[文章编号] 1003－4684(2026)02-0067-05

一种背景游走的行为识别网络

程庆贺[1]，田铎泽[2]，巫世锋[3]，胡 燕[1,2]，钟 珞[1,2]，陈 云[4]

(1 湖北孝感美珈职业学院信息技术学院，湖北 孝感 432017；
2 武汉理工大学计算机与人工智能学院，湖北 武汉 430070；
3 中乾立源工程咨询有限公司，湖北 武汉 430070；4 武汉生物工程学院，湖北 武汉 430415)

[摘 要] 提出一种针对行为识别数据场景增强的背景随机游走(BP-RWE)数据增广算法来增加数据样本的多样性。使用行为本身的背景语义做像素变换，并考虑了场景中实体的帧间时空一致性，让随机游走模式做几何变换，从而增强识别模型对场景中复杂动作的适应性。在两个视频行为识别标准数据集 UCF101 和 HMDB51 上进行了对比、消融和可视化等实验。在对比实验中，比基准(Baseline)方法的准确率分别提升了 1.00%和 2.48%。

[关键词] 行为识别；依赖性；随机游走；数据增广；适应性

[中图分类号] TP389.1 [文献标识码] A

行为识别技术有着重要的研究价值和实际应用意义，在智能监控[1]、AI 机器人[2]、人机交互[3]、虚拟现实技术[4]等领域中具有广阔的应用前景。其任务目的是判断一段人类行为视频的动作类别。

本文对现有数据集利用干扰信息增强数据的多样性和场景的复杂性，并根据网络的输入信息增加复杂度条件，利用特征间的语义关联和相互补充改进网络结构，提高模型对复杂场景中识别动作的鲁棒性。利用深度神经网络提高了识别模型的鲁棒性和运算效率，大大节省了人力时间。

双流卷积网络(two-stream convolutional network，TS)是深度学习在视频行为识别领域的重要网络结构。Two- Stream CNN[5]方法的提出标志着深度学习正式应用到行为识别领域，此网络结合了行为视频中的时空两个维度的信息，并使用多任务学习，比传统方法处理的训练数据量更大。Fusion[6]针对双流网络结构不能充分利用视频的时间维度，提出空间融合和时间融合方式，学习行为动作空间和时间维度之间的特征关系。3D 卷积网络(3D convolutional network，C3D)结构比 2D-CNN 增加了时间维度，改善了双流网络结构无法输入长距离时空信息的缺点，Tran 等人受到动作特征的时间维度启发，首次提出 C3D 网络(3D convolution，C3D)[7]。Hara 等人验证 3D 卷积网络能够在现有行为数据集上进行完全收敛的训练[8]，并且使 ResNet[9]结构达到 152 个卷积层。2019 年，Feichtenhofer 等人将双流网络和 3D 卷积网络结合，提出了 SlowFast 网络[10]，其综合了双流网络结构和 3D 卷积网络的优势，在效率和准确率方面都体现了强大的竞争力，因此作为本文的基准方法(Baseline)。

上述模型仍面临：1)场景偏差会对行为识别模型造成不可预测的干扰，使得动作对象局部信息丢失从而降低识别准确度；2)现有数据集中存在训练数据缺少多样性，模型不能够对真实场景中的动作有鲁棒性的识别；3)随着场景干扰信息的加入，网络的注意力会重新分配，训练难度增加。

1 基于背景随机游走的行为识别

1.1 随机游走擦除

本文提出的随机游走擦除算法，在同一个视频连续帧间的擦除区域之间添加随机游走时空关联信息，每一帧都会相对于前一帧在一定范围内移动，可以更准确地增加视频中场景的动态干扰数据。

算法思路如下：

输入：待处理视频。

输出：经过随机游走擦除数据增广后的视频帧序列。

Step1：视频在每一轮训练中使用视频解码器切割成 N 帧，每帧图像大小为(W× H)。

Step2：每个视频随机初始化一个擦除区域与视频帧的面积比 S_e 和擦除区域本身纵横比 r_e。

[收稿日期] 2023－07－27
[基金项目] 国家自然科学基金(62271361)；湖北省自然科学基金(2021CFB513)
[第一作者] 程庆贺(1995－)，男，河南夏邑人，工学硕士，湖北孝感美珈职业学院讲师，研究方向为行为识别、行人再辨识。
[通信作者] 陈 云(1990－)，女，湖北武汉人，工学硕士，武汉生物工程学院讲师，研究方向为行为识别、目标检测。

Step3：随机初始化擦除区域的长度和宽度，得到同一个 epoch 中同一个视频的擦除区域固定大小和形状，计算如公式(1)和(2)。

Step4：对于每个视频序列的第一帧，随机初始化擦除区域的坐标 x_0 和 y_0，计算如公式(3)和(4)。

Step5：根据前一帧的位置，随机初始化一个角度，然后相对于前一帧进行移动，计算如公式(5)。

Step6：对于剩余的 N-1 帧，使用随机游走来模拟擦除块移动的未知性，引入步幅 s 以使擦除块的移动控制在一定的范围内，计算如公式(6)。

Step7：为每个帧替换擦除区域的像素值。

$$h = \sqrt{S_e * r_e} \tag{1}$$

$$w = \sqrt{S_e / r_e} \tag{2}$$

$$x_0 = random(0, W - w) \tag{3}$$

$$y_0 = random(0, H - h) \tag{4}$$

$$\theta = random(0, 360) \tag{5}$$

$$s = random(0, k) \tag{6}$$

1.2 背景补丁替换

在视频动作识别任务中，动作前景识别的准确率容易受到背景信息的干扰，模型容易过拟合训练集，对新数据的泛化能力差。因此，本文探索了一种背景信息替换的像素变换策略来消除这种负面影响，如图 1 所示，作为背景补丁替换的区域称为背景补丁，即图中的 Ipatch。

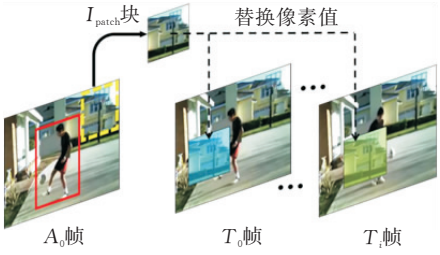


图 1 背景补丁替换算法

算法思路如下：

输入：具有 N 帧的视频。

输出：视频每帧对应擦除区域的像素值被 Ipatch 的像素值覆盖填充。

Step1：给定一个具有 N 帧的视频，对视频帧进行前后景检测。

Step2：取每个视频能检测到 Actor 的第一帧来选择背景区域，用 A_0 表示。

Step3：将 A_0 帧中的前景区域定义为 SDetect，在其之外的其他区域被视为背景区域。

Step4：使用补丁区域与视频帧的面积比 S_e 和补丁区域本身纵横比 r_e，在 A_0 帧的背景区域随机初始化补丁的大小和位置，计算如公式(7)，(8)，(9)和(10)。

Step5：获得的背景补丁图像命名为 Ipatch，大小为 $(w_p \times h_p)$。

Step6：回到本视频 N 帧的第一帧进行背景补丁擦除操作。

Step7：每帧对应擦除区域的像素值被 Ipatch 的像素值覆盖填充。

$$h_p = \sqrt{S_e * r_e} \tag{7}$$

$$w_p = \sqrt{S_e * r_e} \tag{8}$$

$$x_p = \text{random}(0, W - w_p) \tag{9}$$

$$y_p = \text{random}(0, H - h_p) \tag{10}$$

综上分析可知，随机游走擦除策略增加了一定区域内移动的时空信息，背景补丁替换算法增加了背景语义干扰的像素填充，两个策略组成了背景补丁-随机游走的数据增广方式，简称为背景随机游走。

2 模型架构

本文基于双流 3D-CNN 框架的快慢时空网络方法作为基线(Baseline)，基于 SlowFast 架构的思想，使用两个路径，其中“慢”路径用于捕获视频稀疏采样的帧，提供视频中静态语义信息；“快”路径用于学习视频的密集采样帧中快速变化的动作，输入是高帧率的采样(为慢通道的 α 倍，即输入的采样数为 αT)，并且其卷积通道(卷积核的个数决定)通常设置为慢路径的 β 倍(一般 β 设置为 1/8)。快路径主要处理视频中快速变化的动作，处理空间信息的能力较弱，静态的空间信息可以由慢路径来学习，二者通过侧向连接进行融合，如图 2 所示。

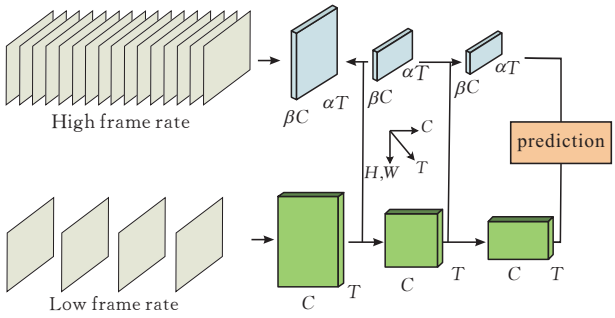


图 2 快慢时空网络框架

此处 3D 卷积网络的动态路径的输出结果通过侧向连接送入静态路径，但是，两个路径通道的数据样本形状是不同的，快通道的特征张量形状是(αT，W×H，βC)，慢通道则是(T，W×H，C)，这就需要在融合前对快通道的输出形状进行变换。本文采用沿着时间轴进行 5×1×1 的 3D 卷积，跨度为 α，输出的通道数为 2βC。

2.1 计算网络实现

快慢时空网络中的快路径相对于慢路径仅使用一部分的 β 个通道(本文与 Baseline 保持一致，β=1/8)，因此快速通道是轻量级的。横向连接将它们

融合在一起，然后执行全局平均池化，并通过连接两个路径的结果送入一个全连接分类层使用 Softmax 来分类行为类别。

交叉熵主要用于度量两个概率分布间的差异，行为识别任务中常使用交叉熵函数作为损失函数，在损失函数值较大时学习率快速更新，损失函数值较小时，学习率慢速更新以细致学习行为特征，因此本文使用交叉熵（crossentropy，CE）损失函数来计算 Loss。L_{CE}的计算公式如下：

$$L_{CE}=\sum_{i=1}^{N}y^{(i)}\log\hat{y}^{(i)}+(1-y^{(i)})\log(1-\hat{y}^{(i)})$$

2.2　数据集

本文采用两个广泛应用的基准数据集来评估方法，UCF101[11] 和 HMDB51[12]。UCF101 包含 101 个不同的人类行为动作类别，包含来自 YouTube 平台的 13320 个人类行为视频，每个源视频通常对应一个动作类型。HMDB51 数据集包含 51 个动作类别中的 6849 个视频。

2.3　评价指标

本文使用到的评价指标主要包括：分类准确率、注意力可视化、分类效果可视化和收敛效率。

2.4　参数设置

在训练阶段，本文使用的主干网络是 ResNet-50，批量大小为 16，初始学习率为 0.025，采用随机梯度下降（SGD）作为优化器，动量为 0.9，权重衰减为 1e-5。对每个输入视频从 64 帧原始剪辑中稀疏采样，时间步长为 τ，沿着时间轴随机采样视频帧。根据快慢时空网络的设计，两条路径的输入分别为 T 帧和 αT 帧，其中 α 和 τ 值通过模型效果来选择。选取双流 3D-CNN 的 8×8 框架作为 Backbone 网络。在测试阶段沿着视频的时间轴均匀地采样 10 个片段，并将较短的空间侧缩放到 256 像素，每个片段有 3 个视图，通过平均所有片段的预测分数得到最终预测。

3　实验结果与分析

3.1　消融实验

表 1 展示了随机游走擦除算法（RWE）的消融实验。首先从表格整体可以看出，与仅使用 Baseline 算法相比，在基线方法中加入图像中的经典随机擦除算法（RE）和加入本文提出的随机游走擦除算法（RWE），都不同程度上提升了识别准确率，验证了复杂场景数据不足将在一定程度上限制动作识别的准确性。同时，本文提出的 RWE 具有更高的准确率，表明它比 RE 方法更适合模拟视频中场景干扰区域的时空信息。此外，表格的第二部分还对 RWE 中的步幅 β 进行了参数比较实验。实验结果表明，当 $k=1$ 时 RWE 算法的效果最好，与视频场景中干扰物体的移动速度更为一致，比 Baseline 算法在 UCF101 数据集上的准确率 Top-1 和 Top-5 提升了 0.50% 和 0.24%，在 HMDB51 数据集上 Top-1 和 Top-5 准确率提升了 1.86% 和 0.55%。

表 1　随机游走算法在标准数据集上的消融实验

方法	UCF101 数据集		HMDB51 数据集	
	Top-1	Top-5	Top-1	Top-5
Baseline	95.72	99.42	75.54	94.54
Baseline+RE	96.03	99.53	76.67	94.54
Baseline+RWE($k=0$)	96.15	99.57	76.85	94.61
Baseline+RWE($k=1$)	96.22	99.66	77.40	95.09
Baseline+RWE($k=2$)	96.12	99.56	76.92	94.96

表 2 展示了背景补丁替换算法（BP）在数据集 UCF101 和 HMDB51 上的消融实验。首先将基线 Baseline 方法与传统图像中使用的随机擦除（RE）算法和本文提出的背景补丁替换算法（BP）进行了比较，实验结果表明：在行为识别任务中加入替换像素值的擦除算法是有效的。其次将 BP 算法与 RE 算法的实验数据进行对比，相比于 RE 算法的随机替换像素值，本文提出的背景补丁替换算法使用的背景信息替换像素值的方法更适用于行为识别领域，其加入了背景信息的干扰，增强了复杂场景中识别动作的鲁棒性，Top-1 和 Top-5 准确率在 UCF101 数据集上比 Baseline 提高了 0.45% 和 0.16%，在 HMDB51 数据集上提高了 1.24% 和 0.04%。

表 2　背景补丁替换算法在标准数据集上的消融实验

方法	UCF101 数据集		HMDB51 数据集	
	Top-1	Top-5	Top-1	Top-5
Baseline	95.72	99.42	75.54	94.54
Baseline+RE	96.03	99.53	76.67	94.54
Baseline+BP	96.17	99.58	76.78	94.58

表 3 展示了本文提出的背景随机游走算法（BP-RWE）每个模块和结合的准确率的消融实验。

表 3　背景随机游走行为识别模型消融实验

方法	UCF101 数据集		HMDB51 数据集	
	Top-1	Top-5	Top-1	Top-5
Baseline	95.72	99.42	75.54	94.54
Baseline+RWE	96.22	99.66	77.40	95.09
Baseline+BP	96.17	99.58	76.78	94.54
Baseline+RWE+BP	96.27	99.71	77.82	95.14

实验数据表明，两个模块相比于基线方法其准确率都提升了，特别是两个模块的结合得到了最佳的准确率，比 Baseline 算法在 UCF101 数据集上提升了 Top-1 0.55%的准确率和 Top-5 0.29%的准确率，在 HMDB51 数据集上提升了 Top-1 2.35%的准

确率和 Top-5 0.60%的准确率，体现了方法组合的优越性。

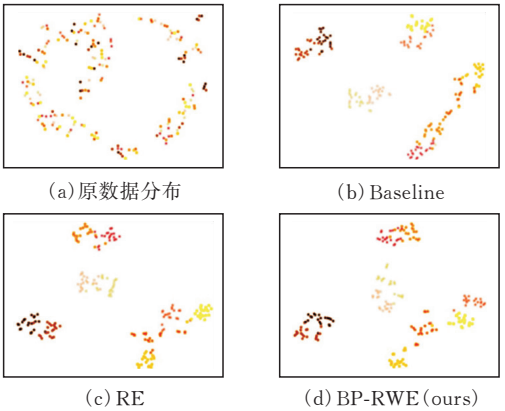
3.2 分类效果可视化实验

为了验证 BP-RWE 数据增广方法来增强识别模型分类的能力，使用 t-SNE 方法将经过模型训练和分类后特征向量，进行统一降维处理。为了验证方法的分类和聚类效果，选取了 UCF101 数据集中识别模型最易混淆的五组类别中的随机数据，进行 t-SNE 分类效果可视化。这些类别中有动作和场景相似(如化眼妆和化唇妆)、场景相似(在篮球场踢足球和打篮球)、动作相似(皮划艇和赛艇)等情况。具体的类别如表 4 的(a)和(b)所示。

表 4 UCF101 数据集中易混淆类别

类别名称(a)	类别名称(b)
Apply Eye Makeup	Apply Lipstick
Basketball	Soccer Juggling
Field Hockey Penalty	Soccer Penalty
Rafting	Rowing
Soccer Juggling	Soccer Penalty

在 UCF101 数据集上，t-SNE 分类效果如图 3 所示。1)是数据的原始分布，可以看到是杂乱无章的分布；2)是 Baseline 方法的分类效果可视化图，可以看到相似类的特征点交叉在一起，并且右下四个类分类很不明显，特征点比较混乱；3)是使用了随机擦除算法(RE)的分类效果可视化图，可以看到特征点分散的情况得到改善，但是右下四个类别仍交叉在一起；4)是使用文章提出的背景随机游走算法(BP-RWE)分类效果可视化，可以看到相似类的特征点分离开，每个类别成团状聚集，体现了算法分类能力的改善，图中 4 个算法的分类效果可视化表明：对动作数据增加干扰的数据增广，可以增强识别模型识别动作的鲁棒性，具有更强的分类能力。

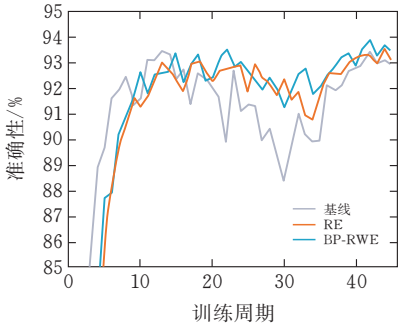


(a)原数据分布 (b) Baseline (c) RE (d) BP-RWE(ours)

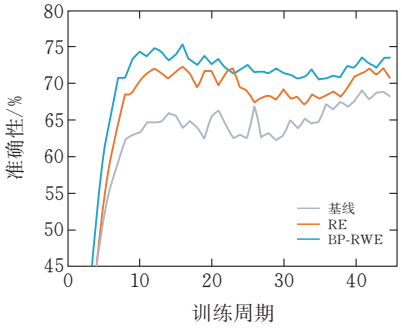
图 3 t-SNE 分类效果可视化

3.3 收敛效率实验

为了关注模型在准确率、时间、算力等方面的性能，图 4 显示了在训练期间模型在两个标准数据集上的准确度。图 4a 是在 UCF101 数据集上 Baseline、RE 和 BP-RWE 模型的训练期间的准确率对比，从图中可以看到：相比于基线方法，后两者的准确率达到收敛时更高并且更稳定，从曲线可以看出 BP-RWE 的模型在训练阶段波动更小，准确率更高。图 4b 是在 HMDB51 数据集上 Baseline、RE 和 BP-RWE 模型的训练期间的准确率对比，相比于 Baseline，后两者比 UCF101 数据集上的效果明显更好，说明了加入场景干扰数据在 HMDB51 数据集上效果要更显著。



(a)UCR101



(b)HMDB51

图 4 在两个标准数据集训练期间的 Top-1 准确度

4 结束语

本文主要关注于对背景随机游走数据增广在行为识别任务上可以取得的效果，对提出的背景随机游走算法(BP-RWE)每个模块准确率进行消融实验。实验数据表明，两个模块相比于基线方法都提升了准确率，特别是两个模块的结合得到了最佳的准确率，比 Baseline 算法在 UCF101 数据集上提升了 Top-1 0.55%的准确率和 Top-5 0.29%的准确率，在 HMDB51 数据集上提升了 Top-1 2.35%的准确率和 Top-5 0.60%的准确率。

[参 考 文 献]

[1] 段志霞，赵娜，宋洁.基于微积分控制的智能灌溉动态监控系统设计[J].农机化研究，2024(05)：206-209.

[2] 包振山，丁义龙，张文博.具备认知能力的行为模仿机

器人系统[J].应用科学学报，2022，40(01)：12.
[3] 陶建华，龚江涛，高楠，等.面向虚实融合的人机交互[J].中国图象图形学报，2023，28(06)：1513-1542.
[4] 朱天乐，姜忠鼎.一种基于视线数据可视化的虚拟现实教学系统[J].计算机应用与软件，2023，40(05)：175-183.
[5] SIMONYAN K，ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Proceedings of the Advances in Neural Information Processing Systems，Montreal，Quebec，Canada，December 8. 2014：568-576.
[6] FEICHTENHOFER C，PINZ A，ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Las Vegas，NV，USA，June 26-July 1. IEEE，2016：1933-1941.
[7] TRAN D，BOURDEY L，FERGUS R，et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision，Santiago，Chile，December 13. IEEE，2015：4489-4497.
[8] HARA K，KATAOA H，SATOH Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Salt Lake City，UT，USA，June 18-22. IEEE，2018：6546-6555.
[9] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，Las Vegas，NV，USA，June 26-July 1. IEEE，2016：770-778.
[10] FEICHTENHOFER C，FAN H，MALIK J，et al. Slowfast networks for video recognition[C]// Proceedings of the IEEE International Conference on Computer Vision，Seoul，South Korea，Oct 27-Nov 3. IEEE，2019：6202-6211.
[11] SOOMMRO K，ZAMIR A R，SHAH M. UCF101：A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402，2012.
[12] KUEHNE H，JHUANG H，GARROTE E，et al. HMDB：a large video database for human motion recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE，2011：2556-2563.

A Behavior Recognition Network Based on Background Walk

CHENG Qinghe[1]，TIAN Duoze[2]，WU Shifeng[3]，HU Yan[1,2]，ZHONG Luo[1,2]，CHEN Yun[4]
(1 *College of Information Tech.*，*Meiga Polytechnic Institute of Hubei*，*Xiaogan* 432017，*China*；
2 *College of Computer & Artificial Intelligence*，*Wuhan Univ. of Tech.*，*Wuhan* 430070，*China*；
3 *Engineering Consulting Company of Zhongqian Liyuan*，*Wuhan* 430070，*China*；
4 *Wuhan University of Bioengineering*，*Wuhan* 430415，*China*)

Abstract：It proposes a background randomly traveled (BP-RWE) data to increase the wide formation algorithm for the enhancement of acting data scenarios to increase the diversity of data samples. It uses the background semantics of the behavior itself to make pixel transformation，and consider the intercourse between time and space of the entity in the scene，so that the random walking mode does geometric transformation，thereby enhancing the adaptability of the complex movement of the identification model to the complex actions in the scene. Comparing，melted and visualized experiments were conducted on the two video behavior recognition standard data set UCF101 and HMDB51. In the comparison experiments，the accuracy of the Baseline method increased by 1.00％and 2.48％，respectively.

Keywords：behavior recognition；dependence；random walk；data expansion；adaptability

[责任编校：闫　品]