文章编号:1671-4598(2026)04-0173-09 DOI:10.16526/j.cnki.11-4762/tp.2026.04.022 中图分类号:TP183;TP391.41 文献标识码:A

基于双流网络的人体行为特征识别研究

任伟建[1,2]，窦 卓[1]，霍凤财[1,2]，任 璐[3]，张永丰[4]，孙勤江[5]
(1. 东北石油大学 电气信息工程学院，黑龙江 大庆 163318；
2. 黑龙江省网络化与智能控制重点实验室，黑龙江 大庆 163318；
3. 海洋石油工程股份有限公司，天津 300450；
4. 大庆油田有限责任公司 第二采油厂 规划设计研究所，黑龙江 大庆 163318；
5. 中海石油（中国）有限公司 天津分公司，天津 300450)

摘要：人体行为特征识别作为计算机视觉领域的一个重要研究方向，在实际生活中有着广泛的应用，其研究方法可以分为基于传统方法和基于深度学习的方法；双流网络作为基于深度学习方法中较为经典的网络，其将视频序列分为时间和空间两种特征的思想为研究者们提供了丰富的思路；从基于 3D-CNN 的双流网络、融合 LSTM 的双流网络、基于图卷积的双流网络和引入注意力机制的双流网络 4 个方面分别探讨双流网络的研究现状，分析各类方法的局限性，梳理双流网络发展的关键节点并总结每种方法的优缺点及应用场景；列举常用数据集；概述人体行为特征识别的应用；指出目前面临的挑战并对未来进行展望。

关键词：人体行为特征识别；深度学习；双流网络；注意力；骨骼

Study on Human Action Feature Recognition Based on Two-Stream Convolutional Networks

REN Weijian[1,2], DOU Zhuo[1], HUO Fengcai[1,2], REN Lu[3], ZHANG Yongfeng[4], SUN Qinjiang[5]
(1. Department of Electrical Information Engineering, Northeast Petroleum University, Daqing 163318, China;
2. Heilongjiang Provincial Key Laboratory of Networking and Intelligent Control, Northeast Petroleum University, Daqing 163318, China;
3. Offshore Oil Engineering Co., Ltd., Tianjin 300450, China;
4. Institute of Planning and Design of the Second Oil Production Plant, Daqing Oilfield Co., Ltd., Daqing 163318, China;
5. Tianjin Branch, China National Offshore Oil Corporation, Tianjin 300450, China)

Abstract: As an important research direction in the field of computer vision, human action feature recognition is widely applied in real life. Its research methods can be divided into traditional methods and deep learning-based methods. Two-stream convolutional network is taken as a classic deep learning-based network, which divides video sequences into temporal and spatial features, providing researchers with a rich idea. From four aspects of two-stream networks: 3D convolutional neural network (3D-CNN) -based, long short-term memory (LSTM) -fused, graph convolution-based, and introduction of attention mechanism, this paper discusses the current research status of two-stream networks respectively, analyzes the limitations of various methods, reviews key milestones in the development of two-stream networks, summarizes the advantages, disadvantages, and application scenarios of different methods, presents commonly used datasets, offers an overview of practical applications in human action feature recognition, identifies current challenges, and provides an prospect for future research.

Keywords: human action recognition; deep learning; two-stream networks; attention mechanism; skeleton

收稿日期:2025-04-17； **修回日期**:2025-05-23。
基金项目:国家自然科学基金资助项目(61933007,61873058);河北省自然科学基金面上项目(D2022107001)。
作者简介:任伟建(1963-),女,博士,教授。
引用格式:任伟建,窦 卓,霍凤财,等.基于双流网络的人体行为特征识别研究[J].计算机测量与控制,2026,34(4):173-181.

0 引言

随着互联网技术的快速发展，视频数据的规模呈现出指数级增长。传统人工处理的方式已难以应对如此庞大的数据量，因此视频理解技术逐渐兴起。其中，视频行为识别作为视频理解中的核心任务之一，已成为计算机视觉领域的研究热点，通过手机录制以及摄像头等设备采集的视频，对视频中人体的行为进行检测、识别、自动分析具有重要的意义[1]。

视频监控设备广泛应用于公共场所，提高了人们生活的便利性和安全性，增强了社会稳定性。在医疗护理方面，行为识别技术可实现为老人或患者等护理对象提供护理工作[2]等，实时监测老人或患者的活动状态，识别摔倒、长时间静止等异常行为，从而及时提供干预和护理服务。在智能安全监控领域，用于识别施工作业人员违规行为[3]、煤矿工人不安全行为[4]等，例如未佩戴安全帽、误入高危区域或未按规范使用设备等行为，这些行为往往是安全事故的诱因，及时识别和处理可有效降低事故风险、提升作业效率，推动现场智能化、规范化管理，对保障生命安全、实现风险可控具有重要意义。在社会治安领域，可检测打架斗殴暴力事件[5]，识别地铁、电梯上乘客的危险行为[6-7]等，从而提升城市公共安全管理能力。

从特征提取的方式来划分，人体行为识别的相关研究主要分为两大类：传统方法和深度学习的方法。传统方法通常采用人工设计并提取特征的方法进行行为识别，根据特定的行为特点，例如轮廓特征、运动轨迹、时空兴趣点或人体关节点等进行特征提取，将得到的特征向量作为输入，采用不同的特征编码，送入分类器完成行为识别。这类方法具有结构简单、计算开销小、易于解释等优点，但在特征提取的灵活性和适应性方面存在明显不足，尤其在处理复杂背景、多样动作或跨场景行为时，识别性能受到较大限制。

基于深度学习的方法避免了手动设计特征的过程，通过深度神经网络对待测对象采用端到端的学习方式，自动学习数据中更具判别性的高层次特征表示，再将其输入到分类器中进行行为类别的预测，并选择概率值最大的类别完成识别。与传统方法相比，深度学习在特征表达能力和识别精度方面具有显著优势，尤其适用于数据复杂、行为多样的实际场景。然而，深度学习方法也面临一些挑战，如对大量标注数据的依赖性强、训练过程计算资源消耗大等，限制了其在资源受限或实际应用中的直接部署。

双流网络作为深度学习方法中的一个重要基础结构，通过结合空间流网络和时间流网络对复杂动作信息进行完整描述，可以克服单一流神经网络在处理动态场景时的局限性。这种时空特征的融合对于人体行为特征识别任务来说具有重要意义，在提高识别准确率方面发挥了关键作用。

基于双流网络的人体行为特征识别方法是本文研究的主要内容。本文第一节对双流网络的结构进行介绍，并根据其主干网络演化的类型进行划分，将其分为基于3D-CNN 的双流网络、融合 LSTM 的双流网络、基于图卷积的双流网络和引入注意力机制的双流网络，对上述网络分别展开综述，将采用不同主干的双流网络与其他人体行为特征识别方法进行对比，梳理双流网络在行为特征识别领域发展的重要节点并阐述每种方法的优缺点及应用场景。第二节列举常用的行为识别数据集，第三节概述人体行为特征识别的应用，第四节对未来的研究进行展望。

1 双流网络及其主干演化

2014 年文献［8］提出双流卷积神经网络（Two-Stream ConvNets，two-stream convolutional networks）。双流网络是指对空间流网络和时间流网络（即两个子网络）的处理，空间流网络以单帧 RGB 图像为输入，通过二维卷积神经网络（CNN，convolutional neural network）提取图像场景和目标特征；时间流网络以多帧光流图像作为输入，使用 CNN 提取运动信息；将这两个流得到的信息进行融合，从而得出最终的行为分类，其网络模型如图 1 所示。

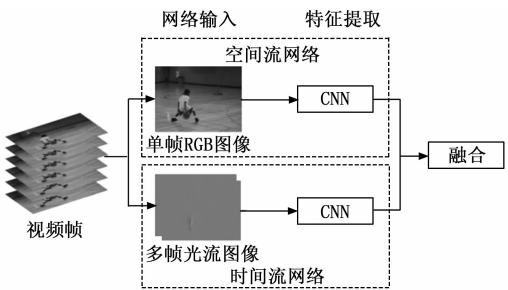

图 1 双流网络模型

从图 1 中看到，空间流网络的输入是单帧图像，经过 CNN 提取静态的空间特征。多帧光流图片是时间流网络的输入，其中光流图片是通过光流计算得到的，光流计算是指分析连续帧之间的像素变化来估计图像中每个像素的运动。两个子网络通过分别对 RGB 图像和光流图像进行特征提取得到静态信息和运动信息，最后进行类别得分融合得到行为分类的结果，但其中光流提取的过程是独立于网络之外的，而且复杂的背景、视角的变化和身体部位的遮挡等因素会对人体行为特征识别造成影响，此外，不同的人做出相同的动作会有差异，不

同的动作之间又存在相似性，这都对特征识别的精度提出了更高的要求。因此，研究者们需要进一步探索并优化双流网络人体行为特征识别技术。

1.1　基于 3D-CNN 的双流网络

传统双流网络主要通过时间流网络获得视频中的运动信息，空间流网络无法直接捕捉时间维度的相关性。3D 卷积神经网络作为视频人体行为特征识别研究的常用架构，是在时间维度上对 CNN 再进行卷积操作，有助于提取更多时间维度上的信息。学者们将双流网络的两个子网络换成 3D-CNN，使得空间流网络不仅可以处理静态的空间特征还可以捕捉时间维度的特征，当时间流网络提取的运动信息不准确时，空间流网络捕捉的时间维度特征可以作为补充，提升整体性能。网络可以从每一帧的 RGB 数据中直接提取运动信息，而不完全依赖于时间流网络提取的运动信息。典型代表是双流膨胀 3D 卷积网络[9]（I3D，two-stream inflated 3D convolutional network），其网络模型如图 2 所示。

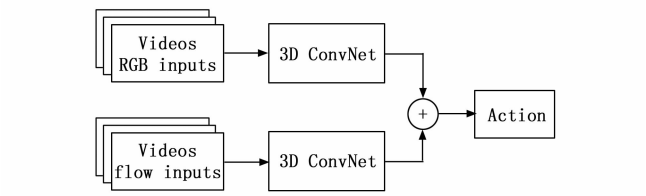

图 2　双流膨胀 3D 卷积网络模型

文献［10］将 3D 卷积神经网络和门控循环单元引入到双流卷积神经网络中，在空间流和时间流中分别使用 3D 卷积提取时空信息，将提取到的时空特征融合并形成有时间顺序的时空特征流，然后时空特征流输入到门控循环单元中学习时间维度的长序列特征，使模型充分利用视频的时间维度信息。文献［11］利用 I3D 对视频进行特征提取，特征提取后，通过广义回归神经网络（GRNN，general regression neural network）对视频的时空特征进行回归处理，输出异常行为的概率评分，实现对异常行为的识别。文献［12］对空间流采用 3D 残差网络，提取信息丰富的 RGB 图像，提高对小幅度动作的识别率，时间流采用 C3D[13] 网络，提取能消除场景信息的光流图特征。针对时间流网络中光流提取复杂且计算效率不高的问题，文献［14］提出一种基于 R（2＋1）D 的端到端双流 CNN，通过引入 PWC-Net[15]，将视频的 RGB 图像序列生成光流图像作为时间流网络的输入，空间流网络依然以视频的 RGB 图像序列作为输入。随着研究的深入，学者们发现在网络中加入过多的 3D 卷积核会使网络变得复杂，因此文献［16］以文献［14］提出的网络结构为基础，提出基于通道剪枝的双流—非局部时空残差卷积神经网络，网络采用双流架构，对两个子网络采用通道剪枝方案，从而降低复杂度，并且还引入非局部模块来更好地学习时空依赖关系，提高网络识别精度。

将双流网络的两个子网络换成 3D-CNN 确实可以提高网络的准确率，但 3D-CNN 对时空特征的提取是基于局部的帧间变化，而在运动模糊或快速运动场景这类极端情况下，RGB 数据可能会出现运动模糊或物体快速移动的情况，3D-CNN 提取到的时空特征可能并不完整或准确，对于一些细微的动作，尽管较小的卷积核可以捕捉细粒度信息，但由于特征图经过多层卷积和下采样，像素级的细节可能丢失。另外，网络的计算量和参数规模会增加，尤其在处理长时间序列的视频数据时，计算成本大幅上升。如果任务仅涉及较少的时序信息，或者对于时序的依赖性较低，使用 3D-CNN 可能导致计算资源浪费，无法充分利用其优势。

1.2　融合 LSTM 的双流网络

循环神经网络（RNN，recurrent neural network）包含了前一刻的信息，因此网络可以保留对前面内容的记忆，对时间序列数据进行建模，捕捉序列中的时序关系和上下文信息。由于 RNN 存在梯度消失和爆炸的缺点，文献［17］提出了长短期记忆网络（LSTM，long short-term memory），使用 3 个不同的门实现对信息的保存和遗忘。LSTM 具有记忆单元和门控机制，能够根据历史状态调整当前的预测并对不同长度的视频序列进行灵活地建模和处理，因此常被用来与双流网络结合使用。针对特定场合下，由于网络无法有效利用时间维度信息而导致人体行为识别准确率不高的问题，文献［18］提出深层次残差长短期双流网络模型，如图 3 所示。网络的输入仍然是 RGB 帧和连续光流帧不变，空间流网络和时间流网络均使用 ResNet 提取低层特征信息，然后分别将提取的特征信息作为 LSTM 的输入，有效学习空间特征和光流特征的时间序列信息，利用其门控机制筛选关键时序信息，增强对异常片段边界的敏感度，并且通过多种加权融合策略加强模型识别效果。

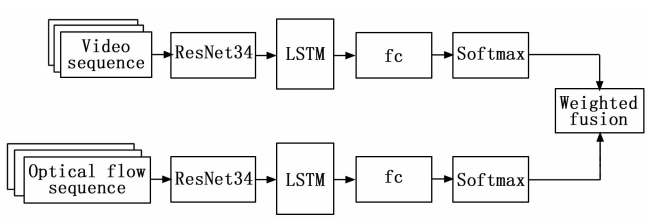

图 3　深层次残差长短期双流网络模型

文献［19］将双流网络提取的特征作为 Bi-LSTM 的输入，以增强时序特征的提取，强化关键行为信息的提取，克服传统双流网络在时空特征融合时未能充分考

虑时序连续性的问题，从而提高行为识别的准确度。针对视频帧内部和跨视频帧之间的时空变化，文献［20］提出了一种双流循环神经网络，将 CNN 与 LSTM 结合在一起，有效地建模空间和时间信息。通过 CNN 从每个视频帧中提取空间特征，通过 LSTM 从视频帧序列中提取时间特征，从而组成新的双流网络架构，该网络依据 LSTM 的记忆属性，可以有效地发现和建模视频流中的短期和长期时空模式，从而提升行为识别的性能。针对现有方法中抗背景干扰能力差和准确率低等问题，文献［21］采用平均融合的方式提出结合 3 种流（空间流、时间流和时空显著性流）的 STS-ALSTM 模型，空间流和时间流的输入依然分别是 RGB 和光流帧不变，使用通过基于地理距离的分割方法生成的时空显著性图作为时空显著流的输入，减少背景干扰，突出视频中物体的前景信息，在 STS-ALSTM 模型中，3 种流分别通过独立的 CNN 提取出相应的特征，提取的特征被输入到 3 个独立的 LSTM 中，捕捉连续帧之间的时序依赖关系。但是，该模型的融合结果仅是通过对不同流的预测进行平均处理得到，没有充分利用不同流之间的互补知识。对此，文献［22］将时空显著性流改成空间显著流和时间显著流，提出包含空间流、时间流、空间显著性流和时间显著性流的多任务学习四流网络，采用多任务学习的 LSTM 方法，提取到的四流特征向量被输入到多任务学习的 LSTM 中，将每个特征通过 LSTM 进行分类视为一个独立任务，并为每个任务计算相应的损失。通过 LSTM 对四种不同的长期依赖关系进行建模，通过多任务学习融合不同的线索，共享不同流之间的互补知识。

融合循环神经网络的方法通过对视频序列的建模来提升视频行为识别任务的性能，但其在实际应用与理论设计中仍存在局限性。LSTM 的递归结构要求每个时间步的计算依赖于前一个时间步，无法实现像 CNN 那样的高效并行化。在视频处理任务中，特别是在实时应用或对大规模视频数据的处理中，并行处理是提升性能的关键，因此融合循环神经网络方法的低并行性成为一个瓶颈。

1.3 基于图卷积的双流网络

与使用 RGB 图像进行行为特征识别的方法相比，利用人体骨骼信息进行特征识别能有效克服复杂背景、光照变化和外貌变化等因素的影响，因此，学者们将骨骼信息融合到双流网络中，将骨骼数据作为双流网络的输入。目前常见的骨骼数据表示方法包括基于序列的关节点坐标表示、图像化编码方法（如关节距离图）以及基于图结构的建模方式。基于坐标序列的方法能够直接反映动作随时间的变化，但难以捕捉空间结构的全局关系。由于 CNN 擅长处理具有空间规则性的数据，而骨架序列具有不规则性，为满足 CNN 的输入要求，图像化编码方法通过构造二维图像使骨骼信息适配传统 CNN 架构，研究者们通常将骨骼数据图像化处理，从矢量序列转换为 2D 伪图像。图像中骨骼每个关节的坐标被视为每个像素的 3 个通道，然后使用卷积核提取伪图像的特征[23]。

在处理简单动作序列时，可以采用基于坐标序列的方法，在处理动作幅度大、关节间耦合强或多人交互等复杂行为时，图结构建模相较于其他方法更具优势，而在资源受限或实时性要求较高的场景下，图像化方法凭借其高效的特征编码方式更易部署。因此，选择合适的骨骼数据表示方式需综合考虑行为特征、应用需求与模型复杂度等因素。

有学者尝试将骨骼数据作为双流网络变体的输入，例如文献［24］针对时空特征挖掘不充分导致精度不够的问题，提出了一种基于 CNN-LSTM 的双流卷积模型，将 CNN 网络与 LSTM 网络并联结合，对人体骨架空间运动姿态分别进行静态与动态特征提取，最后两种特征融合并通过 Softmax 进行分类识别，但其模型训练量大且参数多，不利于实际应用中部署。

近年来，图卷积网络（GCN，graph convolutional network）广受关注，该网络通过整合邻居节点信息来更新目标节点信息，从而有效处理具有复杂拓扑结构的数据类型，利用这一特点，GCN 能够深度挖掘图结构数据中的潜在特征与内在关系。而骨骼数据本身就可以当作是一个自然的拓扑图数据结构，关节点和骨头可以看作是图的节点和边，因此，学者们尝试将骨骼数据作为输入并将 GCN 和双流网络结合用于人体行为特征识别。文献［25］提出一种基于人体骨骼和场景图像的双流模型，使用 ResNet101 处理场景图像提取深层次的视觉特征，骨架数据通过时空图卷积提取运动特征，将场景信息和骨骼信息互补融合。还有一些基于骨骼数据的行为特征识别方法不只局限于双流，大多采用多流网络分别进行训练，使训练成本增加，在时域上采用大尺度卷积，导致聚合大量冗余信息。针对以上问题，文献［26］提出运动融合模块，以关节流和骨骼流作为双流输入，在特征层面将关节流和骨骼流的运动信息进行融合，避免额外的训练，用多尺度时间卷积提取时域信息，有效减少冗余信息，提高计算效率。文献［27］搭建由关节点信息流和骨骼信息流构成的双流 GCN 模型，如图 4 所示，为相距较远但关系密切的动作点创建连接，并提出骨骼的长度和方向对人体行为识别起到重要作用的观点。其中，关节点信息流的输入是人体骨架的关节点坐标信息，骨骼信息流的输入是通过额外处理

得到骨骼的长度和方向信息，两种输入均使用 GCN 进行特征提取，有效缓解识别速度不高或识别精度偏低的问题。文献［28］提出了一种多流轻量级语义图卷积的行为识别方法。该方法将骨架信息划分为骨长流、关节流和细粒度流三种不同的数据流表示形式，使用自适应图卷积对融合语义信息的数据流提取空间特征，还利用具有不同卷积核大小和膨胀率的多尺度时域卷积对时间维度特征进行建模，采用加权融合策略处理各流分类结果。

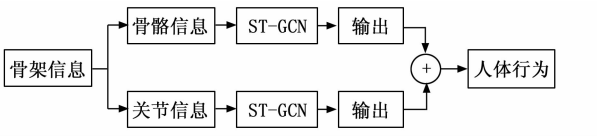

图 4　双流 GCN 模型

融合骨骼信息的双流网络需要复杂的网络结构设计，包括如何有效地融合时空特征，融合不当可能导致特征冗余或信息丢失。文献［27］所提出的观点虽然能够提高关键点的关联性，但关节点和骨骼的特征在网络早期是分开的，无法充分共享信息。只将双流网络与 GCN 结合无法突出重要特征，在实时应用中，GCN 的推理时间通常比 CNN 长，在需要快速响应的场景中还有待改进。

1.4　引入注意力机制的双流网络

前面所述 3 种方法存在依赖手工制作的遍历规则或图拓扑来表述关节点之间的依赖关系，不适于处理序列的长期依赖性和时空关节的全局相关性。注意力是指人选择性地关注较为重要的信息，注意力机制与人类的注意力类似，在行为识别任务中，它使模型能够在大量输入数据中聚焦于关键特征，为不同位置分配注意权重，提取出重要的信息，忽略不重要的部分。引入注意力机制的双流网络能够突出重要的时空特征，使模型更专注于视频中关键的帧和区域，对输入数据不同部分的关注程度进行动态调整，通过可视化权重，有利于理解模型的决策过程，也能够对长时间跨度的序列进行建模。

传统双流网络在处理视频序列时，往往难以有效建模时间上的依赖关系，因而在识别对时序特征要求较高的行为时，表现出一定的局限性。文献［29］使用时间移位思想和注意力机制构建了一个双流网络结构，利用时间移位对视频中的时序关系建模，通过在通道和空间维度应用注意力机制，用于改善因通道信息在时间轴上移动导致的空间特征学习能力下降的问题，使网络关注视频帧中的重要局部细节，有助于提升网络特征提取的能力。实验表明算法能够提高对时序依赖较大的行为和近似行为的辨识能力。针对传统双流卷积网络存在时空相关信息难学习、空间和时间特征缺乏充分融合等缺点，文献［30］提出 STF-ResNet 网络模型，网络结构如图 5 所示。该网络通过残差结构来融合空间和时间特征，这种结构可以补偿高层特征的损失，并通过低层特征增强网络的学习能力。在融合前加入 CBAM 注意力模块[31]，从通道维度和空间维度进一步过滤动作特征，能够突出特征的关键区域，避免在融合时被低权重的冗余特征干扰。经注意力处理的特征图带有注意力权重引导，可以用于加权融合，而非盲目拼接。

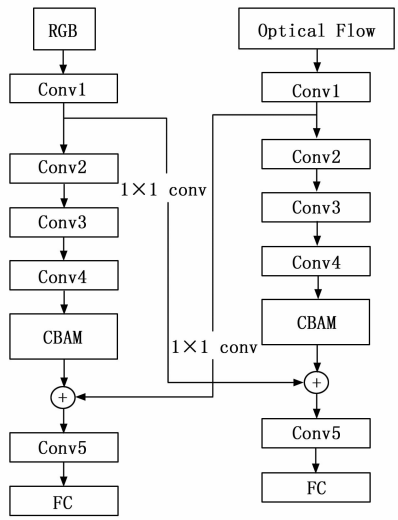

图 5　STF-ResNet 网络模型

文献［32］提出基于多传感器渐进特征的双流卷积神经网络，其中特征增强通道注意力模块强调传感器重要数据的特征表示，注意力特征融合模块在不同模态和多阶段步态特征之间实现灵活的特征融合。文献［33］提出一种基于双流网络的短期动作时空注意模块。该模块能够关注不同短期行动之间在重要性和速度上的差异，通过提取视频中不同短期动作在重要性和速度上的差异，然后对特征进行融合，旨在丰富时空特征，提高动作识别性能。然而，该方法的双流结构虽然能够捕捉不同时间尺度的信息，但在融合时仍然依赖于简单的加权平均，未能充分利用不同时间尺度之间的互补性。文献［34］提出一种双流自适应注意力图卷积网络，设计了能自适应平衡权重的多阶邻接矩阵融合方法和多尺度时空自注意力模块，使模型聚焦于更加重要的邻域并增强模型的特征提取能力，但该方法在融合时仍然依赖于简单的加权平均，未能充分利用数据流之间的潜在分布差异。文献［35］提出一种改进的双流视觉 Transformer 行为识别模型。首先，为增强模型对长时序列数据的处理能力，使用分段采样方法；其次嵌入注意力模块来增强模型特征表示能力；最后，为增大类间差异并减小类内差异，提出新的损失函数。使用双流网络捕获时空信息容易忽略视频中时空信息之间的强互补性和相关

性，对此，文献［36］提出具有时空交互学习模块的双流网络，时空交互学习模块利用空间流和时间流之间的交替共注意力机制来学习空间特征和时间特征之间的相关性，建立两个流之间的交互连接，生成优化的空间注意力特征和时间注意力特征。

在处理大规模的视频数据时，尽管注意力机制可以减少对不重要部分的计算，但引入后需要更多的超参数调优，增加了模型的复杂度和调试的难度。复杂的注意力模块设计虽然增强了特征表达能力，但模块间的耦合性增高，加剧了模型调试与优化的难度。在处理长时间的视频时，模型还会面临信息衰减的问题，尤其是在关注特定时间段时，早期信息可能被后续信息所掩盖。现有方法在处理复杂动作或长时动作时，可能会忽略全局时间信息或局部细节信息，例如文献［33］的方法在处理复杂动作或长时动作时，可能会忽略全局时间信息，导致对长时动作的建模能力不足。文献［34］的方法在处理复杂动作时，尤其是在动作变化较快的情况下，可能会忽略局部细节信息。

本文将采用不同主干的双流网络与其他人体行为特征识别方法进行对比，如表 1 所示。尽管不同方法对同一场景中的同一行为可能给出不同的判定结果，但在典型数据集上采用统一的性能评估指标，仍可作为衡量各算法识别准确性的重要依据。表 1 中的准确率是指模型预测结果与真实标签完全一致的比例。对于 NTU RGB＋D 120 数据集，表中的准确率是在 Cross-Subject（X-Sub）条件下，即按照人物 ID 来划分训练集和测试集计算得到的，对于 UCF101 和 Kinetics 这类标准多分类数据集，则按照官方划分方式计算。

表 1 与其他人体行为特征识别方法的对比

网络类型	模型名称	模态输入	数据集	准确率/%	适用场景
单流网络	3DCCA[37]	RGB	UCF101	90.9	基础识别任务
双流网络（3D-CNN）	I3D[9]	RGB＋Flow	Kinetics-400	71.1	复杂识别任务
双流网络（LSTM）	TS-LSTM NET[20]	RGB＋Flow	UCF101	93.1	长时序动作建模
双流网络（GCN）	MS-SGN[28]	RGB＋骨骼	NTU RGB＋D 120	83.4	跨模态行为识别
双流网络（注意力机制）	ViTSN[35]	RGB＋Flow	UCF101	96.1	细粒度行为识别
多模态网络	M-Mixer[38] VATT[39]	RGB＋深度＋红外 RGB＋音频＋文本	NTU RGB＋D 120 Kinetics-400	92.7 82.1	多源融合场景

本文梳理双流网络在人体行为特征识别领域的发展关键节点，如图 6 所示。

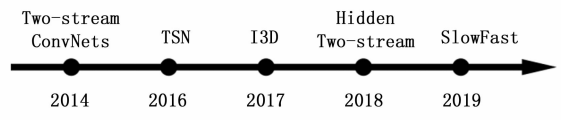

图 6 双流网络发展关键节点

Two-stream ConvNets[8]能够同时提取空间和时间信息，将外观和运动分开建模，实现了深度学习在视频行为分类上的突破，但其需要额外计算光流，对长时序关系建模不足，主要用于基础的行为识别与早期视频分析。

TSN[40]引入稀疏采样策略，实现了长视频长时程依赖建模，在视频级监督下训练，显著提升准确率，但其依赖光流计算成本，时段划分可能遗漏细微动作变化。该网络适用于长视频行为识别、需要综合全局信息的场景，如体育比赛分析。

I3D[9]使用 3D 卷积网络学习时空特征，在大规模数据集上性能优异。该网络模型庞大、计算量高，训练和推理开销大，对大数据依赖性强，适合高精度行为识别，需要充分算力的离线分析。

Hidden Two-stream ConvNets[41]能够实现端到端训练，无需光流预处理，结构轻量，但其时序建模能力不如显式光流分支，模型可以直接在视频流上运行，适合边缘计算与移动端部署。

SlowFast[42]网络同时捕捉粗粒度与细粒度时序特征，提升分类和检测精度。模型复杂度高、计算成本高，实时性差，训练资源需求高，适合需要识别细微异常行为的视频监控。

2 数据集

为了训练和评估双流网络模型，研究人员常使用的数据集如表 2 所示。这些数据集涵盖了各种不同行为的视频样本，为网络模型提供了丰富多样的训练素材，以便识别特定的人体行为。通过利用这些数据集，研究人

表 2 常用数据集

数据集	年份	视频片段数	类别数	来源
HMDB51[43]	2011	7 000	51	电影、公共数据库
UCF101[44]	2012	13 320	101	视频网站
Sports-1M[45]	2014	1 133 158	487	视频网站
ActivityNet[46]	2015	28 000	200	视频网站
Charades[47]	2016	9 848	157	用户拍摄
Kinetics-600[48]	2017	495 547	600	视频网站
AVA[49]	2018	57 600	80	视频网站
NTU RGB＋D 120[50]	2019	114 480	120	用户拍摄
HVU[51]	2020	572 000	3 142	视频网站

员得以深入探讨算法的创新与优化，从而使人体行为识别的准确率不断提升。

3　人体行为特征识别的应用

目前，人体行为特征识别技术已在公共安全、医疗健康、工业生产等多个实际场景中发挥着越来越重要的作用。在公共安全方面，高密度人群场景下，人体行为特征识别技术通过结合不同的特征数据，实时检测人员的异常行为[52]。在医疗健康方面，能够监测精神患者的异常行为[53]或辅助康复训练，例如对患者的康复动作进行识别[54]。在工业生产方面，实现对井下人员的不安全行为识别与预警[55]。随着深度学习的不断发展与多模态融合等技术的引入，人体行为特征识别技术将渐渐拓宽其应用范围。

4　展望

通过对现有研究的梳理，本文对基于双流网络的人体行为特征识别研究进行展望具体如下：

1）避免复杂的光流提取。光流的提取会浪费大量的计算资源，未来的研究可以从以下 3 个方向展开：(1) 对时间通道的信息进行更高效的处理，以挖掘特征图间的时间关联；(2) 通过隐式模拟光流提取过程，实现对时序特征的建模；(3) 使用骨骼数据、深度数据等代替光流数据，多模态数据具有更丰富的信息，不同的数据之间具有互补性。使用多模态数据可以提高特征识别的准确性和鲁棒性，但这通常需要对数据进行大量的标记以学习有效的特征表示，这对数据的采集和标记提出了挑战。

2）细粒度识别。目前仍存在类内差异性和类间相似性、视频中行为难以定义且范围难以界定、遮挡等问题，未来可尝试引入图像超分辨率方法的理念，通过对已提取特征在通道维度上进行融合与增强，从而实现对视频帧序列的分辨率重建。也可以通过增强主体特征进一步对人体局部细微行为进行研究，设计具有增强主体信息功能的特征提取模块，使网络在特征提取阶段增强对视频中主要人物和物体的感知，从而提高获取视频信息的能力，引入注意力机制的方法成为趋势，可以考虑高阶注意力[56]、交叉注意力[57]、CoTAttention[58]等热门模块，但需要注意模型的复杂度，如何优化注意力仍然是研究热点。

3）精简模型。随着双流网络识别准确度逐渐升高的同时，网络的结构也变得复杂，实时性不佳，剪枝虽然可以降低网络的复杂度，但这通常会降低网络的精度。所以，如何设计网络在不影响准确率的前提下还可以更轻量化具有较高的研究价值。在未来的研究中，可以考虑知识蒸馏[59]的方式，通过将大型、性能强的教师模型中学到的知识转移给小型的学生模型，使学生模型在保持较高准确率的同时减少计算成本。双流网络优化方向主要是改进两路特征的提取能力和融合问题，虽然可以通过一些方法简化模型，但需要注意的是模型的复杂度问题无法完全避免。

5　结束语

本文对基于双流网络的人体行为特征识别进行了综述，根据双流网络的主干网络类型进行划分，从基于 3D-CNN 的双流网络、融合 LSTM 的双流网络、基于图卷积的双流网络和引入注意力机制的双流网络四个方面分别展开综述并分析了各网络的局限性。将采用不同主干的双流网络与其他人体行为特征识别方法进行对比，展示了改进后的网络在不同场景下的性能表现，并梳理双流网络在行为特征识别领域发展的重要节点，对每个节点的优缺点及应用场景进行总结。本文还列举常用的行为识别数据集并概述人体行为特征识别的应用，指出目前的问题以及对未来的研究进行展望。

参考文献：

[1] 邓淼磊，高振东，李　磊，等. 基于深度学习的人体行为识别综述 [J]. 计算机工程与应用，2022，58 (13)：14－26.

[2] WU X，ZHANG H，KONG C，et al. Lidar-based 3D human pose estimation and action recognition for medical scenes [J]. IEEE Sensors Journal，2024，24 (9)：15531－15539.

[3] 范冰倩，董秉聿，王　彪，等. 基于深度学习的地铁施工作业人员不安全行为识别与应用 [J]. 中国安全科学学报，2023，33 (1)：41－47.

[4] 饶天荣，潘　涛，徐会军. 基于交叉注意力机制的煤矿井下不安全行为识别 [J]. 工矿自动化，2022，48 (10)：48－54.

[5] 张晓蓉，李　伟，石　岩，等. 基于时空信息可信融合的视频监控暴力检测算法 [J]. 计算机应用，2023，43 (s2)：65－71.

[6] 章宇翔，李先旺，贺德强，等. 基于改进的多算法融合地铁站内乘客行为识别 [J]. 铁道科学与工程学报，2023，20 (11)：4096－4106.

[7] 杨学存，李杰华，陈丽媛，等. 基于人体骨架的扶梯乘客异常行为识别方法 [J]. 安全与环境学报，2024，24 (2)：636－643.

[8] SIMONYAN K，ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [C] //Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS). Cambridge，USA：MIT Press，2014：568－576.

[9] CARREIRA J，ZISSERMAN A. Quo vadis，action recognition? A new model and the kinetics dataset [C] //2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)，Piscataway，USA：IEEE，2017：4724－4733.

[10] 陈 颖，来兴雪，周志全，等. 基于 3D 双流卷积神经网络和 GRU 网络的人体行为识别 [J]. 计算机应用与软件，2020，37 (5)：164－168.

[11] 刘良鑫，林勉芬，钟良泉，等. 基于 3D 双流卷积神经网络的异常行为检测 [J]. 计算机系统应用，2021，30 (5)：120－127.

[12] 欧阳黎，林彤尧，程 莺，等. 基于时空双流 3D 残差网络的服务动作识别 [J]. 计算机应用与软件，2023，40 (6)：112－117.

[13] TRAN D，BOURDEV L，FERGUS R，et al. Learning spatiotemporal features with 3D convolutional networks [C] //2015 IEEE International Conference on Computer Vision (ICCV). Piscataway，USA：IEEE，2015：4489－4497.

[14] HUANG M，QIAN H，HAN Y，et al. R (2＋1) D based two-stream CNN for human activities recognition in videos [C] //2021 40th Chinese Control Conference (CCC). Piscataway，USA：IEEE，2021：7932－7937.

[15] SUN D，YANG X，LIU M，et al. PWC-Net：CNNs for optical flow using pyramid，warping，and cost volume [C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，USA：IEEE，2018：8934－8943.

[16] 钱惠敏，陈 实，皇甫晓瑛. 基于双流-非局部时空残差卷积神经网络的人体行为识别 [J]. 电子与信息学报，2024，46 (3)：1100－1108.

[17] HOCHREITER S，SCHMIDHUBER J. Long short-term memory [J]. Neural Computation，1997，9 (8)：1735－1780.

[18] 张仁路，高丙朋. 基于时序时空双流卷积的异常行为识别 [J]. 现代电子技术，2023，46 (3)：81－87.

[19] 马 莉，王 卓，代新冠，等. 基于双流 CNN 与 Bi-LSTM 的施工人员不安全行为轻量级识别模型 [J]. 西安科技大学学报，2022，42 (4)：809－817.

[20] HASSAN E. Learning video actions in two stream recurrent neural network [J]. Pattern Recognition Letters，2021，151：200－208.

[21] LIU Z，LI Z，WANG R，et al. Spatiotemporal saliency-based multi-stream networks with attention-aware LSTM for action recognition [J]. Neural Computing and Applications，2020，32：14593－14602.

[22] ZONG M，WANG R，MA Y，et al. Spatial and temporal saliency based four-stream network with multi-task learning for action recognition [J]. Applied Soft Computing，2023，132：109884.

[23] 卢 健，李萱峰，赵 博，等. 骨骼信息的人体行为识别综述 [J]. 中国图象图形学报，2023，28 (12)：3651－3669.

[24] 高治军，顾巧瑜，陈 平，等. 基于 CNN-LSTM 双流融合网络的危险行为识别 [J]. 数据采集与处理，2023，38 (1)：132－140.

[25] XU Q，ZHENG W，SONG Y，et al. Scene image and human skeleton-based dual-stream human action recognition [J]. Pattern Recognition Letters，2021，148：136－145.

[26] 卢先领，杨嘉琦. 时空关联的 Transformer 骨架行为识别 [J]. 信号处理，2024，40 (4)：766－775.

[27] 王宪伦，王广宇，孙宇轩. 基于双流图卷积网络的人体行为识别算法 [J]. 传感器与微系统，2023，42 (7)：140－143.

[28] 刘锁兰，王 炎，王洪元，等. 基于多流语义图卷积网络的人体行为识别 [J]. 计算机工程，2024，50 (8)：64－74.

[29] 张红颖，安 征. 基于改进双流时空网络的人体行为识别 [J]. 光学精密工程，2021，29 (2)：420－429.

[30] YAO W，CHEN C，CHENG R. Yoga action recognition based on STF-ResNet [C] //2023 IEEE 3rd International Conference on Power，Electronics and Computer Applications (ICPECA). Piscataway，USA：IEEE，2023：556－560.

[31] WOO S，PARK J，LEE J Y，et al. CBAM：convolutional block attention module [C] //15th European Conference on Computer Vision (ECCV). Berlin，Germany：Springer，2018：3－19.

[32] QIN L，GUO M，ZHOU K，et al. Gait recognition based on two-stream CNNs with multisensor progressive feature fusion [J]. IEEE Sensors Journal，2024，24 (8)：13676－13685.

[33] LIU T. Short-term action learning for video action recognition [J]. IEEE Access，2024，12：30867－30875.

[34] 杜启亮，向照夷，田联房，等. 用于动作识别的双流自适应注意力图卷积网络 [J]. 华南理工大学学报 (自然科学版)，2022，50 (12)：20－29.

[35] 雷永升，丁 锰，沈 尧，等. 基于改进双流视觉 Transformer 的行为识别模型 [J]. 计算机科学，2024，51 (7)：229－235.

[36] LIU T，MA Y，YANG W，et al. Spatial-temporal interaction learning based two-stream network for action recognition [J]. Information Sciences，2022，606：864－876.

[37] ZHAO H，LIU J，WANG W J. Research on human behav-

ior recognition in video based on 3DCCA [J]. Multimedia Tools and Applications, 2023, 82: 20251 - 20268.

[38] LEE S, WOO S, NUGROHO A M, et al. Modality mixer exploiting complementary information for multi-modal action recognition [J]. Computer Vision and Image Understanding, 2025, 256: 104358.

[39] AKBARI H, YUAN L, QIAN R, et al. VATT : transformers for multimodal self-supervised learning from raw video, audio and text [C] //Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS). Cambridge, USA: MIT Press, 2021: 24206 - 24221.

[40] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition [C] // Proceedings of the European Conference on Computer Vision (ECCV), Berlin, Germany: Springer, 2016: 20 - 36.

[41] ZHU Y, LAN Z, NEWSAM S, et al. Hidden two-stream convolutional networks for action recognition [C] //14th Asian Conference on Computer Vision (ACCV), Berlin, Germany: Springer, 2018: 363 - 378.

[42] FEICHTENHOFER C, FAN H, MALIK J, et al. SlowFast networks for video Recognition [C] // 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, USA: IEEE, 2019: 6201 - 6210.

[43] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition [C] // 2011 IEEE/CVF International Conference on Computer Vision (ICCV), Piscataway, USA: IEEE, 2011: 2556 - 2563.

[44] SOOMRO 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.

[45] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks [C] //2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway, USA: IEEE, 2014: 1725 - 1732.

[46] HEILBRON F C, ESCORCIA V, GHANEM B et al. ActivityNet: a large-scale video benchmark for human activity understanding [C] // 2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway, USA: IEEE, 2015: 961 - 970.

[47] SIGURDSSON G A, VAROL G, WANG X, et al. Hollywood in homes: crowdsourcing data collection for activity understanding [C] // Proceedings of the European Conference on Computer Vision (ECCV), Berlin, Germany: Springer, 2016: 510 - 526.

[48] CARREIRA J, NOLAND E, BANKI-HORVATH A, et al. A short note about Kinetics-600 [J]. Arxiv: 1808.01340, 2018.

[49] GU C, SUN C, ROSS D A, et al. AVA: a video dataset of spatio-temporally localized atomic visual actions [C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway, USA: IEEE, 2018: 6047 - 6056.

[50] LIU J, SHAHROUDY A, PEREZ M, et al. NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42 (10): 2684 - 2701.

[51] DIBA A, FAYYAZ M, SHARMA V, et al. Holistic large scale video understanding [C] //Proceedings of the European Conference on Computer Vision (ECCV), Berlin, Germany: Springer, 2020: 593 - 610.

[52] HAMID A A, MONADJEMI S A, SHOUSHTARIAN B. ABDviaMSIFAT: abnormal crowd behavior detection utilizing a multi-source information fusion technique [J]. IEEE Access, 2024, 13: 75000 - 75019.

[53] SONG H, KANG J, KIM T. Real-time abnormal behavior recognition for patient monitoring in hospitals [C] //2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Piscataway, USA: IEEE, 2024: 1 - 8.

[54] 吴冬梅，白　凡，宋婉莹. 分层残差结构的时空图网络多目标在线康复动作识别 [J]. 计算机应用与软件，2024, 41 (11): 199 - 205.

[55] 李雯静，刘　鑫. 基于深度学习的井下人员不安全行为识别与预警系统研究 [J]. 金属矿山，2023 (3): 177 - 184.

[56] 王增强，张文强，张　良. 引入高阶注意力机制的人体行为识别 [J]. 信号处理，2020, 36 (8): 1272 - 1279.

[57] LEE J C, LEE D G. ESC-ZSAR: expanded semantics from categories with cross-attention for zero-shot action recognition [J]. Expert Systems With Applications, 2024, 255 (PD): 124786.

[58] LI Y, YAO T, PAN Y, et al. Contextual transformer networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45 (2): 1489 - 1500.

[59] QUAN Z, CHEN Q, LI Y, et al. ARCTIC: a knowledge distillation approach via attention-based relation matching and activation region constraint for RGB-to-infrared videos action recognition [J]. Computer Vision and Image Understanding, 2023, 237: 103853.