DOI:10.3969/j.issn.1005-202X.2024.01.001

医学影像物理

基于视频的动作智能识别在医学中的应用

黄新瑞[1],黄河颂[2],黄渝川[2],陈美凝[2],范馨月[2],伊鸣[3]
1.北京大学基础医学院生物化学与生物物理学系,北京 100191;2.北京大学基础医学院,北京 100191;3.北京大学神经科学研究所,北京 100191

【摘要】基于视频的动作智能识别是计算机视觉领域的一项具有挑战性的研究。本文回顾了基于视频动作智能识别最先进的方法并进行分析、比较和讨论,主要介绍基于手工制作特征的机器学习方法、基于自动特征提取的深度学习方法和基于多信息融合方法。同时介绍近十年来该技术在医学中流行的重要应用及其相关局限性,并分享关于此技术未来应用以改善人类健康的跨学科观点。

【关键词】动作识别;计算机视觉;特征表示;机器学习;深度学习;综述
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2024)01-0001-07

Medical application of video-based intelligent action recognition

HUANG Xinrui[1], HUANG Hesong[2], HUANG Yuchuan[2], CHEN Meining[2], FAN Xinyue[2], YI Ming[3]
1. Department of Biochemistry and Biophysics, School of Basic Medical Sciences, Peking University, Beijing 100191, China; 2. School of Basic Medical Sciences, Peking University, Beijing 100191, China; 3. Institute of Neuroscience, Peking University, Beijing 100191, China

Abstract: Video-based intelligent action recognition remains challenging in the field of computer vision. The review analyzes the state-of-the-art methods of video-based intelligent action recognition, including machine learning methods with handcrafted features, deep learning methods with automatically extracted features, and multi-information fusion methods. In addition, the important medical applications and limitations of these technologies in the past decade are introduced, and the interdisciplinary views on the future application to improve human health are also shared.

Keywords: action recognition; computer vision; feature representation; machine learning; deep learning; review

前 言

因人体动作中蕴含很多重要的生理、病理信息,近年来对人体动作信息的有效提取已引起医学领域研究人员的广泛关注。动作智能识别是计算机视觉的一个重要探索领域,从1980年开始,提出基于图像和/或视频数据的动作识别的不同研究,但在医学研究和临床实施方面仍处于相对初级阶段[1-2]。了解这些潜在应用,利用临床医生、工程师和数据科学家的跨专业知识,使该技术能在测量评估患者病情方面使治疗师、患者或医疗保健系统受益非常重要。基于视频的无标记动作捕捉使用相机录制的二维/三维视频自动跟踪人体运动,避免在跟踪和分析人体运动期间放置标记,摆脱身体标记对运动带来的限制,能够以更自然的方式捕捉环境中更逼真的人体运动[3]。与成熟的医疗系统和可穿戴设备相比,操作更简单,成本更低,且基于视频运动追踪适合远程监测,在远程个人健康状况评估、临床随访方面更便利。尽管研究人员一直致力于捕捉完整人体动作并实现精确识别,但人体动作具有时间和空间上的双重复杂性以及完整可变性,使得该技术在医学应用中仍面临许多挑战[4]。本文综述回顾了基于视频的动作识别方法,包括各种传统机器学习技术以及深度学习技术,讨论基于视频的动作智能识别在医学中的最新研究进展和应用,通过比较技术特点、分析挑战和提出相关解决建议,以便进一步探索挖掘此技术未来在医疗保健领域的潜在好处。

【收稿日期】2023-10-06
【基金项目】国家自然科学基金(61901008,32271053);北京大学医学部教育教学研究课题(2022YB17);北京大学医学部2023年暑期本科生科研项目;北京市自然科学基金-海淀原始创新联合基金重点研究专题项目(L222016)
【作者简介】黄新瑞,博士,讲师,研究方向:生物医学多模态成像及其数据图像处理,E-mail: huangxr@pku.edu.cn
【通信作者】伊鸣,博士,研究员,研究方向:疼痛与认知相关的神经机制和行为学,E-mail: mingyi@hsc.pku.edu.cn

1 基于视频动作识别的智能系统

医学中的视频动作智能识别是一种典型的针对特定场景中获取的视频数据的监督学习任务[5]，如图1所示。它首先收集和注释感兴趣任务的视频数据，然后将标记数据集分为训练集和测试集；数据特征可由用户手动定义和选择（即手工特征提取），或使用深度学习算法自动提取；然后将训练集特征输入到所选的机器学习模型中进行模型搭建；执行交叉验证以确保模型准确性，使用几个性能指标来评估最终模型；最后将动作识别模型用于对医学研究和临床实践中的视频数据进行预测应用。

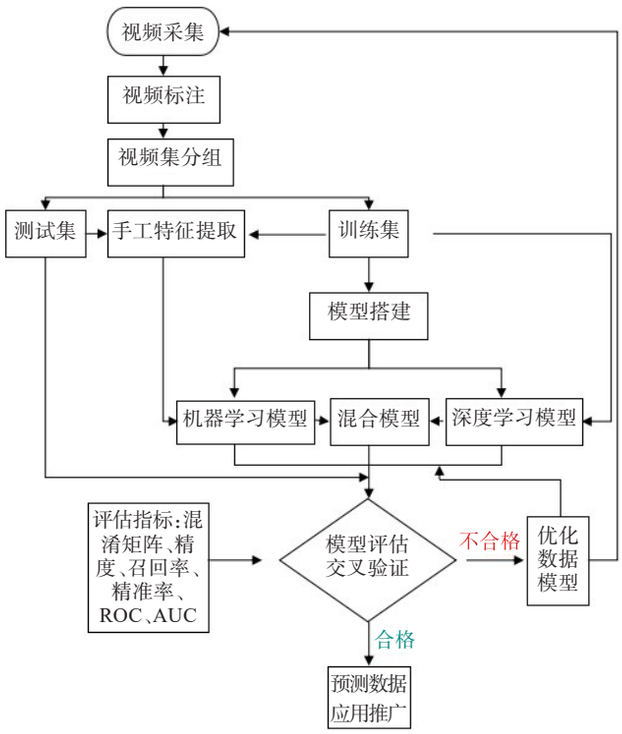

图1 医学中基于视频的动作智能识别系统总体框架图

Figure 1 Overall framework of video-based intelligent action recognition system in medicine

2 基于视频动作识别的智能方法

2.1 基于手工特征表示的机器学习方法

2.1.1 用于动作识别的手工特征表示 在传统的机器学习中，工程师和数据科学家必须手动制作特征提取器识别有用的特征。为了识别重要特征实现强大的医学功能，必须密切和临床医生配合获取相当多的领域专业知识。基于手工的特征分为基于整体特征表示和基于局部特征表示[6]。(1)基于整体特征表示：将视频中感兴趣区域视为一个整体，其中所有像素都被用来计算描述符，可以从感兴趣的区域表示有区别的全局信息，然后将其用于动作表征。整体表示方法考虑人体的整体结构，基于来自轮廓、边缘、光流等信息来表示动作，无需定位身体各个部位，由于仅使用全局信息，整体方法不仅计算简单，而且高效且有效。因此，这种方法对于可能包含背景杂乱、相机运动和遮挡的视频非常重要，但同时对噪声、背景杂乱以及遮挡和视角的变化也很敏感。此外，整体方法的准确性高度依赖于运动目标的检测和分割预处理；(2)基于局部特征表示：倾向于捕获视频中局部区域的形状和运动信息的重要特征，根据时空和尺度的变化来自主表示动作，可将感兴趣的特定局部特征与从杂乱背景中的多个运动或对象计算的特征分开。许多研究人员已经使用检测视频中的兴趣点并通过时域形成轨迹来表示动作特征或运动学参数（如速度和加速度），例如从视频中自动跟踪人体的解剖标志（所谓的关键点）包括脚踝、膝盖、臀部、手腕、肘部、肩膀、脚（如脚后跟、大脚趾和小脚趾）、手（如每个手指的指尖和3个关节）和面部（如耳朵、眼睛、鼻子和嘴巴）等。局部表示方法可以降低识别人类行为动作的计算复杂性，但在某些情况下又无法准确识别人类的行为。因此，结合局部和整体的方法进行手工特征表示可能会有帮助。

2.1.2 用于动作识别的机器学习算法 医学相关研究中涉及的动作识别机器学习算法用于根据输入特征预测结果以做出决策或诊断，算法的选择取决于许多因素，例如特征数据类型、数据大小以及处理数据的可用资源。具体算法[7]包括：(1)朴素贝叶斯（Native Bayes, NB），基于贝叶斯定理的概率分类器，通过最大似然估计的通用决策规则来预测类别；(2)线性判别分析（Linear Discriminant Analysis, LDA），用于识别划分两个或多个类特征的线性组合；(3)二次判别分析（Quadratic Discriminant Analysis, QDA），假设每个类都具有高斯分布并认为每个类都有一个单独的协方差矩阵，从而进行非线性判别分析；(4)逻辑回归（Logistic Regression, LR），探索独立特征和类标签之间的相关性，通过将数据拟合到逻辑曲线来查找分类的可能性，如果类标签由两个以上的类组成，则可以使用多项逻辑回归；(5)支持向量机（Support Vector Machine, SVM），通过在高维特征空间中创建一个超平面，以最大余量精确地分离训练数据，包括线性分类器和基于核函数的非线性分类器；(6)K最近邻（K-Nearest Neighbor, KNN），存储所有训练数据，根据相似性度量对测试数据进行分类，KNN中的K值表示可以参与多数投票过程的最近邻居的数量；(7)决策树（Decision Tree, DT），它的每个节点要么是决策节点，要么是叶节点，预测从根节点开始，通过比较属性值并根据比较进行分支，最终结果是代表特征向量分类的叶节点；(8)随机森林（Random Forests,

RF)，是一种由DT集合组成的集成学习技术，RF中的每个DT都从训练特征向量的随机样本中学习，并在决定分割节点时使用特征子集，RF中的泛化误差高度依赖于树的数量以及它们之间的相关性；(9)AdaBoost，通过将多个弱分类器(如DT)与生成类标签的未加权特征向量相结合，构建一个鲁棒的分类器来提高性能。如果出现任何错误分类，它会提高训练数据的权重；接下来，使用不同的权重构建下一个分类器，错误分类的训练数据的权重得到提升，并且重复此过程；来自所有分类器的预测通过多数投票的方式被组合以做出最终预测；(10)LogitBoost，是一种从AdaBoost扩展而来的集成学习算法，以解决其局限性，例如对噪声和异常值的敏感性，它基于以线性方式修改损失函数的二项对数似然。相比之下，AdaBoost使用指数损失，以指数方式修改损失函数；(11)XGBoost，是一种高效且可扩展的梯度提升技术，包括用于管理稀疏数据的最先进的树学习算法、用于管理近似树学习中实例权重的加权分位数方法、用于快速模型探索的并行和分布式计算；(12)对数线性化高斯混合网络(Log-Linearization Gaussian Mixed Network, LLGMN)，是一种前馈神经网络，可以估计分类的后验概率，该网络包含3层，最后一层的输出被视为每个类别的后验概率，通过学习权重系数，对数线性化高斯混合形成被集成到神经网络中，从而允许评估给定数据集的概率分布；(13)偏最小二乘回归(Partial Least Squares Regression, PLSR)，是一种统计方法，通过尽可能最小地揭示两个矩阵的协方差来揭示两个矩阵之间的关系。

2.2 基于自动特征提取的深度学习方法

最近的研究表明，基于手工制作特征的机器学习方法并不适合所有类型的数据集。基于深度学习的方法能够自动处理原始图像和视频数据把低级输入数据转换为中级或高级特征表示，以进行特征提取、描述和分类[6, 8]。这些方法中经常采用可训练的过滤器和基于多层的模型来进行动作识别[9]：(1)二维卷积神经网络(Two Dimentional Convolutional Neural Networks, 2D CNN)，它通常具有3个主要的神经层：卷积层、池化层和全连接层，卷积层将输入图与K个内核进行卷积以提供K特征图，然后对K特征图进行非线性激活并在池化层进行池化，学习到的特征作为全连接层的输入，用于执行分类任务[10]。较低卷积层提取简单特征，而较高卷积层通过在每层使用滤波器提取复杂特征。用2D CNN捕获运动信息，特征图仅提取二维空间信息，需融合另一个模型来捕获时间信息；(2)循环神经网络(Recurrent Neural Network, RNN)，与具有前馈连接的2D CNN不同，它能将前一个时间步中输入的激活反馈回网络以影响当前输入与输出的关系，从而对顺序行为进行建模。一种为时间序列建模而设计的RNN称为长短期记忆模型(Long Short Term Memory, LSTM)[5]，由称为记忆块的特殊单元组成，并位于循环隐藏层中，通过两个并行卷积网络提取基本空间特征，然后将这些特征用作LSTM模块的输入，通过自连接使用LSTM内部的互相关来提取空间和时间信息以及时间关系。LSTM根据3个元素有效地更新其当前记忆向量：当前帧、先前记忆向量和对象的先前位置；(3)三维卷积单流网络，在网络中使用三维卷积核，通过对多个相邻视频帧应用三维滤波器，直接从多个视频帧中提取时空特征；(4)三维双流/多流网络，通过使用空间流和时间流来编码结构和光流信息，两个流最后通过类别分数进行融合；(5)视觉转换器(Vision Transformer)，是一种无卷积方法，一般组件通常基于两个主要元素：图像的线性投影和包含多层感知器(Multi-Layer Perceptron, MLP)神经网络模型和自注意力机制的编码器变压器。基于注意力机制的无卷积方法是人类动作识别的新趋势，具有高效的计算性能[11]。

2.3 基于多信息融合的方法

深度学习的方法能够从原始数据中自动学习特征，这在一定程度上减少传统机器学习方法对手工制作特征的检测器和描述符的需求，但基于深度学习的方法仍然存在一些需要考虑的局限性[12-13]。因此，有研究采用不同特征融合、综合多种模态数据(例如视频、可穿戴设备、传感器数据等)、组合多种网络架构的策略，或在不同的动作识别任务中采用迁移学习方法使用预训练模型来加速训练过程，以便在合理的硬件要求下提高动作识别系统的性能(表1)。在过去的10年中，已经发布了几种不同的人体姿势估计和运动跟踪开源算法(例如OpenPose[14]、DeepLabCut[15]、Leap Motion[16]、MediaPipe[17])，在这些算法提供的预训练网络基础上可以跟踪为各种医学研究或临床需求定制的新视频训练新网络[18-20]。

3 基于视频动作智能识别在医学中的应用

3.1 婴幼儿生长发育评估

早期发现非典型发育对于先天性运动障碍[例如脑瘫(Cerebral Palsy, CP)]和神经发育障碍[例如自闭症谱系障碍(Autistic Spectrum Disorder, ASD)]的诊断至关重要，可以确保及时获得早期干预服务以改善运动结果(例如协调能力、姿势能力)和其他发展领域(例如社交、语言)。Prechtl[21]提出全身运动评估(General Movement Assessment, GMA)作为预测高危婴儿脑瘫的宝贵工具，该评估是由经过培训的专业人员对婴幼儿视频进行评分，既耗时又受主观的影响[7]。基于视频动作识别智能方法的出现为评估基于运动的临床疾病

表1 基于视频动作智能识别3种方法的比较

Table 1 Comparison among 3 video-based intelligent action recognition methods

方法	优点	缺点
机器学习	用于训练模型的特征是明确已知的；训练所需的数据较少；计算时间和内存使用量也较少；理解模型、分析和可视化功能简单且明确。	通常需要有效且准确的特征检测器和提取方法；由于特征维度高，可能需要特征选择和降维方法；通常需要数据预处理和标准化以获得显著的性能；识别能力通常较低，管理类间和类内问题效率低下。
深度学习	不需要专业知识即可获得合适的特征；特征不是手动设计的，能够直接从原始数据中学习特征；深度神经网络可以提取深层高级特征，更适合复杂任务；可从模型获得层次和平移不变特征解决类间和类内问题；并不是强制性的需要数据预处理和归一化来实现高性能。	需要收集海量数据，会缺乏合适数据集；模型训练需要大量数据以避免过度拟合；耗时，需要高级强大计算系统来加速训练；需考虑模型泛化能力问题。
融合方法	可从不同角度表示特征；数据集由多种模态组成；利用多源信息融合产生的结果更准确。	模型搭建会更复杂，挑战更大；需要处理大量数据集，训练时间可能更长；对计算机资源的需求会显著增加。

预测因子提供令人兴奋的潜力，这种方法没有标记不会干扰自然行为并在其运动中引入不确定的伪影，这对婴幼儿研究非常重要，在这些人群中，将标记附着在身体上可能是无法忍受的，尤其是ASD等非典型人群。不同的工作尝试通过基于视频提取的姿势估计分析结果与专家从相同视频经典GMA产生的结果进行比较（表2），此类研究已成功地根据婴儿视频记录的自发运动评估来预测脑瘫，其性能与脑瘫风险标准化测量相当[22]。通过使用视频记录来实施低成本、自动、客观的替代方案来检测脑瘫风险，从而解决传统GMA缺点，将有助于为儿科医生和全科医生开发一种脑瘫筛查工具。

3.2 成人神经系统疾病的临床运动评估

基于视频的无标记运动跟踪可以在多种背景下提供细粒度、客观的衡量标准，在临床中的应用正在扩大，但仍处于起步阶段，最常用于识别症状或检测疾病人群与健康人群之间运动模式的差异，目前的研究多是评估帕金森（Parkinson's Disease, PD），造成这种趋势的一个可能原因是PD有明显且明确的身体体征和症状以及异常运动，例如PD的特征是运动时出现震颤、运动迟缓和僵硬，涉及头部、手臂、腿部或整个身体的不自主运动[23]。以往对精神疾病患者运动检测的研究均采用模糊运动方法，而较少强调精确的动作和运动模式，基于视频无标记系统检测和分析患者运动模式，可用于全身运动，例如步行和跑步、精细运动，伸手和抓握，以及测量行动的全局特征等[24]。最近的许多研究使用运动跟踪姿势估计来评估PD患者的运动障碍，发现与临床标准评估相似或优于临床标准评估（表2）[22]。

3.3 健康监测和康复训练中的应用

医学研究和技术的发展显著提高了患者的生活质量，医务人员的更高要求促使研究人员尝试不同的技术来改进在紧急情况下必不可少的医疗保健监测方法。在医疗卫生领域，人体肢体的活动完成度一直以来都是脑卒中患者和骨折患者健康恢复程度的评价标准[25]（表2）。个性化康复训练需要医生实时监督指导并及时调整训练任务或训练方案，但受医护人员数量及精力的限制使得其难以有效实施。基于视频动作捕捉识别技术使患者和医疗专业人员无论是否处于同一地理位置都能够很好进行交互，系统可设置实时分配给每个患者的康复计划的管理和监控模块，使得管理康复过程的电子病历成为可能。通过计算机系统获取康复训练动作的视频信息将动作的执行情况进行分类识别并反馈给医生，能够有效地减少医护人员的工作量，实现智慧医疗。

4 基于视频动作智能识别在医学应用中的挑战

在广泛实施针对人类健康的基于视频的动作识别智能系统中，需要处理相当多的挑战和约束，主要是应用程序限制和实施障碍（重点是在临床环境中的实施）[51]。

4.1 应用程序限制

数据采集挑战：记录设备有一定的局限性，常见视频记录设备的采样率通常约为30 Hz，可能无法捕获高速或高频率下发生运动的准确运动学特征，记录设备的光圈和快门速度也会影响图像质量并引入模糊；视点的变化，大多数方法假设动作是从固定的视点执行的，当需要跟踪一个或多个解剖位置时会发生遮挡，如果从两个或多个独立的视点观察，相同的动作可能会显得不同，由此收集的数据可能分属不同的类别；杂乱的背景，对基于颜色和基于区域的分割方法有很大影响，会使识别清晰人体形状特征变得更加困难；采集场景和时间方面，考虑到成功训练模型所需的视频时长不可预测和动作的复杂性，可能需要为某些应用设置特定于现场的精度标准。

缺乏可靠和可行性数据：为了充分训练智能模型，需要大量且多样化的数据集。标记数据是监督学习的

表2　基于视频的动作智能识别方法在医学中的应用

Table 2　Medical applications of video-based intelligent action recognition methods

方法	文献	主要思想	应用	数据	评估
机器学习	文献[26]	光流特征+SVM分类	CP	CP婴儿15例;无CP婴儿67例	10倍交叉验证准确率93%
	文献[27]	食指轨迹特征+SVM分类	PD	晚期患者13例;健康对照6例	症状级别分类准确度88%;对照和患者准确度95%
	文献[28]	大位移光流+婴儿轮廓特征+多个分类器(LR、AdaBoost、LogitBoost和RF)+留一法交叉验证	CP	典型婴儿(98例);非典型婴儿(29例)	AdaBoost的GMA准确率85.83%;RF对CP分类准确率92.13%
	文献[29]	关节轨迹特征+RF分类	PD	患者9例	准确率71.4%
	文献[30]	手部轨迹特征+KNN、RF和SVM分类	PD	患者60例	核SVM准确率为89.7%
	文献[31]	姿势的直方图特征+KNN、LDA和集成分类器+留一法交叉验证	CP	MINI-RGBD的12个序列	集成分类器准确率91.67%
	文献[32]	婴儿轮廓特征+LLGMN分类	CP	21名婴儿的4个类别的数据	正常与异常准确率90.2%;区分4种的准确率83.1%
	文献[33]	关节轨迹特征+NB	神经运动疾病	高风险婴儿19例;健康婴儿85例	NB评分在高风险与健康组有显著差异
	文献[34]	身体关键点轨迹特征+RF、LDA、LR、SVM、XGBoost等分类	PD	患者729例	RF实现50%平衡准确度及95%不偏差1例准确度
	文献[19]	新生儿身体关键点轨迹+DT、NB、LDA、SVM、LogitBoost、KNN、RF分类	神经发育	正常19例;单调性运动25例	SVM分类准确率78%
	文献[35]	手、脚、腿轨迹特征+RF分类	PD	患者628例	症状等级分类平衡准确度45%,可接受准确度86%
深度学习	文献[36]	Keras VGG19迁移学习预训练模型+LSTM+10倍交叉验证	CP	CP 80例;正常135例	分类准确率65.1%
	文献[37]	由自动编码器和全连接神经网络组成基于周期性运动的网络	PD	患者121例	识别运动迟缓F分数0.7778
	文献[38]	3D CNN对行走分类	脑卒中	患者206例	准确率86.3%
融合方法	文献[39]	视频光流及轨迹特征+运动传感器数据特征+SVM分类	CP	CP婴儿14例;无CP婴儿64例	准确率87%
	文献[40]	视频和传感器数据频率特征+PLSR交叉验证	CP	CP婴儿14例;无CP婴儿64例	准确率91%
	文献[41]	视频片段、身体关键点轨迹+SVM、LSTM分别组合进行定时起立测试	PD	DBS术后PD患者24例	LSTM准确率93.1%
	文献[42]	婴儿全身轨迹+离散小波变换特征+SVM、RF、AdaBoost与XGBoost堆叠集成的分类器stacking	CP	正常婴儿60例,异常婴儿60例	准确率93.3%
	文献[43]	直方图特征HOJO2D、HOJD2D以及两者融合特征+多种网络架构分类	CP	MINI-RGBD的12个序列	FCNet网络最佳准确率91.67%
	文献[44]	CNN对原始视频分段+手部轮廓光流特征+NB、LR、SVM分类	PD	患者20例;对照15例	SVM预测运动迟缓精度0.8;NB预测PD准确度0.67
	文献[45]	多种局部特征+MLP、NB、SVM分类器;关节姿势序列+LSTM	ASD	有或无ASD的儿童108例	光流直方图特征与MLP组合最佳,分类准确率79.28%
	文献[46]	手指运动关键点轨迹+嵌入时空注意力机制的三流细粒度CNN	PD	患者48例;健康对照11例	分类准确率72.4%,可接受准确率98.3%
	文献[47]	骨骼关键点轨迹+CNN-LSTM深度神经网络架构	ASD	ASD 169例;典型儿童68例	分类准确率80.9%
	文献[48]	婴儿关节轨迹特征+基于时空注意力的模型	CP	烦躁或非烦躁婴儿235例	模型ROC-AUC得分81.87%
	文献[49]	身体关键点轨迹步态特征+卷积神经网络	脑卒中	患者8例	时间的绝对准确度和精度在(0.04±0.11) s内
	文献[50]	肩部、头部姿势信息和反应持续时间等特征+深度学习框架PyTorch预测	ASD	ASD 29例;典型儿童1例	ASD预测与临床诊断的一致性达93.3%

CP:脑瘫;PD:帕金森;ASD:自闭症谱系障碍

关键，由于其耗时且劳动密集的性质，手动数据注释成为该过程中的主要障碍之一。给定动作的样本数量分布可能不均匀，如大多数训练数据集都偏向于健康的运动模式。缺乏多样性的图像集（例如服装、照明、视点、与临床诊断相关的异常动作）。类间差异，不同的人以自己的方式执行不同的动作，这些动作有时彼此之间的相似性非常低，例如，行走方法可能在步幅或速度上有所不同。类内相似性，属于不同类的动作可能看起来相似，例如慢跑和跑步。使用有限的训练数据训练的模型在测试视频有较大差异的应用中可能表现不佳。缺少可公开访问的标记数据集。

特征设计及建模技术挑战：只有在考虑模型所使用的数据集之后才能准确比较模型的性能。在未来的研究中引入用于注释的半监督和无监督学习算法可以稍微促进加速并降低数据准备过程的成本。如何将无监督和半监督技术与监督技术结合使用以改善整体结果还有待观察。另外，需要测试模型的敏感性、特异性、可靠性和可行性：首先需建立标准有效性或构建动作估计测量和年龄同步、临床医生通用的临床测量金标准之间的有效性。这可以通过多种方式来实现，包括（但不限于）与三维运动捕捉、具有经过验证的准确性的可穿戴设备、专家临床评级和/或评估、或者甚至可能的其他动作识别算法进行比较。然后根据患者和医生已完成并提交的可用视频的数量及满意度调查情况，评估新动作识别模型的可行性和可接受性及可解释性。未来的工作应侧重于使用运动学金标准工具和临床标准评估进一步验证，这些研究中应包括具有各种不同运动模式的其他患者群体，以便开发广泛适用的算法。

4.2 实施障碍

系统用户友好性：目前缺乏用于动作智能识别的即插即用设备；系统配置复杂，如需要多相机校准或较长的设置拍摄时间；结果延迟，无法让许多用户获得近乎实时的可解释的数据结果；有编程和培训要求，一些现有的动作识别系统对于具有基本技术专业知识的用户来说容易安装和使用，然而对于没有技术背景的临床医生和研究人员来说仍然令人望而生畏。需要任何数量的编程或大量培训的技术不太可能在临床环境中广泛使用。

结果衡量挑战：在某些情况下，用户希望使用运动数据来改善临床或与健康相关的决策，但目前尚不清楚哪些运动参数将导致改善的结果（例如用户可能表示有兴趣测量“步行”，但不确定哪些具体步态参数与其研究或临床干预最相关）。因此，人们希望收集运动学数据，但如何使用这些数据还没有明确定义。同样，在临床评估的情况下，需要与相关临床和转化结果有明确的联系，即用户应该了解哪些输出指标很重要。

有限的硬件基础设施：如上所述，人体运动跟踪的动作识别的一些应用需要大量的计算能力。一些临床和研究环境不太可能访问及时执行所需应用程序所需的硬件（例如高性能并行图形处理单元）。

技术挑战：许多有望对临床或人类健康产生影响的技术在完全开发之前就已经可用，这可能会导致软件出现错误和频繁更新，从而损害用户之间的信任和信誉。反过来，这可能会加剧一些临床和研究界对采用新技术的犹豫，特别是旨在补充甚至取代人类医学专家评估的人工智能技术中。

5 总结与展望

本文全面概述了基于计算机视觉的人类动作识别研究的最新方法，包括基于手工制作的传统机器学习方法、基于深度学习的方法和基于多信息融合的方法，总结各种方法的优缺点，并得到以下结论：用合适的数据来捕捉动作可能有助于提高动作识别的性能；多信息融合可以有效提高性能；具有注意力机制的变压器网络方法是基于视频动作智能识别的新趋势。综合分析近十年的基于视频的动作智能识别在改善人类健康方面的典型应用研究，揭示此技术在检测和识别疾病体征和症状方面的潜在用途。针对现有的一些技术存在忽略数据可变性、参数数量较多、消耗大量资源、难以在实时嵌入式设备中实现等缺点，开发基于视频的简化实验设置的动作智能识别系统并将其集成到一个用户友好且临床医生可以准确分析的平台中，可以推广此技术在临床测量中的使用。随着数据增强方法的发展和不同迁移学习方法的应用，预计为特定运动中不同类型动作的学习模型准备的视频数据库将会增加，应在全球范围内利用海量数据存储和连接，以使医疗服务提供者和患者受益。

这篇综述虽不能声称涵盖所有已发表的基于视频动作智能识别的文章，但它可能是了解该领域当前趋势和挑战一个很好的起点，希望本文工作能够激励临床医生和当地人工智能相关部门之间合作来探索动作识别应用的未知领域。

【参考文献】

[1] Marr D, Vaina L. Representation and recognition of the movement of shapes[C]//Proceedings of the Royal Society of London Series B, Containing Papers of a Biological Character Royal Society (Great Britain), 1982: 501-524.
[2] Hester CF, Casasent D. Multivariant technique for multiclass pattern recognition[J]. Appl Opt, 1980, 19(11): 1758-1761.
[3] Corazza S, Mündermann L, Gambaretto E, et al. Markerless motion capture through visual hull, articulated ICP and subject specific model generation[J]. Int J Comput Vis, 2010, 87(1): 156-169.
[4] Arshad MH, Bilal M, Gani A. Human activity recognition: review, taxonomy and open challenges[J]. Sensors, 2022, 22(17): 6463.
[5] Host K, Ivašić-Kos M. An overview of human action recognition in

sports based on computer vision[J]. Heliyon, 2022, 8(6): e09633.
[6] Saif S, Tehseen S, Kausar S. A survey of the techniques for the identification and classification of human actions from visual data[J]. Sensors, 2018, 18(11): 3979.
[7] Irshad MT, Nisar MA, Gouverneur P, et al. AI approaches towards Prechtl's assessment of general movements: a systematic literature review[J]. Sensors, 2020, 20(18): 5321.
[8] Al-Faris M, Chiverton J, Ndzi D, et al. A review on computer vision-based methods for human action recognition[J]. J Imaging, 2020, 6 (6): 46.
[9] Sarma D, Bhuyan MK. Methods, databases and recent advancement of vision-based hand gesture recognition for HCI systems: a review [J]. SN Comput Sci, 2021, 2(6): 436.
[10] Kozma R, Alippi C, Choe Y, et al. Artificial intelligence in the age of neural networks and brain computing[M]. London: Academic Press, 2018.
[11] Moutik O, Sekkat H, Tigani S, et al. Convolutional neural networks or vision transformers: who will win the race for action recognitions in visual data?[J]. Sensors, 2023, 23(2): 734.
[12] Xiao X, Xu D, Wan W. Overview: video recognition from handcrafted method to deep learning method[C]//2016 International Conference on Audio, Language and Image Processing (ICALIP), 2016.
[13] Georgiou T, Liu Y, Chen W, et al. A survey of traditional and deep learning-based feature descriptors for high dimensional data in computer vision[J]. Int J Multimed Info Retr, 2020, 9(3): 135-170.
[14] Cao Z, Hidalgo G, Simon T, et al. OpenPose: realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Trans Pattern Anal Mach Intell, 2021, 43(1): 172-186.
[15] Mathis A, Mamidanna P, Cury KM, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning[J]. Nat Neurosci, 2018, 21(9): 1281-1289.
[16] Van Schaik JE, Dominici N. Motion tracking in developmental research: methods, considerations, and applications[J]. Prog Brain Res, 2020, 254(7): 89-111.
[17] Quiñonez Y, Lizarraga C, Aguayo R. Machine learning solutions with MediaPipe[C]//11th International Conference on Software Process Improvement (CIMPS), 2022.
[18] Huang X, Chen X, Shang X, et al. Image-recognition-based system for precise hand function evaluation[J]. Displays, 2023, 78(3): 102409.
[19] 伊鸣, 黄新瑞, 韩彤妍, 等. 新生儿运动发育评估系统, 方法, 装置及存储介质: ZL202210622070.5[P]. 2023-02-28.
Yi M, Huang XR, Han TY, et al. System, method, device and storage media for neonatal motor development assessment: ZL202210622070.5[P]. 2023-02-28.
[20] 伊鸣, 黄新瑞, 周非非, 等. 基于图像识别的手功能精准评估系统及方法: ZL202111281915.0[P]. 2022-12-09.
Yi M, Huang XR, Zhou FF, et al. Accurate evaluation system and method of hand function based on image recognition: ZL202111281915.0[P]. 2022-12-09.
[21] Prechtl HF. General movement assessment as a method of developmental neurology: new paradigms and their consequences[J]. Dev Med Child Neurol, 2001, 43 (12): 836-842.
[22] Stenum J, Cherry-Allen KM, Pyles CO, et al. Applications of pose estimation in human health and performance across the lifespan[J]. Sensors, 2021, 21(21): 7315.
[23] Poewe W, Seppi K, Tanner CM, et al. Parkinson disease[J]. Nat Rev Dis Primers, 2017, 3(1): 17013.
[24] Lam WW, Tang YM, Fong KN. A systematic review of the applications of markerless motion capture (MMC) technology for clinical measurement in rehabilitation[J]. J Neuroeng Rehabil, 2023, 20(1): 57.
[25] Alarcón-Aldana AC, Callejas-Cuervo M, Bo AP. Upper limb physical rehabilitation using serious video games and motion capture systems: a systematic review[J]. Sensors, 2020, 20(21): 5989.
[26] Stahl A, Schellewald C, Stavdahl Ø, et al. An optical flow-based method to predict infantile cerebral palsy[J]. IEEE Trans Neural Syst Rehabil Eng, 2012, 20(4): 605-614.
[27] Khan T, Nyholm D, Westin J, et al. A computer vision framework for finger-tapping evaluation in Parkinson's disease[J]. Artif Intell Med, 2014, 60(1): 27-40.
[28] Orlandi S, Raghuram K, Smith CR, et al. Detection of atypical and typical infant movements using computer-based video analysis[C]// 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018: 3598-3601.
[29] Li MH, Mestre TA, Fox SH, et al. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation [J]. J Neuroeng Rehabil, 2018, 15(1): 97.
[30] Liu Y, Chen J, Hu C, et al. Vision-based method for automatic quantification of Parkinsonian bradykinesia[J]. IEEE Trans Neural Syst Rehabilitation Eng, 2019, 27(10): 1952-1961.
[31] McCay KD, Ho ES, Marcroft C, et al. Establishing pose based features using histograms for the detection of abnormal infant movements[C]// 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.
[32] Tsuji T, Nakashima S, Hayashi H, et al. Markerless measurement and evaluation of general movements in infants[J]. Sci Rep, 2020, 10(1): 1422.
[33] Chambers C, Seethapathi N, Saluja R, et al. Computer vision to automatically assess infant neuromotor risk[J]. IEEE Trans Neural Syst Rehabilitation Eng, 2020, 28(11): 2431-2442.
[34] Rupprechter S, Morinan G, Peng Y, et al. A clinically interpretable computer-vision based method for quantifying gait in Parkinson's disease[J]. Sensors, 2021, 21(16): 5437.
[35] Morinan G, Dushin Y, Sarapata G, et al. Computer vision quantification of whole-body Parkinsonian bradykinesia using a large multi-site population[J]. npj Parkinsons Dis, 2023, 9(1): 10.
[36] Schmidt W, Regan M, Fahey M, et al. General movement assessment by machine learning: why is it so difficult?[J]. J Med Artif Intell, 2019, 2(7): 15.
[37] Lin B, Luo W, Luo Z, et al. Bradykinesia recognition in Parkinson's disease *via* single RGB video[J]. ACM Trans Knowl Discov Data, 2020, 14(2): 16.
[38] Lee JT, Park E, Jung TD. Machine learning-based classification of dependence in ambulation in stroke patients using smartphone video data[J]. J Pers Med, 2021, 11(11): 1080.
[39] Rahmati H, Aamo OM, Stavdahl Ø, et al. Video-based early cerebral palsy prediction using motion segmentation [C]//36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2014: 3779-3783.
[40] Rahmati H, Martens H, Aamo OM, et al. Frequency analysis and feature reduction method for prediction of cerebral palsy in young infants[J]. IEEE Trans Neural Syst Rehabil Eng, 2016, 24(11): 1225-1234.
[41] Li T, Chen J, Hu C, et al. Automatic timed up-and-go sub-task segmentation for Parkinson's disease patients using video-based activity classification[J]. IEEE Trans Neural Syst Rehabilitation Eng, 2018, 26(11): 2189-2199.
[42] Dai X, Wang S, Li H, et al. Image-assisted discrimination method for neurodevelopmental disorders in infants based on multi-feature fusion and ensemble learning[M]. Springer: Cham, 2019.
[43] McCay KD, Ho ES, Shum HP, et al. Abnormal infant movements classification with deep learning on pose-based features[J]. IEEE Access, 2020, 8(3): 51582-51592.
[44] Williams S, Relton SD, Fang H, et al. Supervised classification of bradykinesia in Parkinson's disease from smartphone videos[J]. Artif Intell Med, 2020, 110(11): 101966.
[45] Negin F, Ozyer B, Agahian S, et al. Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders [J]. Neurocomputing, 2021, 446(7): 145-155.
[46] Li H, Shao X, Zhang C, et al. Automated assessment of Parkinsonian finger-tapping tests through a vision-based fine-grained classification model[J]. Neurocomputing, 2021, 441(6): 260-271.
[47] Kojovic N, Natraj S, Mohanty SP, et al. Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children[J]. Sci Rep, 2021, 11(1): 15069.
[48] Nguyen-Thai B, Le V, Morgan C, et al. A spatio-temporal attention-based model for infant movement assessment from videos[J]. IEEE J Biomed Health Inform, 2021, 25(10): 3911-3920.
[49] Lonini L, Moon Y, Embry K, et al. Video-based pose estimation for gait analysis in stroke survivors during clinical assessments: a proof-of-concept study[J]. Digit Biomark, 2022, 6(1): 9-18.
[50] Song C, Wang S, Chen M, et al. A multimodal discrimination method for the response to name behavior of autistic children based on human pose tracking and head pose estimation[J]. Displays, 2023, 76(1): 102360.
[51] Stenum J, Cherry-Allen KM, Pyles CO, et al. Applications of pose estimation in human health and performance across the lifespan[J]. Sensors, 2021, 1(21): 7315.

（编辑：陈丽霞）