

基于深度学习的多特征融合人体动作识别算法

徐海宁¹, 王彦坤^{2,3†}, 樊勇⁴, 文奴²

(1. 中国通信建设集团设计院有限公司数字化咨询中心, 河南 郑州 450052; 2. 深圳职业技术大学物联网研究院, 广东 深圳 518055; 3. 深圳大学建筑与城市规划学院智慧城市研究院, 广东 深圳 518061; 4. 深圳职业技术大学人工智能学院, 广东 深圳 518055)

摘要:针对传统特征提取方式计算量大, 识别率低等问题, 提出了一种基于深度学习的多特征融合人体动作识别算法。方法首先提取原始深度视频序列, 接着对每帧图像旋转特定角度补充更多视角下的动作信息, 提出自适应时间模型将深度视频序列划分为若干子动作。通过累积每帧图像之间能量变化较大的部分形成动态能量图, 累积能量变化较小的部分形成静态能量图, 统一称之为运动能量图。引入多通道卷积神经网络提取运动能量图中的动态特征和静态特征; 引入自适应矩估计调整神经网络训练中每个参数的学习率, 提高模型训练的效率和稳定性; 引入 L2 范数正则化, 减小模型复杂度并防止过拟合。在公开的数据集 UTD-MHAD、MSR Action3D 和 MSR DailyActivity3D 数据库上测试, 算法能够有效地提取动作的运动信息并取得较好的识别结果, 与主流方法比有一定的提升。

关键词: 自适应时间模型; 运动能量图; 多通道卷积神经网络; 自适应矩估计; L2 正则化

中图分类号: TP391.41

文献标识码: A

Multi-feature Fusion Algorithm for Human Action Recognition Based on Deep Learning

XU Haining¹, WANG Yankun^{2,3†}, FAN Yong⁴, WEN Nu²

(1. Digital Consultation Centre, China Communications Construction Group Design Institute Co., Ltd., Zhengzhou, Henan 450052, China; 2. Shenzhen Polytechnic University & Institute of Internet of Things, Shenzhen, Guangdong 518055, China; 3. College of Architecture and Urban Planning, Smart Cities Research Institute, Shenzhen University, Shenzhen, Guangdong 518061, China; 4. Shenzhen Polytechnic University & Institute of Artificial Intelligence, Shenzhen, Guangdong 518055, China)

Abstract: To address the issues of high computational complexity and low recognition rates associated with traditional feature extraction methods, a deep learning-based multi-feature fusion algorithm for human action recognition is proposed. The method begins by extracting the original depth video sequence, followed by rotating each frame image by a specific angle to supplement action information from multiple viewpoints. An adaptive temporal model is proposed to divide the depth video sequence into several sub-actions. Dynamic energy maps are formed by accumulating portions with relatively larger energy changes between each frame of image, while static energy maps are formed by accumulating portions with relatively little energy changes. Both are collectively referred to as motion energy maps. A multi-channel convolutional neural network is introduced to extract dynamic and static features from the motion energy maps. Adaptive Moment Estimation is used to adjust the learning rate of each parameter during neural network training, improving the efficiency and stability of model training. L2 norm regularization is introduced to reduce model complexity and prevent overfitting. Tested on the public datasets of

收稿日期: 2024-03-12

基金项目: 国家自然科学基金青年基金资助项目(42001389, 42201496); 广东省普通高校重点领域专项(2022ZDZX3071); 自然资源部城市国土资源监测与仿真重点实验室开放基金资助课题(KF-2022-07-024); 深圳职业技术学院博士后出站后期资助项目(6021271017K, 6023271011K); 深圳职业技术学院项目(6022312062K, 6023310002K)

作者简介: 徐海宁(1989—), 男, 河南浚县人, 硕士, 高级工程师, 研究方向: 模式识别和图像处理。

† 通信联系人, E-mail: yankun.wang@whu.edu.cn

UTD-MHAD, MSR Action3D and MSR DailyActivity3D, the algorithm can effectively extract the motion information of actions and achieve good recognition results, which is a certain improvement compared with the mainstream methods.

Key words: adaptive temporal model; motion energy map; multi-channel convolutional neural network; adaptive moment estimation; L2 regularization

随着 AI 技术的发展,人体动作识别广泛应用在视频监控、安全驾驶和人机交互等方面。早期,大多数工作都专注于 RGB 视频序列^[1-3]。传统的基于 RGB 视频的人体动作识别有天然劣势,首先,RGB 图像对光照和背景非常敏感,变化的光线条件和背景会影响到图像中的像素值。其次,RGB 图像是二维的,即使在连续帧中捕获到运动,也无法直接获取关于深度和三维结构的信息。最后,处理大规模 RGB 图像数据集通常需要手动设计复杂的算法。

与传统的 RGB 摄像机不同,深度摄像机能够捕获每个像素的距离信息,从而提供场景的三维深度信息。本文方法基于 Kinect 深度摄像机,Kinect 内置时间飞行深度传感器,能够准确捕捉场景中物体的三维深度信息,精确地识别人体姿势和动作;另外,Kinect 还配备了 RGB 摄像头,用于捕捉场景的彩色图像,在获取深度信息的同时,还能够获取场景的视觉信息,从而实现更丰富的交互体验。在低光照或光照条件快速变化的情况下,Kinect 仍然能够准确地捕捉物体的轮廓和深度信息,而不会出现色彩失真的问题。Kinect 的环境适应性使得其在各种复杂场景中都具有广泛的应用潜力。目前,基于深度摄像机的研究主要集中在深度图像^[4-12]、骨骼数据^[13-15]以及多种数据融合^[16-18]等方面。

1 相关工作

对于骨骼数据的人体动作识别,文献[13]提出一种辅助特征细化头(FR-Head)方法,通过时空数据解耦和特征细化两部分来获得骨骼数据的表示,在特征空间中动态发现和校准模糊样本。文献[14]提出 MAMP 框架预测人体关节的相应时序运动,通过运动信息充当经验性的语义,指导掩码过程,促使模型更好地关注语义丰富的时序区域。对于经典深度图像的人体动作识别,文献[4]将视频序列分成若干子动作,对子动作和整体动作做帧间差分得到深度运动图(DMM),提取 DMM 的多

层 HOG 特征进行动作识别。文献[6]将深度视频序列通过旋转方式获取多角度视频序列,通过能量累积获取 DMM,通过提取深度运动图的 HOG 特征进行动作识别。

近些年,人体动作识别相关研究逐渐转向卷积神经网络(CNN)。文献[5]提取空间特征和时间特征,使用三维卷积神经网络和激活层对动作进行识别。文献[7]将双流卷积神经网络扩展为三维卷积神经网络,以更好地捕捉显著的运动信息。文献[8]提出两种视频超分辨率方法,用于生成高分辨率视频,然后将这些视频输入到空间流和时间流中,用 CNN 进行动作类别分类。文献[9]从概率的角度分析了 CNN 中的时空融合。文献[10]提出一种通用的特征级时序金字塔网络,用于对速度变化的动作进行建模。文献[11]提出一种随机均值缩放正则化方法,以解决 CNN 中的过拟合问题。

经典方法依赖手工设计的特征,往往需要研究者的经验知识,并且在处理大数据集时,识别速度和识别率难以保证。传统 CNN 中卷积层在进行特征提取时,每个神经元只能看到输入数据中局部区域的信息,导致局部感受野的限制。CNN 中的参数共享机制使得模型具有一定的平移不变性,但在处理一些复杂的任务时,可能需要更灵活的参数设置以适应不同的数据特征。另外,传统方法在训练深层 CNN 时,可能会遇到梯度消失或梯度爆炸问题,导致模型的训练变得困难,甚至无法收敛。

本文提出了一种多通道卷积神经网络人体动作识别算法。首先通过旋转三维坐标系,得到多视角深度图像序列;提出基于运动能量的自适应时间模型,将多视角深度图像序列分割成多个子动作,提取完整动作和子动作的动态能量图(DEM)和静态能量图(SEM),称之为运动能量图(MEM)。受多通道卷积神经网络(MC-CNN)在 RGB 图像分类和目标检测方面的启发,融合多视角 MEM 作为 MC-CNN 的输入,通过多个卷积层和多个全连接层对动作进行分类,最后引入自适应矩估计和 L2 范数正则化对神经网络进行参数优化。本文算法框架如图 1 所示。

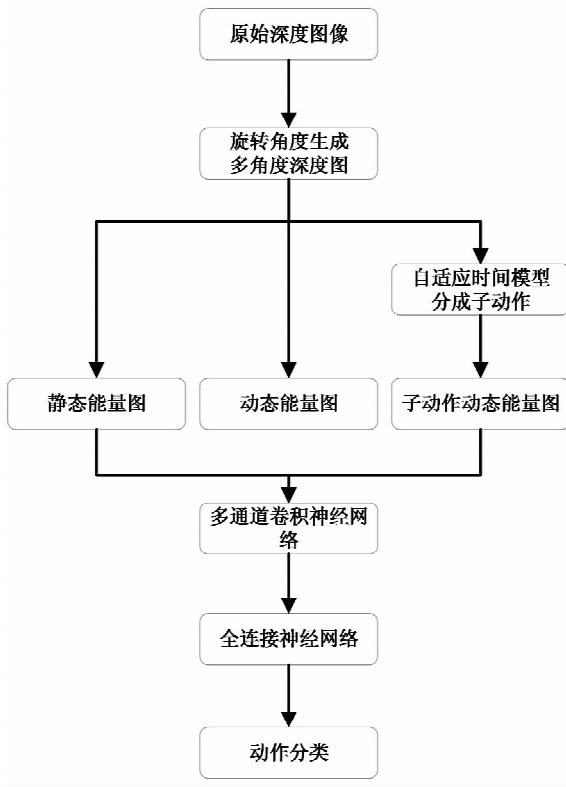
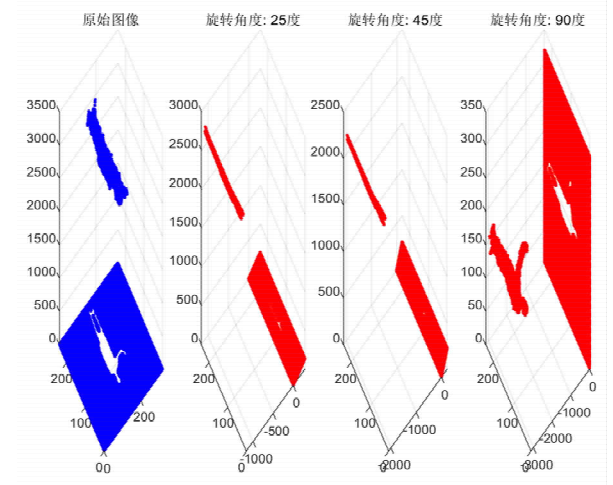


图 1 本文算法框架

其中 T_x 和 T_y 表示转置矩阵, $[X_1, Y_1, Z_1]$ 和 $[X_2, Y_2, Z_2]$ 为新视角下的坐标。通过该方式可以得到视频序列的多视角深度视频序列。图 2 为 UTD-MHAD 数据集 swipt_left 动作绕 Y 轴分别旋转 25° , 45° 和 90° 后的深度图像示意图。

图 2 原始深度图像绕 Y 轴旋转 25° , 45° , 90° 后得到的多视角深度图像(动作 swipt_left)

2 动作表示

2.1 运动能量图(MEM)

针对数据集样本数量相对较小,网络模型参数难以充分优化,且在训练过程中容易发生过拟合的现象,本文方法通过旋转视角增加样本数据量。假设深度视频序列 $A = \{I^i\}$, $I^i = \{X, Y, Z\}$, 其中 A 表示视频序列, I^i 表示第 i 帧图像, X 轴正方向为屏幕右方, Y 轴正方向为屏幕上方, Z 轴正方向为垂直屏幕向外, $\{X, Y, Z\}$ 为深度图像 I^i 的三维信息,将深度图像绕 X 轴旋转 α 度,可以得到上下偏离 X 轴 α 角度的深度图像;将深度图像绕 Y 轴旋转 β 角度,可以得到左右偏离 Y 轴 β 角度的深度图像,公式如下:

$$\begin{cases} [X_1, Y_1, Z_1]^T = T_x [X, Y, Z]^T \\ T_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & \sin \alpha \\ 0 & -\sin \alpha & \cos \alpha \end{bmatrix} \end{cases} \quad (1)$$

$$\begin{cases} [X_2, Y_2, Z_2]^T = T_y [X, Y, Z]^T \\ T_y = \begin{bmatrix} \cos \beta & 0 & -\sin \beta \\ 0 & 1 & 0 \\ \sin \beta & 0 & \cos \beta \end{bmatrix} \end{cases} \quad (2)$$

在得到多视角深度视频序列后,计算相邻两帧之间的深度数据差异,通过累积相邻帧之间能量变化较大部分形成 DEM,通过累积相邻帧之间能量变化较小的部分形成 SEM,具体如下:

$$DEM_v^i = \sum_{i=2}^n |I_v^i - I_v^{i-1}| \geq \delta \quad (3)$$

$$SEM_v^i = \sum_{i=2}^n |I_v^i - I_v^{i-1}| \leq \epsilon \quad (4)$$

其中, v 表示某视角, i 代表帧数, δ 和 ϵ 为相应阈值, DEM_v^i 为 v 视角下深度视频序列前 i 帧的动态能量图, SEM_v^i 为 v 视角下深度视频序列前 i 帧的静态能量图。动态能量图反映视频序列中动作变化较大的部分,静态能量图则反应的是视频序列中整体呈现相对静止的部分。

由于动作之间的相似性,如 pick up throw 动作后半部分与 right arm throw 整体动作相似性极大,如果对整体动作提取特征势必引起动作混淆。本文提出基于运动能量的自适应时间模型,具体如下:

$$DEM_v^i / DEM_v^n = K / L \quad (5)$$

其中, DEM_v^n 为 v 视角下整个深度视频的动态能量图, L 为子动作个数, K 为小于 L 的自然数,模型均匀地将标准化运动能量轴细分为若干时间段,通过帧索引将动作划分为相关子动作,如图 3 所示。

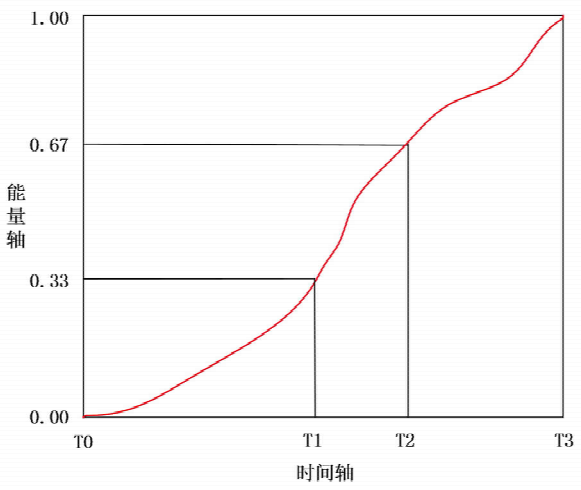


图 3 基于能量的自适应时间模型

在子动作划分的同时可以得到子动作 DEM, 子动作 DEM 能够反映相似动作前后间的差别, 对最终识别人体动作有着较大的影响。通过上述方法, 可以得到动作某个视角下的 4 个 DEM 和 1 个 SEM, 统一称之为 MEM, 如图 4 所示。

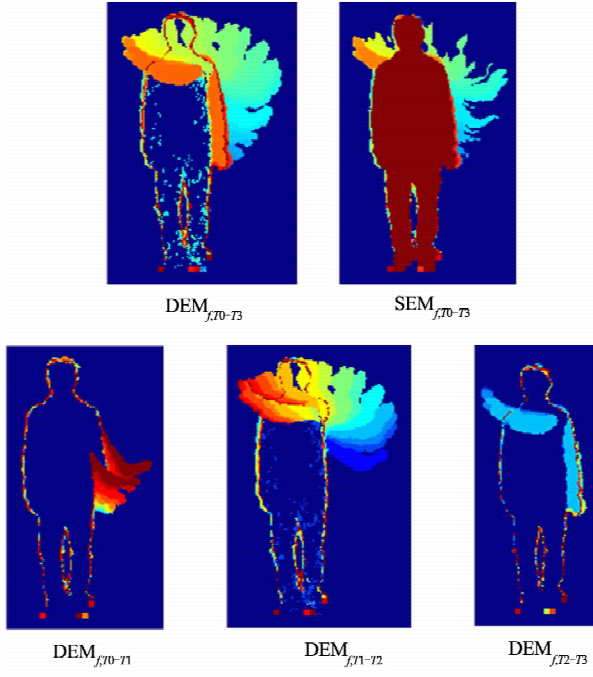


图 4 swipt_left 动作原始深度视频序列得到的运动能量图

2.2 多通道卷积神经网络 (MC-CNN)

多通道卷积神经网络 (MC-CNN) 的优势在于其能够在图像处理任务中更全面地捕获和学习多层次的特征表示。如本文多通道图像数据, 其中每个通道代表同一动作不同视角下的动态特征和静态特征。

令通道数为 C , 对每个通道做卷积运算, 卷积操作的数学表达式为:

$$Z_{i,j,k} = \sum_{m=0}^{k_w-1} \sum_{n=0}^{k_h-1} \text{MEM}_{i+m,j+n,k} \times K_{m,n,k} \quad (6)$$

其中, k 表示某个通道, i, j 为坐标, $Z_{i,j,k}$ 表示单通道卷积操作的结果, $\text{MEM}_{i+m,j+n,k}$ 是输入运动能量图相应坐标的像素值, $K_{m,n,k}$ 是卷积核的参数, K_w, K_h 为卷积核的维度。

对于每个坐标元素, 将所有通道的单通道卷积操作结果进行加权求和, 得到多通道卷积结果:

$$Y_{i,j} = \sum_{k=1}^C W_k \times Z_{i,j,k} \quad (7)$$

其中, $Y_{i,j}$ 是多通道卷积输出的结果, W_k 为 k 通道的权重参数, 在训练过程中, 通过反向传播算法更新权重, 可使训练数据上的损失函数最小化。根据梯度下降规则, 令 η 为学习率, L 为网络的损失函数, 权重调整过程如下:

$$W_k \leftarrow W_k - \eta \frac{\partial L}{\partial W_k} \quad (8)$$

通过权重的学习, 对多通道赋予不同的权重, 加权的结果 $Y_{i,j}$ 作用于 ReLU 激活函数, 产生最终输出 $\text{Output}_{i,j}$, 公式如下:

$$\begin{cases} \text{Output}_{i,j} = \text{ReLU}(Y_{i,j}) \\ \text{ReLU}(x) = \max(0, x) \end{cases} \quad (9)$$

ReLU 激活函数的非零梯度有助于缓解梯度消失问题, 使得网络更易于训练。

在卷积操作后, 引入全连接神经网络 (FCNN) 对动作进行分类, $T = \{T_1, T_2, T_3, \dots, T_n\}$ 表示样本卷积运算后得到的特征向量, 其中 n 表示样本数量, 特征向量为 FCNN 的输入。FCNN 可以包含多个隐藏层, 对于第 i 层隐藏层, 计算公式如下:

$$Z_i^{(l)} = \sum_{j=1}^{N^{(l-1)}} (\omega_{ij}^{(l)} \times a_j^{(l-1)}) + b_i^{(l)} \quad (10)$$

其中, $N^{(l-1)}$ 为上一层的神经元数量, $\omega_{ij}^{(l)}$ 是连接第 $(l-1)$ 层第 j 个神经元到第 l 层第 i 个神经元的权重, $b_i^{(l)}$ 是第 l 层第 i 个神经元偏差, 通过 ReLU 激活函数对其进行非线性变换, 最后一个隐藏层的输出作为输入传递给输出层。对于输出层, 计算方式与隐藏层相似, 对于第 i 个神经元, 计算加权输入:

$$Z_i^{(L)} = \sum_{j=1}^{N^{(L-1)}} (\omega_{ij}^{(L)} \times a_j^{(L-1)}) + b_i^{(L)} \quad (11)$$

其中, L 是输出的索引, $N^{(L-1)}$ 是倒数第二层的神经元数量, 最后, 通过输出层的激活函数 Softmax 进行非线性变换:

$$\begin{cases} a_i^{(L)} = \text{Softmax}(z_i^{(L)}) \\ \text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^N e^{z_j}} \end{cases} \quad (12)$$

其中, N 是类别数, z_i 是原始输出的第 i 个元素, Softmax 函数对网络的原始输出进行归一化并将其转化为一个概率分布。

2.3 网络优化

在 MC-CNN 过程中引入 Adam 算法,在训练过程中通过梯度一阶矩估计(动量项)和二阶矩估计(RMSprop 项)自适应调整每个参数的学习率。

首先,与传统的固定学习率相比,Adam 算法更加灵活,它可以在训练过程中根据参数的梯度状况动态调整学习率,从而更好地适应不同参数的更新速度。其次,由于 Adam 基于梯度的一阶矩和二阶矩进行调整,能够更好地处理稀疏梯度,从而避免出现训练过程中的学习率过大或过小的情况,提高了模型的稳定性和鲁棒性。最后,Adam 算法通常能够加速模型的收敛速度,通过合理地调整学习率,它可以有效地平衡参数更新的速度,避免出现梯度下降过慢或过快的情况,从而使得模型能够更快地收敛到局部最优解。具体过程如下:

令初始学习率为 α ,通过公式:

$$\alpha_t = \frac{\alpha}{\sqrt{t}} \quad (13)$$

得到自适应学习率 α_t ,其中 t 为迭代步数。网络引入动量项,动量项可以被视为梯度的滑动平均,有助于平稳化参数更新过程,动量项的计算公式为:

$$m_t = \beta_1 \times m_{t-1} + (1 - \beta_1) \times g_t \quad (14)$$

其中, β_1 为指数衰减率, g_t 为当前步的梯度。网络引入 RMSprop 项衡量了梯度平方的滑动平均,用于处理梯度的方差。在训练中,高方差的维度得到较小的学习率,而低方差的维度得到较大的学习率,有助于更有效地进行参数更新,计算公式为:

$$v_t = \beta_2 \times v_{t-1} + (1 - \beta_2) \times g_t^2 \quad (15)$$

其中, β_2 为指数衰减率。

在 MC-CNN 过程中引入 L2 正则化控制模型权重大小,通过向损失函数添加一个正则化项,以防止过拟合。公式如下:

$$\text{损失函数} = \text{原始损失} + \frac{\lambda}{2} (W_k^2 + W_{fc}^2) \quad (16)$$

其中, λ 是正则化参数, W_k^2 表示卷积核权重矩阵的 Frobenius 范数的平方, W_{fc}^2 表示全连接层权

重矩阵的 Frobenius 范数的平方。

3 实验与结果分析

3.1 实验参数设定

多通道卷积神经网络输入层选用原始视角、绕 X 轴 90° 和绕 Y 轴 90° 三个视角,采用 3 个子动作来补充动作信息,每个视角下 5 个 MEM,通道数为 15。设置三个卷积层,每个卷积层后跟批量归一化层、ReLU 激活层和最大池化层。设置三个全连接层,全连接层后跟 ReLU 激活层。输出层采用全连接层、softmax 激活函数和分类层的组合方式。采用 Adam 优化器,分段常数学习率调整,初始学习率为 0.001,添加 L2 正则化项,正则化参数为 0.001。

3.2 UTD-MHAD 数据集

UTD-MHAD 数据集包含 27 个动作,分别为 1. swipe_left, 2. swipe_right, 3. wave, 4. clap, 5. throw, 6. arm_cross, 7. basketball_shoot, 8. draw_x, 9. draw_circle_CW, 10. draw_circle_CCW, 11. draw_triangle, 12. bow-ling, 13. boxing, 14. baseball_swing, 15. tennis_swing, 16. arm_curl, 17. tennis_serve, 18. push, 19. knock, 20. catch, 21. pickup_throw, 22. jog, 23. walk, 24. sit2stand, 25. stand2sit, 26. lunge 和 27. squat。数据集由 8 个受试者重复动作 4 次。为了同主流方法进行对比,实验设置和文献[16-18]保持一致,对整个数据集的 27 个动作统一进行实验,选用每人第 1、3 次动作作为训练样本,第 2、4 次动作作为测试样本。最终本数据集测试识别率为 98.14%,混淆矩阵如图 5 所示。

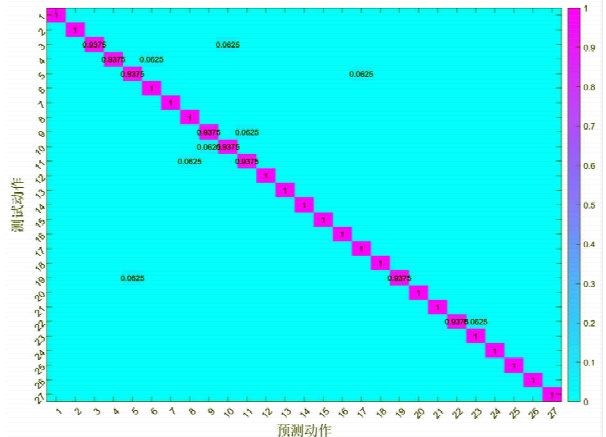


图 5 UTD-MHAD 数据集实验结果混淆矩阵

通过混淆矩阵可以看出本文方法在 19 个动作上都取得了 100% 的识别率,对于一些相似性较大

的动作,依然存在误识别。表 1 列出了本文方法与主流方法在识别率上的对比,本文方法在仅使用 DEM 的卷积神经网络时,识别率为 94.47%;在使用多视角 DEM 和 SEM 的多通道卷积神经网络时,识别率为 95.57%;在融合多视角子动作 MEM 时,识别率提升至 98.14%,本文方法识别率明显高于已知主流方法的识别率。

表 1 本文方法与主流方法识别率对比	
方法	识别率 (%)
光流与深度运动图融合 ^[18]	94.86
加权乘积融合法 ^[16]	96.49
SFI+SI ^[17]	97.20
本文方法(DEM)	94.47
本文方法(多视角 DEM+SEM)	95.57
本文方法(多视角 MEM)	98.14

3.3 MSR Action3D 数据集

MSR Action3D 数据集包含 20 种动作,动作名称如表 2 所示。数据集由 10 个受试者重复动作 2 到 3 次。为了证明本文方法的有效性,按照主流方法^[4,6]的分组方式,将动作分为三组,将 20 个动作中较为相似的动作分到一组,分组情况如表 2 所示。

表 3 本文方法与主流方法识别率对比													
方法	实验一				实验二				交叉测试				平均
	AS1	AS2	AS3	平均	AS1	AS2	AS3	平均	AS1	AS2	AS3	平均	
MHPC ^[6]	97.00	98.2	96.5	97.2	100.00	100.00	99.10	99.70	91.70	95.10	93.90	95.79	96.80
TDM-SPHOG ^[4]	96.70	98.00	98.00	97.60	98.60	98.70	100.00	99.10	95.30	92.90	96.40	94.90	97.18
本文方法(DEM)	98.00	96.73	98.00	97.58	98.65	100.00	100.00	99.55	96.23	92.92	98.21	95.79	97.64
本文方法(多视角 DEM+SEM)	98.00	97.39	98.00	97.80	98.65	100.00	100.00	99.55	96.23	92.92	98.21	95.79	97.71
本文方法(多视角+子动作)	100.00	98.00	98.00	98.67	100.00	100.00	100.00	100.00	97.17	94.69	99.11	96.99	98.55

为了进一步证明本文方法的有效性,实验对数据集 20 个动作按照交叉测试条件进行实验四,表 4 列出了本文算法与主流算法在识别率上的对比,在融合多视角子动作 MEM 时,识别率达到了 96.36%。

表 4 本文方法与主流方法识别率对比	
方法	识别率 (%)
SNV ^[12]	93.45
MHPC ^[6]	93.82
本文方法(DEM)	93.45
本文方法(多视角 DEM+SEM)	94.91
本文方法(多视角 MEM)	96.36

表 2 MSR Action3D 数据集的分组方式		
Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal Wave (2)	High Wave (1)	High Throw (6)
Hammer (3)	Hand Catch (4)	Forward Kick (14)
Forward Punch (5)	Draw X (7)	Side Kick (15)
High Throw (6)	Draw Tick (8)	Jogging (16)
Hand Clap (10)	Draw Circle (9)	Tennis Swing (17)
Bend (13)	Hands Wave (11)	Tennis Serve (18)
Tennis Serve (18)	Forward Kick (14)	Golf Swing (19)
Pickup Throw (20)	Side Boxing (12)	Pickup Throw (20)

每一组动作分别进行 3 种实验:实验一中每个人每个动作的第 1 次作为训练样本,第 2 和第 3 次作为测试样本;实验二中,每个人每个动作的第 1 和第 2 次是训练样本,第 3 次作为测试样本;在交叉测试组,人物 1、3、5、7、9 的动作数据当作训练样本,2、4、6、8、10 的动作数据作为测试样本。由于不同人之间,即使是同一动作也可能有不同的速度和幅度,所以交叉测试相对于实验一和实验二,更具有挑战性。表 3 总结了本文方法与主流方法在 MSR Action3D 数据集上的实验结果。可以看出本文方法在几乎所有的测试组均得到了最高识别率。

3.4 MSR DailyActivity3D 数据集

MSR DailyActivity3D 数据集包含 16 种动作:1. drink, 2. eat, 3. read book, 4. call cell-phone, 5. write, 6. use laptop, 7. vacuum clean, 8. cheer up, 9. sit still, 10. toss paper, 11. play game, 12. lie down, 13. walk, 14. play guitar, 15. stand up 和 16. sit down。数据集由 10 个受试者分别在站立和坐姿下进行动作两次。与其他两个数据集相比,该数据集中大多数动作涉及人与物的互动。为了同主流方法进行对比,实验设置和文献^[12,16]保持一致,表 5 列出了本文算法与主流算法在识别率上的对比。

表 5 本文方法与主流方法识别率对比	
方法	识别率 (%)
SNV ^[12]	86.25
加权乘积融合法 ^[16]	86.88
本文方法 (DEM)	85.63
本文方法 (多视角 DEM+SEM)	87.50
本文方法 (多视角 MEM)	90.63

本文方法在融合多视角子动作 MEM 时,识别率提升至 90.63%。由于该数据集中动作均为日常生活,大多数动作涉及人与物体的相互作用,导致识别率较前两个数据集较低。

4 结论

本文方法融合多视角动态特征和静态特征,通过多通道卷积神经网络实现端到端的人体动作识别。通过旋转坐标系获取多视角深度视频序列,在多视角深度视频序列基础上提取 MEM,通过多通道卷积神经网络得到最终动作分类结果。在三个公开数据集上测试,并与主流方法进行了比较。实验结果表明,本文方法在公开数据集上明显优于主流方法。未来可以融合骨骼数据进行动作识别,利用其数据量小、对遮挡不敏感的特性,通过多通道卷积神经网络或结合图卷积神经网络获取多模态特征,从而进一步提高人体动作识别的速度和准确率。

参考文献

[1] ZHAO H, ZHAO J S, LAI Y K, et al. High-fidelity human avatars from a single RGB camera[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022;15883—15892.

[2] HU T, SARKAR K, LIU L J, et al. EgoRenderer: rendering human avatars from egocentric camera images [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021; 14508—14518.

[3] WEN Y L, PAN H, YANG L, et al. Hierarchical temporal transformer for 3D hand pose estimation and action recognition from dgocentric RGB videos[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023; 21243—21253.

[4] 徐海宁,陈恩庆,梁成武. 三维动作识别时空特征提取方法

[J]. 计算机应用, 2016, 36(2);568—573+579.

[5] SANCHEZ A, LOPEZ S, FUENTES D, et al. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information[J]. Applied Intelligence, Multimedia Tools and Applications, 2022(81); 24119—24143.

[6] 刘婷婷,李玉鹏,张良. 多视角深度运动图的人体行为识别[J]. 中国图象图形学报,2019,24(3): 400—409.

[7] ZONG M, WANG R, Chen X, et al. Motion saliency based multi-stream multiplier resnets for action recognition[J]. Image and Vision Computing, 2021;107;119—135.

[8] ZHANG H C, LIU D, XIONG Z W. Two-stream action recognition oriented video super resolution [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019; 8798—8807.

[9] ZHOU Y Z, SUN X Y, LUO C, et al. Spatiotemporal fusion in 3D CNNs: a probabilistic view[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2020; 9826—9835.

[10] YANG C Y, XU Y H, SHI J P, et al. Temporal pyramid network for action recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2020;588—597.

[11] KIN J, Cha S, WEE D, et al. Regularization on spatio-temporally smoothed feature for action recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2020; 12100—12109.

[12] YANG X D, TIAN Y L. Super normal vector for human activity recognition with depth cameras[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(5); 1028—1039.

[13] XIANG W M, LI C, ZHOU Y X, et al. Learning discriminative representations for skeleton based action recognition [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023; 10608—10617.

[14] MAO Y Y, DENG J J, ZHOU W G, et al. Masked motion predictors are strong 3D action representation learners[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023;10181—10191.

[15] LIU Y N, ZHANG H, LI Y Q, et al. Skeleton-based human action recognition via large-kernel attention graph convolutional network[J]. IEEE Transactions on Visualization and Computer Graphics, 2023; 2575—2585.

[16] 李元祥,谢林柏. 结合 RGB-D 视频和卷积神经网络的行为识别算法[J]. 计算机与数字工程, 2020, 48 (12); 3052—3058.

[17] AHMAD Z,KHAN N. Human action recognition using deep multilevel multimodal (M2) fusion of depth and inertial sensors[J]. IEEE Sensors Journal,2020,20(3);1445—1455.

[18] 马亚彤,王松,刘英芳. 融合多模态数据的人体动作识别方法研究[J]. 计算机工程. 2022,48(9);180—188.