文章编号：2096-1472(2021)-09-24-03
DOI:10.19644/j.cnki.issn2096-1472.2021.09.006

基于深度学习的自动扶梯视频人体动作识别

汪 威，胡旭晓，吴跃成，丁楠楠，王 佳

(浙江理工大学机械与自动控制学院，浙江 杭州 310018)
✉WW1909@126.com; huxuxiao@zju.edu.cn; wuyuecheng@126.com;
3296362443@qq.com; wangjiajia1118@163.com

摘 要：在自动扶梯场景下的视频人体动作识别中，视频数据源不稳定，如遮挡、多视角、光照、低分辨率、动态背景以及背景混乱等均导致动作分类及检测不准确。针对这些问题，提出使用基于改进的SlowFast网络的人体动作识别方法，以更好地捕获视频连续帧中隐藏的时间和空间信息。通过与R(2+1)D卷积网络模型的识别准确率进行对比，改进的SlowFast网络模型在视频中的动作分类和检测方面都表现了很好的性能，能够有效地解决自动扶梯场景下的人体动作识别问题。

关键词：人体动作识别；单流三维卷积神经网络；慢速路径；快速路径；改进的SlowFast
中图分类号：TP249 **文献标识码**：A

Human Motion Recognition in Escalator Video based on Deep Learning

WANG Wei, HU Xuxiao, WU Yuecheng, DING Nannan, WANG Jia

(*School of Machinery and Automatic Control*, *Zhejiang University of Technology*, *Hangzhou* 310018, *China*)
✉WW1909@126.com; huxuxiao@zju.edu.cn; wuyuecheng@126.com; 3296362443@qq.com; wangjiajia1118@163.com

Abstract: In human motion recognition in escalator scene video, the instability of the video data source, such as occlusion, multiple viewing angles, illumination, low resolution, dynamic background, and background confusion, leads to inaccurate motion classification and detection. Aiming at these problems, this paper proposes to use a human motion recognition method based on the improved SlowFast network to better capture the temporal and spatial information hidden in the continuous video frames. Compared with the recognition accuracy of the R (2+1) D convolutional network model, the improved SlowFast network model has achieved better performance in motion classification and detection in videos, and can effectively solve the problem of Human body motion recognition in escalator scene.

Keywords: human motion recognition; single stream 3-D convolutional neural network; slow path; fast path; improved SlowFast

1 引言(Introduction)

自动扶梯是空间开放性运输工具，活动空间相对较大，导致伤害的因素比较多[1]。台阶是持续运动的，乘客进入或者离开台阶区域时运行状态的改变容易使其站立不稳，发生跌倒危险；在乘客越界后自动扶梯与墙壁交叉处产生的“剪切”将严重威胁乘客安全[2]；此外，乘客逆行、携带大件物品等都容易发生意外伤害。自动扶梯人体动作识别的主要目标是判断一段视频中人的动作的类别，主要识别判断危险动作类别，比如身体部位越过安全线、头部外探、下蹲、跌倒、逆行、手提行李箱等大件物品等，保障乘客的人身与财产安全。近年来，基于深度学习网络模型的端到端方法实现了特征提取和分类的无缝连接[3]。本文基于深度学习的方法实现自动扶梯视频中的人体动作识别，对自动扶梯乘客危险动作进行实时监测预警。

2 单流三维卷积神经网络(Single stream 3D convolution neural network)

2.1 三维卷积

单流三维卷积神经网络使用时间卷积来识别视频中人类

基金项目：浙江省基础公益研究计划(LGF19E050005)。

行为，利用在大规模监控视频数据集上训练的深度三维卷积网络进行时空特征学习。三维卷积网络比二维卷积网络更适于时空特征学习，在所有层中均具有3×3×3卷积核的同类架构是三维卷积网络性能最佳的架构之一[4]。与二维卷积网络相比，由于三维卷积和三维池化操作，三维卷积网络能够对时间信息进行建模。在三维卷积网络中，卷积和池化操作是在时间上进行的，而在二维卷积网络中，卷积和池化操作仅在空间上进行。二维卷积网络在每次卷积操作之后立即丢失输入信号的时间信息，只有三维卷积才能保留输入信号的时间信息，从而产生输出量。

2.2 R(2+1)D卷积

将三维卷积滤波器分解为单独的空间和时间分量会显著提高准确性。基于三维卷积，研究设计了一个新的时空卷积块“R(2+1)D”[5]，它将3D卷积显式分解为两个独立且连续的运算，即2D空间卷积和1D时间卷积。用一个大小为$N_{i-1}\times 1\times d\times d$的$M_i$2D卷积滤波器和一个大小为$M_i\times t\times 1\times 1$的$N_i$时间卷积滤波器组成的(2+1)D块替换了大小为$N_{i-1}\times t\times d\times d$的$N_i$3D卷积滤波器。第一个优点是这两个操作之间的附加非线性整流。与在相同数量的参数下使用完整3D卷积的网络相比，这有效地使非线性数量增加了一倍，从而使网络模型能够表示更复杂的函数。第二个潜在的好处是分解有助于优化。

3 改进的SlowFast网络(Improved SlowFast network)

3.1 网络原理

一种著名的视频识别体系结构是双流设计[6]，但其提出的观念并没有探索时间轴的影响，其两个流采用相同的主干结构。

运动是方向的时空对应物，但并非所有的时空方向都具有相同的可能性。慢动作比快动作更有可能运动，如果所有时空方向的可能性都不相同，那么就没有理由像基于时空卷积的视频识别方法中所说明的那样，对空间和时间进行对称处理。对于人体动作识别，SlowFast网络[7]不额外捕获光流或近似光流特征，而是用帧的刷新速度来区分空间和时间关系，分别处理空间结构和时间事件。视频场景中的帧通常包含两个不同的部分：不怎么变化或者缓慢变化的静态区域和正在发生变化的动态区域。在视觉内容的范畴空间语义往往发展缓慢，例如，挥手在挥手动作的跨度上不会改变自己作为“手”的身份，一个人即使可以从走路切换到跑步，也始终处于“人”的范畴。因此，动作分析中语义的识别，如颜色、纹理、光线等可以相对缓慢地刷新。另一方面，正在执行的动作可以比主体身份变化快得多，例如拍手、挥手、颤抖、走路或跳跃，于是我们迅速地去刷新动作帧，但是不改变执行动作人的身份信息。利用快速刷新帧(高时间分辨率)对潜在的快速变化运动进行有效建模是一种理想的方法。

3.2 网络结构

SlowFast网络可以描述为在两个不同帧率下运行的单一流架构，可以进行端到端的网络训练。其网络结构原理图如图1所示。

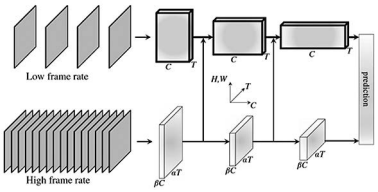

图1 SlowFast网络结构

Fig.1 SlowFast network structure

SlowFast网络主要包含两个网络分支：一个低帧，低时序分辨率的慢速路径；一个高帧，高时序分辨率的快速路径。快速路径的时序分辨率为慢速路径的α倍数，通道数为慢速路径的β倍数(如1/8)。最后，进行横向连接融合两个路径。

(1)慢速路径(Slow pathway)

慢速路径输入为低帧率数据，主要捕获空间语义信息，以低帧率和缓慢的刷新速度运行。慢速路径可以是任何卷积模型，其输入源视频剪辑作为一个时空量。慢速路径在输入帧上有一个大的时间步伐τ，原始输入视频$T\times\tau$帧，以步伐τ进行采集，采集到T帧图像送入慢速通道训练。

(2)快速路径(Fast pathway)

快速路径输入为高帧率数据，主要捕获时序动作信息，以高帧率和快速的刷新速度运行。尽管快速路径在时间维度刷新很快，但是在整个网络中，其只占用了20%的计算量，通道数很少，是一个轻量级子网络。快速路径对空间信息的捕获能力较弱，但能捕获到对动作识别有用的信息。快速路径与慢速路径平行，是另一个卷积模型。快速路径在时序方向使用步伐比较小的方式进行采样，步伐表示为τ/α，这里$\alpha>1$，表示快速路径与慢速路径帧率的比值。这两条路径在同一输入视频源上进行剪辑操作(但步伐不一样)。快速路径采样$\alpha\times T$帧，比慢速路径密度大。

(3)横向连接(Lateral connections)

两条路径的信息是融合的，在融合之前，其中一条路径并不会意识到另一条路径所习得的信息。每一个“阶段”在两条路径之间附加一个横向连接[8]，对于ResNets[9]，这些横向连接的部分分别位于pool1、res2、res3与res4层之后。两种路径的时间维度是不一样的，需要对它们进行一个转换后才能进行匹配，使用单向连接的方式，融合快速路径的特征到慢速路径。最后，对于每个路径的输出，将两个混合的特征向量串联起来作为全连通分类器层的输入。

3.3 网络结构的改进

(1)进一步减少轻量级快速路径的空间容量

快速路径在空间维度上没有特殊处理。因此，其空间建模能力应低于慢速路径，需要减少快速路径对空间的捕获能力，同时增加其对时间的捕获能力。结合降低输入空间分辨率和去除颜色信息等方式，最大化降低快速路径的空间容量来实现轻量化。

(2)对时态卷积的优化应用

在慢速路径中，从conv1层到res3层本质上都是使用二维卷积核。通过实验发现，如果在早期的网络层使用带时序的

卷积核会降低准确率。当目标移动比较快、时间步长比较大时，如果时间感受野比较小，就没有办法把动作连贯起来，除非空间感受野足够大，否则在一个时间感受野内几乎没有相关性。因此，我们只在res4层和res5层中使用非退化的时态卷积。

4 实验与结果分析(Experiment and result analysis)

4.1 数据集与实验环境

按照UCF101[10]公共数据集，将一个人体动作类的剪辑分为25 个组，每个组包含4—7 个剪辑，每一组剪辑具有一些共同的特征，例如背景或乘客。针对身体部位越过安全线、头部外探、下蹲、跌倒、逆行、手提行李箱等大件物品等危险动作类别，采集动作序列视频数据作为自动扶梯人体动作模型库标准，划分出训练集和测试集。

利用楼梯场景下人体动作数据集进行预训练，进一步提高训练模型针对我们预设几种人体动作的识别准确率。其中楼梯场景下的人体动作类别与自动扶梯场景下需进行识别的人体动作类别一致。部分自动扶梯场景下人体动作数据集视频帧如图2所示。

图2 自动扶梯场景下人体动作数据集视频帧示例

Fig.2 Video frame examples of human motion data set in escalator scenes

此次实验在Ubuntu 16.04操作系统下进行，处理器型号为Intel i7-9750H，显卡型号为NVIDIA GTX1660ti，深度学习平台使用PyTorch框架搭建。网络训练的初始学习率设置为0.01，每进行10 次迭代学习率除以10，网络训练的周期设置为300，一次训练所选取的样本数设置为16。以原始图像数据的方式加载数据，把视频先切割成每帧图片，然后加载训练。使用训练集进行训练，并使用测试集进行测试。

4.2 实验过程

针对R(2+1)D网络训练，将网络设置为18 层，输入的视频帧被缩放为128×170的大小，然后通过随机裁剪大小为112×112的窗口方式来生成每个剪辑。在训练时，从视频中随机采样$L=16$ 个连续帧，并对视频进行时间抖动。批量归一化应用于所有卷积层。

针对SlowFast网络训练，慢速路径的主干网络选择3D ResNet-50结构，从输入的64 帧图像中，使用时间步长$\tau=16$ 稀疏采样的方式，采集$T=4$帧图像作为慢速路径的输入。快速路径的时间步长$\tau/\alpha=2(\alpha=8)$以及采样$\alpha\times T=32$ 帧$(\beta=1/8)$图像，在整个网络的时序维度上都没有进行下采样，尽可能保持时间逼真度。横向连接从快速路径到慢速路径使用一个卷积层进行融合。慢速路径的特征形状表示为$\{T,S^2,C\}$，快速路径的特征形状表示为$\{\alpha T,S^2,\beta C\}$。慢速路径的特征形状不进行改变，主要调整快速路径输出特征的形状，让其能和慢速路径进行匹配。

4.3 实验结果与对比分析

针对网络训练所得到的网络模型，R(2+1)D网络模型与改进的SlowFast网络模型的最终训练效果比较如表1所示。

表1 网络模型效果表现比较

Tab.1 Comparison of network model performance

网络模型	预训练	top-1	top-5
R(2+1)D	楼梯场景数据集	68.73	80.65
SlowFast4×16，R50	楼梯场景数据集	75.80	93.40

使用R(2+1)D模型的RGB网络流在自动扶梯数据集上达到了80.65%的识别准确率。以视频切割帧的方式进行模型训练的部分识别测试结果截图，如图3所示。

图3 R(2+1)D网络正确识别视频帧示例

Fig.3 Video frame examples of correct recognition in R(2+1)D network

R(2+1)D模型以视频切割帧的方式进行模型训练的部分错误识别测试结果截图，如图4所示。

图4 R(2+1)D网络错误识别视频帧示例

Fig.4 Video frame examples of error recognition in R(2+1)D network

使用改进的SlowFast网络模型在自动扶梯数据集上达到了93.4%的识别准确率。以视频切割帧的方式进行模型训练的部分识别测试结果截图，如图5所示。

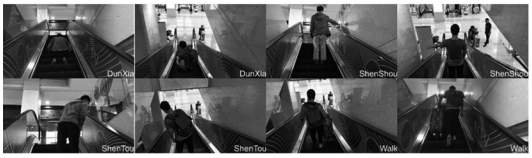

图5 SlowFast网络正确识别视频帧示例

Fig.5 Video frame examples of correct recognition in SlowFast network

针对不同的人做同一类动作，即使同一个人做同一类动作，由于个体差异、动作快慢、环境及背景等不同，以及不同类的动作可能表现出很相似的特征[3]，R(2+1)D模型在视频中的表现可能会产生很大误差。通过实验对比，改进的SlowFast网络对于动作的类内差异性和类间相似性表现出了相对于R(2+1)D模型更加优异的性能，大大提高了识别准确率，并且达到了更好的实时性要求。

5 结论(Conclusion)

本文根据自动扶梯场景下人体危险动作类别识别监测的需要，考虑到时间轴这一特殊的维度，研究设计了一种架构，该架构对比了沿时间轴的速度，它可为视频动作分类和检测提供更优异的准确性与更好的识别速度。通过与R(2+1)D网络模型的对比分析，改进的SlowFast网络能有效地解决自动扶梯场景下的人体动作识别问题，并且能够满足实时性要求，一定程度上促进了对视频识别的进一步研究。

参考文献(References)

[1] 杨冠宝.基于全景视觉的自动扶梯节能及智能监控系统[D].杭州:浙江工业大学,2011.

[2] 陈旻.浅析自动扶梯及自动人行道中的“剪切”危险[J].机电技术,2009,32(04):104-107.

[3] 罗会兰,童康,孔繁胜.基于深度学习的视频中人体动作识别进展综述[J].电子学报,2019,47(05):1162-1173.

[4] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]// MORTENSEN E, FIDLER S. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015:4489-4497.

[5] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]// MORTENSEN E. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018:6450-6459.

[6] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 1(4):568-576.

[7] FEICHTENHOFER C, FAN H, MALIK J, et al. SlowFast networks for video recognition[C]// MORTENSEN E. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South):IEEE, 2019:6201-6210.

[8] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]// MORTENSEN E. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017:936-944.

[9] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// MORTENSEN E, SAENKO K. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, 2016:770-778.

[10] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. Computer Science, 2012, 3(12):2-9.

作者简介：

汪　威(1997-)，男，硕士生.研究领域：图像处理，计算机视觉.

胡旭晓(1965-)，男，博士，教授.研究领域：图像处理，机器视觉.

吴跃成(1966-)，男，博士，副教授.研究领域：人机交互.

丁楠楠(1996-)，男，硕士生.研究领域：图像处理.

王　佳(1998-)，女，硕士生.研究领域：故障诊断算法研究.

(上接第31页)

[7] 李慧,马小平,胡云,等.融合社会网络与信任度的个性化推荐方法研究[J].计算机应用研究,2014(03):808-810.

[8] HEATON J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning[J]. Genet Program Evolvable Mach, 2018(19):305-307.

[9] GU J X, WANG Z H. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77(01): 329-353.

[10] LIU H, WANG Y, PENG Q, et al. Hybrid neural recommendation with joint deep representation learning of ratings and reviews[J]. Neurocomputing, 2020, 374(1):77-85.

[11] HUANG Z H, SHAN G X, CHENG J J, et al. TRec: An efficient recommendation system for hunting passengers with deep neural networks[J]. Neural Computing and Applications, 2019, 31(1):209-222.

[12] 邓存彬,虞慧群,范贵生.融合动态协同过滤和深度学习的推荐算法[J].计算机科学,2019,046(008):28-34.

[13] 吴国栋,宋福根,涂立静,等.基于改进CNN的局部相似性预测推荐模型[J].计算机工程与科学,2019,041(006):1071-1077.

[14] 杨洋,邸一得,刘俊晖,等.基于张量分解的排序学习在个性化标签推荐中的研究[J].计算机科学,2020,47(S2):525-529.

[15] HAN J, ZHENG L, XU Y, et al. Adaptive deep modeling of users and items using side information for recommendation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019(99):1-12.

[16] 赵文涛,任行学.融合标签信息和时间效应的矩阵分解推荐算法[J].信息与控制,2020,49(4):472-477.

[17] BANERJEE S, BANJARE P, PAL B, et al. A multistep priority-based ranking for top-N recommendation using social and tag information[J]. J Ambient Intell Human Comput, 2021(12):2509-2525.

作者简介：

郑东霞(1978-)，女，硕士，副教授.研究领域：机器学习，推荐系统.