

基于梯度特征与 3D 卷积的视频 不同跨度行为识别

戴立新[†], 赵德中, 孙琦, 胡春成, 魏赛, 王亮

(国家电投集团江苏电力有限公司无锡苏州分公司, 江苏 苏州 215000)

摘 要:为了优化视频不同跨度行为识别性能,提出了基于梯度特征与 3D 卷积的视频不同跨度行为识别方法。捕捉视频中的动态信息变化,计算梯度特征值并映射到梯度空间。建立了基于 3D 卷积的多时长特征融合模型,融合不同时间节点下的梯度特征值。结合人体骨骼行为概率、标签隶属度对比以及 0 阶和 1 阶梯度深度特征值,实现视频中的多种行为识别。实验结果表明,所提方法在各种视频环境和行为种类下,识别准确率高达 95.2%,误识别率和漏识别率分别低至 3.1% 和 5.4%,推理时间仅为 37 ms/帧,证明了该方法具有较高的实际应用价值。

关键词:多特征融合; 监控视频; 梯度特征; 像素变化率序列; 梯度空间

中图分类号: TN552

文献标识码: A

Video Different Span Behavior Recognition Based on Gradient Feature and 3D Convolution

DAI Lixin[†], ZHAO Dezhong, SUN Qi, HU Chuncheng, WEI Sai, WANG Liang

(Wuxi Suzhou Branch of State Power Investment Group Jiangsu Electric

Power Co., Ltd., Suzhou, Jiangsu 215000, China)

Abstract: In order to optimize the performance of video behavior recognition across different spans, a video behavior recognition method based on gradient features and 3D convolution is proposed. The dynamic information changes in the video had been captured, and the gradient feature values had been calculated and mapped into the gradient space. A multi-duration feature fusion model based on 3D convolution had been established to fuse the gradient feature values at different time points. By combining the probabilities of human skeletal actions, the comparison of label membership degrees, and the 0th-order and 1st-order gradient depth feature values, multiple types of behavior recognition in the video had been achieved. The experimental results show that the proposed method has a recognition accuracy of up to 95.2% in various video environments and behavior types, with a low false recognition rate and missed recognition rate of 3.1% and 5.4%, respectively. The inference time is only 37 ms/frame, proving that the method has high practical application value.

Key words: multi feature fusion; surveillance video; gradient features; pixel change rate sequence; gradient space

行为识别技术作为机器视觉的关键组成部分,其重要性源于现代社会对高效智能化监控与分析的强烈需求。在安防监控、人机交互界面、医疗健

康监测以及自动驾驶技术等广泛领域内,视频、音频及多种传感器数据的采集与处理技术的飞速进步,为行为识别技术的广泛应用提供了坚实的基

收稿日期:2025-02-12

作者简介:戴立新(1968—),男,江苏常熟人,本科,高级工程师,研究方向:电气工程及其自动化。

[†]通信联系人, E-mail: gongminyan256@163.com

础。然而,人类行为的复杂多变及其对环境的高度依赖性,使得同一种行为在不同跨度下可能展现出截然不同的特征,包括行走速度的变化、姿态的调整等,这些特征差异增加了行为识别的挑战性。

目前,为了实现行为识别,相关学者采用了数据采集、深度学习等方法。蒲瞻星等^[1]给出了一种小样本的视频行为识别算法,对图像进行分割处理,提取图像的视频帧数,并将图像进行组合排列,通过队列计算图像卷积特征并转换为像素参数,按照参数对比实现行为模式识别。该方法对较为复杂的人体变动模式处理效果较差,不能很好地捕捉动作特征,只能识别一些简单的行为;Feng 等^[2]提出细粒度灵长类行为识别的渐进式深度学习框架。用一个以区域为中心的深度卷积神经网络和一个标记的细粒度识别数据集,建立一种渐进的注意力训练模型,并给出区域聚焦图像生成器,通过生成器来识别信息丰富的局部区域,完成行为识别。该方法受复杂背景影响,精度有待进一步提升;Hu 等^[3]给出一种基于深度学习的行为识别方法。建立深度学习的神经网络模型,通过模型提取图像和视频的帧数特征,并将特征与人体骨骼数据相融合得到对比点位对比,完成行为识别。人体的动作变化相对复杂,只以骨骼数据变动作为参照会影响识别的精准性;刘云萍^[4]给出基于神经网络监控图像异常行为识别方法。将监控视频图像进行分割,提取图像帧并建立数值阵列,按照阵列特征对比实现高效行为识别。该方法图像特征帧运算量较大,容易出现误识现象。

由于现有行为识别方法在处理复杂多变的人类行为及环境依赖性方面存在的局限性,包括处理效果受限、精度有待提升以及识别精准性受影响等问题。因此,提出了一种基于梯度特征与 3D 卷积的视频不同跨度行为识别。通过捕捉视频中不同跨度下的动态信息变化并映射到梯度空间,结合 3D 卷积的多时长特征融合模型,实现了对不同跨度动态行为特征的高效提取。此外,通过引入人体骨骼行为概率、标签隶属度对比以及梯度深度特征值,进一步提高了不同跨度下行为识别的准确性和鲁棒性。本文所提方法不仅解决了现有技术的不足,还在不同跨度的视频环境和行为种类下表现出高识别精准度,具有较高的实际应用价值,为行为识别技术的发展提供了新的思路和方法。

1 基于像素动态变化率的视频帧梯度特征提取

在场景转换的时间节点上,行为特征会因为环境的变化而展现出截然不同的特征,从而导致行为跨度的增大,对行为识别方法构成挑战。因此,引入像素变化率的梯度特征提取,通过捕捉到不同跨度下视频帧之间的动态差异,将这些差异量化为梯度特征,从而实现对不同跨度下行为特征的精确捕捉。

像素变化率^[5]是一个能够描述视频中运动信息的变量,当对象运动越剧烈且持续时间越长时,其在像素变化序列中数值越大。基于该特点构建视频前景图像 D_i 的变化比率序列,得到视频的原始区域为 DD_i :

$$DD_i = |D_{i+1} - D_i| + |D_i - D_{i-1}| \quad (1)$$

式中, D_{i+1} 、 D_{i-1} 为前景点和后景点^[6]。

根据式(2)将图像分别放置在水平和垂直方向上进行投影,计算点位序列。

$$PCRM(x, y) = \frac{1}{\max(x, y)} \sum_{i=2}^L DD_i(x, y) \quad (2)$$

式中, (x, y) 为视频原始区域, x, y 为像素点的坐标; L 为视频帧数; i 为像素点。

为了提高多种行为识别精度,将不同跨度下的运动视频可分解为多幅静态图像,以此提供关于多种行为的方向、位移等运动属性,表达式为:

$$MHI(x, y, t) = \begin{cases} \zeta & \text{if } D(x, y, t) = 1 \\ \max[0, MHI(x, y, t-1) - 1] & \text{else} \end{cases} \quad (3)$$

式中, ζ 为时间窗口的大小,用于控制静态图像分解的时间范围; $D(x, y, t)$ 为第 t 帧的前景图像; t 为静态图分解的时间间隔。

对视频进行静态图像分解后,求得图像中每个点的运动方向 $\theta(x, y)$:

$$\theta(x, y) = \arctan \frac{(dMHI/dy)}{(dMHI/dx)} \quad (4)$$

式中, $(dMHI/dy)$ 为纵向分解^[7]的梯度数值; $(dMHI/dx)$ 为横向分解的梯度数值。将 $\theta(x, y)$ 按照横向和纵向进行量化^[8],得到落入每个方向的梯度数量,形成运动方向直方图。

为了保证梯度映射过程中,视频点和图像点之间的对应关系,对运动梯度方向进行改进,表达

式为：

$$F(k) = \text{MHI}(x,y,t) \sum_{\theta(x,y) \in \text{bin}(k)} \theta(x,y),$$
$$1 < k < K \tag{5}$$

$$F'(k) = \left| F(k) + F(k + \frac{K}{2}) \right|, 1 \leq k \leq \frac{K}{2}$$
$$\tag{6}$$

式中， $F(k)$ 为正梯度空间； $F'(k)$ 为负梯度空间； K 、 k 分别为正梯度和负梯度的量化阈值，由视频动态范围归一化后确定。

通过上述公式可输出视频中不同跨度下的运动点的梯度特征值，将该值作为后续行为识别的重要参照，可提升识别的准确性。

2 3D 卷积神经网络框架下多时长视频特征融合

在处理复杂多变的监控视频时，仅依靠梯度特征难以全面覆盖视频中不同跨度下行为的复杂性和多样性。同时，长时间视频处理中累积效应会引发特征冗余，影响识别精度。因此，构建基于 3D 卷积的多时长特征融合模型，通过整合不同跨度下的梯度特征，提升对不同跨度下行为特征的高效提取与精准识别。

在常规的卷积神经网络中，卷积操作是为了便于提取区域内的空间特征值，设视频单帧图像为 $I \in \mathbb{R}^{H \times W}$ ，可提取到特征值 f 为：

$$f = F'(k) \sum_{\Delta y=1}^S \sum_{\Delta x=1}^S \omega(\Delta x, \Delta y) v(x + \Delta x, y + \Delta y)$$
$$\tag{7}$$

式中， Δx 、 Δy 分别为特征变化量； S 为特征权值； ω 、 v 为特征权重和时间跨度参数。

然而，在处理视频时，不仅需要提取空间特征，还需要考虑时间特征。而 3D 卷积操作与 2D 卷积操作^[9]相比，2D 卷积操作主要用于处理二维图像数据，其卷积核在二维空间上滑动以提取特征。3D 卷积操作则用于处理三维数据，包括视频序列，其卷积核在三维空间(包括时间维度)上滑动以提取时空特征。同时，3D 卷积操作能捕捉到视频中的动态信息，包括物体的运动轨迹和速度变化等，这对于行为识别等任务至关重要。此外，3D 卷积操作还可以减少计算冗余，提高处理速度。因此，在多时长特征融合模型中，采用 3D 卷积多特征融合方法对视频中不同跨度的动态行为特征进行融合，实现全特征高效提取。融合结构示意如图 1 所示。

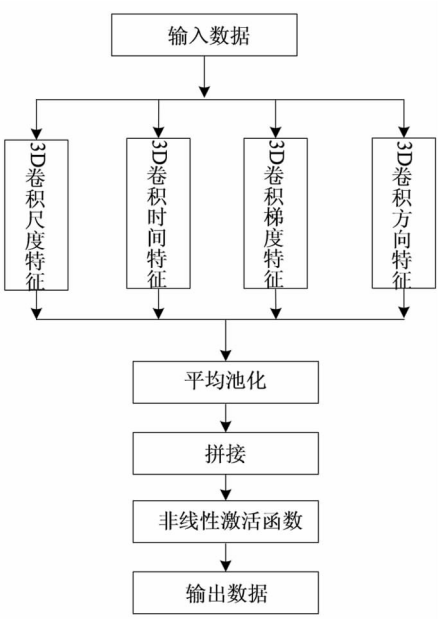


图 1 多时长特征融合结构

由图 1 可知，首先，将视频划分为多个时间窗口，每个时间窗口包含一系列连续的视频帧。这种分割不仅有助于模型处理视频的时序性，还通过数据的局部独立性减轻了累积效应。其次，在每个时间窗口内，运用 3D 卷积操作提取时空特征，包括物体的形状、颜色、运动轨迹和速度等，同时调整卷积核大小和步长来优化特征提取，减少冗余。随后，为了降低特征维度和减少计算量，在 3D 卷积层后添加了池化层，通过平均池化降低特征维度和计算量，同时保留关键特征，进一步缓解累积效应。此外，为了进一步增强模型的表达能力和泛化能力，引入非线性激活函数。通过非线性激活函数增强模型的表达能力和泛化能力，进一步减轻累积效应。最后，将融合后的特征输入到全连接层中，以进行最终的分类。全连接层可以学习特征之间的复杂关系，并输出最终的识别结果。

引入非线性激活函数得到多时长特征融合结果为：

$$x' = G_{\text{ReLU}}[f_{\text{avg}}(f)] \tag{8}$$

式中， $f_{\text{avg}}()$ 为非线性激活函数； $G_{\text{ReLU}}()$ 为动态行为的信息映射函数。通过上述过程可将视频中不同跨度的多种行为特征进行融合。

3 不同跨度下视频中多样化行为的识别

多时长特征融合模型在面对复杂多变的视频时，仍面临对不同跨度行为特征区分能力不足的挑战。因此，引入不同跨度下视频中不同行为的识别

方法,通过结合深度梯度特征、人体骨骼行为概率及标签隶属度对比,从而精确辨识视频中不同跨度下的多样行为,有效提升了模型的识别效能及准确性。

为保证对视频中不同跨度下的多种行为识别准确性和有效性,以上述过程求得的多时长特征融合结果 x' 为初始输入参照,通过求解视频中不同梯度值在邻域内的 0 阶和 1 阶梯度的深度特征值,并结合人体骨骼行为概率和标签隶属度对比实现监控视频的多种行为识别。

0 阶梯度特征值 F_0 为:

$$F_0 = \sum_{r=1}^n f_{d_0}(r)$$

(9)

1 阶梯度特征值 F_1 为:

$$F_1 = \sum_{r=1}^n f_{d_0}(r) f_{d_1}(x' + a_1)$$

(10)

式中, a_1 为视频中邻近点 r 与多时长特征融合结果 x' 之间位移; n 为邻近点数量; f_{d_0} 、 f_{d_1} 分别为 0 阶梯度和 1 阶梯度的位置元素。

通常情况下,视频图像分为深度图序列和浅度图序列,深度图一般为主视图,行为动作的变化在深度图 DMM_c 中表现会更为明显。根据该特点,为了避免极小值和极大值对行为识别判定的影响,将 DMM_c 分布在 $[0,1]$ 中,可得到深度特征值 F_{D1} 的表达式为:

$$F_{D1} = \{F_0, F_1\}$$

(11)

设人体骨骼特征值为 F_s ,将 F_s 与 F_{D1} 相融合并输入到分类器中,分配对应权重 μ_1 、 μ_2 ,通过隶属度函数计算得到隶属度最大时,不同行为对应的标签:

$$\lg P(l_c | F_{D1}) = \mu_1 P_1(l_c | F_s) + \mu_2 P_2(l_c | F_{D1})$$

(12)

式中, P_1 、 P_2 为按照行为为检验的后验概率; l_c 为行为特征隶属度。通过上述公式即可计算视频中多种行为隶属度参数,根据隶属度和行为标签对比实现高效识别。

4 实验与分析

4.1 实验设置

为验证文中提出基于梯度特征与 3D 卷积的视频不同跨度行为识别应用的有效性,采用 HMDB51 数据集作为实验对象,该数据集来源于电影、公共数据库以及 YouTube 等网络视频库。数据库包含有 6849 段样本,分为 51 类,每类至少包

含有 101 段样本,可满足测试条件需求。实验在配备 Intel Core i7 处理器、NVIDIA GTX 1080 Ti 的 GPU、64 GB 内存的计算机上进行,操作系统为 Ubuntu 20.04,使用的深度学习框架为 TensorFlow 2.17.0,编程语言为 Python 3.13.1。

在数据集中选择两段视频,并以中间帧的前后 8 帧为测试序列,输入到实验平台中进行测试。为了便于观察和分析,以其中某一帧图像为例进行行为识别对比,如图 2 所示。



案例 1



案例 2

图 2 测试图像

实验相关参数如表 1 所示。

表 1 实验相关参数示意

参数	数值
像素	1280×720 P
帧率	30~100 fps
焦距	20~50 mm
码率	500 bps

为验证所提方法在处理不同跨度行为时的相似性和差异性识别能力,从 HMDB51 数据集中精心挑选了包含不同跨度行为的视频样本,确保所选样本覆盖了广泛的动作类别和场景变化。使用所提方法对预处理后的视频进行特征提取(如图 3 所示),引入相似度度量方法和差异性评估指标,用于量化不同行为特征之间的相似性和差异性。并与文献[9]的 2D 卷积神经网络对比,实验结果如图 4 所示。



图 3 实验过程

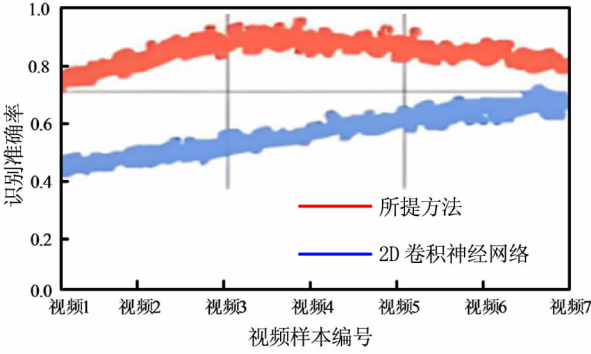


图 4 不同视频样本上的识别准确率

由图 4 分析可知,所提方法在处理不同跨度行为时展现出了出色的相似性和差异性识别能力。与 2D 卷积神经网络相比,该方法在 HMDB51 数据集上挑选的视频样本中,不仅准确识别了广泛的

动作类别和场景变化,而且在量化行为特征之间的相似性和差异性方面表现更优。这是由于所提方法能更有效地捕捉视频中的时序信息和动态特征,从而准确区分行为跨度大小不同的视频样本,同时保持对相似动作的敏感度,进一步验证了其在视频行为识别领域的有效性和先进性。

4.2 结果分析

与文献[1]的小样本视频行为识别算法、文献[3]的基于深度学习算法的行为识别研究进行对比分析,得到实验结果如图 5、图 6 所示。

从图 5 中可以看出,所提方法无论是基于哪种案例得到行为识别结果都是最为精准度,对画面中出现的所有行为都能实现全覆盖识别,精准度较高,并且没有出现误识别的现象;反观另外两种方法,在人体行为较为复杂的案例 1 中,识别精准度较低,针对行为较为复杂存在重叠情况时,不能很好地提取人体的行为特征,识别误差较大。相比于图 5 的测试结果,图 6 中包含的行为类别相对较少,另外两种方法在该环境下呈现的结果相对较好,但从整体识别精准度来看,所提方法精准度更高。综合对比可得,所提识别方法的综合实用性能更强,漏识、误识现象出现概率较低,性能表现优异。

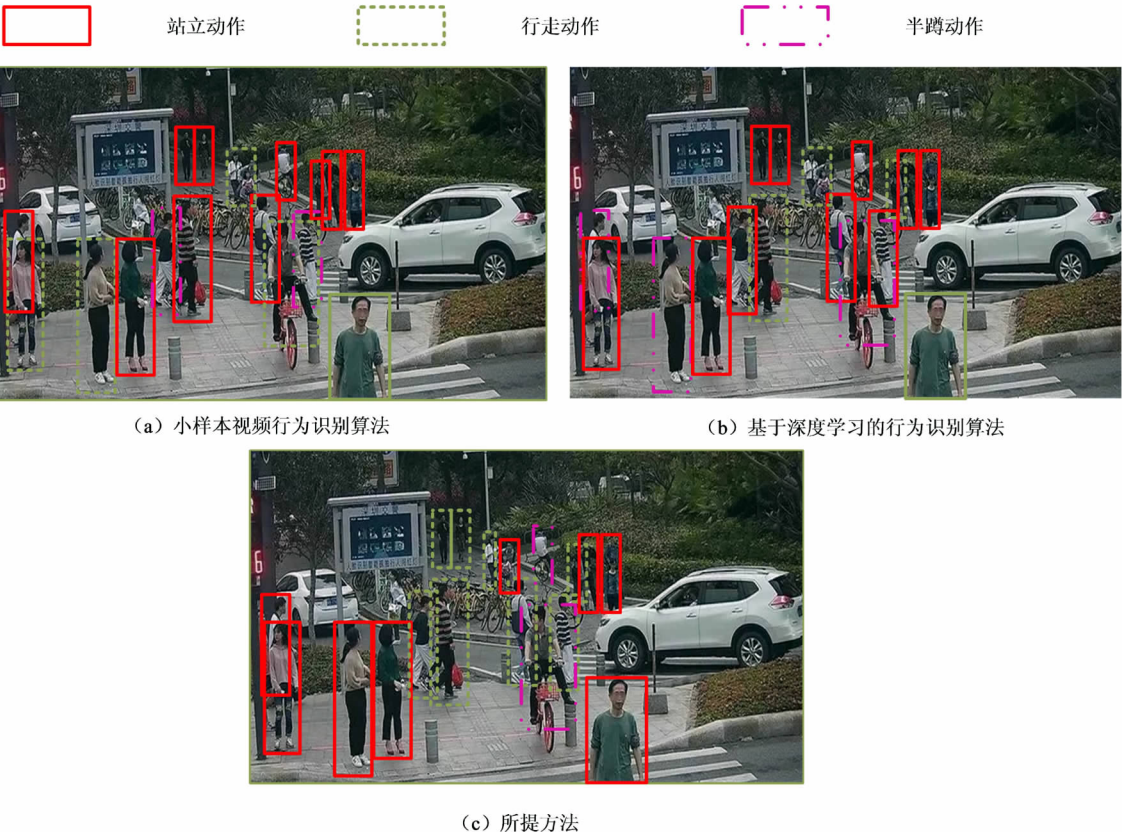


图 5 案例 1 行为识别结果对比

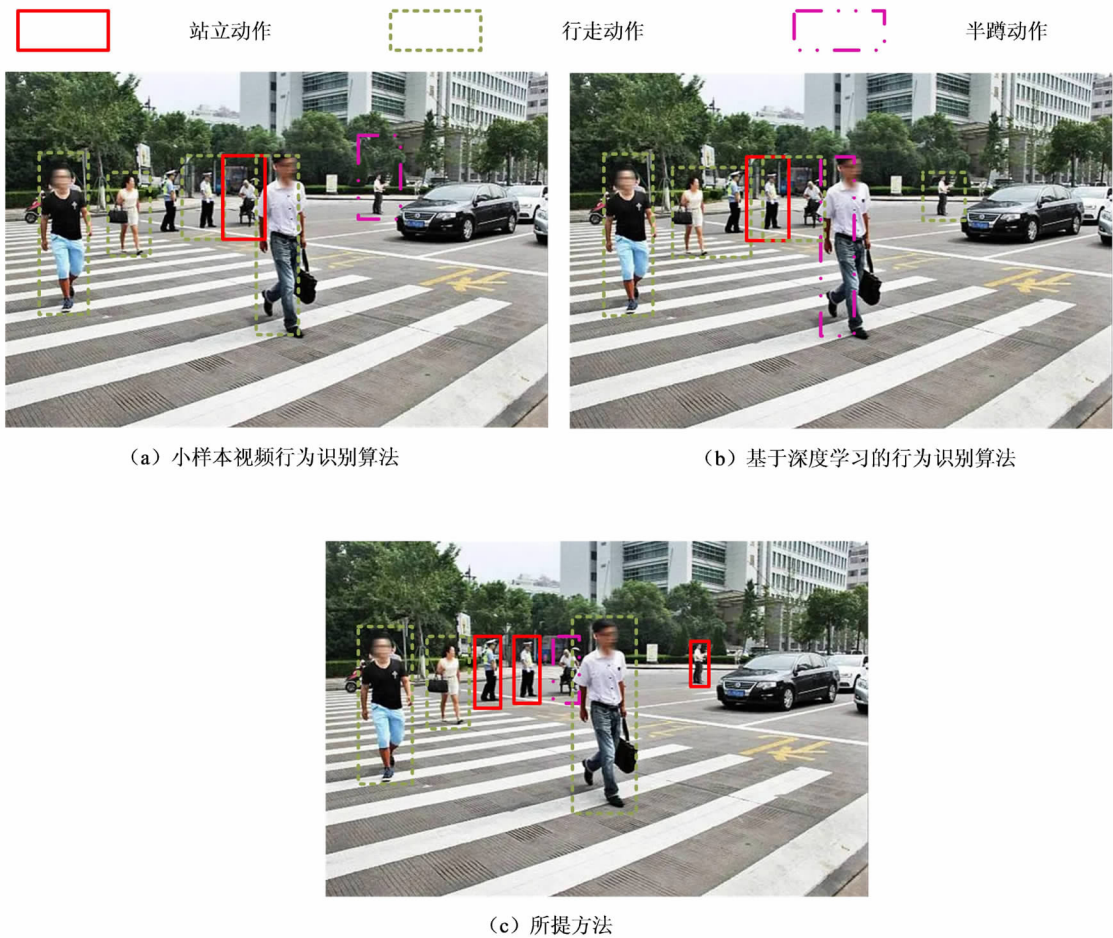


图 6 案例 2 行为识别结果对比

为进一步验证所提方法识别的有效性,通过混淆矩阵输出基于不同行为种类,结果如图 7 所示。

从图 7 中可以看出,所提方法在群体行为识别任务中的准确率是最高的,无论是基于哪种行为种类得到的识别准确率基本在 90% 以上,针对上肢推、下肢推等一些特征表现不明显的动作也能实现高效识别;反观另外两种方法的识别准确率表现较差,在行走、半蹲、伸展等有大幅度变化动作上的识别结果优异,但针对动作幅度较小的上肢推、下肢推、上肢拉以及下肢拉动作的识别效果很差。

为了更全面地评估所提方法的性能,除了识别准确率外,还引入了误识别率(False Positive Rate, FPR)、漏识别率(False Negative Rate, FNR)和推理时间等性能指标。实验结果如表 2 所示。

由表 2 分析可知,所提方法在识别准确率、误识别率、漏识别率以及推理时间等性能指标上均表

现出色。具体而言,所提方法的识别准确率达 95.2%,远超小样本视频行为识别算法和基于深度学习的行为识别算法,表明其在行为识别任务中具有更高的准确性。同时,所提方法的误识别率和漏识别率分别仅为 3.1% 和 5.4%,远低于其他两种方法,进一步证明了其在减少误判和漏判方面的优势。此外,所提方法的推理时间也较短,仅为 37 ms/帧,这有助于其在实际应用中实现更快的响应速度。综上所述,所提方法在行为识别性能上表现优异,具有较高的实际应用价值。

表 2 三种行为识别方法性能指标对比				
方法	识别准确率/%	误识别率/%	漏识别率/%	推理时间/(ms/帧)
小样本视频行为识别算法	82.4	12.4	18.3	45
基于深度学习的行为识别算法	87.9	10.5	12.8	52
所提方法	95.2	3.1	5.4	37

行为种类	行走	77.8	13.3	0.1	0.4	2.4	9.9	14.7	0.7
	站立	10.5	69.2	2.7	2.5	1.6	8.9	8.2	2
	上肢推动作	10.8	0.6	89.7	1.7	0.9	8.2	3.9	3.3
	下肢推动作	6.2	2.8	3.4	86.1	0.6	1.8	6.5	1
	上肢拉动作	3.3	3.7	0.8	3.8	84.4	1.5	1	2
	下肢拉动作	0.7	5.1	0.5	1.5	5.9	78.8	4.1	0.8
	半蹲动作	0.2	2.3	2.8	2.7	1.0	2	66.2	1.2
	伸展动作	0.4	3	0	1.3	3.2	2.9	1.4	90.4
行为种类									

(a)小样本视频行为识别算法矩阵输出结果

行为种类	行走	82.7	1	1	0.1	10.8	0.3	0.5	0.6
	站立	0.2	65.9	0.1	1	0.1	1.8	0.5	1.8
	上肢推动作	0.3	2	73.3	2	0.2	4.4	2	11.2
	下肢推动作	1.5	3	0.2	69.9	2	3.9	1	6.3
	上肢拉动作	2	5	11.2	3	68.7	1.5	1	2.2
	下肢拉动作	4	6	2	5	6	72.1	3	1.7
	半蹲动作	4	10	4	8	9	12.5	88.5	0.5
	伸展动作	3	6.1	8	10.9	3	3.5	3	77.1
行为种类									

(b)基于深度学习的行为识别方法的矩阵输出结果

行为种类	行走	90.6	0.9	0.3	1.1	0.8	0	2	2.4
	站立	0.6	91.4	1.3	0.1	1.1	0.1	0.3	2
	上肢推动作	0.7	0.8	92.8	0.5	0.9	0.2	1.7	1.5
	下肢推动作	1.3	1.5	0.9	95.4	0.6	1	1	1
	上肢拉动作	1.8	1.9	1.2	0.9	93.3	1.5	1	0.5
	下肢拉动作	2.2	0.2	0.5	1.5	0.5	92.2	0.2	0.4
	半蹲动作	2.2	1.6	1.9	0.2	1.0	2	93.7	0.1
	伸展动作	0.6	1.7	1.1	0.3	0.8	3	0.1	92.1
行为种类									

(c)所提方法的矩阵输出结果

图 7 三种方法矩阵输出对比

征值。将梯度特征值与区域特征值相融合,该步骤可以提高后续多种行为识别的效率,通过特征提取的方式预先判定行为所属分类。在此基础上,再对每个动态行为特征值都分配对应权重,将人体骨骼特征值以向量的形式输入,通过标签对比实现有效识别。实验结果表明,所提方法的行为识别准确率达 95.2%,较其他方法提升显著。同时,该方法误识别率和漏识别率分别仅为 3.1%和 5.4%,表明该方法有效减少了误判和漏判。此外,其推理时间仅为 37 ms/帧,确保了在实际应用中的快速响应,增强了其实用性。综上所述,该方法在视频不同跨度的行为识别任务中表现卓越,具备较高的应用价值和推广潜力。未来,将继续研究更高效的特征提取与融合技术,以期进一步提升行为识别的精度和稳定性。

参考文献

[1] 蒲瞻星,葛永新. 基于多特征融合的小样本视频行为识别算法[J]. 计算机学报,2023,46(3):594—608.

[2] FENG J, LUO H, FANG D. A progressive deep learning framework for fine-grained primate behavior recognition[J]. Applied Animal Behaviour Science, 2023, 269:106099.

[3] HU K, JIN J, ZHENG F, et al. Overview of behavior recognition based on deep learning[J]. Artificial Intelligence Review: An International Science and Engineering Journal, 2023,56(3):1833—1865.

[4] 刘云萍. 基于神经网络的视频监控图像异常行为识别方法[J]. 数字通信世界,2023(7):71—73.

[5] 高阳峰,雷涛,杜晓刚,等. 基于像素距离图和四维动态卷积网络的密集人群计数与定位方法[J]. 计算机应用,2024,44(7):2233—2242.

[6] 杨德贵,许道峰. 基于时频域特征融合的 IR-UWB 穿墙雷达人体行为识别方法[J]. 系统工程与电子技术,2024,46(3):849—858.

[7] 李光,刘丕亮,张雪松. 基于骨架平衡的 3D 人体异常行为识别方法仿真[J]. 计算机仿真,2024,41(2):492—495+521.

[8] 冉宪宇. 基于惯性传感器的人体行为识别系统设计[J]. 微型电脑应用,2023,39(1):152—154+158.

[9] 胡正平,王昕宇,董佳伟,等. 时空多尺度关联特征融合的二维卷积网络细粒度动作识别模型[J]. 高技术通讯,2024,34(6):590—601.

5 结 论

提出了一种基于梯度特征与 3D 卷积的视频不同跨度行为识别。以视频中不同时间点下的图像帧数为提取帧数,形成图像帧序列,并计算在不同时间点和不同区域中动态行为的方向和梯度特