+ Concat +Sumpool+Linear+Dropout	36.9	56.8

从表3,4可看出,TS-LSTM网络要比传统LSTM网络识别效果好,并且随着Concat层、Sumpool层、Linear层以及Dropout层的增加,网络的精度也在不断增加。

实验3 时空融合策略

双流网络时空融合方式的确立是保证识别精度和识别效率的关键,本文通过式(21)的计算方式将时间流和空间流网络得出的预测值融合。为了找出针对与本文背景下精度最优的λ取值,将该问题描述成一个简单的最优解问题,使用粒子群算法求解。CS、CV、Top-1和Top-5精度随着λ变化的精度曲线如图5所示。由图5结果可以得出,当$\lambda=0$即全时间流效果要比$\lambda=1$即全空间流的效果好,并且随着λ取值的变化,识别精度也相应有所提高,这说明时空双流网络融合的效果与不同网络模型的融合占比有关,使用双流卷积模型要比单一模型效果好,并且当λ接近1/3时识别效率最高。最终平均CS精度为87.9%,平均CV精度为92.5%;平均Top-1精度为37.8%,平均Top-5精度为59.8%。

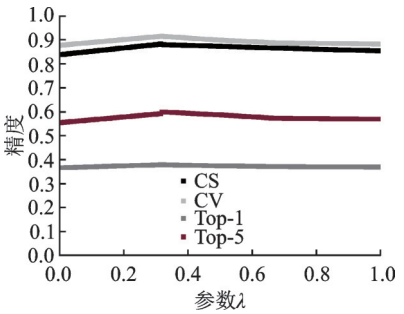

图5 CS/CV/Top-1/Top-5精度曲线

Fig.5 CS/CV/Top-1/Top-5 accuracy curves

为突显本文对于相似动作的识别效率,现选取7种相似动作并绘制出CV精度混淆矩阵,如图6所示。图6结果表明,这几种相似动作的混淆率仅在0.01~0.2之间,说明本文算法具有很好的识别效果。

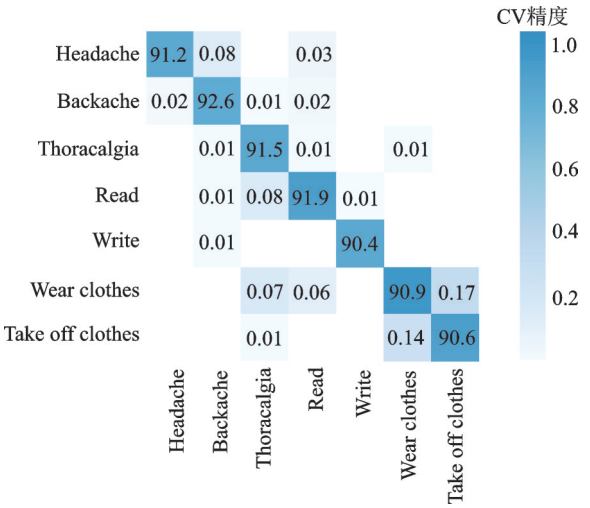

图6 相似动作混淆矩阵

Fig.6 Similar action confusion matrix

实验4 同类算法对比

为进一步证明本文识别算法的优越性,现与其他主流人体行为检测算法进行对比,主要是在NTU RGB+D数据集下进行不同算法的CS和CV的精度对比,结果如表5所示。从表5结果可

表5 不同算法精度对比

Table 5 Accuracy comparison of different algorithms

算法	CS精度/%	CV精度/%
RES-LSTM[10]	76.5	87.8
Two stream-LSTM[11]	77.1	85.1
ATT-RNN[12]	80.7	88.4
IND-RNN[13]	81.8	91.0
ST-AGCN[14]	86.4	92.1
HCN[15]	86.5	91.1
CNN-TS LSTM	87.9	92.5

以看出,本文提出的方法精度较高,因而在人体行为识别领域内具有一定的参考价值。

3 结束语

在总结现有人体骨架行为识别方法的基础上,本文对传统的双流卷积模型进行改进,提出了一种基于CNN-LSTM的双流卷积危险行为识别模型。该模型将CNN网络与LSTM网络并联,其中CNN网络作为空间流,将人体骨架空间运动姿态信息分为静态与动态进行分别提取;在时间流中,利用改进的TS-LSTM网络,增强了人体骨架时序特征的提取能力。通过实验验证该模型能够有效对危险动作进行识别。美中不足的是该模型训练量大且参数多,在未来的工作中争取以更简洁的结构及更少的参数量达到更高的精确度。

参 考 文 献:

[1] 张晓龙,王庆伟,李尚滨.基于强化学习的多模态场景人体危险行为识别方法[J].应用科学学报,2021,39(4): 605-614.
ZHANG Xiaolong, WANG Qingwei, LI Shangbin. Recognition method of human dangerous behavior in multimodal scenes using reinforcement learning[J]. Journal of Applied Sciences,2021,39(4): 605-614.

[2] 丁培甫,詹玲超,胡天敏,等.基于OpenPose的行人异常姿态研究[J].数字技术与应用,2019,37(2): 107-109.
DING Peifu, ZHAN Lingchao, HU Tianmin, et al. Research on pedestrian abnormal pose based on openpose[J]. Digital Technology and Application,2019,37(2): 107-109.

[3] DING W, DING C, LI G, et al. Skeleton-based square grid for human action recognition with 3D convolutional neural network [J]. IEEE Access, 2021, 9: 54078-54089.

[4] JIA J G, ZHOU Y F, HAO X W, et al. Two-stream temporal convolutional networks for skeleton-based human action recognition[J]. Journal of Computer Science and Technology, 2020, 35(3): 538-550.

[5] 田志强,邓春华,张俊雯.基于骨骼时序散度特征的人体行为识别算法[J].计算机应用,2021(5): 1450-1457.
TIAN Zhiqiang, DENG Chunhua, ZHANG Junwen. Human behavior recognition algorithm based on skeletal temporal divergence feature[J]. Journal of Computer Applications, 2021(5): 1450-1457.

[6] WEN Y H, GAO L, FU H, et al. Graph CNNs with motif and variable temporal block for skeleton-based action recognition [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI, 2019, 33: 8989-8996.

[7] SU T T, SUN H Z, MA C M, et al. Research on human behavior recognition based on recurrent neural networks[J]. Journal of Tianjin Normal University (Natural Science Edition), 2018,38(6): 58-62, 76.

[8] PENG W, SHI J, VARANKA T, et al. Rethinking the ST-GCNs for 3D skeleton-based human action recognition[J]. Neurocomputing (Amsterdam), 2021,454: 45-53.

[9] WANG J, LIU Z, WU Y, et al. Learning actionlet ensemble for 3D human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 914-927.

[10] ZHANG S Y, YANG Y, XIAO J, et al. Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks[J]. IEEE Transactions on Multimedia, 2018, 20(9): 2330-2343.

[11] LIU J, WANG G, DUAN L Y, et al. Skeleton-based human action recognition with global context-aware attention LSTM networks[J]. IEEE Transactions on Image Processing, 2018, 27(4): 1586-1599.

[12] ZHANG P, XUE J, LAN C, et al. Adding attentiveness to the neurons in recurrent neural networks[C]//Proceedings of the European Conference on Computer Vision. Beijing, China: [s.n.], 2018: 135-151.

[13] LI S, LI W Q, COOK C, et al. Independently recurrent neural network (Ind-RNN): Building a longer and deeper RNN[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 5457-5466.

[14] 曹毅,刘晨,黄子龙,等.时空自适应图卷积神经网络的骨架行为识别[J].华中科技大学学报(自然科学版),2020,48(11): 5-10.
CAO Yi, LIU Chen, HUANG Zilong, et al. Skeleton-based action recognition based on spatio-temporal adaptive graph convolutional neural-network[J].Journal of Huazhong University of Science and Technology(Natural Science Edition), 2020, 48(11): 5-10.

[15] LI C, ZHONG Q, XIE D, et al. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[C]//Proceedings of International Joint Conference on Artificial Intelligence. Stockholm, Sweden: [s.n.], 2018: 786-792.

作者简介:

高治军(1978-),男,博士,教授,研究方向:信号检测与处理等,E-mail: gzj@sjzu.edu.cn。

顾巧瑜(1997-),**通信作者**,男,硕士研究生,研究方向:数字图像处理,E-mail: gqysyjz@163.com。

陈平(1983-),男,博士,教授,研究方向:信号检测与图像处理等。

韩忠华(1977-),男,博士,教授,研究方向:计算机应用等。

(编辑:王静)