复杂场景下的人体行为识别研究新进展

雷 庆[1,2,3] 陈锻生[1] 李绍滋[2,3]
(华侨大学计算机学院 厦门 361021)[1]
(厦门大学信息科学与技术学院智能科学与技术系 厦门 361005)[2]
(厦门大学福建省仿脑系统重点实验室 厦门 361005)[3]

摘 要 人体行为识别是计算机视觉的研究难点和热点,主流的研究框架包括行为特征提取、人体行为表示和识别算法3个方面,目前简单场景下的人体简单动作的识别已基本得到解决,而复杂场景下的行为识别仍面临很多困难。对近几年人体行为识别的发展做了比较详细的研究,从人体行为识别的研究范畴、特征提取以及行为模型等方面综述了目前复杂场景下人体行为识别的研究方法。与已有的相关综述文献不同的是,文中结合了近三年国内外人体行为识别领域中新的研究热点和成果,如姿态特征的提取和表示、基于稀疏编码和卷积神经网络的人体行为表示方法等。最后阐述了该领域目前存在的困难以及可能的发展趋向。

关键词 人体行为识别,行为特征提取,行为表示,计算机视觉
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.001

Advances on Human Action Recognition in Realistic Scenes

LEI Qing[1,2,3] CHEN Duan-sheng[1] LI Shao-zi[2,3]
(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)[1]
(Cognitive Science Department, School of Information Science and Technology, Xiamen University, Xiamen 361005, China)[2]
(Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen University, Xiamen 361005, China)[3]

Abstract Human action recognition has become a hot and difficult spot currently in the domain of computer vision. The framework of mainstream methods includes visual feature detection, action representation and action classification. Action recognition in simple scenes has been implemented at present. This paper introduced in detail the research of human action recognition in realistic scenes from perspectives of research scope, feature detection, and action modeling. Unlike several recent published researches, we analyzed the state-of-the-arts and advances of this field, such as pose estimation, sparse coding based or deep learning based human action representation etc. Finally, the problems, difficulties as well as possible solutions were discussed.

Keywords Human action recognition, Visual feature detection, Action representation, Computer vision

1 引言

视觉是人类感知外部世界、获取环境信息的重要途径,相对于听觉、触觉和嗅觉等其他感觉器官,它具有独特的时空特性。计算机视觉就是使计算机获得与人类似的视觉感知能力,能够自动对环境中的物体以及它们的活动进行分析和识别,由此衍生出计算机视觉领域中两个备受关注的研究方向:物体识别和人体运动分析。目前,人体运动分析已成为计算机视觉领域中最活跃的研究主题之一,它的核心是利用计算机视觉技术对图像序列进行分析,识别出人的动作,通过连续的跟踪并结合上下文环境对其行为进行推理和描述。

人体运动分析包含人体检测、目标分类和跟踪、动作识别和高层行为理解等内容,在智能监控、视频检索、智能人机接口、智能家居环境和身份鉴别等领域有着广泛的应用前景[1,2]。(1)智能监控系统。传统的人工监控系统已不能适应社会发展的需要,在数量巨大的监控视频数据面前,单纯依靠人来处理监视内容难以构成真正安全的系统。将计算机视觉技术应用到监控系统中,让计算机对监视视频中的内容进行自动理解、识别出异常行为并发出报警,这已成为新一代的智能监控技术。(2)基于内容的视频存储和检索。传统的视频检索主要依靠用户对其标注的文字或数字,受主观影响大且存在错误,如果计算机能够自动对视频的内容进行理解,那么基于内容的视频检索将大大地改进检索的效果。由于人在现实世界的主导地位,人体大多成为了视频记录的主要对象,因此对人体动作的自动识别将有助于提高对视频进行自动标注的准确性。(3)智能人机接口。在高级用户接口的应用领域,我们希望未来的人机接口方式能够更好地理解人和机器之间的交互,视觉信息可以作为自然语言和声音之外的有效补充

到稿日期:2013-06-25 返修日期:2013-08-16 本文受国家自然科学基金(60873179),高等学校博士学科点专项科研基金(20090121110032),中央高校基本科研业务费专项资金(11QZR04)资助。

雷 庆(1980—),女,博士生,讲师,主要研究方向为图像/视频检索、人体运动分析,E-mail:leiqing@hqu.edu.cn;陈锻生(1959—),男,博士,教授,主要研究方向为计算机视觉与模式识别、机器学习与数据挖掘;李绍滋(1963—),男,博士,教授,主要研究方向为计算机视觉与模式识别。

来实现更加智能的人机交互方式，因此对人体运动进行分析，实现对人体动作的自动识别和理解的计算机视觉技术是实现智能人机接口的重要前提。(4)智能家居环境。在进入高龄化社会的国家，对独居老人的监护问题已成为突出的社会问题，建立智能的家居环境能够缓解这一社会矛盾。通过对老人的日常生活进行监控，对其行为进行自动的分析和理解，实时地发现摔倒或被盗等异常行为，从而为包括老人、残疾人等在内的特殊群体提供有效的监护。相关的应用还包括对家用电器的智能控制，例如识别出人体的睡眠状态从而关闭电视机电源，这将有效地节省电源。

目前人体行为识别的研究对象按照其复杂程度分为4个层次：姿势(gestures)、动作(actions)、交互动作(interactions)和群体行为(group activities)。姿势是指身体部位的简单移动，例如“伸手”、“抬腿”，是构成人体动作的基本部件。动作是由单人的多个姿势按顺序组成的，例如“走”、“挥手”和“推”等。交互动作指发生在两个人或者人和物体之间的动作，例如“两人打架”、“偷包”等。而群体行为则指发生在多人或多个人-物构成的群体中的行为，例如“游行”、“团体会议”和“群体斗殴”等。简单的动作识别对已经分割好的、仅包含单人且单个动作的视频进行分析，将视频分类到已定义的动作类别中，而更一般的任务是对输入视频中发生的若干人体行为进行连续识别，检测出每个动作的起点和终点。因此根据要解决的问题和任务的不同，动作识别可以分为：动作分类和动作检测。动作分类通常的处理方法是从视频中抽取出特征，并根据特征匹配的结果分配相应的动作标签。而动作检测不仅需要识别发生了什么类别的动作，还需要确定动作发生的时间和空间位置，由此可以看出，动作检测更具挑战性。

而根据动作识别即所处理的视频所处的录制环境，可以将视频数据分为：受控环境视频、受限环境视频和开放环境视频。受控环境的典型商业应用是电影制作中建立的人体运动捕捉系统，它往往给人体的关节和四肢贴上标签，为更好地检测出标签所标识的身体部位，可以对光照进行调节并且放置多个摄像头为3D重建取得足够多的视角，目前在受控环境下已经能够识别出人脸表情的变化以及手指的运动。受限环境往往存在一定的假设，例如假设人体完全可见或者有合适的照明条件，典型的受限环境为摄像头位置固定和参数已知的视频监控系统，由于摄像头静止，可以运用背景减除技术来获得人体形状模板，并将这些模板进一步用于识别人体动作，然而在这一环境下开发的动作识别方法必须对人体大小、衣着以及动作的细微变化具备鲁棒性。开放环境下的视频只能对录制条件施加很少的限制，例如电视、电影、体育节目、音乐视频等，在这样的真实视频数据下进行动作识别的难点和挑战包括：视角、尺度和光照变化、人体和目标的部分遮挡，以及复杂背景的影响等。近十年来，人体运动分析受到人们的持续关注，从2001年到2012年在国际权威期刊IJCV、TPAMI、CVIU、IVC和顶级学术会议ICCV、CVPR、ECCV上发表了近1000篇相关学术论文。

2 人体动作识别的特征提取

目前采用的人体动作特征提取方法主要分为全局特征方法和局部特征方法，全局特征方法往往利用背景减除或人体检测技术对人体的位置和运动进行分割，并借助跟踪技术将得到的人体位置和速度等参数用来表示人体运动，采用的模型包括整体模型和人体部位模型。局部特征方法则往往定义某种显著性函数，并通过对视频的密集计算找到显著性最大的时空位置及其所在的尺度，对该位置的邻域范围进行特征描述得到局部特征描述子。常用的特征主要分为以下4类(见表1)：表观特征、运动特征、姿态特征和时空兴趣点。

表1 人体行为识别特征提取方法分类

类别		形式	代表性例子
全局特征	整体模型	表观特征	运动能量/历史图 MEI/MHI[4]，3D时空立方体[5]
		运动特征	光流 optical flow[6]
		混合特征	梯度直方图/光流直方图 HOG/HOF[10]
	人体部位模型	姿态特征	形变部位模型 DPM[12]
局部特征	显著性位置提取	时空兴趣点	3D Harris[18]，3D SIFT[20]

2.1 整体模型

整体模型首先从视频中检测出人体位置，利用动作中人体位置信息的变化，学习出一个与身体部位无关的整体运动模型。整体模型不需要对身体各个部位进行检测，而是用整体的结构和动态信息来表示动作。其实现方法是抽取以人体为中心的感兴趣区域ROI，比如完整的人体剪影(Silhouette)、以人体为中心的团块(Blob)或者不规则区域(region)，抽取该区域的大小、颜色、轮廓和形状等特征来对人体动作进行表征。整体模型所采用的底层特征分为两类：表观特征和运动特征。

2.1.1 表观特征

表观特征是利用人体几何结构、轮廓等信息来估计运动目标每个时刻的静止姿势，利用得到的姿态序列来描述人体运动，通常采用减背景或差分图像方法来检测运动区域。例如，Yamato等[3]最早提出利用剪影图像来表示动作，通过对剪影图像区域划分网格并计算每个单元的前景和背景像素的比例，得到一个被称为单词的姿势表示，使用隐马尔科夫模型(HMM)对一组单词表示的姿势序列建立动作模型来识别网球运动中的动作。Bobick和Davis[4]对图像序列中的相邻图像进行差分，得到运动能量图MEI和运动历史图MHI，并将其用来表示动作。运动能量图是通过对相邻图像进行差分并二值化，然后叠加起来得到的运动区域图像，而对运动区域按发生的时间顺序进行加权(时间顺序越靠后权值越高)得到运动历史图MHI，然后从MHI中提取出基于矩的特征来表示动作。Blank等[5]提出一种基于剪影图像的时空形状动作表示方法，亦即将使用减除背景得到的剪影图像沿时间轴组合起来形成一个3D立方体，从3D XYT时空立方体中抽取出一组特征对动作进行表征。典型的时空识别方法通过训练为每个动作类别建立一个3D XYT时空立方体模型，识别阶段采用模板匹配的方法将测试视频的3D时空立方体与每个动作模型进行比较，从形状和外观上测量它们之间的相似度，将新的视频分到相似度最高的动作类别中。

在摄像头固定的情况下，可以利用背景减除技术来轻松获得人体剪影和轮廓等形状信息，因此表观特征在视觉监控系统应用中非常流行。然而在复杂场景和摄像头运动的情况下，准确的剪影和轮廓难以获取，而且在人体被遮挡的情况下，无法获得准确的人体外观。

2.1.2 运动特征

运动特征方法不考虑人体结构的任何形状信息，而是直

接从图像序列中提取出目标运动信息(如运动方向、轨迹、位置、速度等)来表征运动状态。运动特征提取的最典型的代表是光流法,在假设相邻两幅图像中对应两点的灰度不变的情况下,计算空间运动物体表面上像素点运动产生的瞬时速度场来对运动进行表征。光流法在没有背景区域的任何先验知识条件下就能实现对运动目标的检测和跟踪,但计算量较大。

Polana 和 Nelson[6] 提出了一种人体跟踪框架,对以人体为中心的区域进行网格划分,并计算以网格为单元的光流幅值来表示动作,将这一框架运用于周期性运动的动作表示取得了很好的效果,例如走、跑、游泳等。Efros 等[7] 为了解决低精度运动图像的光流信息易受噪声干扰的问题,提出一种新的描述运动的光流特征,即模糊光流特征(Blurred Optical Flow)。该方法先计算前景目标的光流场,然后将光流场映射到 4 个通道,通过多次高斯滤波泛化 4 个通道生成动作模板。该方法的计算量随着类别数的增加而急剧增大,很难实现在线实时的行为识别。此外,Dalal 等[8] 将微分光流直方图和梯度方向直方图同时作为输入特征来分析人体的行为,并取得了很好的实验结果。

基于运动特征的目标检测适用于中远距离视觉和能见度低的情况,在这些条件下表观特征往往难以很好地对运动进行表征。然而,大多数的光流计算方法相当复杂,且抗噪性能差,如果没有特别的硬件装置则不能被应用于全帧视频流的实时处理。

2.1.3 *表观和运动相结合*

表观特征和运动特征各有优势,可以相互补充。Wang 和 Mori[9] 同时使用全局特征模板和模糊光流特征,采用一种改进的判别式模型来识别视频序列中的人的行为。实验表明他们的结果优于大部分的方法。Laptev 和 Perez[10] 将梯度直方图所代表的表观特征和光流直方图(histograms of optical flow)所代表的运动特征融合起来,在真实视频中取得了较好的效果。

2.2 人体部位模型

采用人体部位模型进行动作识别的方法主要依靠身体部位的位置以及运动信息,相关研究主要关注于关节位置的变化以及身体部位的运动轨迹,或者利用人体运动学模型中的身体标志点来识别动作。这种方法往往借鉴神经物理学中有关生物运动的视觉感知知识,例如 Johansson[11] 利用贴在人体关节上亮点的运动(moving light displays,MLD)来研究整个人体的运动。这种方法比较简单,但对人体外观进行了限制,因此其应用领域有很大的局限性。人体部位模型适用于真实视频中的动作识别,但以人体定位为前提,因此其本质上依赖于人体检测的结果。受到当时尚未成熟的人体检测技术的限制,特别是在真实视频中的检测效果比较差,在后续的十多年时间里人体行为识别的研究方法大多都绕过人体检测这一难题,转向底层的基于表观或运动的局部特征表示方法。

然而人体动作最直接的定义是由一组关节位置的运动变化而组成的姿态序列,随着近年来人体检测技术的发展,基于姿态提取和表示的身体部位模型重新受到研究者们的关注,此外基于姿态特征的行为识别方法具有以下 3 个优点:1)视角和表观不变性。基于姿态特征的行为表示受人体行为类内变化的影响较小,特别是三维骨架姿态模型能够对视角和表观的变化保持不变。2)高层语义特征。姿态特征比底层的表观和运动特征包含更高层次的语义,因此有助于后续的分类识别,得到更好的识别效果。3)适应于部分遮挡。基于姿态提取和表示的三维骨架模型还可以适应于遮挡造成的部分身体部位不可见的情况,有效地识别出行为类别并估计出遮挡部位的位置。

Felzenszwalb 等[12] 提出了一种基于形变部位模型(Deformable Part Model,DPM)的目标检测方法,DPM 本质上是一个星型结构的部位模型,包括一个全局检测器(或者称之为根节点滤波器,root filter)、一组部位检测器和一个考虑部位检测器和全局检测器之间的几何关系的部位形变模型。检测时,DPM 在图像中的某个位置和尺度下的值等于全局检测器的值加上各部位检测器和的最大值,然后减去各部位的形变代价。全局检测器和各部位检测器值的计算方式为滤波器(一组权重)和子窗口特征向量的点积。Angela Yao 等[13] 提出了一种将动作识别和姿态估计结合起来共同求解的框架,将姿态估计表示为在一组基于特定类别的流形上的优化问题,将基于表观特征的 2D 动作识别中计算得到的动作类别的可信度作为先验分布帮助简化姿态估计的优化过程,而估计出的姿态通过提取 5 种类型的关节几何信息特征来帮助进行 3D 的动作识别。姿态估计采用 26 个关节(每个关节 3 个方向的旋转角度)和 1 个根节点(3 个维度的位置和 3 个维度的平移)的人体骨架模型来表示姿态,计算姿态中关节点之间的几何关系(关节点距离、平面特征、归一的平面特征、速度特征、归一的速度特征),将其作为姿态特征来进行 3D 动作识别。Bangpeng Yao 等[14] 通过对人-物交互行为中物体和人物姿势间相互上下文的建模来识别静态图像中人-物交互行为,以克服交互行为中的物体部分可见以及人体部位被部分遮挡的情况。其提出一种包含图像证据、身体部位、物体、人物姿势和动作几个层次的相互上下文条件随机场模型,用于模拟物体和身体部位之间以及身体部位之间的空间关系,并提出一种最大间隔学习算法来获取模型的参数和权重。Packer 等[15] 提出了一个结合基于姿态特征和基于表观特征的两种行为识别方法的框架,同时实现了人体行为的分类识别以及行为对象的检测。利用最近提出的一种姿态跟踪器的输出结果及其测量方法,将姿态轨迹表示为一组有意义的由动态实例值和间隔组成的序列,在动态实例值中加入视觉分量将一个人体动作表示为一种视觉运动轨迹,并训练出一个用于模拟行为对象位置的潜在结构支持向量机。该模型能够对运动轻微变化保持鲁棒性,并能够适应于运动轨迹长度的变化。Yao[16] 建立了统一的实验环境,对表观特征和姿态特征在人体行为识别中的性能进行了对比,实验结果表明姿态特征的识别效果明显优于底层表观特征,甚至在复杂场景中噪音点的严重干扰下,姿态估计也能够对行为识别起到很好的促进效果。

2.3 局部特征方法

局部特征抽取视频中局部区域的形状和运动信息,不需要人体位置或者身体部位的任何先验知识,提供了针对时空位置和尺度以及场景中的复杂背景和多种运动情况下的事件的独立表示,局部特征可以从视频中直接提取,因此避免了运动分割或者人体检测等其他预处理方法造成的错误。实现方法是首先定义以某种显著性函数表示的特征检测算子,对视频中的时空位置和所处的尺度进行计算,找到使得显著性函数值最大的位置和尺度,然后采用某种特征描述方式计算以显著性位置为中心的局部邻域的形状和运动特征,得到特征

描述算子(目前流行的特征描述方式包括单尺度/多尺度的高阶导数、光流直方图、时空梯度直方图等),最后建立动作分类器进行行动作识别。

Laptev[17,18]将二维图像中的 Harris 角点检测技术扩展到三维时空领域中,从视频中检测出丰富的代表时空事件的兴趣点,建立以兴趣点为中心的时空立方体并抽取光流直方图和梯度直方图的联合特征 HOG/HOF 对运动进行表征,最后建立基于 SVM 的动作分类器对动作进行分类。Dollar[19]提出了基于 Gabor 滤波器和 Gaussian 滤波器的时空兴趣点检测方法,首先使用高斯滤波器在空间域上对图像进行滤波,然后使用一维的 Gabor 滤波器在时间域上作用于图像序列,并提出了基于 Cuboids 的特征描述算子,采用 PCA 算法对特征进行降维,最后采用基于 χ^2 距离的最近邻分类器进行动作识别。Scovanner 等人[20]提出了一种三维时空梯度方向直方图特征描述子 3D SIFT,它可以看作是经典的尺度不变特征变换描述算子(2D SIFT)从静态图像到视频序列的扩展,由于能够更好地适应缩放、旋转变换以及噪声带来的影响,采用 3D SIFT 的特征描述方式能够准确地捕捉到视频数据的时空特性本质。Oikonomopoulos[21]提出了基于熵的时空显著性区域检测算子,用以计算给定时空位置为中心的圆柱体的熵,为了获得更多稳定的兴趣点以及动作的稀疏表示,对兴趣点进行聚类并通过阈值来选取局部最大候选值。Willems[22]将图像的团块特征检测中常用的 Hessian 矩阵检测算子扩展到时空域,得到 Hessian3D 检测算子,将 3D Hessian 矩阵的行列式值作为显著性测量值,旨在获得更密集的、尺度不变的和更高计算效率的兴趣点检测算子。

对上述几种特征包含的语义层次及其在复杂场景下存在的背景噪音、摄像机运动等影响下的鲁棒性进行了比较,结果如表 2 所列。

表 2 人体行为识别特征比较

特征	语义层次	复杂场景下的挑战				
		类内变化	背景噪音	光照变化	视角变化	部分遮挡
表观特征	底层	低	低	低	低	低
运动特征	底层/中层(运动轨迹)	中	低	中	低	低
3D 姿态特征	中层	高	低	高	高	高
时空兴趣点	底层	高	低	高	低	低

3 人体行为表示方法

目前复杂场景下的人体行为表示方法主要包括:结合上下文的时空特征袋表示、人体行为的稀疏表示、基于深度学习的行为表示。

3.1 结合上下文的时空特征袋表示

特征袋是在局部特征中流行的动作表示方法,来源于文档检索中以无序单词集合来表示文本,词袋模型使用单词的词频分布来描述文本,并在自然语言理解和信息检索领域得到了广泛的应用。在特征袋表示方法中,首先利用特征检测算子从序列中找出一组显著性位置(例如兴趣点),然后采用某种特征描述方式(例如 HOG/HOF/3DSIFT)对每个显著性位置的邻域范围进行描述并得到一组特征集合,使用聚类算法对训练数据集中提取出的特征集合进行聚类,将生成的聚类中心看作时空单词 $w_i=\{f_1,\cdots,f_m\}$,m 为特征维数,f_i 表示时空单词的第 i 个特征分量。所有时空单词组成的集合 $V=\{w_1,\cdots,w_n\}$ 称为时空码本,其中 n 为聚类中心的个数。对于不同的动作视频,从训练集中按照上述的步骤训练出对应于不同动作类别的时空码本,在后续的动作识别过程中通过计算兴趣点的特征与时空单词的距离实现对兴趣点的分类,并进一步完成动作分类。

词袋方法中对动作进行表示的直方图只包含特征描述子类别(visual words)在视频中出现的统计值,丢弃了这些特征描述子类别之间的时间和空间关系。随着研究的深入,特征之间的时空关联被证明对真实场景下的交互动作例如人-人交互以及人-物交互的识别具有重要的意义。针对词袋模型存在的缺点,最近的研究工作考虑了特征之间的时空关系,建立了一些包含更多语义的动作模型。Laptev[9]对时空立方体进行划分,计算不同划分方式下的几种特征描述子,将这些特征融合起来放入多核的 SVM 分类器中进行分类识别,采用贪心算法为每个动作类别寻找最优的特征类别以及组合方式。SUN[23]提出了一种对时空上下文信息进行层次建模的方法,首先采用 SIFT 算法对视频进行兴趣点检测并进一步抽取出兴趣点的运动轨迹,然后抽取了兴趣点运动轨迹的 3 层次的时空上下文信息,最后建立了基于 MKL 的多核学习非线性 SVM 分类器对特征进行分类。Gupta[24]提出了一种集成不同感知元素的识别人-物交互行为的贝叶斯方法,将场所、对象、动作和对象反应 4 个不同的感知元素结合起来,将对特定类别动作的语义理解转化为一组对感知元素的功能限制以及元素之间的空间位置限制,并施加到每一种感知元素上,得到了比只依靠单个元素进行识别更好的识别效果。

3.2 人体行为的稀疏表示

稀疏表示理论来源于信号处理领域的压缩传感理论[25-28],该理论指出若信号的某个变换域是稀疏的或可压缩的,则可以利用与变换矩阵非相干的测量矩阵将变换稀疏线性投影为低维观测向量,通过进一步求解稀疏优化问题就能够从低维观测向量精确地重建原始信号。其中"稀疏"表示的含义是指将输入信号表示为超完备词典中一组基向量的线性组合,其中大部分基向量的系数都为 0。近年来,稀疏表示在信号处理和计算机视觉领域引起了广大研究者的关注,而且已经被证明为高维信号的获取、表示和压缩提供了一种强有力的工具。最近的压缩传感理论研究表明,利用信号稀疏表示的先验条件,可以在测量矩阵满足约束等距性条件下重构原信号,并证明了若信号足够稀疏,重构算法可通过 1 范数最小化来求解。基于该理论,稀疏表示已被成功应用于运动和数据分割、图像去噪和修复与图像分类等。大部分应用都是以稀疏性作为先验条件,并得到了很好的结果。

Wrigth 等[29]将不同表情、光照以及遮挡和掩饰下的人脸正面图像的自动识别投射为多元线性回归模型中的模式分类问题,通过 1 范数最小化求解得到的稀疏表示来克服人脸识别中的两个关键难题:遮挡和噪声。本文主要研究通过超完备词典学习获得的稀疏表示在视频中人体动作的分类问题中的有效运用。提出的方法还能够适用于解决其他模式识别问题,如人脸识别、人体检测和其他物体识别等。Guha[30]研究了将通过超完备词典学习得到的稀疏表示运用于人体行为识别问题中进行行为表示的有效性;提出了共享型(shared)、特定类别型(class-specific)以及连接型(concatenated) 3 种词典建立方法,并为每一种方法设计了相应的分类算法。实验表明,基于稀疏编码的行为表示和分类方法在识别效果上明显

优于传统的基于向量量化的词袋模型。Castrodad[31]等提出了两层的稀疏编码方法用于人体行为识别，第一层为每个动作类别学习出一个词典并得到该类训练样本相应的稀疏编码，然后将第一层学习得到的一组特定动作类别的词典集合输入到第二层中，通过对词典类间关系的学习得到一组新的词典。实验结果表明，两层的稀疏表示比单层的稀疏编码能取得更好的效果。

稀疏表示相对于词袋模型的改进在于稀疏编码允许将特征映射到少数几个不同的视觉单词上，从而有效地降低了近似误差并得到了更紧致的词典。然而并没有解决词袋模型的本质问题，如仅仅依靠特征在局部表观上的相似性，忽略了特征之间的时空关系；且仍然依靠 K-means 聚类算法对词典的学习进行初始化，其采用的均值-偏移更新方法往往导致对特征空间的不均匀划分。

3.3 基于深度学习的人体行为模型

近年来针对传统浅层结构机器学习存在的数据表示缺乏判别能力和有效语义等问题，在人脑认知过程的深度架构的启发下，研究者提出了结合多层非线性映射与无监督学习的深度学习[32]建立特征的层次结构来获取更有效的数据表示。深度学习的目的是从低层次的特征中学习出特征之间的层次关系从而得到高层次的特征。典型的深度学习结构包括：卷积神经网络 CNN、深度置信网 DBN 和自动编码器（AutoEncoder）等。CNN 是包含卷积和下采样两种操作的一种卷积网络，用于描述数据的后验分布提供对模式分类的能力。目前卷积神经网络已广泛运用于自然语言处理和计算机视觉的物体识别、图像分割、场景标注等领域中，通过在原始图像上交替运用滤波器和局部近邻汇聚操作得到复杂特征的层次结构，能够在视觉物体识别中取得很好的效果。

Ji 等[33]对深度学习中的 2D CNN 模型进行扩展，提出了一种监控环境中进行人体动作识别的 3D CNN 模型。首先对于输入的图像序列建立 5 种不同的特征通道（灰度、x 方向梯度、y 方向梯度、x 方向光流、y 方向光流），然后在不同的特征通道上计算 3D 卷积，抽取图像的时空特征以得到特征图像，接下来对特征图像进行降采样以得到新的特征图像，重复卷积和降采样两个步骤，最后将得到的不同通道上的一维特征串联起来形成最后的特征表示。提出的模型在真实场景下的人体动作识别中得到了很好的效果。Le 等[34]采用无监督的特征学习方式直接从视频数据中学习出特征，具体提出了一种扩展的独立子空间分析算法（Independent Subspace Analysis，ISA），从未标记的视频数据中学习出具有不变性的时空特征，采用了层叠（stacking）和卷积（convolution）两种深度学习策略来学习数据的层次化表示并取得了非常好的效果。Farabet 等[35]提出对原始像素点训练得到的多尺度卷积网络抽取密集的图像像素点的特征向量的方法，特征向量是对以每个像素点为中心的多尺度区域中信息的编码。该方法减轻了对特征引擎设计的依赖，能够导出捕捉了图像中像素点的纹理、形状和场景的有力特征表示；此外还提出了一种自动选取场景表示的最优子集部件选取方法。该方法在 Sift Flow 和 the Barcelona Dataset 两个数据集上取得了超过所有已发表方法的准确率，并且在计算时间上更快，对一幅 320×240 的图像进行标注（包括特征提取）的时间不到 1 秒。

综上所述，结合上下文的时空特征袋行为表示是对传统词袋模型的改进，将词袋模型中丢弃的特征之间的时空关系引入到模型中，然而并未从本质上解决词袋模型存在的特征量化误差大以及简单地运用 K-Means 聚类算法导致对特征空间的不对称划分等问题。基于稀疏编码的行为表示将特征映射放宽到可以投影到码本中多于 1 个的基向量上，相比于词袋模型减轻了特征的量化误差，然而词典的学习采用随机方式或仍采用 K-Means 初始化，此外这两种方法都使用底层特征表示数据。基于卷积神经网络的行为模型不依赖于特征引擎的设计，采用层叠和卷积两种技术学习出深层次的特征架构，能够对一定的姿势、光照变化以及背景噪声具备鲁棒性。

上述 3 种模型在真实场景下的复杂行为数据集 UCF Sports、Youtube 和 Hollywood2 上取得的最新识别效果如表 3 所列。这 3 个数据集包含了从大量的电影或个人视频中收集的数据，它们的共同点是摄像机位置不固定且发生运动，动作发生的场景不同，同类动作的类内变化较大，因此都存在较大的挑战。

表 3 行为表示模型在复杂数据集上的识别效果比较

模型	代表方法	数据集		
		UCF-Sports[36]	Youtube[37]	Hollywood2[38]
上下文时空特征袋模型	HOG/HOF[10]	78.10%	*58.14%	45.20%
基于稀疏编码的行为表示	SM[31]	87.60%	88.18%	*52.6%
基于深度学习的行为表示	ISA[34]	86.50%	76.50%	50.80%

注：* 表示将论文方法应用于未报告数据集中得到的平均分类准确率

4 存在的困难及发展的可能趋向

由于人本身是一个复杂的非刚性物体，当前的人体动作识别受到人体外表、姿势、动作、衣着的个体差异、视角变化和摄像机运动、光照变化、遮挡和复杂背景的影响，给底层的动作识别造成很大的困难。此外人体行为的含义往往依赖于行为涉及的对象，这又使得必须对动作进行深层次的理解。

4.1 存在的困难

4.1.1 类内变化和类间差异

相同的动作在不同个体上可能表现出差异，构成差异的因素包括：人体的衣着和尺度等表观上的差异、动作执行的速度等。表观和速度上的不同，分别对动作的空间域和时间域造成变化，一个鲁棒的人体动作识别算法应尽量避免受到类内变化的影响，具备较强的类内泛化能力，同时又能够从数据中抽取出有判别力的信息从而具备类间判别能力。此外随着动作类别的增加，类别之间会出现更多的重叠，例如排球的扣球和网球的正手发球，这两种动作包含相同的姿势，从而为识别造成困难，此时对场景信息进行分析能够更有效地帮助提高识别结果。

4.1.2 复杂场景

复杂场景下的人体行为识别往往受到背景混杂、光照变化、遮挡等环境因素的影响。在复杂的动态环境中人体的某些部位可能被障碍物或其他人体遮挡，或者由于照明的变化使得人体的可见程度下降，使得人体的外观产生很大变化，从而对人体的检测和跟踪造成了很大的困难。此外视频图像的分辨率也在很大程度上影响着动作检测和识别的结果。好的识别方法应该根据任务的不同，能够从数据中提取出对后续识别任务有帮助的目标信息和环境信息。

4.1.3 摄像环境(摄像机移动/视角变化)

遮挡造成的影响可以通过引入多个不同视角的摄像头来解决,动作识别可以通过对多视角的观察值进行 3D 重建来实现,然而这一方法需要在学习阶段提供大量的训练样本进行训练,无法达到实时的要求。此外摄像头运动还可能造成人体尺度的变化,使得固定尺度下抽取出的人体动作特征难以准确地对人体动作进行表征,从而降低动作识别的准确率。基于 3D 姿态估计的行为识别能够很好地适应视角变化,解决身体部分遮挡下的行为识别问题,目前这一领域的研究已取得初步成果。

4.1.4 大数据应用环境

人体行为识别在智能监控系统、基于内容的视频检索中存在巨大的应用前景,仅以 YouTube 为例,据统计每分钟有 20 小时的视频量被上传,然而已有的识别方法无法满足这些应用中实时性和准确性的需求。在数量巨大的视频数据中,如何针对不同的任务从海量的数据中抽取出有用的信息,如何建立鲁棒的快速的行为检测算法,仍尚未解决。

4.2 发展的可能趋向

4.2.1 基于深度学习的数据表示

简单场景下的人体行为识别往往采用手动设计的特征引擎进行行为特征提取,而特征往往是与任务高度相关的,在不同任务下很难预先知道哪些特征是重要的、哪些是不重要的,因此传统的特征提取方法很难获取到鲁棒的行为特征。近年来,研究者针对传统浅层结构学习算法的不足,提出了结合多层非线性映射与无监督学习的深度学习[32]来完成复杂函数的逼近,并建立特征的层次结构来获取更有效的数据表示[33,34]。利用深度学习可以从低层次的特征中学习出特征之间的层次关系从而得到高层次的特征,这将有助于简化后续的识别算法并得到更准确的识别结果。

4.2.2 建立有效的行为模型

目前主流的行为表示方法主要针对词袋模型进行改进,提出了层次化的词典学习以及结合上下文关系的码本建立等方法[39,40],然而这些改进并不能从本质上解决词袋模型存在的对噪音点敏感、量化误差大等缺点。最近,研究者尝试将稀疏表示理论应用到计算机视觉的相关问题中,并在人脸识别以及简单场景下的人体行为识别中取得了较好的效果[29,30]。然而复杂场景下的行为识别还需要考虑类间码本之间的相互关系,选出更具判别力的基向量,从而建立更深层次的超完备词典来表示行为。

4.2.3 对场景的理解

场景信息对人-人交互、人-物交互行为以及行为语义的高层理解起着至为关键的作用,现有的研究方法大多是通过目标检测、姿态估计、行为建模 3 个步骤[41,42],利用已有的目标识别技术从静态图像或视频中检测出物体位置,并采用形变部位模型表示人体动作,最后使用条件随机场或动态贝叶斯网络等概率模型对交互行为进行建模。该方法的缺点是模型复杂、参数多且训练时间长,如何从场景中提取出有用的信息,并以什么方式更好地融入到模型中是一个值得关注的问题。

4.2.4 多摄像机的特征融合

多摄像机环境能够解决单摄像机系统中存在的视频特征与三维模型对应的歧义问题,在多视角环境下能够提供深度信息[43,44]并通过准确恢复三维模型的参数来帮助解决遮挡问题。将形状、纹理、颜色等表观特征,光流、速度、运动轨迹等运动特征和深度信息等多种特征融合起来是解决复杂人体行为识别的一个重要途径。此外,多摄像机环境下的三维姿态估计也已成为目前人体行为识别研究的一个热点。

结束语 本文结合近年来国内外人体行为识别领域中新的研究热点和成果,对目前主流的人体行为识别方法所采用的一般步骤进行了总结。从人体行为识别的研究范畴、特征提取以及行为模型等方面综述了目前复杂场景下人体行为识别的研究方法,最后阐述了该领域目前存在的困难以及可能的发展趋向。

参 考 文 献

[1] 徐光祐,曹媛媛. 动作识别与行为理解综述[J]. 中国图象图形学报,2009,14(2):189-195

[2] 黎洪松,李达. 人体运动分析研究的若干新进展[J]. 模式识别与人工能,2009,22(1):70-78

[3] Yamato J,Ohya J,Ishii K. Recognizing human action in time-sequential images using hidden Markov model[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition. 1992:379-385

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

[5] Blank M,Gorelick L,Shechtman E,et al. Actions as space-time shapes[C]// Proceedings of the International Conference On Computer Vision (ICCV'05). 2005:1395-1402

[6] Polana R, Nelson R C. Detection and recognition of periodic, nonrigid motion[J]. International Journal of Computer Vision (IJCV),1997,23(3):261-282

[7] Efros A A,Berg A C,Mori G,et al. Recognizing action at a distance [C]// Proceedings of the International Conference on Computer Vision (ICCV'03). 2003:726-733

[8] Dalal N,Triggs B. Histograms of oriented gradients for human detection[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005,1:886-893

[9] Wang Yang,Mori G. Learning a discriminative hidden part model for human action recognition[C]// Advances in Neural Information Processing Systems (NIPS). 2008,21:1721-1728

[10] Laptev I,Marszałek M,Cordelia Schmid,et al. Learning realistic human actions from movies[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08). 2008:1-8

[11] Johansson G. Visual Perception of Biological Motion and a Model for its Analysis[J]. Perception and Psychophysics, 1973, 14(2):210-211

[12] Felzenszwalb P F, Girshick R B, McAllester D. Cascade Object Detection with Deformable Part Models[C]// Computer Vision and Pattern Recognition (CVPR). 2010:2241-2248

[13] Yao A,Gall J,Gool L V. Coupled Action Recognition and Pose Estimation from Multiple Views[J]. International Journal of Computer Vision (IJCV),2012,100(1):16-37

[14] Yao Bang-peng, Li Fei-fei. Modeling mutual context of object and human pose in human-object interaction activities[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),2012,34(9):1691-1703

[15] Packer B,Saenko K,Koller D. A combined pose,object,and feature model for action understanding[C]// Computer Vision and

Pattern Recognition (CVPR). 2012:1378-1385
[16] Yao A, Gall J, Fanelli G, et al. Does Human Action Recognition Benefit from Pose Estimation? [C]// Proceedings of the British Machine Vision Conference. BMVA Press, 2011:1-11
[17] Laptev I, Caputo B, Schuldt C, et al. Local velocity-adapted motion events for spatio-temporal recognition [J]. Computer Vision and Image Understanding (CVIU), 2007, 108(3):207-229
[18] Laptev I, Lindeberg T. Space-time interest points [C]// Proceedings of the International Conference on Computer Vision (ICCV'03). Nice, France, 2003, 1:432-439
[19] Dollar P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]// Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2005:65-72
[20] Scovanner P, Ali S, Shah M. A 3-dimensional SIFT descriptor and its application to action recognition[C]// Proceedings of the International Conference on Multimedia (MultiMedia'07). Augsburg, Germany, 2007:357-360
[21] Oikonomopoulos A, Patras I, Pantic M. Spatio-temporal salient points for visual recognition of human actions[J]. IEEE Transactions on Systems Man And Cybernetics (SMC), 2006, 36(3):710-719
[22] Willems G, Tuytelaars T, Van Gool L J. An efficient dense and scaleinvariant spatio-temporal interest point detector[C]// Proceedings of the European Conference on Computer Vision (ECCV'08). 2008:650-663
[23] Sun Ju, Wu Xiao, Yan Shui-cheng, et al. Hierarchical spatio-temporal context modeling for action recognition[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'2009). 2009:1-8
[24] Gupta A, Kembhavi A, Davis L S. Observing human-object interactions: using spatial and functional compatibility for recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2009, 31(10):1775-1789
[25] Candes E J, Wakin M B. An introduction to compressive sampling[J]. IEEE Signal Processing Magzine, 2008, 25(2):21-30
[26] Wright J, Ma Y, Mairal J, et al. Sparse Representation for Computer Vision and Pattern Recognition [J]. Proceeding of the IEEE, 2010, 98(6):1031-1044
[27] Davenport M A, Duarte M F, Eldar Y C, et al. Introduction to compressed sensing[OL]. 2011. http://www.dfg-spp1324.de/download/preprints/preprint093.pdf
[28] 焦李成,杨淑媛,刘芳,等. 压缩感知回顾与展望[J]. 电子学报, 2010, 39(7):1651-1662
[29] Wright J, Yang A Y, Ganesh A, et al. Robust face recognition via sparse representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2):210-227
[30] Guha T, Ward R K. Learning Sparse Representations for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(8):1576 - 1588
[31] Castrodad A, Sapiro G, Castrodad A, et al. Sparse Modeling of Human Actions from Motion Imagery[J]. International Journal of Computer Vision, 2012, 100(1):1-15
[32] Bengio Y. Learning Deep Architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1):1-127
[33] Ji Shui-wang, Xu Wei, Yang Ming, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1):221-231
[34] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis[C]// Computer Vision and Pattern Recognition (CVPR). 2011:3361-3368
[35] Farabet C, Couprie C, Najman L, et al. Learning Hierarchical Features for Scene Labeling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. Preprints, 2013, 35(8):1915-1929
[36] Rodriguez M, Ahmed J, Shah M. Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Anchorage, Alaska, UCF Sports, 2008:1-8
[37] Liu Jin-gen, Luo Jie-bo, Shah M. Recognizing Realistic Actions from Videos "in the Wild"[J]. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). Miami, 2009
[38] Marzalek M, Laptev I, Schmid C. Actions in context [C]// CVPR. 2009:2929-2936
[39] Gilbert A, Illingworth J, Bowden R. Action Recognition Using Mined Hierarchical Compound Features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5):883-897
[40] Roshtkhari M J, Levine M D. A Multi-Scale Hierarchical Codebook Method for Human Action Recognition in Videos Using a Single Example[C]// Proc. of the conference on computer and robot vision (CRV). 2012:182-189
[41] Yao Bang-peng, Li Fei-fei. Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(9):1691-1703
[42] Desai C, Ramanan D. Detecting Actions, Poses, and Objects with Relational Phraselets[C]// European Conference on Computer Vision. 2012:158-172
[43] Shotton J, Fitzgibbon A W, Cook M, et al. Real-time human pose recognition in parts from single depth images [J]. Machine Learning for Computer Vision, 2013, 411:193-135
[44] Wang Jiang, Liu Zi-cheng, Wu Ying, et al. Mining actionlet ensemble for action recognition with depth cameras[R]. Microsoft Research, 2012
[45] Turaga P, Veeraraghavan A, Chellappa R. Unsupervised view and rate invariant clustering of video sequences[J]. Computer Vision and Image Understanding (CVIU), 2009, 113(3):353-371
[46] Rodriguez M D, Ahmed J, Shah M. Action MACH: a spatio-temporal maximum average correlation height filter for action recognition[C]// Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08). Anchorage, 2008:1-8
[47] Brand M. Coupled hidden Markov models for modeling interacting processes[J]. Daa, 1997
[48] Nguyen N T, Phung D Q, Venkatesh S, et al. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov models[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005, 2:955-960
[49] Park S, Aggarwal J K. A hierarchical Bayesian network for event recognition of human actions and interactions[J]. Multimedia Systems, 2004, 10(2):164-179
[50] Muncaster J, Ma Y. Activity recognition using dynamic Bayesian networks with automatic state selection[C]// IEEE Workshop on Motion and Video Computing (WMVC). 2007:30-37