Computer Science and Application 计算机科学与应用, 2026, 16(3), 703-708
Published Online March 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.163096

Hans汉斯

基于LSTM多特征融合人体动作识别算法

隋龙飞，薛欢欢

嘉兴职业技术学院互联网学院，浙江 嘉兴

收稿日期：2026年2月11日；录用日期：2026年3月9日；发布日期：2026年3月18日

摘　要

为解决现有人体动作识别方法依赖单一模态数据、复杂场景下识别精度不足的问题，提出一种基于LSTM的多特征融合人体动作识别算法。首先，通过Kinect深度相机获取人体20个关键关节点的3D坐标信息，提取关节点的位置、运动速度及加速度等多维度特征；其次，通过关键帧定位与降采样处理优化特征数据，降低计算复杂度；最后，将整合后的多模态特征输入LSTM模型，利用其时空建模能力实现动作分类。为验证算法性能，在Kinetics-Skeleton、NTU-RGB (X-Sub)及NTU-RGB + D (X-View)三个公开数据集上进行实验，结果表明：算法对前20类动作的识别准确率均达到98%以上，其中NTU-RGB + D (X-View)数据集上部分动作类别准确率超99%；混淆矩阵分析显示模型分类一致性良好，具备较强的抗干扰能力与场景适应性。该算法通过多特征融合策略充分挖掘人体动作的时空关联信息，为复杂环境下的高精度动作识别提供了有效解决方案。

关键词

人体动作识别，LSTM，多特征融合，关键帧定位，骨架关节点，多模态数据

Human Action Recognition Algorithm Based on LSTM with Multi-Feature Fusion

Longfei Sui, Huanhuan Xue

School of Internet, Jiaxing Vocational and Technical College, Jiaxing Zhejiang

Received: February 11, 2026; accepted: March 9, 2026; published: March 18, 2026

Abstract

To address the issues of existing human action recognition methods relying on single-modal data and insufficient recognition accuracy in complex scenarios, a human action recognition algorithm based on LSTM with multi-feature fusion is proposed. Firstly, 3D coordinate information of 20 key human joint points is acquired using a Kinect depth camera, and multi-dimensional features such

文章引用: 隋龙飞, 薛欢欢. 基于 LSTM 多特征融合人体动作识别算法[J]. 计算机科学与应用, 2026, 16(3): 703-708.
DOI: 10.12677/csa.2026.163096

as the position, motion velocity, and acceleration of the joint points are extracted. Secondly, key frame localization and downsampling are applied to optimize the feature data and reduce computational complexity. Finally, the integrated multi-modal features are input into the LSTM model, which leverages its spatiotemporal modeling capability to achieve action classification. To verify the algorithm's performance, experiments are conducted on three public datasets: Kinetics-Skeleton, NTU-RGB (X-Sub), and NTU-RGB + D (X-View). The results demonstrate that the algorithm achieves an accuracy of over 98% for the top 20 action categories on all three datasets, with some categories exceeding 99% accuracy on the NTU-RGB + D (X-View) dataset. Confusion matrix analysis shows that the model exhibits good classification consistency, along with strong anti-interference ability and scene adaptability. By fully exploring the spatiotemporal correlation information of human actions through a multi-feature fusion strategy, the algorithm provides an effective solution for high-precision action recognition in complex environments.

Keywords

Human Action Recognition, LSTM, Multi-Feature Fusion, Key Frame Localization, Skeleton Joint Points, Multi-Modal Data

1. 引言

人体动作识别的核心目标在于对人体的运动状态、肢体姿态及行为模式进行精准解析与分类，最终实现对人体动作的自动化辨识与深度语义理解[1] [2]。在该领域的现有研究中，各类方法均存在一定的局限性，尚未能很好地兼顾识别精度、环境适应性与实时性能。文献[3]提出一种基于 Koopman 理论的非衰减稳定特征值归一化方法，通过参数化高阶池化技术将非线性动力学系统转化为线性形式，借助动力学矩阵对人体动作的动态特征进行有效捕捉，进而提升识别准确率。但该方法在复杂场景下的适配能力较弱，尤其当面临多样化、高复杂度的动作识别任务时，其性能易出现明显衰减。文献[4]通过构建 FP-NET 网络训练人体图像数据集，并通过加入回归模块和特征融合模块，提升了正面姿态估计的准确性。FP-NET 能够从任意角度的人体图像中有效提取正面姿态，实现对人体动作图像的准确识别。但在人体被其他物体严重遮挡时，FP-NET 可能无法准确估计出被遮挡部分的关键点位置。文献[5]利用 Transformer 网络进行时序建模，在单模态和跨模态下以自监督方式区分实例，同时引入雷达组合图来增强数据密度，解决了雷达数据稀疏性问题，实现对人体动作的有效识别。文献[6]则采用热释电红外传感器作为感知器件，将人体散发的红外辐射信号转化为可处理的电信号，通过提取信号中与人体动作对应的特征参数，与预设动作模式库进行比对匹配，从而完成动作类型判别。但该类传感器的抗干扰能力欠佳，当人体与周围环境温度趋于一致时，难以有效捕获红外辐射信号，导致动作变化识别出现偏差。综合上述研究现状可见，当前多数人体动作识别方法普遍依赖单一模态数据开展研究，这一局限在很大程度上制约了算法的识别性能与适用场景范围。为突破这一技术瓶颈，本文聚焦多模态人体动作识别技术，提出一种基于 LSTM 多特征融合人体动作识别算法，旨在提升复杂环境下人体动作识别的精准度。

2. 相关工作

2.1. 人体骨架动作特征提取

人体骨架序列通常由连续的人体骨架帧组成，每个人体骨架帧是由一系列关节点的 3D 坐标以及坐

标的位置关系所构成的集合。定义人体骨架序列 $RT=\{RT_1,RT_2,RT_3,\cdots,RT_n\}$ 其中 n 为人体固件帧数等于30， RT_i 为人体骨架序列中第 i 的骨架，每一帧的人体骨架 RT_i 都有一些列的人体骨架主要关节点组成，即 $RT_i=\{R_1,R_2,R_3,\cdots,R_t\},1\le i\le n$ 其中 n 为人体骨架帧中关节点的数目。其中第 i 个关节点 $R_i=(x_i,y_i,h_i)$ 其中 (x_i,y_i) 为骨骼点位置信息， h_i 为关节点的深度信息。

2.2. 人体骨骼间关节数据特征

使用 Kinect 深度相机拍摄人体动作视频时，会同步生成人体的 20 个主要骨骼点空间坐标信息[7]，把 20 个节点按照真实人体结构进行连接，再根据各关节点的坐标信息，可以对 Kinect 深度相机捕获到的坐标信息进行三维重组，可视化后如图 1 所示。

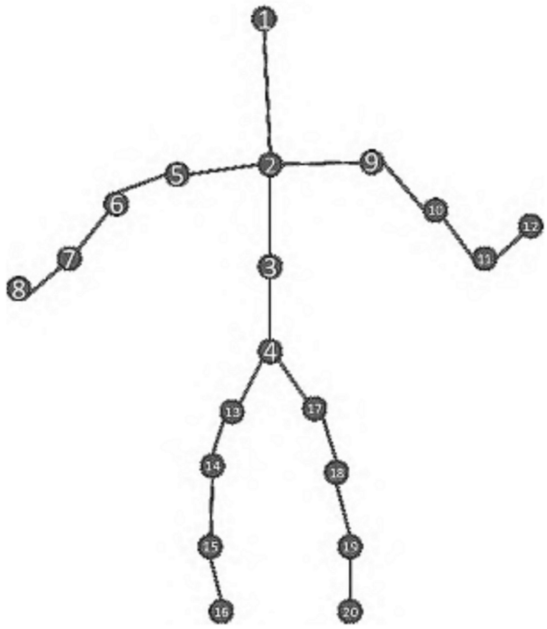

Figure 1. Human skeleton diagram
图 1. 人体骨架图

人体骨骼关节点有 20 个，主要包括躯干关节点和四肢关节点。以第 i 帧第 k 个关节点坐标 (x_i^k,y_i^k,z_i^k) 和 $i+1$ 帧第 k 个关节点 $(x_{i+1}^k,y_{i+1}^k,z_{i+1}^k)$，$i\le 19$，$k\le 30$ 为例，$\boldsymbol{a}_i^k=(a_{i_x}^k,a_{i_y}^k,a_{i_z}^k)$，$a_{i_x}^k=x_{i+1}^k-x_i^k$，$a_{i_y}^k=y_{i+1}^k-y_i^k$，$z_{i_x}^k=z_{i+1}^k-z_i^k$，向量的模为 $\left|\boldsymbol{a}_i^k\right|=\sqrt{{a_{i_x}^k}^2+{a_{i_y}^k}^2+{a_{i_z}^k}^2}$，$\boldsymbol{\alpha}$ 与 x、y、z 轴的余弦分别是 $\cos a=\dfrac{a_{i_x}^k}{\left|\boldsymbol{a}_i^k\right|}$，$\cos b=\dfrac{a_{i_y}^k}{\left|\boldsymbol{a}_i^k\right|}$，$\cos c=\dfrac{a_{i_z}^k}{\left|\boldsymbol{a}_i^k\right|}$，$\varphi_i^k=(\cos a,\cos b,\cos c)$，$\varphi=\{\varphi_i^k\}$，$i\le 29$，$k\le 20$。

2.3. 基于人体动作特征关键帧定位

人体在运动过程中，骨架序列中关节点会发生快速移动，在人体骨架序列全局空间建模的基础上，通过精确计算人体骨架序列关节点的运动速度和加速度特性[8]，能够快速定位人体动作中的关键帧，其详细实计算程如下。连续采集该人体视频序列 f 帧，将第 i 个骨架中第 p 个关节点的第一帧和最后一帧三

维坐标表示为$\left(x_{i1}^{p},y_{i1}^{p},z_{i1}^{p}\right)$和$\left(x_{in}^{p},y_{in}^{p},z_{in}^{p}\right)$，该关节点的运动速度 v 的计算公式如下：$v_x=\left(x_{in}^{p}-x_{i1}^{p}\right)/f$，$v_y=\left(y_{in}^{p}-y_{i1}^{p}\right)/f$，$v_z=\left(z_{in}^{p}-z_{i1}^{p}\right)/f$，$v_i^p=\sqrt{v_x^2+v_y^2+v_z^2}$，$1\le i\le 30$，$1\le p\le 20$，$v_x,v_y,v_z$ 分别表示人体骨架中第 p 个关节点在三个坐标方向的速度 f 为帧率，速度特征 $v=\left\{v_i^p\right\}$，$1\le i\le 30$，$1\le p\le 20$。加速的特征 $\partial=\left\{\left(v_i^p\right)\right\}$，$1\le i\le 30$，$1\le p\le 20$。最终的特征图 $\beta_{m\times n}=\{\varphi,v,\partial\}$，$m,n$ 表示行数和列数。

卷积后人体动作的多模态数据维度巨大[9]，直接应用计算和分类会增加复杂度。通过降维采样将特征图数量不变但计算量将大幅降低[10]，降为采样后的数据为 $\varepsilon_{m\times n}=\left(1/W_1W_2\right)\sum_i^{W_1}\sum_i^{W_2}\beta_{m\times W_1+j,n\times 2+i}$，式中：$W_1,W_2$ 表示人体动作多模态数据特征降采样尺度。将降采样层接入全连接层，将这些分散的局部特征进一步整合，形成更具代表性的人体动作多模态数据全局特征图。$t_i=f\left(\sum_{k=1}^{s}\varepsilon_k\times w_{ij}+p_{ij}\right)$，其中 $f,\varepsilon_k,w_{ij},p_{ij}$ 分别表示激活函数，第 k 个降采样后的人体特征图，层之间的权重和偏置项。s 表示全连接的神经元个数。输出层使用 Softmax 分类器 $\delta=\mathrm{e}^{t_i}/\sum_{i=1}^{\lambda}\mathrm{e}^{t_i}$。

3. 基于 LSTM 多特征融合人体动作识别算法

将人体动作多模态数据特征图 δ 输入 LSTM 模型中，结合人体动作的时空性，实现多模态人体动作精准识别[11]。设置 LSTM 模型的输入门、遗忘门、输出门依次是 i_t,f_t,o_t，$f_t=\sigma\left(w_f\left[h_{t-1};x_t\right]+b_f\right)$，$i_t=\sigma\left(w_i\left[h_{t-1};x_t\right]+b_i\right)$，$\tilde{C}_t=\tanh\left(w_C\left[h_{t-1};x_t\right]+b_C\right)$，$C_t=f_t\odot C_{t-1}+i_t\odot\tilde{C}_t$，$o_t=\sigma\left(w_o\left[h_{t-1};x_t\right]+b_o\right)$，$h_t=o_t\odot\tanh\left(C_t\right)$，$x_t$，$h_{t-1}$，$C_{t-1}$，$w_*$，$b_*$，$C_t$，$\tilde{C}_t$，$h_t$ 分别表示第 t 时间步批量输入，第 $t-1$ 时间步隐藏状态，第 $t-1$ 时间步细胞状态，权重矩阵，偏置向量，候选细胞状态矩阵，第 t 时间步细胞状态，第 t 时间步隐藏状态。

将 LSTM 模型输出门的信息所代表的动作分类结果中出现次数最多的作为最终识别结果。

4. 实验分析

本节分别在 Kinetics-Skeleton 数据集、NTU-RGB (X-Sub)数据集和 NTU-RGB + D (X-View)数据集上使用 LSTM 多特征融合人体动作识别算法，得到每个数据集前 10 个类别与文献 10、11 的准确率的对比表格，如表 1~3 所示。

Table 1. Classification accuracy of the proposed method on the Kinetics-Skeleton dataset
表 1. 本文方法在 Kinetics-Skeleton 数据集上的分类准确率

Class	1	2	3	4	5	6	7	8	9	10
本文方法	0.987	0.999	0.995	0.992	0.983	0.983	0.981	0.997	0.992	0.994
文献 10	0.878	0.899	0.985	0.892	0.893	0.893	0.881	0.899	0.899	0.912
文献 11	0.889	0.899	0.899	0.894	0.888	0.894	0.896	0.899	0.899	0.896

Table 2. Classification accuracy of the proposed method on the NTU-RGB (X-Sub) dataset
表 2. 本文方法在 NTU-RGB (X-Sub)数据集上的分类准确率

Class	1	2	3	4	5	6	7	8	9	10
本文方法	0.992	0.983	0.986	0.987	0.989	0.996	0.984	0.99	0.992	0.981
文献 10	0.895	0.888	0.889	0.897	0.899	0.896	0.888	0.898	0.899	0.891
文献 11	0.902	0.893	0.891	0.909	0.909	0.906	0.896	0.892	0.904	0.899

Table 3. Classification accuracy of the proposed method on the NTU-RGB + D (X-View) dataset
表 3. 本文方法在 NTU-RGB + D (X-View)数据集上的分类准确率

Class	1	2	3	4	5	6	7	8	9	10
本文方法	0.982	0.99	0.981	0.998	0.985	0.993	0.986	0.99	0.991	0.984
文献 10	0.912	0.903	0.901	0.919	0.919	0.916	0.906	0.902	0.914	0.909
文献 11	0.909	0.906	0.909	0.908	0.902	0.908	0.892	0.894	0.891	0.897

在以上 3 个数据集上前十个分类实验结果表明，本文提出的基于 LSTM 多模态融合人体动作识别算法能够较为准确地识别出视频序列中的动作类别，且部分类别的识别准确率高达 99%以上。本文算法具有较高的骨架动作识别准确率。

本文使用 NTU-RGB + D 数据集，由 10 个受试者重复动作 10 次动作。为了同主流方法进行对比，实验设置和文献[12] [13]保持一致，对整个数据集的前 20 个动作统一进行实验测试，选用每人第 1~5 次动作作为训练样本，第 6~10 次动作作为测试样本。最终本数据集测试识别率为 98.23%，混淆矩阵如图 2 所示。

预测＼真实	站立	行走	跑步	跳跃	深蹲	弯腰	拍手	挥手	踢腿	鞠躬	坐下	站起	爬楼梯	下楼梯	俯卧撑	仰卧起坐	伸展	转身	跳跃摸高	单脚站立
站立	0.988	0	0.002	0	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0	0
行走	0	0.998	0	0	0.003	0.005	0.008	0	0	0	0	0	0	0.002	0	0.006	0.008	0	0.003	0
跑步	0.004	0	0.994	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.004
跳跃	0	0	0.001	0.992	0.001	0	0	0.001	0	0	0.004	0	0	0.002	0.005	0	0	0.005	0	0
深蹲	0	0	0	0.004	0.984	0	0	0	0	0.001	0	0.001	0	0	0.004	0.009	0	0	0	0
弯腰	0	0.001	0	0	0	0.984	0	0.002	0	0	0	0	0	0	0	0	0	0	0	0
拍手	0.004	0	0	0.003	0	0	0.982	0	0	0	0	0	0.002	0	0	0	0	0	0	0
挥手	0	0	0	0	0	0	0	0.997	0	0	0.005	0	0	0	0	0	0	0	0	0.003
踢腿	0	0	0	0	0	0.003	0.004	0	0.992	0	0	0	0.001	0	0.006	0	0	0	0.005	0.003
鞠躬	0	0	0	0	0	0.004	0	0	0.005	0.994	0	0.001	0.002	0	0	0	0	0	0	0
坐下	0	0	0	0	0.009	0	0	0	0.001	0.002	0.981	0	0	0.006	0	0	0	0.002	0.004	0
站起	0	0	0	0	0	0	0	0	0	0	0	0.998	0	0	0	0	0	0.001	0	0
爬楼梯	0	0	0	0	0	0	0.003	0	0	0.003	0	0	0.996	0	0	0	0	0	0	0
下楼梯	0	0	0	0	0	0	0	0	0	0	0	0	0	0.985	0	0	0.004	0	0	0
俯卧撑	0	0	0	0	0.003	0	0.003	0	0	0	0	0	0	0	0.984	0.001	0	0	0	0
仰卧起坐	0	0	0	0	0	0	0	0.001	0	0	0.005	0	0	0	0	0.984	0.002	0	0	0
伸展	0	0.001	0	0	0	0.004	0	0	0.002	0	0	0	0	0.005	0	0	0.986	0	0	0
转身	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.99	0	0
跳跃摸高	0	0.001	0	0	0	0	0	0	0	0	0.005	0	0	0	0	0	0	0.002	0.989	0.005
单脚站立	0.004	0	0.003	0.001	0	0	0	0	0	0	0	0	0	0	0	0.001	0	0	0	0.986

Figure 2. Confusion matrix of experimental results on NTU-RGB + D dataset
图 2. NTU-RGB + D 数据集实验结果混淆矩阵

5. 结束语

本文提出了基于 LSTM 多模态融合人体动作识别算法。该方法采用骨骼点的三维信息和对应关节点的速度和加速度进行特征提取和相关性融合，使模型对不同动作具有更高的识别精度。本文模型分别在 Kinetics-Skeleton 数据集、NTU-RGB (X-Sub)数据集和 NTU-RGB + D (X-View)数据集上进行大量实验，并与多个主流模型进行对比验证了其先进性。其中，对于三个数据集本文方法均达到了 98%以上。实验结果表明本文所提方法能够提高动作识别性能。

参考文献

[1] 王杨, 许佳炜, 王傲, 等. 基于 CSI 实例标准化的域泛化人体动作识别模型[J]. 通信学报, 2024, 45(6): 196-209.

[2] Yoshikawa, Y., Shigeto, Y., Shimbo, M. and Takeuchi, A. (2023) Action Class Relation Detection and Classification across Multiple Video Datasets. *Pattern Recognition Letters*, **173**, 93-100. https://doi.org/10.1016/j.patrec.2023.08.002

[3] 叶典, 邱卫根, 张立臣, 等. 基于 2S-LSGCN 的人体动作识别[J]. 计算机工程与设计, 2022, 43(2): 510-516.

[4] 陈路飞, 张勇, 唐永正, 等. FP-Net: 基于任意角度单幅人体图像的正面姿态估计[J]. 计算机辅助设计与图形学学报, 2022, 34(10): 1604-1612.

[5] Chen, Y. and Cheng, K. (2024) BICLR: Radar-Camera-Based Cross-Modal Bi-Contrastive Learning for Human Motion

Recognition. *IEEE Sensors Journal*, **24**, 4102-4119. https://doi.org/10.1109/jsen.2023.3344789

[6] 徐晓冰, 左涛涛, 孙百顺, 等. 基于热释电红外传感器的人体动作识别方法[J]. 红外与激光工程, 2022, 51(4): 391-398.

[7] 王琳玮, 邵星灵, 杨卫. 基于惯性传感器的球形机器人位姿控制系统及实验研究[J]. 中国测试, 2020, 46(3): 123-127.

[8] 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.

[9] 孙浩, 何宏, 汪焰兵, 等. 基于运动特征的骨骼行为识别方法[J]. 计算机工程与设计, 2024, 45(6): 1836-1842.

[10] 余金锁, 卢先领. 基于分割注意力的特征融合 CNN-Bi-LSTM 人体行为识别算法[J]. 电子测量与仪器学报, 2022, 36(2): 89-95.

[11] 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.

[12] 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.

[13] Ahmad, Z. and Khan, N. (2020) Human Action Recognition Using Deep Multilevel Multimodal (M^2) Fusion of Depth and Inertial Sensors. *IEEE Sensors Journal*, **20**, 1445-1455. https://doi.org/10.1109/jsen.2019.2947446