计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：人体行为识别方法研究综述
作者：梁绪，李文新，张航宁
DOI：10.19734/j.issn.1001-3695.2021.07.0350
收稿日期：2021-07-29
网络首发日期：2021-11-17
引用格式：梁绪，李文新，张航宁．人体行为识别方法研究综述[J/OL]．计算机应用研究．https://doi.org/10.19734/j.issn.1001-3695.2021.07.0350

人体行为识别方法研究综述

梁　绪，李文新，张航宁

(兰州空间技术物理研究所，兰州 730000)

摘　要：随着计算机视觉不断发展，人体行为识别在视频监控、视频检索和人机交互等诸多领域中展现出其广泛的应用前景和研究价值。人体行为识别涉及到对图像内容的理解，由于人体姿势的复杂多样和和背景的遮挡因素导致实际应用的进展缓慢。全面回顾了人体行为识别的发展历程，深入探究了该领域的研究方法，包括传统手工提取特征的方法和基于深度学习的方法，以及最近十分热门的基于图卷积网络(GCN)的方法，并按照所使用的数据类型对这些方法进行了系统的梳理。此外，针对不同的数据类型，分别介绍了一些热门的行为识别数据集，对比分析了各类方法在这些数据集上的性能。最后，对全文进行了概括总结，并对未来人体行为识别的研究方向进行了展望。

关键词：计算机视觉；人体行为识别；深度学习；图卷积网络；数据集
中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2021.07.0350

Review of research on human action recognition methods

Liang Xu, Li Wenxin, Zhang Hangning
(*Lanzhou Institute of Physics*, *CAST*, *Lanzhou* 730000, *China*)

Abstract: With the rapid development of computer vision, human action recognition has shown its wide application prospects and research value in many fields such as video surveillance, video retrieval, and human-computer interaction. Human action recognition involves the understanding of image content, and the progress of practical applications is slow due to the complexity and diversity of human postures and the occlusion factors of the background. This paper comprehensively reviews the development of human action recognition, and deeply explores the research methods in this field, including traditional manual feature extraction methods and deep learning-based methods, as well as the recently popular graph convolutional network (GCN) -based method. And these methods are systematically combed according to the data types they use. In addition, for different data types, some popular action recognition datasets are introduced, and the performance of various methods on these data sets is compared and analyzed. Finally, the review is summarized, and the future research direction of human action recognition is prospected.

Key words: computer vision; human behavior recognition; deep learning; GCN; data set

0　引言

近年来，随着机器学习和人工智能的发展，计算机视觉也取得了飞速进展，并开始应用到不同领域，给人类的生活带来极大的改变。随着我国人口老龄化的推进和“三孩政策”的施行，以及短视频行业的飞速发展，人体行为识别作为计算机视觉的子课题，在智能家庭监护和视频信息检索等方面体现出了广泛的应用前景和研究价值。人体行为识别涉及到对图像内容的理解，由于人体姿势的多样性和复杂性，遮挡和背景杂乱等的混杂因素，相较于仅仅对图像内物体的识别或者检测来说，更加具有难度和挑战性。

人体行为识别的关键是提取出健壮性的行为特征，与图像空间中的特征不同，视频中人的行为特征不仅需要描述了人在图像空间中的外观，而且还必须提取外观和姿势的变化，即从二维空间特征扩展到三维时空特征。近年来，已经提出了许多基于 RGB 数据的人体行为识别方法，包括传统的手工提取特征的方法[1~4]和基于深度学习的方法[5~9]。随着一些深度传感器的应用，例如微软的 Kinect 设备(目前已经推出了性能卓越的第三代)，许多研究者也开始利用深度数据进行人体行为识别的研究[10~15]，这是因为深度数据对背景环境更具有鲁棒性。最近也有一些研究者对这些人体行为识别的方法展开了调研。但是，他们的研究只是侧重具体的某一方面，例如基于深度数据的方法[16,17]、基于深度学习的方法[18~20]和基于 3D 卷积的方法[21]等。而且，最近已经开发出了许多新的行为识别方法，如基于图卷积神经网络的方法[22~26]。因此，对这些新的方法进行深入的调研是非常有必要的。

这篇文献对人体行为识别的方法进行了全面的综述，并按照所使用的数据类型对这些方法进行了系统的梳理，深入探究了最新的一些研究方法，并将其进行了归纳整理(如图 1 所示)。同时，根据不同的数据类型，分别介绍了一些热门的行为识别数据集，并对这些数据集上的一些经典方法的性能进行了对比分析。最后，还对人体行为识别未来的研究方向进行了展望。这份文献引用了大量收录在重要的学术期刊和计算机视觉会议中的文献，具有很高的认可度和引用率，如 TPAMI、CVIU、CVPR、ICCV、ECCV 等。这篇文献为那些从事人体行为识别研究或对人体行为识别感兴趣的人提供了很好的参考价值。

1　基于 RGB 数据的方法

早期的研究都是基于 RGB 数据进行展开的。一些研究者使用传统的方法，利用机器学习来手动提取行为特征，并选择合适的分类算法进行识别；随着深度学习在图像识别领

收稿日期：2021-07-29；**修回日期：**2021-10-19

作者简介：梁绪(1996-)，男，甘肃兰州人，硕士研究生，主要研究方向为计算机视觉、人体行为识别(18993896708@163.com)；李文新(1966-)，男，甘肃临洮人，研究员，博导，博士，主要研究方向为嵌入式系统及软件设计、系统测试与仿真；张航宁(1997-)，男，山西运城人，硕士研究生，主要研究方向为机械臂视觉伺服控制器.

域取得巨大的进展，研究者们又将深度学习应用于人体行为识别中。

1.1　传统方法

传统方法是手工提取能够代表视频中人体运动的时间和空间变化的行为特征，主要包括基于时空体积的方法、基于时空兴趣点(STIP)的方法和基于轨迹的方法等。这些方法主要采用经典机器学习分类方法来进行人体行为识别，如 BOOST、SVM 和概率图模型等。

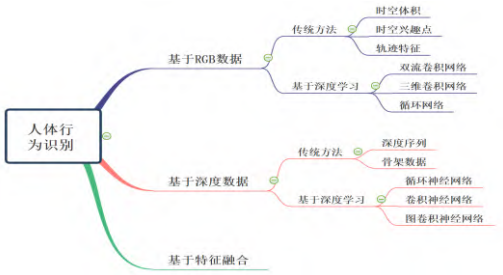

图 1　人体行为识别方法分类

Fig. 1　Classification of human action recognition methods

1) 基于时空体积

基于时空体积的方法主要是模板匹配技术，但与图像处理中的对象识别不同，它们使用三维时空模板进行人体行为识别。这些方法的核心就是构建一个合理的人体行为模板，并基于此模板执行有效的匹配。Bobick 等人[1]最早采用轮廓来描述人体的运动信息，提出用运动能量图(MEI)和运动历史图(MHI)表示行为特征，图 2 展示了 3 类不同行为的 MEI 和 MHI。

Zhang 等人[27]使用极坐标在 MHI 中划分人体的中心区域，并使用基于尺度不变特征转换(SIFT)的运动上下文(MC)描述符来表示行为。Klaeser 等人[2]将图像的梯度(HOG)特征的直方图扩展到时空维度，并使用 3 维的 HOG 特征来描述视频中的人类行为。Somasundaram 等人[28]利用稀疏表示和

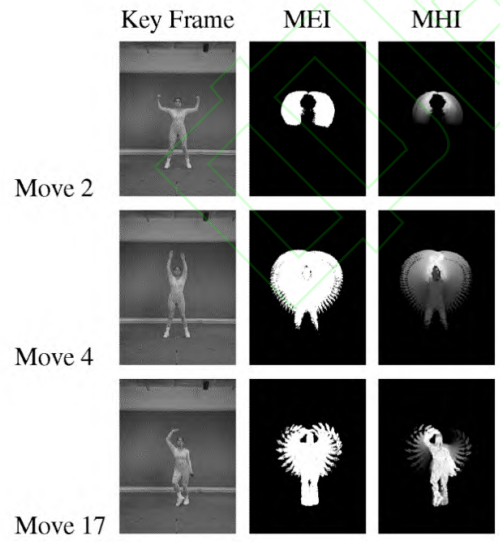

图 2　不同行为的 MEI 和 MHI[1]

Fig. 2　MEI and MHI of different action[1]

字典学习的方法计算视频在时间和空间维度中的相似性，用最显著区域上的时空描述符来表示行为。Patel 等人[29]利用运动目标检测和分割，提取出分割对象的 HOG，并融合目标的速度、位移及区域特征来描述行为。

当相机固定时，这些方法可以使用背景减法技术获取形状信息，如人类剪影和轮廓。然而，在复杂的场景和摄像机移动的情况下，很难获得精确的轮廓形状。而且，在人体被遮挡的情况下，很难识别出精确的人体外观。此外，要确定同一场景中的多个操作，大多数方法都使用滑动窗口，但此方法的计算代价很高。

2) 基于 STIP

基于 STIP 的方法广泛用于行为识别，从视频中提取运动变化的关键区域来表示行为。与图像中目标检测的局部特征类似，STIP 方法必须确定要使用的关键区域检测方法，使用哪个特征矢量来描述关键区域，以及使用哪种分类算法。因此，这些方法大部分是从应用于图像的目标检测方法扩展到视频中的。

STIP 中的“时空兴趣点”通常是指在时空维度中变化最显著的位置[30]，如图 3 所示。经典的 STIP 方法包括 3D-Harris 时空特征点[31]和其改进的技术[32]，主要思想是将特征检测技术从 2D 图像扩展到 3D 时空域，然后计算特征描述符，并学习表示人类行为的可视化字典。此外，Nguyen 等人[3]提出了一种基于时空注意力机制的关键区域提取方法，构建了视觉字典和行为特征；Peng 等人[33]根据局部时空特征和视觉字典构造，对以往的方法进行了回顾和对比，并提出了一种简单而有效的混合表示方法，从而构建出更准确、更高效的行为识别系统；Nazir 等人[34]集成了 3D-Harris 时空特征和 3D-SIFT 检测方法，以提取视频的关键区域，并使用传统的视觉单词直方图来表示人体行为。

基于时空特征的方法引起了许多研究者的注意。主要优点是，此类方法不需要预处理，如背景分割或人体检测。局部特征具有尺度和旋转不变性，在光照变化下稳定，对遮挡

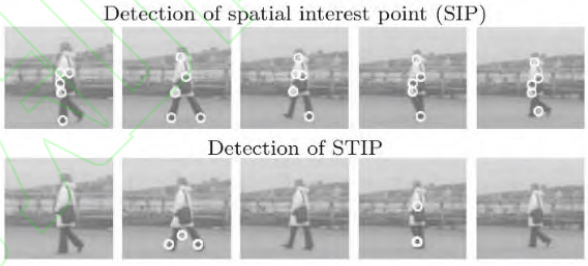

图 3　检测出的时空兴趣点(STIP)[30]

Fig. 3　Detected space-time points of interest (STIP)[30]

的鲁棒性优于其他方法。但时空特征点易受摄像机视角变化的影响。在背景运动和摄像机运动的情况下，基于局部时空特征的方法会产生许多背景特征点，对目标的运动估计会产生较大的误差。

3) 基于轨迹特征

基于轨迹的特征是利用人体骨架中关键点或关节的轨迹来表示行为。这类方法最成功的是 Wang 等人提出的密集轨迹方法(DT)[35]及其改进的方法(IDT)[4]。如图 4 所示，在视频帧中，采样密集的点云，通过光流方法来跟踪这些特征点，计算运动轨迹，并沿着轨迹提取出更有效的运动物体的边界信息(MBH)来描述人体行为。许多研究者也尝试对 IDT 算法进行改进，Gaidon 等人[36]利用分裂聚类分析局部运动轨迹，并利用聚类结果代表不同的运动水平计算人体行为特征；并

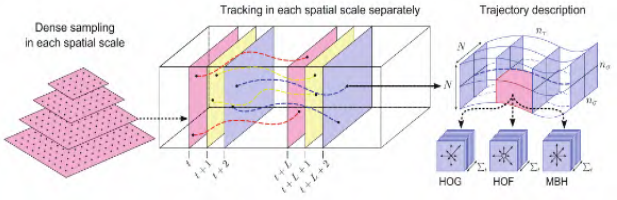

图 4　密集轨迹算法[35]

Fig. 4　Dense trajectory algorithm[35]

利用聚类结果代表不同的运动水平计算人体行为特征；Wang 等人[37]又将人体检测结果融合到 IDT 特征中，用于减少背景轨迹的干扰；Peng 等人[38]基于 IDT 特征使用堆叠 Fisher 向量来表示人的行为，这是改进 IDT 较为成功的方法

之一；Xia 等人[39]对 IDT 的光流算法进行了扩展，并设计了一种多特征融合的描述子来表示行为。

基于轨迹的行为识别方法的主要优点是可以用来分析人体的局部运动信息，而且大多数方法都可以克服视角的变化。然而，这种方法需要一个精确的二维或三维人体骨架模型，并需要精确跟踪关键点。然而，人体建模与跟踪本身仍然是计算机视觉领域的一个具有挑战性的问题。

1.2　基于深度学习的方法

随着深度学习在图像识别领域的不断演进，研究者们也尝试将深度学习应用于人体行为识别。但对于视频来说，卷积神经网络(CNN)仅仅是进行 2D 空间上的特征提取与分类，在时间方面的特征如何提取并与空间特征进行融合，研究者们提出了不同的想法，大致分成了三个流派分支，分别是基于双流网络(Two-Stream)、基于三维卷积网络(C3D)和基于长短时记忆网络(LSTM)的方法。

1) 基于双流卷积网络

在双流卷积网络[5]中，如图 5 所示，光流信息是由图像序列计算得到的。在模型训练过程中，使用图像和光流序列作为两个卷积神经网络(CNNs)的输入，分别提取空间特征和时间特征。特征的融合发生在网络的最后一个分类层。该双流网络的输入为单帧 RGB 图像和堆叠的光流图像，网络采用二维图像卷积。

一些研究者尝试对双流网络进行了性能上的改进。Wang 等人[6]对双流卷积网络的输入、卷积网络结构和训练策略进行了详细的讨论，并提出了一种时间段网络(Temporal Segment Networks, TSN)来进一步改进双流卷积网络的结果。

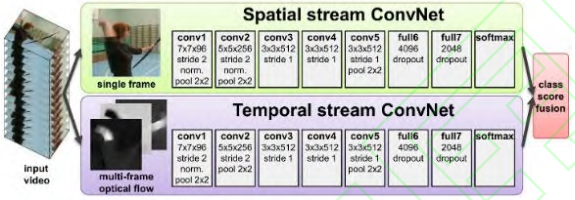

图 5　双流卷积网络[5]

Fig. 5　Two-stream convolutional network[5]

最近，Wang 等人[40]对 TSN 又做了进一步改进，提出了时态差异网络(Temporal Difference Networks, TDN)的视频架构，利用捕获多尺度的时间信息来进行有效的行为识别。TDN 采用了两级差分建模范式，对于局部运动建模，连续帧上的时间差异用于为 2D CNN 提供更精细的运动模式，而对于全局运动建模，跨段的时间差异被结合以捕获运动特征激发的远程结构。Feichtenhofer 等人[41]将时空信息融合的过程从最初的分类层，转移到网络的中间层；Lan 等人[42]将 CNN 提取到的局部标签聚合成全局标签。Zhou 等人[43]设计了一种时间关系网络(Temporal Relation Network, TRN)，用于学习和推理在多个时间尺度下视频帧之间的时间相关性。Feichtenhofer 等人还提出了用于视频识别的 SlowFast 网络[44]，该网络设计了一个低帧率的慢速路径和高帧率的快速路径分别用于提取空间特征和时间特征，在行为识别方面实现的强大的性能。文献[45]采用基于二进制密集 SIFT 流的双流 CNN 代替光流，避免了光流对时间特征的影响。Liu 等人[46]用深度网络模型代替现有模型，将网络结构进行了改进，优化了网络性能。Du 等人[47]提出了一种基于预训练卷积神经网络(CNNs)的双流深度特征提取框架，并采用线性动力系统(LDSs)的方法，在双流体系结构进行人体行为识别的研究。

2) 基于三维卷积网络

和双流网络相比，三维卷积网络[7]则将视频视为三维时空结构，采用三维卷积方法学习人体行为特征，其卷积核和池化核也相应地从 2D 扩展到了 3D，而且其网络结构更加简单，如图 6 所示，运行效率也比双流网络快很多。

同样，也有许多研究者基于三维卷积网络的思想，试图将不同的二维卷积网络扩展到三维时空结构中，以学习和识别人体行为特征。Tran 等人[48]将 3D 卷积滤波器分解为单独的空间和时间分量，提出了一个新的时空卷积块“R(2+1)D”模型。Carreira 等人[8]将 Inception-V1 的网络结构从二维扩

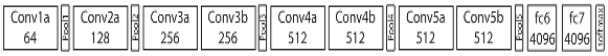

图 6　三维卷积网络[7]

Fig. 6　3D convolutional network[7]

展到三维，提出了用于行为识别的双流膨胀三维卷积网络模型(I3D)。Diba 等人[49]对 DenseNet 进行了扩展，提出了时间三维卷积网络。Qiu 等人[50]针对 C3D 存在所需内存大和计算成本高的问题，提出了用二维空间卷积和一维时间卷积来模拟 3D 卷积的想法(P3D)，并将这种设计整合到一个深度残差学习框架中。Du 等人[51]通过大量实验探索更优的 3D 网络结构，并提出了一种深度三维残差卷积神经网络(R3D)。文献[52]将 2D 卷积和 3D 卷积结合在一起，去除了相邻帧之间存在大量冗余信息，极大地提升了算法的运算性能。Crasto 等人[53]利用两种学习方法来训练一个标准的 3D 卷积网络，在 RGB 帧上运动流，避免了光流计算。Kim 等人[54]提出了一种轻量级的用于人体行为识别的弱监督时间注意 3D 网络(TA3DNet)，以弱监督的方式训练时间注意模块，极大地减少了输入帧的数量。文献[55]在 3D 卷积中嵌入注意机制来克服视频编解码技术造成的模糊特征，证明了卷积核具有信道依赖性。Kumawat 等人[56]提出时空短时傅里叶变换(STFT)块，通过减少网络中参数的数量来提升特征学习的能力，有效的避免了算法过拟合。文献[57]利用残差结构和注意机制对现有的 3D 卷积模型进行了改进，提出了注意残差 3D 网络(AR3D)，加强了人体行为特征的提取。

3) 基于长短时记忆网络

另一种重要的人体行为识别方法涉及到使用 LSTM 和 CNN，如图 7 所示。与使用各种卷积时间特征池架构对行为进行建模的双流和三维卷积网络不同，基于 LSTM 的方法将视频视为有序的帧序列，人的行为可以通过每一帧的特征变化来表示。

Donahue 等人[58]提出了长期循环卷积网络来将可变长度的视频的帧序列映射到可变长度的输出(例如行为描述文本)。此外，Ng 等人[9]提出了一种递归神经网络来识别人体行为，它将 LSTM 细胞与底层 CNN 的输出连接起来；Aghaei 等人[59]设计了卷积注意 LSTM 网络，在每个 LSTM 层之后添加一个稀疏层来克服过拟合，并将注意力机制应用于卷积神经网络，对 LSTM 权值和每层输出进行剪枝处理，使得网络结构朝着更深层次的方向发展。文献[60]将基于注

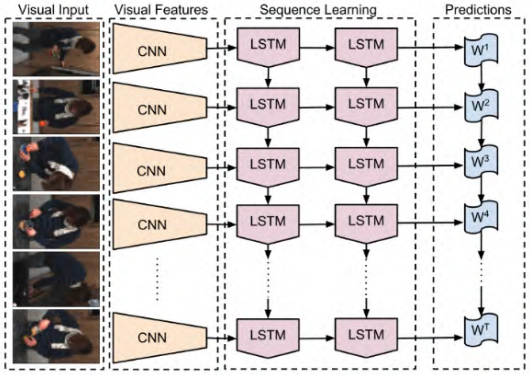

图 7　LSTM+CNN 模型[58]

Fig. 7　LSTM+CNN model[58]

意力机制的卷积长短时记忆神经网络与传统的双流卷积

进行结合，实现了对视频数据中运动信息的非线性特征更好的学习，对局部显著特征及其空间关系更好的利用。

2　基于深度数据的方法

随着微软 Kinect 等一些深度传感器的性能不断提升，由结构光传感器生成的深度图像对于光照条件的变化更具鲁棒性，使用深度相机还可以更容易地从杂乱的背景中减去前景，从而忽略来自杂乱背景的混乱纹理。于是，研究者们又开始将人体行为识别的研究和深度数据结合起来，试图解决背景遮挡、光照变化等干扰因素对人体行为识别的影响。此外，还有一些研究者们同样受到深度学习的启发，将深度学习和骨架数据联系起来，也取得不错的效果。

2.1　传统方法

基于深度数据的传统方法原理和前文 1.1 节所提到的基于 RGB 数据的传统方法大致相同，都是手工提取能够描述人体行为的特征，然后选择合适的分类算法进行人体行为识别。基于深度数据的传统方法主要分为基于深度序列和基于骨架数据的方法以及特征融合的方法。

1) 基于深度序列

基于深度序列的方法主要使用人体深度图中的运动变化来描述行为。在 RGBD 视频中，深度数据可以看做是由深度信息组成的时空结构。从图 8 中可以看出，人体行为的特征表示就是在这种时空结构中进行特征提取的过程，选择具有深度变化的外观和运动信息来描述行为。

Yang 等人[10]基于深度图序列构造了一个超常向量特征(Super Normal Vector, SNV)来表示行为。Oreifej 等人[61]将 HOG 特征扩展到时空深度结构，提出了四维法向量的方向直方图特征来表示三维时空深度结构的外观信息。Rahmani 等人[62]提出了一种基于深度曲面主方向的行为表示方法，根据主方向旋转视频的视角，计算与视角无关的行为特征表示，并且使用主分量方向直方图来表示行为。

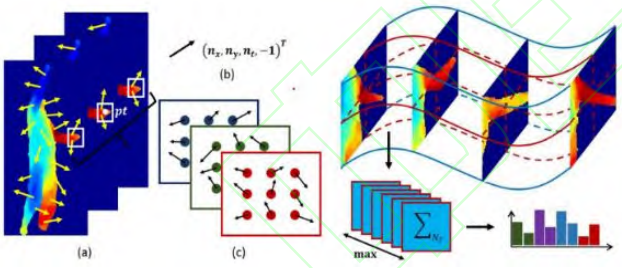

图 8　基于深度序列的方法[10]

Fig. 8　Method based on depth sequence[10]

上述方法都使用外观信息来描述深度数据中的人体行为。也有一些研究者尝试从深度信息中计算运动数据来表示行为。Yang 等人[11]提出了深度运动图(DMM)，从正视图、侧视图和俯视图投影和压缩时空深度结构，形成三个运动历史图。然后用 HOG 特征表示这些运动历史图，得到的特征串联起来描述行为。Chen 等人[12]则使用局部二进值模式特征代替 HOG 来描述基于 DMM 的人类行为。Chen 等人[63]还分析了正视、侧视、俯视方向的时空深度结构，提取时空兴趣点的运动轨迹形状和边界直方图特征，并使用每个视图中的密集样本点和关节点来描述行为。此外，Miao 等人[64]使用离散余弦变化来压缩深度图，并使用变换系数来构造行为特征。Bulbul 等人[65]提出了一种使用 3D 时空梯度自相关(STACOG) 算法的深度图序列行为识别框架，用多个 DMM 序列输入 STACOG 框架中计算得到的自相关特征向量代替从 DMM 中直接获得的特征向量。Ji 等人[66]提出一个简单高效的基于深度图序列的人体行为建模框架，设计了一种深度方向梯度向量(DOGV) 的帧级特征，用于捕捉短时期内的外观和运动。

2) 基于骨架数据

从深度数据中可以快速、准确地估计出人体骨架[13]。基于骨架的行为识别方法是利用深度数据的另一个热门的研究领域。如图 9 所示，基于人体骨架序列的方法利用视频中各帧之间的人体骨架节点的变化来描述行为，包括骨架节点的位置和外观变化。

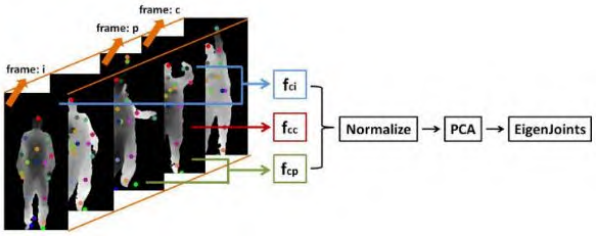

图 9　基于骨架数据的方法[69]

Fig. 9　Method based on skeleton data[69]

Xia 等人[67]使用 3D 关节点直方图来表示人体的姿势，并通过离散的隐马尔可夫模型(HMM)对人体行为进行建模。Keceli 等人[13]利用 Kinect 传感器获取深度和人体骨架信息，然后根据骨架关节点的角度和位移信息提取人体行为特征。Gowayyed 等人[68]利用方向位移直方图(HOD)描述骨架节点的轨迹，从前部、侧面和顶部视图提取出 HOD 特征，形成三维 HOD 特征。Yang 等人[69]提出了特征关节方法，使用累积运动能量(AME)函数选择视频帧和更多信息关节点来模拟行为。Pazhoumanddar 等人[70]利用最长公共子序列(LCS)算法从骨架相对运动轨迹中选择高分辨能力特征来描述相关行为。Nguyen 等人[71]提出了两种不同的最大信息量关节的选择方案，自适应最大信息量关节数和固定最大信息量关节数，并设计了一种新的基于联合速度的时间协变特征描述子。

综上所述，利用骨架节点轨迹的方法可以从不同角度得到骨架节点之间的对应关系。因此，在透视图转换的情况下可以提取出更健壮的行为特征。然而，这些方法的性能取决于人体姿态估计的结果。当场景中出现遮挡时，会导致骨架节点估计缺失或错误，从而影响行为识别结果。

2.2　基于深度学习的方法

基于深度学习的方法在 RGB 数据中取得了很好的效果，性能远超于传统的手工提取特征的方法。另一方面，深度骨架序列具有丰富的空间信息和时域信息。于是，有许多研究者也尝试将深度学习和骨架数据结合起来，用于人体行为识别。

1) 基于循环神经网络

循环神经网络(RNN)是一种处理序列数据非常有效的方法，可以将上一时刻的输出作为当前时刻的输入来形成其结构内部的递归连接。此外，LSTM 和门控循环单元(GRU)等模型，在 RNN 内部引入了门空单元和线性记忆单元解决了梯度消失问题和长时建模等问题。于是，研究者们试图利用 RNN 进行人体行为识别的研究。

Liu 等人[14]提出了一种用于三维人体行为识别的时空 LSTM 模型，该模型将 RNN 扩展到时空域，分析与行为相关的隐藏信息源。Li 等人[72]提出了 RNN 树(RNN-t)的模型，这是一种自适应学习框架，使用多个循环神经网络(RNNs)构成一个树状的层次，有效地解决了单一网络难以处理的细粒度行为类的难题。Wang 等人[73]提出了一个新颖的双流 RNN 结构来为骨架数据建模时域和空域特征，有效解决了利用 RNN 处理原始骨架的方法时，忽略骨架关节空间构型的问题。文献[74]提出了全局上下文感知注意 LSTM (GCA-LSTM)，用于三维行为识别，该网络能够结合全局上下文信息来选择性地聚焦行为序列中的信息关节。Lee 等人[75]提出了一种基于骨架的集成时态滑动 LSTM (TS-LSTM) 网络，将骨架转换为

另一个坐标系，提升了缩放、旋转和平移的鲁棒性，然后从中提取显著的运动特征。Song 等人[76]基于 LSTM 提出了一种时空注意模型，,该模型选择性地关注每一帧中骨架的判别性关节，并对不同帧的输出给予不同程度的关注，从骨架数据中探索人体行为识别和检测的时空鉴别特征。文献[77]针对骨架节点在三维空间中的复杂变化，提出了记忆注意网络(MANs)对骨架节点进行时空重标定，并利用时间注意重标定模块(TARM)和时空卷积模块(STCM)对 MANs 进行了部署。Zhang 等人[78]提出了一种简单而有效的元素注意门(EleAttG)，它可以方便地添加到 RNN 的任意神经元中，使 RNN 的神经元具有注意能力，能够自适应地调节模型的输入。文献[79]提出了独立循环神经网络(IndRNN)，有效解决了 RNN 梯度消失和爆炸的问题，并且支持网络学习长期依赖关系，有效地扩展的 RNN 的网络层数。Zhang 等人[80]还提出了一种基于学习的数据驱动方式自动确定行为过程中的虚拟观测视点的视图自适应方案，在很大程度上消除了视图变化的影响，使得网络能够专注于对特定行为特征的学习，从而获得更好的性能。此外，还有研究者尝试用无监督的方式学习行为特征[81]。

2) 基于卷积神经网络

与 RNN 不同，CNN 模型凭借其优秀的高级信息提取能力，能够高效、轻松地学习高级语义线索。而且，有大量的研究证明 CNN 在 RGB 数据中已经取得了不错的行为识别效果。于是，许多研究者也试图将 CNN 应用于骨架数据中。为了满足神经网络输入的需要，将三维骨架序列数据从矢量序列转换为伪图像。然而，要同时表达空间和时间信息并不容易，因此很多研究者将骨架关节编码成多个二维伪图像，然后输入 CNN 学习有用的特征[82,83]。

Ke 等人[15]将骨架序列转换为基于图像的表征，利用 CNN 进行时空信息学习，允许对骨架序列进行全局长期时间建模，并提出一种多任务学习网络(MTLN)来联合处理所有时间步长的特征向量来学习骨架序列的空间结构和时间信息。Bo 等人[84]将三维骨架视频映射到彩色图像，提出了一种多尺度深度卷积神经网络(CNN)来进行行为识别，该网络可以增强模型的时间频率调整能力。Li 等人[85]提出了一种端到端卷积共现特征学习框架，并引用全局空间聚合方案来学习不同层次上的共现特征。文献[86]提出了基于骨架序列三维坐标的剪辑表示方法和多任务卷积神经网络(MTCNN)特征学习算法来探索骨架序列的时空信息。文献[87]提出了一种利用几何代数从骨架序列中学习形状-运动表示方法，在关注孤立关节的坐标的同时也考虑了关节之间的空间关系，全面地描述骨架行为。Caetano 等人[88]基于运动信息提出 SkeleMotion 的表示方法，该方法通过显式计算关节运动的幅度和方向值来编码时间动态信息。文献[89] 提出了树结构参考关节图像(TSRJI)模型，这种新的骨架图像表示方法结合参考关节和树型结构骨架的优点，有效地学习了骨架关节之间空间关系。文献[90]设计了一种 CNN 融合模型，用于识别可持续智能家居中的骨架人体行为，通过灰度值编码将每个三维骨架序列的时空信息输入 CNN 融合模型中进行骨架行为识别。

3) 基于图卷积神经网络

图卷积神经网络(GCN)是一种能对图数据进行深度学习的方法。人体 3D 骨架数据是自然的拓扑图，顶点表示关节，边表示连接关节的肢节，因此可以用图卷积网络来发掘骨架之间的空间联系，将图卷积操作拓展到时域上，就能同时发掘空间和时间特征。因此，越来越多的研究者将 GCN 用到骨架行为识别研究中。

Yan 等人[22]首次提出了一种基于骨架行为识别的时空图卷积网络(ST-GCN)的新模型，如图 10 所示，该网络首先将人的关节作为时空图的顶点，将人体连通性和时间作为图的边；然后使用标准 Softmax 分类器来将 ST-GCN 上获取的高级特征图划分为对应的类别。这项工作让更多人关注到使用 GCN 进行骨架行为识别的优越性，因此最近出现了许多相关工作。

Shi 等人[23]将骨架数据表示为基于自然人体中关节和骨骼之间的运动学依赖性的有向无环图。并设计出一种新颖的有向图神经网络专门用于提取关节、骨骼及其关系的信息，

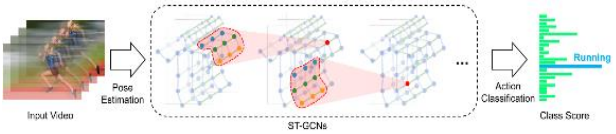

图 10　时空图卷积网络(ST-GCN)模型[22]

Fig. 10　Spatial-temporal graph convolutional network (ST-GCN) model[22]

根据提取的特征进行预测。Li 等人[91]设计了 A-link 推理模块，可以直接从行为中捕获特定于行为的潜在依赖关系，并扩展了现有的骨架图来表示高阶依赖关系，然后将两种类型的连接组合成一个广义的骨架图，进一步提出了行为结构图卷积网络(AS-GCN)，它将行为结构图卷积和时间卷积作为基本构建块，以学习空间和时间行为识别功能。Shi 等人[24]提出了一种新颖的双流自适应图卷积网络(2s-AGCN)，用于基于骨架的行为识别，模型中的图的拓扑既可以通过 BP 算法统一学习，也可以通过端到端的方式单独学习。这种数据驱动的方法增加了图构建模型的灵活性，使其具有更强的通用性，以适应不同的数据样本。Si 等人[92]设计了一种新的注意力增强图卷积 LSTM 网络(AGC-LSTM)，用于从骨架数据中识别人体行为。AGC-LSTM 不仅可以捕获空间形态和时间动态的判别特征，还可以探索时空域的共现关系。Song 等人[93]提出了一种多流图卷积网络模型，用于探索分布在所有骨架关节上的足够多的判别特征，该模型被称为丰富激活 GCN (RA-GCN)，所激活的关节明显比传统方法多，进一步提升了模型的鲁棒性。Zhang 等人[94]提出了一种简单而有效的语义引导神经网络(SGN)，用于基于骨架的行为识别，成功将关节的高级语义(关节类型和帧索引)引入到网络中以增强特征表示能力。Peng 等人[25]对 ST-GCNs 进行了理论和实验分析，提出了捕获全局图的策略，高效地建模输入图序列的表示。此外，全局图策略还可以将图序列降维到欧氏空间，从而引入多尺度时间滤波器来有效地捕获动态信息。张家想等人[26]提出了一种将时空注意力机制和自适应图卷积网络相结合的人体骨架行为识别方案，可以更好的捕捉时空特征和全局上下文信息。Hao 等人[95]提出了一种超图神经网络(Hyper-GNN)来捕获基于骨架的行为识别的时空信息和高阶依赖，通过构建“超边”结构来提取局部和全局结构信息，消除了无关节点带来的噪声影响。Xu 等人[96]设计了一种用于基于骨架的行为识别的多尺度骨架自适应加权图卷积网络(MS-AWGCN)模型，用于提取骨架数据中更丰富的空间特征，并结合图顶点融合策略，将手工绘制的邻接矩阵替换为可学习，自适应地学习潜在的图拓扑结构，最后采用加权学习方法聚合并丰富特征。李扬志等人[97]提出了基于时空注意力图卷积网络(STA-GCN)模型，STA-GCN 包含空间注意机制和时间注意机制，可以同时捕捉空间构造和时间动态的判别特征，探索时空域之间的关系。

3　基于特征融合的方法

对比 RGB 数据和深度数据的行为特征，RGB 数据的优势是外观信息丰富，而深度数据可以更准确地描述人体的运动特征。研究结果[98]表明，基于深度信息的方法可以实现实

时的行为识别，识别性能优于基于 RGB 的方法。因此，一些研究者也尝试使用多特征融合来建模人体行为。

Chaaraoui 等人[99]试图将骨架节点特征和深度序列特征合并，克服遮挡和视角变化导致的骨架特征误差。此外，Li 等人[100]提出了一种基于关节点组的稀疏回归学习方法，将关节和深度特征融合在建模行为中。Althloothi 等人[101]在频域计算了深度信息的球面谐波表示，并通过多核学习方法将其与骨架关节的位置信息融合在一起来建模人体行为。文献[102]将深度信息和骨架信息进行融合，形成了一种新的历史点云轨迹特征。

除了深度序列与骨架特征的融合外，一些研究者还尝试将 RGB 数据中的特征与深度数据中的特征融合。Ni 等人[103]提出了 RGB 数据和深度数据特征的多级融合策略。Jalal 等人[104]将 RGB 数据和深度数据中的时空特征进行了融合。Xu 等人[105]提出一种基于人体骨架和场景图像的双流模型，该模型充分利用了骨架信息在运动表达中的优势和图像在场景表达中的优势，将场景信息与基于时空图卷积的人体骨骼肢体进行互补融合，克服了不稳定的光照条件对人体行为识别算法性能的影响。周雪雪等人[106]提出了基于多模态特征学习的人体行为识别算法，从视频中分别提取 RGB 特征和 3D 骨架特征，并将其进行了融合。

总的来说，基于特征融合的方法试图利用不同数据之间的优势，以获得更稳健的特征表示。因此，在开发大多数基于特征融合的方法时，主要考虑如何将不同数据类型的特征进行有效的融合。常见的方法包括早期融合和晚期融合。前者的融合是在特征层面进行的，被称为特征级融合，作为识别模型的输入；后者在识别模型的输出评分层面进行的融合，被称为决策级融合。然而，大多数基于多模态数据融合的方法往往优于单一数据方法的识别结果。但是，多模态数据融合意味着需要处理的数据量更大，特征维数会更高。这些因素在一定程度上也增加了行为识别算法的计算复杂度。

4　行为识别数据集

随着人们对人体行为识别的不断探索，大量的与行为识别相关的数据集被创建，用于评估和检测算法的性能。本文按照数据集的数据类型，将其划分为 RGB 数据集和深度骨架序列数据集。

4.1　RGB 数据集

早期人体行为识别的研究都是基于 RGB 数据的，用来检验算法性能的数据集也都是 RGB 数据集。表 1 列出了目前一些比较热门的 RGB 数据集的基本信息，包括数据集的数据模态、采集数量和类别数。

表 1　RGB 数据集

Tab. 1　RGB data set

数据集	数据模态	采样数量	类别
KTH[107]	RGB	2391	6
HMDB51[108]	RGB	6766	51
UCF50[109]	RGB	6618	50
UCF101[110]	RGB	13320	101

表 2 对比分析了这些数据集上的一些经典方法的识别性能，其中上标 D 表示的是基于深度学习的方法。从表中可以看出，基于深度学习的方法在各个数据集上均取得了最好的成绩，且基于深度学习的方法性能普遍优于传统方法(例如，传统方法在 HMDB51 和 HMDB51 数据集上精度不超过 70%和 95%；而基于深度学习的方法在 HMDB51 上普遍超过 70%，文献[54]的精度最高达到 83.8%，在 HMDB51 上普遍超过 95%，文献[8]的精度最高达到 98.0%)。从表 2 中看出，基于深度学习的方法在人体行为识别中取得了显著的成绩。

表 2　RGB 数据集上不同算法的性能

Tab. 2　Performance of different algorithms on RGB datasets

方法	年份	KTH	HMDB51	UCF50	UCF101
[4]	2013	—	57.2%	91.2%	85.9%
[38]	2014	—	66.79%	—	—
[33]	2016	—	61.1%	92.3%	87.9%
[37]	2016	—	60.1%	91.7%	86%
[34]	2018	91.82%	—	—	—
[29]	2020	95.85%	54.19%	—	89.32%
[39]	2021	94.88%	69.32%	—	94.66%
[5]^{D}	2014	—	59.4%	—	88%
[7]^{D}	2015	—	—	—	85.2%%
[9]^{D}	2015	—	—	—	77.3%
[40]^{D}	2016	—	76.3%	—	97.4%
[41]^{D}	2016	—	65.4%	—	92.5%
[8]^{D}	2017	—	80.9%	—	**98.0%**
[49]^{D}	2017	—	63.5%	—	93.2%
[50]^{D}	2017	—	—	—	88.6%
[48]^{D}	2018	—	78.1%	—	97.3%
[52]^{D}	2018	—	68.4%	—	93.6%
[53]^{D}	2019	—	75%	—	95.8%
[46]^{D}	2021	98.83%	—	—	—
[47]^{D}	2021	—	62.56%	**96.71%**	96.15%
[54]^{D}	2021	—	**83.8%**	—	97.2%
[59]^{D}	2021	**98.41%**	71.62%	—	95.24%
[60]^{D}	2021	—	69.8%	—	94.6%

表 3 分析对比了各个深度学习模型在 HMDB51 数据集和 UCF101 数据集上的性能。从表 3 中可以看出 3 个深度学习模型在 UCF101 数据集上性能差别不大(双流网络模型最高达到 97.4%[40]，三维卷积模型最高达到 98.0%[8]，LSTM 模型最高达到 95.24%[59])；在 HMDB51 数据集上存在较大差异(双流网络模型最高达到 76.3%[40]，三维卷积模型最高达到 83.8%[54]，LSTM 模型最高达到 71.62%[59])。从整体来看，基于三维卷积的深度学习模型在 3 类深度学习模型中性能最佳。

表 3　RGB 数据集上各个深度学习模型的比较

Tab. 3　Comparison of each deep learning model on RGB data set

方法		年份	HMDB51	UCF101
基于双流网络	[5]^{D}	2014	59.4%	88%
	[40]^{D}	2016	76.3%	97.4%
	[41]^{D}	2016	65.4%	92.5%
	[46]^{D}	2021	—	—
	[47]^{D}	2021	62.56%	96.15%
基于三维卷积	[7]^{D}	2015	—	85.2%%
	[8]^{D}	2017	80.9%	**98.0%**
	[49]^{D}	2017	63.5%	93.2%
	[50]^{D}	2017	—	88.6%
	[48]^{D}	2018	78.1%	97.3%
	[52]^{D}	2018	68.4%	93.6%
	[53]^{D}	2019	75%	95.8%
	[54]^{D}	2021	**83.8%**	97.2%
基于 LSTM	[9]^{D}	2015	—	77.3%
	[59]^{D}	2021	71.62%	95.24%
	[60]^{D}	2021	69.8%	94.6%

4.2　深度骨架序列数据集

随着微软 Kinect 等一些深度传感器的应用，也产生了许多深度骨架序列数据集。表 4 列出了一些常用的深度骨架序列数据集的基本信息，包括数据集的数据模态、采集数量和类别数。

表 4　深度骨架序列数据集

Tab. 4　Deep skeleton sequence data set

数据集	数据模态	采样数量	类别
MSR-Action3D[111]	Depth	567	20
Kinetics[112]	RGB + Depth	300000	400
NTU-RGB+D[113]	RGB + Depth	56880	60
NTU-RGB+D 120[114]	RGB + Depth	114480	120

注：Kinetics 数据集本身只提供 RGB 数据，其深度骨架数据由 Yan 等人[22]提供。

表 5 整理了这些深度骨架序列数据集上的一些最新的研究成果，其中上标 D 表示的是基于深度学习的方法。通过分析表中数据，传统方法主要集中于规模较小的 MSR-Action3D 数据集，并且取得了不错的效果，(在 MSR-Action3D 数据集上几乎全部达到 90%以上，最优的精度达到了 96.2%[98]，与深度学习的方法仅差 1.02%)；而基于深度学习的方法则更适合在规模较大的数据集(在 Kinetics 数据集上达到最高精度 59.8%[26]、NTU-RGB+D 数据集上达到最高精度 96.2%[26]和 NTU-RGB+D 120 数据集上达到最高精度 85.03%[66])，这是因为基于深度学习的方法需要足够多的数据来训练模型，且经过大规模数据集训练后的深度模型在 MSR-Action3D 数据集上也取得了比传统方法更好的成绩(例如，文献[75]在 MSR-Action3D 数据集上达到了 97.22%的识别精度)。

表 5　深度骨架数据集上不同算法的性能

Tab. 5　Performance of different algorithms on depth skeleton data set

方法	年份	MSR-Action3D	Kinetics	NTU-RGB+D	NTU-RGB+D 120
[10]	2014	93.09%	—	—	—
[69]	2014	83.3%	—	—	—
[12]	2015	94.9%	—	—	—
[98]	2016	96.2%	—	—	—
[100]	2016	94.3%	—	—	—
[65]	2021	93.4%	—	—	—
[14]^D	2016	—	—	77.7%	—
[74]^D	2017	—	—	82.8%	—
[75]^D	2017	**97.22%**	—	81.25%	—
[82]^D	2017	—	—	82.31%	—
[84]^D	2017	—	—	90.9%	—
[22]^D	2018	—	52.8%	88.3%	—
[24]^D	2018	—	58.7%	95.1%	—
[79]^D	2018	—	—	87.97%	—
[83]^D	2018	—	—	91.3%	—
[23]^D	2019	—	59.6%	96.1%	—
[80]^D	2019	—	—	95%	—
[87]^D	2019	—	—	90.05%	—
[88]^D	2019	—	—	84.7%	66.9%
[89]^D	2019	—	—	80.3%	67.9%
[91]^D	2019	—	56.5%	94.2%	—
[92]^D	2019	—	—	89.2%	—
[94]^D	2020	—	—	94.5%	81.5%
[25]^D	2021	—	55.2%	95.9%	81.7%
[26]^D	2021	—	**59.8%**	**96.2%**	—
[66]^D	2021	—	—	90.54%	**85.03%**
[90]^D	2021	—	—	95.2%	—
[97]^D	2021	—	57.3%	95%	—

表 6 对比分析了近年来，基于深度骨架数据集上的各个深度学习模型算法的性能。从表 6 中可以看出，这 3 类深度学习模型在 NTU-RGB+D 数据集上表现出的整体性能差别不大(RNN 模型最高达到 95%[80]，CNN 模型最高达到 95.2%[90]，GCN 模型最高达到 96.2%[26])；在 NTU-RGB+D 120 数据集上，最新 CNN 模型的性能优于 GCN 模型(文献[66]识别精度达到 85.03%，文献[25]的识别精度仅有 81.7%，二者相差 3.33%)。此外，随着 Yan 等人[22]为 Kinetics 数据集提供了原始的骨架数据，基于 GCN 的学习模型在 Kinetics 数据集上的性能由最初的 52.8%[22]上升到 59.8%[26]。

表 6　深度骨架数据集上各个深度学习模型的比较

Tab. 6　Comparison of each deep learning model on the deep skeleton data set

方法		年份	Kinetics	NTU-RGB+D	NTU-RGB+D 120
基于 RNN	[14]^D	2016	—	77.7%	—
	[74]^D	2017	—	82.8%	—
	[75]^D	2017	—	81.25%	—
	[79]^D	2018	—	87.97%	—
	[80]^D	2019	—	95%	—
	[82]^D	2017	—	82.31%	—
	[84]^D	2017	—	90.9%	—
	[83]^D	2018	—	91.3%	—
基于 CNN	[87]^D	2019	—	90.05%	—
	[88]^D	2019	—	84.7%	66.9%
	[89]^D	2019	—	80.3%	67.9%
	[66]^D	2021	—	90.54%	**85.03%**
	[90]^D	2021	—	95.2%	—
基于 GCN	[22]^D	2018	52.8%	88.3%	—
	[24]^D	2018	58.7%	95.1%	—
	[23]^D	2019	59.6%	96.1%	—
	[91]^D	2019	56.5%	94.2%	—
	[92]^D	2019	—	89.2%	—
	[94]^D	2020	—	94.5%	81.5%
	[25]^D	2021	55.2%	95.9%	81.7%
	[26]^D	2021	**59.8%**	**96.2%**	—
	[97]^D	2021	57.3%	95%	—

5　结束语

本文全面的梳理了人体行为识别的方法，按照所使用的数据类型对方法进行了系统地归纳和总结，并对各种方法做了相关的分析和讨论，指出了各类方法的优缺点，还介绍了一些主流的人体行为识别数据集。大量的研究结果表明，绝大多数基于深度学习的方法在性能上均优于传统方法。由此可见，未来的行为识别研究方向应该是集中于基于深度学习的方法。但是深度学习算法需要花费大量的时间和大规模的数据集来训练模型，而在视频监控和家庭监护等实际应用中往往需要实时性的进行人体行为识别。如何快速有效的应用到实际工作中则是未来研究工作的重难点。

人体行为识别的关键是提取鲁棒性的行为特征，包括空间特征和时间特征。因此本文对该领域未来的研究方向有以下推测：

a) 基于多流网络的行为识别。双流卷积神经网络可以有效的从视频中的提取时间和空间特征，从而取得了很好的识别效果。但是设计的网络越多，模型会越复杂，进一步导致计算量的增加。因此，如何设计多流网络提取有效的特征是未来研究的重难点之一。

b) 基于特征融合的方法。基于数据融合的方法试图利用不同数据之间的优势，以获得更稳健的特征表示。大多数基于多模态数据特征融合的方法往往优于单一数据特征方法的识别结果。但是，多模态数据特征融合需要处理的数据量更

大，特征维数更高，在一定程度上增加了计算复杂度。因此，选择哪些特征以及采用哪种的融合方法和策略是未来该领域需要重点研究的方向。

c) 基于注意力机制的深度神经网络。从拟合误差来看，大多数情况下，深层网络比浅层网络更有效。此外，循环神经网络可以有效的处理序列数据，LSTM 和 GRU 有效的解决了 RNN 梯度消失的问题，结合注意力机制可以使得网络结构朝着更深层次发展。因此，如何结合注意力机制设计深层网络也是未来人体行为识别领域中的具有挑战性的课题。

d) 基于图卷积神经网络的方法。相较于 RGB 数据，深度骨架数据对复杂场景具有更强的鲁棒性。人体 3D 骨架数据本身就可以被看做为一个自然的拓扑图数据结构，顶点表示关节，边表示连接关节的肢节，GCN 能对图数据进行深度学习。此外，随着 Kinect 等深度传感器的应用，产生了 NTU-RGB+D 等大型深度骨架序列数据集，基于 GCN 的方法将会是未来十分热门的研究方向。

参考文献：

[1] Bobick A F, Davis J W. The recognition of human movement using temporal templates [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23 (3): 257-267.

[2] Klser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients [C]// Proc of the 19th British Machine Vision Conference. [S. l.] : BMVA Press, 2008: 99. 1-99. 10.

[3] Chakraborty B, Holte M B, Moeslund T B, *et al.* Selective spatio-temporal interest points [J]. Computer Vision & Image Understanding, 2012, 116 (3): 396-410.

[4] Wang H, Schmid C. Action recognition with improved trajectories [C]// Proc of IEEE ICCV. Piscataway, NJ: IEEE Press, 2013: 3551-3558.

[5] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C]// Proc of the 27th Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 568-576.

[6] Wang Limin, Xiong Yuanjun, Wang Zhe, *et al.* Temporal segment networks: towards good practices for deep action recognition [C]// Proc of ECCV. Berlin: Springer, 2016: 20-36.

[7] Tran D, Bourdev L, Fergus R, *et al.* Learning spatiotemporal features with 3D convolutional networks [C]// Proc of IEEE ICCV. Piscataway, NJ: IEEE Press, 2015: 4489-4497.

[8] Carreira J, Zisserman A. Quo vadis, action recognition?A new model and the Kinetics dataset [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2017: 4724-4733.

[9] Ng Y H, Hausknecht M, Vijayanarasimhan S, *et al.* Beyond short snippets: deep networks for video classification [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2015: 4694-4702.

[10] Yang Xiaodong, Tian Yingli. Super normal vector for activity recognition using depth sequences [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2014: 804-811.

[11] Yang Xiaodong, Zhang Chenyang, Tian Yingli. Recognizing actions using depth motion maps-based histograms of oriented gradients [C]// Proc of the 20th Acm International Conference on Multimedia. New York: ACM Press, 2012: 1057-1060.

[12] Chen, Jafari R, Kehtarnavaz N. Action recognition from depth sequences using depth motion maps-based local binary patterns [C]// Proc of IEEE WACV, Piscataway, NJ: IEEE Press, 2015: 1092-1099.

[13] Keceli A S, Can A B. Recognition of basic human actions using depth information [J]. International Journal of Pattern Recognition & Artificial Intelligence, 2014, 28 (2): 1450004. 1-1450004. 21.

[14] Liu Jun, Shahroudy A, Xu Dong, *et al.* Spatio-temporal LSTM with trust gates for 3D human action recognition [C]// Proc of ECCV. Berlin: Springer, 2016: 816-833.

[15] Ke Qiuhong, Bennamoun M, An S, *et al.* A new representation of skeleton sequences for 3D action recognition [C]// Proc of IEEE CVPR, IEEE, 2017: 1063-6919.

[16] Wang Lei, Huynh D Q, Koniusz P. A comparative review of recent Kinect-based action recognition algorithms [J]. IEEE Trans on Image Processing, 2019, 29 (1): 15-28.

[17] 孙彬, 孔德慧, 张雯晖, 等. 基于深度图像的人体行为识别综述 [J]. 北京工业大学学报, 2018, 44 (10): 1353-1368. (Sun Bin, Kong Dehui, Zhang Wenhui, *et al.* Survey on human action recognition from depth maps [J]. Journal of Beijing University of Technology, 2018, 44 (10): 1353-1368.)

[18] 钱慧芳, 易剑平, 付云虎. 基于深度学习的人体动作识别综述 [J]. 计算机科学与探索, 2021, 15 (3): 438-455. (Qian Huifang, Yi Jianping, Fu Yunhu. Review of human action recognition based on deep learning [J]. Journal of Frontiers of Computer Science and Technology, 2021, 15 (3): 438-455.)

[19] 赫磊, 邵展鹏, 张剑华, 等. 基于深度学习的行为识别算法综述 [J]. 计算机科学, 2020, 47 (6A): 139-147. (Hao Lei, Shao Zhanpeng, Zhang Jianhua, *et al.* Review of deep learning-based action recognition algorithms [J]. Computer Science, 2020, 47 (6A): 139-147.)

[20] 蔡强, 邓毅彪, 李海生, 等. 基于深度学习的人体行为识别方法综述 [J]. 计算机科学, 2020, 47 (4): 85-93. (Cai Qiang, Deng Yibiao, Li Haisheng, *et al.* Survey on human action recognition based on deep learning [J]. Computer Science, 2020, 47 (4): 85-93.)

[21] 黄海新, 王瑞鹏, 刘孝阳. 基于 3D 卷积的人体行为识别技术综述 [J]. 计算机科学, 2020, 47 (11A): 139-144. (Huang Haixin, Wang Ruipeng, Liu Xiaoyang. Review of human action recognition technology based on 3D convolution [J]. Computer Science, 2020, 47 (11A): 139-144.)

[22] Yan Sijie, Xiong Yuanjun, Lin Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]// Proc of the 32nd American Association for Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 7444-7452.

[23] Shi Lei, Zhang Yifan, Cheng Jian, *et al.* Skeleton-based action recognition with directed graph neural networks [C]// Proc of IEEE/CVF CVPR. Piscataway, NJ: IEEE Press, 2019: 7904-7913.

[24] Shi Lei, Zhang Yifan, Cheng Jian, *et al.* Two-stream adaptive graph convolutional networks for skeleton-based action recognition [C]// Proc of IEEE/CVF CVPR. Piscataway, NJ: IEEE Press, 2019: 12026-12035.

[25] Peng Wei, Shi Jingang, Varanka T, *et al.* Rethinking the ST-GCNs for 3D skeleton-based human action recognition [J]. Neurocomputing, 2021, 454: 45-53.

[26] 张家想, 刘如浩, 金辰曦, 等. 结合时空注意力机制和自适应图卷积网络的骨架行为识别 [J]. 信号处理, 2021, 37 (7): 1226-1234. (Zhang Jiaxiang, Liu Ruhao, Jin Chenxi, *et al.* Skeleton-based action recognition on spatio-temporal attention mechanism and adaptive graph convolutional network [J]. Journal of Signal Processing, 2021, 37 (7): 1226-1234.)

[27] Zhang Ziming, Hu Yiqun, Chan S, *et al.* Motion context: a new representation for human action recognition [C]// Proc of ECCV, Berlin: Springer, 2008: 817-829.

[28] Somasundaram G, Cherian A, Morellas V, *et al.* Action recognition using global spatio-temporal features derived from sparse representations [J]. Computer Vision & Image Understanding, 2014, 123: 1-13.

[29] Patel C I, Labana D, Pandya S, *et al.* Histogram of oriented gradient-

based fusion of features for human action recognition in action video sequences [J]. Sensors, 2020, 20 (24): 7299-7330.

[30] Das Dawn D, Shaikh S H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector [J]. Visual Computer, 2016, 32 (3): 289-306.

[31] Nguyen T V, Song Zheng, Yan Shuicheng. STAP: spatial-temporal attention-aware pooling for action recognition [J]. IEEE Trans on Circuits and Systems for Video Technology, 2015, 25 (1): 77-86.

[32] Laptev I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64 (2-3): 107-123.

[33] Peng Xiaojiang, Wang Limin, Wang Xingxing, *et al.* Bag of visual words and fusion methods for action recognition: comprehensive study and good practice [J]. Computer Vision & Image Understanding, 2016, 150: 109-125.

[34] Nazir S, Yousaf M H, Velastin S A. Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition [J]. Computers & Electrical Engineering, 2018, 72: 660-669.

[35] Heng Wang, Klser A, Schmid C, *et al.* Dense trajectories and motion boundary descriptors for action recognition [J]. International Journal of Computer Vision, 2013, 103 (1): 60-79.

[36] Gaidon A, Harchaoui Z, Schmid C. Activity representation with motion hierarchies [J]. International Journal of Computer Vision, 2014, 107 (3): 219-238.

[37] Wang Heng, Dan O, Verbeek J, *et al.* A robust and efficient video representation for action recognition [J]. International Journal of Computer Vision, 2015, 119 (3): 219-238.

[38] Peng Xiaojiang, Zou Changqing, Yu Qiao, *et al.* Action recognition with stacked fisher vectors [C]// Proc of ECCV, Berlin: Springer, 2014: 581-595.

[39] Xia Limin, Ma Wentao. Human action recognition using high-order feature of optical flows [J/OL]. The Journal of Supercomputing, 2021, 77. (2021-05-10) [2021-07-10]. https://doi. org/10. 1007/978-3-319-14249-4_60.

[40] Wang Limin, Tong Zhan, Ji Bin, *et al.* TDN: temporal difference networks for efficient action recognition [C]// Proc of IEEE/CVF CVPR. Piscataway, NJ: IEEE Press, 2021: 1895-1904.

[41] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2016: 1933-1941.

[42] Lan Zhenzhong, Yi Zhu, Hauptmann A G. Deep local video feature for action recognition [C]// Proc of IEEE CVPRW. Piscataway, NJ: IEEE Press, 2017: 1219-1225.

[43] Zhou Bolei, Andonian A, Oliva A, *et al.* Temporal relational reasoning in videos [C]// Proc of ECCV. Berlin: Springer, 2018: 831-846.

[44] Feichtenhofer C, Fan H, Malik J, *et al.* Slowfast networks for video recognition [C]// Proc of IEEE/CVF ICCV. Piscataway, NJ: IEEE Press, 2019: 6201-6210.

[45] Sang K P, Chung J H, Kang T K, *et al.* Binary dense sift flow based two stream CNN for human action recognition [J/OL]. Multimedia Tools and Applications, 2021. (2021-06-10) [2021-07-10]. https://doi. org/10. 1007/s11042-021-10795-2.

[46] Liu Congcong, Ying Jie, Yang Haima, *et al.* Improved human action recognition approach based on two-stream convolutional neural network model [J]. The Visual Computer, 2021, 37 (6): 1327-1341.

[47] Du Zhouning, Mukaidani H. Linear dynamical systems approach for human action recognition with dual-stream deep features [J/OL]. Applied Intelligence, 2021. (2021-05-03) [2021-07-10]. https://doi. org/10. 1007/s10489-021-02367-6.

[48] Tran D, Wang Heng, Torresani L, *et al.* A closer look at spatiotemporal convolutions for action recognition [C]// Proc of IEEE/CVF CVPR. Piscataway, NJ: IEEE Press, 2018: 6450-6459.

[49] Diba A, Fayyaz M, Sharma V, *et al.* Temporal 3D convnets: new architecture and transfer learning for video classification [EB/OL]. (2017-11-22) [2021-07-10]. https://arxiv. org/pdf/1711. 08200. pdf

[50] Qiu Zhaofan, Yao Ting, Mei Tao. Learning spatio-temporal representation with pseudo-3D residual networks [C]// Proc of IEEE ICCV. Piscataway, NJ: IEEE Press, 2017: 5534-5542.

[51] Tran D, Ray J, Shou Zheng, *et al.* Convnet architecture search for spatiotemporal feature learning [EB/OL]. (2017-08-17) [2021-07-10]. https://arxiv. org/pdf/1708. 05038. pdf

[52] Zolfaghari M, Singh K, Brox T. ECO: efficient convolutional network for online video understanding [C]// Proc of ECCV. Berlin: Springer, 2018: 713-730.

[53] Crasto N, Weinzaepfel P, Alahari K, *et al.* MARS: motion-augmented RGB stream for action recognition [C]// Proc of IEEE/CVF CVPR. Piscataway, NJ: IEEE Press, 2019: 7874-7883.

[54] Kim J, Li Gen, Yun Inyong, *et al.* Weakly-supervised temporal attention 3D network for human action recognition [J]. Pattern Recognition, 2021, 119: Article ID 108068.

[55] Shi Xin, Jiang Haiyang, Lu Yuanyao. A novel channel attention mechanism for human action recognition based on convolutional kernel [J]. Journal of Physics: Conference Series, 2021, 1944 (1): Article ID 012015.

[56] Kumawat S, Verma M, Nakashima Y, *et al.* Depthwise spatio-temporal STFT convolutional neural networks for human action recognition [J/OL]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2021. (2021-04-29) [2021-07-10]. https://doi. org/10. 1109/TPAMI. 2021. 3076522.

[57] Dong Min, Fang Zhenglin, Li Yongfa, *et al.* AR3D: attention residual 3D network for human action recognition [J]. Sensors, 2021, 21 (5): 1656-1669.

[58] Donahue J, Hendricks L A, Guadarrama S, *et al.* Long-term recurrent convolutional networks for visual recognition and description [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2015: 2625-2634.

[59] Aghaei A, Nazari A, Moghaddam M E. Sparse deep LSTMs with convolutional attention for human action recognition [J]. SN Computer Science, 2021, 2 (3): Article ID 151.

[60] 揭志浩, 曾明如, 周鑫恒, 等. 结合 Attention-ConvLSTM 的双流卷积行为识别 [J]. 小型微型计算机系统, 2021, 42 (2): 405-408. (Jie Zhihao, Zeng Mingru, Zhou Xinheng, *et al.* Two stream CNN with Attention-ConvLSTM on human behavior recognition [J]. Journal of Chinese Computer Systems, 2021, 42 (2): 405-408.)

[61] Oreifej O, Liu Zicheng. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2013: 716-723.

[62] Rahmani H, Mahmood A, Du Q H, *et al.* Real time action recognition using histograms of depth gradients and random decision forests [C]// Proc of IEEE WACV. Piscataway, NJ: IEEE Press, 2014: 626-633.

[63] Chen Wenbin, Guo Guodong. TriViews: a general framework to use 3D depth data effectively for action recognition [J]. Journal of Visual Communication and Image Representation, 2014, 26 (1): 182-191.

[64] Miao Jie, Jia Xiaoyi, Mathew R, *et al.* Efficient action recognition from compressed depth maps [C]// Proc of IEEE ICIP. Piscataway, NJ: IEEE Press, 2016: 16-20.

[65] Bulbul M F, Tabussum S, Ali H, *et al.* Exploring 3D human action recognition using STACOG on multi-view depth motion maps sequences

[J]. Sensors, 2021, 21 (11): 3642-3659.

[66] Ji Xiaopeng, Zhao Qinsong, Cheng Jun, *et al.* Exploiting spatio-temporal representation for 3D human action recognition from depth map sequences [J]. Knowledge-Based Systems, 2021, 227: Article ID 107040.

[67] Xia Lu, Chen C C, Aggarwal J K. View invariant human action recognition using histograms of 3D joints [C]// Proc of IEEE CVPRW. Piscataway, NJ: IEEE Press, 2012: 20-27.

[68] Torki M, Gowayyed M A, Hussein M E, *et al.* Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition [C]// Proc of the 23rd International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2013: 1351–1357.

[69] Yang Xiaodong, Tian Yingli. Effective 3D action recognition using eigenjoints [J]. Journal of Visual Communication & Image Representation, 2014, 25 (1): 2-11.

[70] Pazhoumand-Dar H, Lam C P, Masek M. Joint movement similarities for robust 3D action recognition using skeletal data [J]. Journal of Visual Communication & Image Representation, 2015, 30: 10-21.

[71] Nguyen V T, Nguyen T N, Le T L, *et al.* Adaptive most joint selection and covariance descriptions for a robust skeleton-based human action recognition [J]. Multimedia Tools and Applications, 2021, 80 (20): 27757–27783.

[72] Li Wenbo, Wen Longyin, Chang M C, *et al.* Adaptive RNN tree for large-scale human action recognition [C]// Proc of IEEE ICCV. Piscataway, NJ: IEEE Press, 2017: 1453-1461.

[73] Wang Hongsong, Wang Liang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2017: 3633-3642.

[74] Liu Jun, Wang Gang, Hu Ping, *et al.* Global context-aware attention LSTM networks for 3D action recognition [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2017: 3671-3680.

[75] Lee I, Kim D, Kang S, *et al.* Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks [C]// Proc of IEEE ICCV. Piscataway, NJ: IEEE Press, 2017: 1012-1020.

[76] Song Sijie, Lan Cuiling, Xing Junliang, *et al.* Spatio-temporal attention-based LSTM networks for 3D action recognition and detection [J]. IEEE Trans on Image Processing, 2018, 27 (7): 3459-3471.

[77] Li Ce, Xie Chunyu, Zhang Baochang, *et al.* Memory attention networks for skeleton-based action recognition [J/OL]. IEEE Trans on Neural Networks and Learning Systems, 2021. (2021-03-15) [2021-07-10]. https://doi. org/10. 1109/TNNLS. 2021. 3061115

[78] Zhang Pengfei, Xue Jianru, Lan Cuiling, *et al.* EleAtt-RNN: adding attentiveness to neurons in recurrent neural networks [J]. IEEE Trans on Image Processing, 2020, 29: 1061-1073.

[79] Li S, Li W, Cook C, *et al.* Independently recurrent neural network (IndRNN): building a longer and deeper RNN [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2018: 5457-5466.

[80] Zhang Pengfei, Lan Cuiling, Xing Junliang, *et al.* View adaptive neural networks for high performance skeleton-based human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2019, 41 (8): 1963-1978.

[81] Rao Haocong, Xu Shihao, Hu Xiping, *et al.* Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition [J]. Information Sciences, 2021, 569: 90-109.

[82] Ding Zewei, Wang Pichao, Ogunbona P O, *et al.* Investigation of different skeleton features for CNN-based 3D action recognition [C]// Proc of IEEE ICMEW, Piscataway, NJ: IEEE Press, 2017: 617-622.

[83] Xu Yangyang, Cheng Jun, Wang Lei, *et al.* Ensemble one-dimensional convolution neural networks for skeleton-based action recognition [J]. IEEE Signal Processing Letters, 2018, 25 (7): 1044-1048.

[84] Li Bo, Dai Yunchao, Cheng Xuelian, *et al.* Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN [C]// Proc of IEEE ICMEW. Piscataway, NJ: IEEE Press, 2017: 601-604.

[85] Li Chao, Zhong Qiaoyong, Xie Di, *et al.* Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation [C]// Proc of the 27th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 786–792.

[86] Ke Qiuhong, Bennamoun M, An S, *et al.* Learning clip representations for skeleton-based 3d action recognition [J]. IEEE Trans on Image Processing, 2018, 27 (6): 2842-2855.

[87] Li Yanshan, Xia Rongjie, Liu Xing, *et al.* Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition [C]// Proc of IEEE ICME. Piscataway, NJ: IEEE Press, 2019: 1066-1071.

[88] Caetano C, Sena J, F Brémond, *et al.* SkeleMotion: a new representation of skeleton joint sequences based on motion information for 3D action recognition [C]// Proc of the 16th International Conference on Advanced Video and Signal Based Surveillance. Piscataway, NJ: IEEE Press, 2019: 1-8.

[89] Caetano C, Bremond F, Schwartz W R. Skeleton image representation for 3D action recognition based on tree structure and reference joints [C]// Proc of the 32nd SIBGRAPI Conference on Graphics, Patterns and Images. Piscataway, NJ: IEEE Press, 2019: 16-23.

[90] Li Meng, Sun Qiumei. 3D skeletal human action recognition using a CNN fusion model [J]. Mathematical Problems in Engineering, 2021, 2021: Article ID 6650632.

[91] Li Maosen, Chen Siheng, Chen Xu, *et al.* Actional-structural graph convolutional networks for skeleton-based action recognition [C]// Proc of IEEE/CVF CVPR. Piscataway, NJ: IEEE Press, 2019: 3590-3598.

[92] Si Chenyang, Chen Wentao, Wang Wei, *et al.* An attention enhanced graph convolutional lstm network for skeleton-based action recognition [C]// Proc of IEEE/CVF CVPR. Piscataway, NJ: IEEE Press, 2019: 1227-1236.

[93] Song Yifan, Zhang, Wang Liang. Richly activated graph convolutional network for action recognition with incomplete skeletons [C]// Proc of IEEE ICIP. Piscataway, NJ: IEEE Press, 2019: 1-5.

[94] Zhang Pengfei, Lan Cuiling, Zeng Wenjun, *et al.* Semantics-guided neural networks for efficient skeleton-based human action recognition [C]// Proc of IEEE CVPR. Piscataway, NJ: IEEE Press, 2020: 1109-1118.

[95] Hao Xiaoke, Li Jie, Guo Yingchun, *et al.* Hypergraph neural network for skeleton-based action recognition [J]. IEEE Trans on Image Processing, 2021, 30: 2263-2275.

[96] Xu Weiyao, Wu Muqing, Zhu Jie, *et al.* Multi-scale skeleton adaptive weighted GCN for skeleton-based human action recognition in IoT [J]. Applied Soft Computing Journal, 2021, 104: Article ID 107236.

[97] 李扬志, 袁家政, 刘宏哲. 基于时空注意力图卷积网络模型的人体骨架行为识别算法 [J]. 计算机应用, 2021, 41 (7): 1915-1921. (Li Yangzhi, Yuan Jiazheng, Liu Hongzhe. Human skeleton-based action recognition algorithm based on spatiotemporal attention graph convolutional network model [J]. Journal of Computer Applications, 2021, 41 (7): 1915-1921.)

[98] Chen Chen, Liu Kui, Kehtarnavaz N. Real-time human action recognition based on depth motion maps [J]. Journal of Real-Time Image Processing, 2016, 12 (1): 155-163.

[99] Chaaraoui A A, JR Padilla-López, Florez-Revuelta F. Fusion of skeletal and silhouette-based features for human action recognition with RGB-D

devices [C]// Proc of IEEE ICCVW. Piscataway, NJ: IEEE Press, 2013: 91-97.

[100] Li Meng, Leung H, Shum H. Human action recognition via skeletal and depth based feature fusion [C]// Proc of the 9th International Conference on Motion in Games. New York: ACM Press, 2016: 123-132.

[101] Althloothi S, Mahoor M H, Zhang Xiao, *et al.* Human activity recognition using multi-features and multiple kernel learning [J]. Pattern Recognition, 2014, 47 (5): 1800-1812.

[102] Li Donglu, Jahan H, Huang Xiaoyi, *et al.* Human action recognition method based on historical point cloud trajectory characteristics [J/OL]. The Visual Computer, 2021, 37 (6) . (2021-06-02) [2021-07-10]. https://doi. org/10. 1007/s00371-021-02167-6.

[103] Ni B, Pei Yong, Moulin P, *et al.* Multilevel depth and image fusion for human activity detection [J]. IEEE Trans on Cybernetics, 2013, 43 (5): 1383-1394.

[104] Jalal A, Kim Y H, Kim Y J, *et al.* Robust human activity recognition from depth video using spatiotemporal multi-fused features [J]. Pattern Recognition, 2017, 61: 295-308.

[105] Xu Qingyang, Zheng Wanqiang, Song Yong, *et al.* Scene image and human skeleton-based dual-stream human action recognition [J]. Pattern Recognition Letters, 2021, 148: 136-145.

[106] 周雪雪，雷景生，卓佳宁. 基于多模态特征学习的人体行为识别方法 [J]. 计算机系统应用, 2021, 30 (4): 146-152. (Zhou Xuexue, Lei Jingsheng, Zhuo Jianing. Human action recognition algorithm based on multi-modal features learning [J]. Computer Systems & Applications, 2021, 30 (4): 146-152.)

[107] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach [C]// Proc of the 17th International Conference on Pattern Recognition. Piscataway, NJ: IEEE Press, 2004: 32-36.

[108] Kuehne H, Jhuang H, Garrote E, *et al.* HMDB: A large video database for human motion recognition [C]// Proc of IEEE ICCV. Piscataway, NJ: IEEE Press, 2011: 2556-2563.

[109] Reddy K K, Shah M. Recognizing 50 human action categories of web videos [J]. Machine Vision and Applications, 2013, 24 (5): 971-981.

[110] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. (2012) [2021-07-10]. https://arxiv. org/pdf/1212. 0402. pdf

[111] Li Wanqing, Zhang Zhengyou, Liu Zicheng. Action recognition based on a bag of 3D points [C]// Proc of CVPRW. Piscataway, NJ: IEEE Press, 2010: 9-14.

[112] Kay W, Carreira J, Simonyan K, *et al.* The Kinetics Human Action Video Dataset [EB/OL]. (2017) [2021-07-10]. https://arxiv. org/pdf/1705. 06950. pdf.

[113] Shahroudy A, Liu Jun, Ng T, *et al.* NTU RGB+D: a large scale dataset for 3D human activity analysis [J]. IEEE Computer Society, 2016: 1010-1019.