基于视觉的人体动作识别综述

胡 琼[1)] 秦 磊[1)] 黄庆明[1),2)]

[1)](中国科学院计算技术研究所 中国科学院智能信息处理重点实验室 北京 100190)

[2)](中国科学院大学 北京 100190)

摘 要 基于视觉的人体动作识别是图像处理、计算机视觉、模式识别、机器学习、人工智能等多个学科的交叉研究课题,在视频监控、视频检索、人机交互、虚拟现实、医疗看护等领域具有深远的理论研究意义和很强的实用价值.文中从特征提取的方法、动作识别的方法、相关国际竞赛与常用数据库等方面详细阐述该领域目前的研究现状以及研究难点与可能的发展方向.

关键词 计算机视觉;模式识别;视觉特征提取;人体动作识别

中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2013.02512

A Survey on Visual Human Action Recognition

HU Qiong[1)] QIN Lei[1)] HUANG Qing-Ming[1),2)]

[1)](*Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*University of Chinese Academy of Sciences, Beijing* 100190)

Abstract Visual Human Action Recognition is a universal hot topic of image processing, computer vision, pattern recognition, machine learning and artificial intelligence with wide applications in video surveillance, video retrieval, human-computer interaction, virtual reality, health care, etc. In this paper, we analyze the state-of-the-arts and advances of this field from perspectives of feature extraction, action recognition methods as well as benchmark datasets and competitions. In addition, the problems, difficulties, challenges and valuable future directions of human action recognition are presented.

Keywords computer vision; pattern recognition; visual feature extraction; human action recognition

1 引 言

随着视频获取设备和宽带网络的快速普及和发展,视频已成为信息的主要载体,且视频数据的数量呈爆炸式增长,每时每刻都会有大量新的内容产生.面对海量涌现的视频数据,如何去自动获取,分析其中包含的内容就成为一个亟待解决的问题.

大多数视频记录的都是作为社会活动主体的人的活动,不论是从安全、监控、娱乐,还是个人存档的角度,对视频中人体动作识别进行研究具有重要的学术和应用价值[1].基于视觉的人体动作识别要解决的主要问题是通过计算机对传感器(摄像机)采集的原始图像或图像序列数据进行处理和分析,学习并理解其中人的动作和行为[2].一般在运动检测、特征提取的基础之上,通过分析获得人体运动模式,建

收稿日期:2011-09-23;最终修改稿收到日期:2013-10-22.本课题得到国家"九七三"重点基础研究发展规划项目课题(2009CB320906)、国家自然科学基金(61025011,61133003,61035001,61003165)、北京市自然科学基金(4111003)资助.**胡 琼**,女,1986 年生,博士研究生,主要研究方向为计算机视觉与模式识别.E-mail:qionghu2006@gmail.com.**秦 磊**(通信作者),男,1977 年生,博士,副研究员,中国计算机学会(CCF)会员,主要研究方向为计算机视觉与模式识别.E-mail:qinlei@ict.ac.cn.**黄庆明**,男,1965 年生,博士,教授,博士生导师,主要研究领域为多媒体分析、图像处理、计算机视觉、模式识别等.

立视频内容和动作类型描述之间的映射关系，以使计算机能够“看”视频或“理解”视频. 基于视觉的人体动作识别主要包含以下 3 个步骤：(1) 从图像帧中检测运动信息并提取底层特征；(2) 对行为模式或是动作进行建模；(3) 建立底层视觉特征与动作行为类别等高层语义信息之间的对应关系.

早在 20 世纪 70 年代末期，Marr[3] 提出计算机视觉理论，将整个视觉感知过程划分成底层、中层、高层 3 个层次，希望使计算机完全自动地以一种自底向上的方式从 2 维图像序列中恢复 3 维结构信息. 人体动作分析属于其中的高层视觉部分，近年来越来越多的大学、研究所、商业机构投入到该领域的研究中. 国际上的一些计算机视觉方向的权威期刊(如 TPAMI、IJCV、TIP) 和重要的学术会议(如 CVPR、ICCV) 也将人体动作分析与识别作为主题内容之一.

目前，基于视觉的人体动作识别的处理方法大体可分为 3 类：非参数方法、立方体分析方法以及参数化时间序列分析的方法[1]. 非参数方法通常从视频的每一帧中提取某些特征，然后用这些特征与预先存储的模板(template) 进行匹配；立方体分析方法不是基于帧的处理，而是将整段视频数据看作是一个 3 维的时空立方体进行分析；而参数化时间序列分析的方法对运动的动态过程给出一个特定的模型，并通过对训练样本数据的学习获得每一类动作特定的模型参数，其中比较常用的模型包括：隐马尔可夫模型(Hidden Markov Models，HMMS)、线性动态系统(Linear Dynamical Systems，LDSs) 等.

近年来，人体动作识别的研究任务也在逐步地发展，对计算机视觉领域提出了一些新的挑战. 从早期受限条件下(constrained settings) 简单动作的识别逐步转向了对真实自然场景下(videos “in the wild”) 复杂动作的识别；从对单人动作识别的研究自然地过渡到对交互动作甚至是大规模群体动作识别的研究.

本文将分别从动作识别特征、动作识别方法、相关国际竞赛与常用数据库等方面来阐述该领域目前的研究现状以及研究难点与可能的发展方向.

2　动作识别特征

从视频序列中提取出有效的运动特征是人体动作识别中重要的一环，直接影响到动作识别的准确度和鲁棒性，且同一特征对不同类别人体动作的描述能力并不相同. 因此，依据视频质量和应用场景的不同，往往要选用不同类型的特征，这与具体的应用场景以及研究者所关心的动作类别均有关系. 比如：在远景情况下，可以利用目标的运动轨迹进行轨迹分析；而近景情况下，则需利用从图像序列中提取的信息对目标的四肢与躯干进行 2 维或 3 维的建模.

常见的形状、轨迹、光流、局部时空兴趣点等特征可以分为以下 4 类(如表 1 所示)：基于轮廓和形状的静态特征、基于光流或运动信息的动态特征、基于时空立方体的时空特征以及描述性特征.

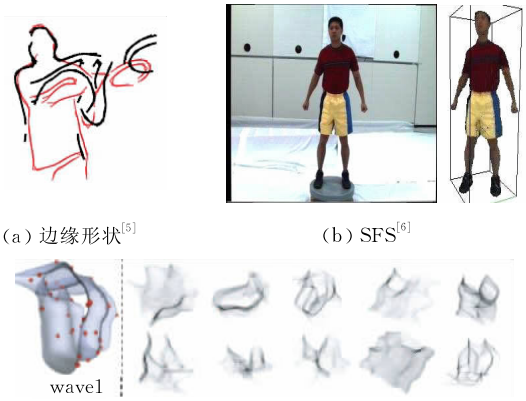
表 1　动作识别中常用特征分类表

类别	形式	代表文章
静态特征	大小、颜色、轮廓、形状、深度	[4～9]
动态特征	光流、速度、速率、方向、轨迹	[10～17]
时空特征	时空形状、时空兴趣点、时空上下文	[18～25]
描述性特征	属性、场景、物体、姿态	[26～28]

2.1　静态特征

静态特征主要描述的是人体目标的尺寸、颜色、边缘[5]、轮廓、形状[6] 和深度[9] 等. 静态特征可以较好地表示出人体目标的整体信息，可为动作识别提供有用线索. 例如：人体轮廓(Contour) 可以表示当前人体目标的基本形状.

Carlsson 等人[5] 通过从动作视频中提取到的关键帧和保存的动作原型之间做形状匹配来完成动作识别，其中，形状信息是以通过 Canny 边缘检测器检测到的边缘数据来表示的(图 1(a)). 这种方法能够容忍图像和样本之间一定程度的形变且能够准确识别不同人体姿态形成的极度相似的形状. Cheung 等人[6] 将传统的仅适用于静态对象的 SFS(Shape from Silhouette) 方法扩展到做刚体运动的对象，并



(a) 边缘形状[5]　(b) SFS[6]　(c) 旋转图像[4]

图 1　几种动作识别中用到的静态特征

进一步将其扩展到铰接体对象上(图 1(b)),用于获取人体各个部分的形状和运动信息,并通过解铰接部件之间简单的运动约束方程来估计出人体关节所在位置,从而实现动作识别的目的. Liu 等人[4]用菲德勒嵌入(Fiedler Embedding)的方法将旋转图像(Spin Image,图 1(c))和局部时空立方体嵌套到同一空间中. 近期,微软剑桥研究院的 Shotton 等人[7]提出从一幅深度图像中快速准确地预测人体关节 3 维时空位置的方法,具有姿态、形体以及衣着不变性等优势,相对于整体骨架最近邻匹配的方法更具通用性.

通过静态特征,可以获取目标的很多有用信息,但边缘与轮廓信息的获取并不容易,尤其在背景复杂,运动对象较多的场景中尤为困难. 因而有很多研究者尝试新的研究思路,不再进行目标分割、目标检测和目标跟踪,而是从视频中直接提取运动信息.

2.2 动态特征

运动信息一直被认为是计算机视觉中非常重要的线索,早在 1973 年 Johansson[11]通过经典的移动光斑实验(Moving Light Display)心理物理学实验,证实了仅通过观察连接在人体关节处的灯光信息,人能够准确识别出走、跑、上楼梯等动作,甚至能够从步态信息中识别出演员的性别和身份,预示了直接通过运动信息进行动作识别的可能性.

在人体动作分析中广泛使用的动态特征包括运动速度、运动方向和轨迹等. 轨迹刻画了目标在空间中的移动路径. 有了轨迹后,目标的运动速度和方向等特征可以很方便地计算出. 文献[10,12,29]中首先进行物体检测、跟踪和分类,然后利用得到的对象轨迹特征对动作模式进行建模.

上述轨迹特征在物体检测、跟踪或是识别效果不理想的情况下极易出错,尤其是在复杂的场景下. 鉴于此,一些研究者开始使用光流特征[15,30]. Efros 等人[15]为了提高低分辨率图像序列中运动信息提取的鲁棒性,构造了新的基于光流的描述子用于在线的动作识别. 其主要思想是把光流场先分为水平、竖直两种通道,然后每个通道再分为左、右两个通道,然后利用高斯滤波器对 4 个通道进行滤波,最后进行归一化得到运动描述子. Wang 等人[30]受到图像分类中的密集采样方法的启发,提出利用密集光流特征来描述视频内容(图 2),并利用运动边界直方图来描述密集光流特征.

然而,光流特征的准确获取本身是一个很棘手

图 2 密集光流特征[30]

的问题,即便是目前最好的光流计算算法,也存在着噪声,同时计算复杂度高,因此研究者开始尝试对特征点进行跟踪. Matikainen 等人[16]发现仅仅是用简单的 KLT 跟踪器,能够实现用比计算光流更少的复杂度比较鲁棒地在较长的时间区间内跟踪一系列特征点. Messing 等人[13]提出一种关键点运动历史的特征用于动作识别,该特征在包含复杂动作的高精度视频上发挥出较大的优势. Matikainen 等人在文献[16]中 trajecton 特征的基础上,提出序列编码图(Sequence Code Maps, SCMs)和相对位置概率(Relative Location Probabilities, RLPs),得到了一种对动作识别更为有效的特征两两之间关系的表述方式[17]. Raptis 等人在图像产生模型假设下,推导出的轨迹特征能捕获动作基元随时间的变化过程[14],同时该特征对一些随机扰动不敏感.

2.3 时空特征

这类特征将一段视频作为一个(x,y,t) 3 维空间中的时空体来分析和处理,即视频图像在时间轴上的级联(如图 3,左侧表示基于整幅图像的时空体,右侧为基于前景 Blob 的表示),然后提取 3 维数据模式,如时空形状、时空立方体、时空兴趣点等,用于动作的描述. 时空特征具有如下优点:(1)通过对立方体的分析,可以获得较长时间的动态特性;(2)联合考虑空间和时间的连续性,特征匹配的复杂度大大降低;(3)对遮挡等事件的处理更加鲁棒有效.

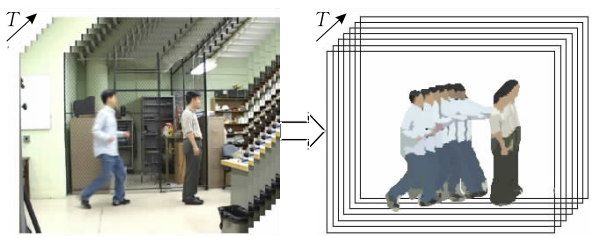


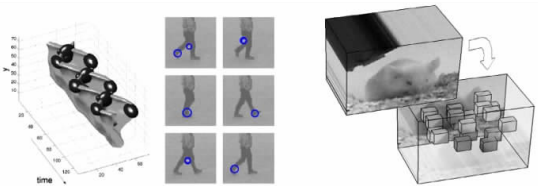
图 3 视频的时空体表示[2]

2.3.1　时空模板

Davis 等人[18]提出了一个时空域模板的概念作为动作模型，他们认为单个静态帧难以给出足够的辨别信息，但是如果把静态帧连接起来形成一个运动序列，可以分辨出其中表现的动作类别，他们将背景减除后的动态序列聚集成一幅静态图像，运动能量图(Motion Energy Image)或运动历史图(Motion History Image)，以此来表征某类动作，并从中提取不变量来进行动作识别. Zhong 等人[19]借鉴静态图像识别中基于滤波响应的纹理分割思想，对视频数据从空间和时间上计算滤波响应，将从滤波响应中提取的特征聚为动作原型，建立动作原型和视频片段之间的共生矩阵，以此来进行异常动作的检测. Gorelick 等人[20]提出了用背景差分块累叠形成一个(x,y,t)空间内的二值时空体，从这个时空体中，通过解 Possion 方程可以提取其 3 维形状描述子作为动作模板来进行识别，因为这种方法需要准确的前景和背景分割，因而仅局限于固定摄像机的应用中. 与 Gorelick 的特征不同，Yuan 等人[21]将人体动作看作是视频中重复出现的时空模式(由时空不变特征构成)，他们的方法从时空体内时空特征的分布模式出发，不依赖背景减除，在处理上更加灵活，但如何高效地在 3 维空间中进行模式的自动搜索和匹配是该方法的局限点.

2.3.2　局部时空兴趣点

基于时空模板的方法往往依赖于预处理(人体轮廓或是剪影的提取)或是模式自动搜索和匹配的精度与速度，因而其有效性受到一定限制. 因此有的研究者在整段视频中寻找局部时空特征来表征动作[22,31]. Laptev[31]将 2 维图像上的 Harris 角点扩展到 3 维空间上. 在 2 维图像上，Harris 角点是在两个方向上都有很大变化的点. 通过对 Harris 角点增加时间约束，检测在时空上都变化剧烈的点，得到 3 维 Harris 角点，达到在时空维度中检测局部结构的目的，如图 4(a)所示. 3 维 Harris 角点检测器要求特征点在 3 个维度上都有剧烈的变化，导致检测到的特征点比较稀疏. Dollar 等人[22]采用可分离的线性滤波器来进行时空域兴趣点的检测，在空域采用 2 维的高斯滤波器，而在时域上采用两个正交的1 维 Gabor 滤波器来检测运动特征，如图 4(b)所示. 这种方法检测到的特征点数量一般比 3 维 Harris 角点的数量要多很多. 在上述研究的基础上，Scovanner 进一步扩展，利用子直方图(sub-histogram)来对局部时空信息进行编码，构造出 3 维的 SIFT 描述子[32]. 也有研究者从学习的角度来提取局部特征，Le 等人[23]利用独立子空间分析从视频数据中无监督地学习局部时空特征.



(a) 3维Harris角点[31]　　(b) 基于Gabor滤波器的时空特征[22]

图 4　局部时空特征

2.3.3　时空上下文特征

基于局部时空兴趣点的方法在动作识别中取得了广泛的应用，但是这类方法没有很好利用局部特征之间的空间几何关系. 所以有研究者提出时空上下文特征，对局部特征之间的空间几何关系进行建模，期望可以进一步提高特征的描述能力[24-25,33].

Wang 等人[24]提出利用多尺度时空上下文特征进行动作识别. 他们首先利用 3 维 Harris 角点检测器来得到局部特征，然后对每个局部特征的空间和时间邻域进行多尺度网格划分，并统计网格中的局部特征分布，最后多个网格的分布相连得到最后的特征. Wu 等人[25]更进一步提出联合时空上下文和表观特征分布来进行动作识别. 他们利用多个高斯混合模型来描述成对特征点之间相对坐标的时空上下文分布关系，且对表观特征也用高斯混合模型进行描述，最后他们用多核学习的方法(Multiple Kernel Learning)把两类特征融合到一起.

2.4　描述性特征

基于静态特征、动态特征和时空特征等底层视觉特征的动作识别方法取得了不错的效果，但是这些方法通常仅通过底层特征直接得到动作的类别，动作视频中丰富的语义信息并没有得到充分的利用. 借鉴物体识别中基于属性方法的成功应用[34-35]，研究者提出利用中层的描述性特征来进行动作识别[26-28,36]. Liu 等人[26]定义了一个动作属性空间，其中的每一维表示一个语义属性，这样每个动作就可以表示成属性空间中的一个点，如图 5 所示. 为了克服人工定义的属性集带来的主观性和不完整性，Liu 等人[26]还提出利用互信息(Mutual Information)方法来从数据中自动学习具有判别力的属性，最后人工定义属性集和数据中学习得到的属性集被综合起来进行动作学习. Yao 等人[27]所利用的描述性特征包括原子动作、物体和姿态. 他们对这些描述性特征

之间的共生统计进行建模，并把共生关系称之为“动作基”.然后一个动作就可以表示成这些“动作基”的子集的加权组合.

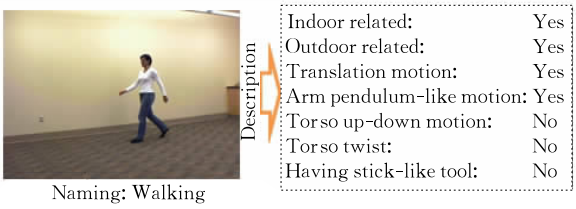


图 5 人体动作的描述性特征表示[26]

综上所述，特征提取方法主要分为 3 类：(1) 基于底层跟踪或姿态估计的方法，提取的特征主要是静态特征和基于运动信息的动态特征，因此提取特征的有效性依赖于目标跟踪和人体姿态估计的准确性.在真实的场景中，由于背景经常比较杂乱，运动目标也比较多，进行准确的目标跟踪和人体姿态估计具有极大的挑战性.导致这类特征的鲁棒性不是很好；(2) 基于图像处理技术直接从图像中提取特征的方法，提取的一般是基于光流的动态特征和时空特征，这类方法提取的一般是对图像或是时空立方体局部运动的描述，因而计算量比较大，易受噪声的干扰，且缺乏对动作行为模式整体性的考虑和全局性的分析；(3) 基于学习方法得到的属性描述，提取的一般是物体、姿态和场景等中层语义特征，这类特征可能对特定场景下的动作识别极为有效，但是因为涉及到人为定义的“动作属性空间”，在真实自然场景下，可能存在着属性空间不完备或是不准确而导致动作识别性能下降的风险.对上述几种动作识别中所用特征的比较，见表 2.

表 2 人体动作识别特征比较

特征	基本技术	鲁棒性	计算复杂度	适合动作类别
静态特征	姿态估计方法	低	低	简单动作
基于运动信息的动态特征	底层跟踪方法	低	中	简单动作
基于光流信息的动态特征	光流计算方法	中	高	简单动作
时空特征	图像处理方法	高	高	连续动作交互动作
描述性特征	机器学习方法	高	高	连续动作交互动作

3 动作识别方法

人体动作识别可以看成时变数据的分类问题，即将测试序列归入特定已知的动作类别中，同时需要处理同类动作运动模式在时间和空间上的类内变化.对于连续动作的识别，通常引入滑动窗口策略对所有可能的子序列进行分类.常见的动作识别方法分为 3 类：基于模板的方法、概率统计的方法、基于语法的方法.

3.1 基于模板的方法

在基于模板的动作识别方法中，先利用一个或一组模板来表示待识别目标的运动，然后将待识别目标的模板与预先存储的已知模板进行比较，根据相似度度量判别动作类别.依据匹配的对象是一个还是一组静态模式可以将该类方法进一步分为模板匹配(Template Matching)和动态时间规整(Dynamic Time Warping).

3.1.1 模板匹配

模板匹配是模式分类中常用的方法之一，是在视频上计算待识别目标的模板和候选视频区域之间的距离.如果距离小于阈值，则认为待识别目标被检测到.该方法可以对单帧图像或是一个图像序列进行识别.在模板匹配算法中对表示模板的特征没有特别要求，常见的颜色、形状和纹理等特征都可以利用.在计算模板与候选区域的距离时，可以利用欧式距离、马氏距离或是经过距离度量学习后的加权距离.

Davis 等人[18]提出利用运动能量图和运动历史图来表示一个图像序列，并利用马氏距离来计算模板之间的距离.很多采用 K-近邻/最近邻分类器的方法实际也是模板匹配的方法[4,15,20,22,37-39]，这种分类器通过计算观测序列的图像描述符与训练序列的图像描述符之间的距离，分类结果为其 K 个近邻训练序列中最常见的动作类型.

基于模板匹配的方法具有计算复杂度低的优点.这类算法的一个难点是如何选择时间间隔，当选取的时间间隔比较小时，存储的模板的数目比较多，且样本和匹配的模板之间的差异比较小，识别效果会相对比较好；反之，当选取的时间间隔比较大时，预先存储的模板数目会较少，且样本和匹配的模板之间的变化较大，识别效果会较差.

3.1.2 动态时间规整

由于同一个动作在不同视频中持续的时间可能并不相同，因此有必要在时间上对动作样本进行规整，典型的方法是动态时间规整.动态时间规整算法利用动态规划原理进行时间规整，可以降低搜寻比对所用的时间.Ji 等人[40]利用动态时间规整方法来计算查询动作和候选动作之间的相似性程度.Jiang 等人[37]更进一步，利用快速动态时间规整方法来自

动确认匹配的片段并计算对齐后两个序列的距离.动态时间规整方法具有算法时间复杂度较低、鲁棒性较高的特点.

3.2　基于概率统计的方法

概率统计模型把动作表示成一个连续的状态序列,每个状态都有自己的表观和动态特征.状态之间的切换规律可以用时间转移函数表示.常用的概率统计模型可以分成产生式模型和判定式模型两大类.

3.2.1　产生式模型

产生式模型估计联合概率分布,计算后验概率,从统计学的角度表示特征和状态之间的关联情况.模型具有灵活性强的特点,可处理数据不完整的情况.常见的产生式模型主要有:高斯混合模型、隐马尔可夫模型(Hidden Markov Model,HMM)、概率潜在语义分析(Probabilistic Latent Semantic Analysis,PLSA)和潜在狄利克雷分配(Latent Dirichlet Allocation,LDA)等[8,41-43].

第1个利用HMM进行人体动作识别的是Yamato等人[44].他们利用一个离散的HMM来表示网球动作.Shi等人[45]针对连续的动作序列,提出一种判定式的马尔可夫模型,利用这种模型可以同时进行动作分割和识别.为了高效地寻找模型的解,他们设计了一个类Viterbi的动态规划算法,可以进行准实时的求解.HMM的局限性在于它只适用于表示具有马尔可夫性的序列动作.基于DBN的方法比基于HMM的方法具有更好的可扩展性.Natarajan等人[8]把一个复杂的动作分解成一个原子动作序列,然后利用动态贝叶斯网络来表示这个动作序列,充分描述原子动作之间的转换关系.Weinland等人[46]把视点作为一个隐含变量加入到HMM中,这样用一个模型就可以建模从任意视点观看的动作.

文献[43,47]采用pLSA的方法自动学习图像序列中隐含的动作类别.直观上,该模型将每个视频序列表示成K类动作特征向量的凸组合,例如:特定视频中视频单词的分布可通过计算基于动作类别向量的一个凸组合得到.视频可以理解成由特定因子以及混合系数构成的一个组合.在特征向量和组合系数都被归一化为概率分布的情况下,该问题变成了一个矩阵的分解问题.然后,通过决定所有视频通用的动作类别向量以及针对某一段视频特定的混合系数来求解该模型.

3.2.2　判别式模型

产生式模型的局限性在于依靠简化的统计假设来计算特征和状态之间的联合概率,不能直接计算条件概率.而判定式模型在给定了特征之后,可以直接计算条件概率分布,估计类别之间的分类面.可以进行多类的识别,通常情况下识别性能比产生式模型稍好.常见的判别式模型包括线性判别分析、支持向量机(SVM)、提升方法(Boosting)、条件随机场(CRF)等.

支持向量机方法是建立在统计学习理论的VC维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力,在模式识别等相关领域获得了广泛的应用,Schuldt将其应用到了动作识别上[48],之后多个研究者在动作识别的过程中都用到了支持向量机[14,16-17,49-50].

Lafferty等人[51]提出条件随机场这种概率图模型,并被迅速而广泛地应用到自然语言理解领域、图像分割、场景分析和动作识别等领域.Natarajan等人[52]将CRF用到了人体动作识别中,首先通过已知动作的Mocap数据从多个角度渲染合成人体姿态并用条件随机场进行表示,其中观测概率通过形状相似度进行计算而转移概率通过光流来计算.在传统CRF的基础上,通过引入一些时间和空间上的约束,得到了增强形状、光流和持续时间等信息的CRF模型(SFD-CRF).Konstantinos等人[53]扩展了隐条件随机场,提出了无限状态隐条件随机场(infinite hidden conditional random fields).这个模型利用了层次Dirichlet过程,不需要预先指定隐含状态的数量,可以自动学习出分类任务的最优隐含状态数量.

近期,Shotton等人[7]将机器学习中的随机决策森林(random decision forests)应用到了动作识别上,随机森林是一个包含多个决策树的分类器,其输出类别由每棵树输出类别的总数而定.它能够处理大量的输入并具有较高的分类准确度.Chris等人[54]利用GentleBoost来进行特征选择,提高了动作识别的准确率.

3.3　基于语法的方法

近年来,语法分析(Syntactic Analysis)的技术逐渐被用于人体动作识别中.基于文法的方法将人体动作描述为一连串的符号,每一个符号代表了动作中的一个原子级的分解.这类方法需要首先识别这些原子动作,然后将人体动作表示为通过一系列生成规则形成的原子动作流.识别的过程主要用到了自然语言处理里面的文法分析技术(Parsing

Techniques).

Nevatia 等人[55]定义了一种事件描述语言(Event Recognition Language,ERL)来表示物理世界中发生的由一系列简单事件/动作构成的时空事件. Ryoo 等人[56]提出利用随机上下文无关语法来进行行为识别. 通过随机上下文无关语法来对行为内部的不确定性和行为结果的多变性进行很好的建模,从而有效地识别这些行为. Si 等人[57]提出利用与或图来表示动作事件. 他们把信息投影和最小描述长度准则统一在一个概率框架下,来学习动作的与或图表示. 这种方法不需要人工标注视频中的事件及事件的开始时间.

目前常用的动作识别方法包括:基于模板的方法、基于概率统计的方法以及基于语法的方法. 模板匹配的方法直观、简单,但是缺乏鲁棒性,一般用于静态姿势或是简单动作的识别中;概率统计的方法是目前主流的方法,应用广泛,存在的问题是需要大量的训练数据来学习模型参数,且对于产生式模型,为了模型求解的便利,一般假设样本是独立同分布的,并假设不同的观测值之间是相互独立的,这样强的独立性假设,往往和数据的实际产生过程不符;基于语法的方法有利于对复杂结构的理解和对先验知识的有效利用,一般可与前两种方法结合. 对上述几种识别方法的比较,见表 3.

表 3 人体动作识别方法比较

识别方法	时间尺度鲁棒性	空间尺度鲁棒性	先验知识与模型假设	计算复杂度	适合动作类别
模板匹配	低	低	低	低	简单动作
动态时间规整	高	低	低	中	简单动作 连续动作
产生式模型	高	高	中	高	简单动作 连续动作
判别式模型	高	高	中低	高	连续动作 交互动作
基于语法	高	依赖于底层描述	高	高	连续动作 交互动作

4 相关国际竞赛与常用数据库

TREC Video Retrieval Evaluation①, TRECVID 是视频检索领域的权威国际评测,由美国国家标准技术局组织实施,从 2001 年开始,每年举行一次. 自 2008 年起,加入了监控视频事件检测(Surveillance Event Detection,SED)的任务,主要是处理采集自英国伦敦 Gatwick 国际机场的约 144 个小时监控视频数据(图 6). 由于机场监控背景复杂,人员密度和流动性大,人体之间遮挡严重,因此其事件检测任务具有极高的难度. 近年来监控视频领域顶级国际期刊和会议上提出的很多人体检测、跟踪和事件分析算法在此数据库上的表现都不尽人意.

图 6 TRECVID SED 数据

International Workshop on Performance Evaluation of Tracking and Surveillance②: PETS 自 2000 年在法国格勒诺布尔市召开第 1 届研讨会以来,一直延续至今. 在 PETS 比赛中,所有的参加者均要求将他们的算法应用于相同的数据库(图 7).

图 7 PETS 数据

Semantic Description of Human Activities③: SDHA 与在土耳其伊斯坦布尔举行的第 20 届模式识别国际会议(ICPR 2010)同时进行,是一项研究人体行为语义描述的竞赛,主要致力于识别真实场景中的人体行为. 为了鼓励开发适合真实场景(如监控系统)的行为识别方法,该竞赛共设置了 3 种不同的挑战任务,针对每一个任务提供了一套完整的训练和测试视频数据(图 8). 这个竞赛的目的是对一段视频中正在进行的所有行为进行语义标注.

图 8 SDHA 比赛任务

① http://trecvid.nist.gov/
② http://pets2010.net/
③ http://cvrc.ece.utexas.edu/SDHA2010/

相关的国际竞赛能够有效地促进本研究领域的发展，同时也需要一些公共数据库来对算法进行评估.因此，包含人体动作或行为的视频数据库的构建对动作识别算法的研究起到了至关重要的作用，表 4 列出了一些动作识别领域常用的公共数据库，大致可以分成以下 4 类：

第 1 类是 KTH[48] 和 Weizmann[20] 等通用动作识别数据库，它们包含表演者在受限场景下执行的一系列简单动作，如："行走"或是"招手"等.

第 2 类是一些针对真实环境，比如机场等，面向特定应用的数据库，表 2 中 Keck gesture[37]、Daily Living[13]、PETS 以及 TRECVID 等均属于此类.PETS 数据中定义的"偷行李"、"打斗"等面向监控应用的动作是该类的典型范例.

第 3 类是一些从新近产生的大量的电影或是个人视频中收集的数据，比如，HOHA[50]、Youtube[58]、WebVideo[59] 数据库等，它们共同的特点是摄像机、场景不固定且同类动作的类内散度比较大，因而极具挑战性.

第 4 类是包含多类别动作的数据集，比如 UCF50[60] 或是 UCF101①.据我们所知，UCF101 是目前最大的动作数据集，包含 101 个动作类别，约 13 000 个视频片段.

表 4　人体动作识别领域常用数据库

动作数据库	网址	引用
KTH	http://www.nada.kth.se/cvap/actions/	[48]
Weizmann	http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html	[20]
CMU	http://www.yanke.org/	[61]
HOHA	http://www.irisa.fr/vista/actions/	[50]
UCF Sports	http://server.cs.ucf.edu/~vision/data.html	[62]
Keck gesture	http://www.umiacs.umd.edu/~zhuolin/Keckgesturedataset.html	[37]
MSR	http://users.eecs.northwestern.edu/~jyu410/index_files/actiondetection.html	[21]
YouTube	http://www.cs.ucf.edu/~liujg/YouTube_Action_dataset.html	[58]
Daily Living	http://www.cs.rochester.edu/~rmessing/uradl	[13]
UT-Interaction	http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html	[39]
TRECVID	http://trecvid.nist.gov/trecvid.data.html	[63]
PETS	http://pets2010.net/	[64]
VIRAT	http://www.viratdata.org/	[65]
WebVideo	http://img.cs.uec.ac.jp/dohang/vrank_result/flv100/video.cgi	[59]
UCF50	http://crcv.ucf.edu/data/UCF50.php	[60]
UCF101	http://crcv.ucf.edu/data/UCF101.php	

5　目前的难点与未来可能的研究方向

人体动作识别主要涉及底层特征的提取和表示，及模式的分类、识别两个阶段，它们构成了人体动作识别的重要步骤.分析这两个子问题，可以发现人体动作识别目前存在的难点和挑战主要有：

(1) 动作的分类和定义

目前，研究者对如何定义人体动作，如何划分动作的层次还没有给出一个明确的准则.状态、动作、行为、事件等不同层次的视频描述之间界限模糊.一个人体动作具体包含哪些姿势，可以分几个阶段执行以及起始时间如何确定都有待研究者进一步明确.

(2) 动作的类内变化及类间变化

对于同一类动作，由于人体尺度的变化，如图 9(a)所示，即使是同一个人做同一个动作，表现在视频图像中的大小也不一样.而手持设备由于摄像机的抖动或是设备的运动，如图 9(b)所示，也会造成一段动作视频不同时刻具有很大的表观差异，难以定位和跟踪人体以及对人体动作进行识别.此外，对象在做某一个动作的过程中，可能会与摄像机具有相同或是相似的运动速度(如滑冰，见图 9(c)).上述诸多因素造成同一动作具有很大的类内散度，因此，一种鲁棒的人体动作识别算法要求能够适应同一种动作类型的变化.而随着动作类别的增多，不同动作姿态之间重叠的可能性也随着表示空间的细分而增加，比如，对于慢跑(jogging)和跑步(running)这两类动作，表现在视频中很多时刻的姿势就是一致的，这使得不同动作类型间散度较小，这也是动作识别算法所面临的一个难题[66].

(3) 动作采集环境以及采集设备的影响

在复杂的场景中，一方面动作的执行者经常被障碍物或者其他对象所遮挡(图 9(d))，另一方面，背景上其他人或对象的运动也会对动作的识别产生干扰(图 9(e)).环境的光照变化容易导致固定的人体动作表观模型失效(图 9(f)).同属于人遛狗这一类动作，由于不同的视角，不同的季节，不同的场地(如草坪、马路等)，视觉表观很不同(图 9(g)).这些环境因素的影响都会导致计算机得到不同的视觉观测，动作识别算法的设计需要能够容忍这些变化的存在.与此同时，视频图像的质量也会对算法产生影

① http://crcv.ucf.edu/data/UCF101.php

图 9 人体动作识别的难点与挑战

响,如一些在高清或标清视频上比较有效的算法在低图像质量的视频上识别效果往往大打折扣.

(4)连续动作的分割和长时视频中动作的识别

人体行为往往由一连串动作构成,动作之间没有明显的边界指示.现有的动作识别方法大多数是对已经从时间域分割好的视频片断来进行分类,而不能识别长时视频中发生的多个动作,也不能定位事件发生的开始帧和结束帧.此外,动作执行的速度会影响动作持续时间的长短,这也对动作识别的鲁棒性产生干扰.

(5)模型的泛化能力

目前在动作识别领域使用比较广泛的基于概率统计的模型,研究者可以利用标注的训练数据、有效的学习模型的参数,但是模型的结构需要人工进行设计.当需要分类的动作类别比较多的时候,如何得到有效且具有较强泛化能力的动作模型值得进一步深入的研究与探讨[67].

(6)大量训练数据的标注

人体动作识别的研究经历了从受限场景中简单动作的识别到网络(如 Youtube Dataset[58])或是电影(如 Holleywood Action Dataset[50])中的动作识别,以及自然生活场景中的人体动作识别(如 Daily Living Dataset[13]). 目前现有数据库中训练集及测试集的数据量有限,如何对大量的视频进行标注是一个问题,如 TRECVid 数据库.利用人工标注的方式是非常费时费力而不可取的,这就需要视频数据的自动标注工具.

从对本领域研究现状的分析可知,人体动作识别可能的发展方向大致可以划分为两类:特征的提取与构造;识别问题本身.

特征的提取对人体动作识别至关重要,目前主流的人体动作识别特征种类繁多,包括静态特征、动态特征、时空特征以及描述性特征,且其中每一类特征都有多种不同的表述形式.尽管如此,在特征提取上仍然有很多问题亟待解决:

(1)底层特征提取

人体动作识别的效果直接取决于从视频或是图像序列中提取到的底层特征的有效性,"好的"特征包含的信息量大,对噪声的干扰比较鲁棒,对不同动作模式具有很强的区分能力.尽管单一的视觉线索,比如颜色、纹理、形状、运动或是深度信息等在特定的场景下能够取得不错的识别效果,但是在真实自然场景识别性能却急剧下降.因此,相对复杂场景特征的鲁棒提取技术值得进一步研究.

(2)特征的表达能力

常见的"视觉词袋"(Bag-of-Words,BoWs)[43]方法最近得到大家的关注,并且在动作识别领域取得了不错的效果.但是人体动作是非常复杂的,肢体在动作过程中在空间和时间上有很多关联和约束关系.而视觉词袋模式是一个非常简化的模型,它只保留了局部特征的出现次数,而完全抛弃了局部特征之间的时空依赖关系.如果把空间和时间约束以恰当的方式引入到视觉词袋模型中,能提高视觉词袋的表达能力.已经有研究者在这方面进行了尝试,但还有待进一步的改进和研究.

(3)特征的比较与评测

在诸多特征"百花齐放"的状态,一种特征可能在某一类人体动作测试数据库上效果显著,但是为了更好地对不同特征进行比较与评测,一方面需要构造比较合理的人体动作公用测试视频库,另一方面,也需要在合理的比较与评测方法上进行研究.

(4)不同特征之间的融合

不同的特征从不同的侧面对人体动作进行描述,特征之间具有很强的互补性.因此融合不同特征可能是解决复杂人体动作识别问题的基本思路之一[68].在特征的融合策略上,多核学习、多任务学习等策略有待进一步探讨.

上面是从特征的角度来看下一步可能的研究方向,而人体动作识别本身就是一个没有很好定义的

问题.所以从面向问题本身的角度,也有很多值得进一步研究和努力的方向：

(1)真实自然场景下的动作识别

早期的人体动作识别仅针对受限环境下(如简单、视角固定、背景静止等)由专门的演员表演的有限类别的动作,常见的有走、跑、拍手、招手等.越来越多的研究者开始关注对人们日常生活中的动作进行识别的问题,如 Daily Living 数据库[13]、机场监控数据及其医院疗养院病人监控等,但由于图像噪声、摄像机的运动、复杂背景及光照变化等影响使得该类问题在现阶段仍然很难解决,有待进一步的研究[69].

(2)群体动作识别

对包含多人交互或是群体运动的动作识别问题,与单人动作不同,这类问题动作模式的类内散度更大,一般需要考虑到不同特征的融合以及动作的分层建模等.融合多种底层的视觉特征可以提高特征的鉴别能力和区分能力,通过分层建模的思想将问题分而治之(divide and conquer),一般包括特征层的建模和原子动作层的建模等[70],提供对一个复杂动作不同层次、不同粒度的描述.各种图结构(Graph Model)及潜藏语义 SVM(Latent-SVM)模型等开始应用于此问题,但是多种视觉特征的融合、动作模型的构建以及模型的优化求解等都不甚完善,还有着广阔的发展空间.

6　结束语

人体动作识别是计算机视觉中的一个重要研究领域,具有广阔的应用前景.本文根据人体动作识别的一般步骤,分析了特征提取、动作识别、评测数据库等 3 个方面的研究成果.在综述最新研究进展的基础上,讨论了目前人体动作识别领域存在的难点与挑战,论述了未来人体动作识别领域可能的研究方向.

参　考　文　献

[1] Turaga P, Chellappa R, Subrahmanian V S, Udrea O. Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(11): 1473-1488

[2] Aggarwal J K, Ryoo M S. Human activity analysis: A review. ACM Computing Surveys, 2011, 43(3): 1-43

[3] Marr D. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. New York, NY, USA: Henry Holt and Co., Inc., 1982

[4] Liu Jin-Gen, Ali S, Shah M. Recognizing human actions using multiple features//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

[5] Carlsson C, Carlsson S, Sullivan S. Action recognition by shape matching to key frames//Proceedings of the Workshop on Models Versus Exemplars in Computer Vision. Colorado, USA, 2001: 1-8

[6] Cheung K M, Baker S, Kanade T. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Madison, WI, USA, 2003: 77-84

[7] Shotton J, Fitzgibbon A, Sharp T, et al. Real-time human pose recognition in parts from a single depth image//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 1297-1304

[8] Natarajan P, Singh V K, Nevatia R. Learning 3D action models from a few 2D videos for view invariant action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 2006-2013

[9] Cheng Zhong-Wei, Qin Lei, Ye Yi-Tuo, et al. Human daily action analysis with multi-view and color-depth data//Proceedings of the 2nd Workshop on Consumer Depth Cameras for Computer Vision Conge. Firenze, Italy, 2012: 52-61

[10] Wang Xiao-Gang, Tieu K, Grimson E. Learning semantic scene models by trajectory analysis//Proceedings of the European Conference on Computer Vision. Graz, Austria, 2006: 110-123

[11] Johansson G. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 1973, 14(2): 201-211

[12] Stauffer C, Grimson W E L. Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22: 747-757

[13] Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 104-111

[14] Raptis M, Soatto S. Tracklet descriptors for action modeling and video analysis//Proceedings of the European Conference on Computer Vision. Heraklion, Crete, Greece, 2010: 577-590

[15] Efros A A, Berg A C, Mori G, Malik J. Recognizing action at a distance//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France, 2003: 726-733

[16] Matikainen P, Hebert M, Sukthankar R. Trajectons: Action recognition through the motion analysis of tracked features// Proceedings of the Workshop on Video-Oriented Object and Event Classification. Kyoto, Japan, 2009: 514-521

[17] Matikainen P, Hebert M, Sukthankar R. Representing pairwise spatial and temporal relations for action recognition// Proceedings of the European Conference on Computer Vision. Heraklion, Crete, Greece, 2010: 508-521

[18] Davis J W, Bobick A F. The representation and recognition of action using temporal templates//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Juan, Puerto Rico, 1997: 928-934

[19] Zhong H, Shi J, Visontai M. Detecting unusual activity in video //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC, USA, 2004: 819-826

[20] Gorelick L, Blank M, Shechtman E, et al. Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2247-2253

[21] Yuan J S, Liu Z C, Wu Y. Discriminative subvolume search for efficient action detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009: 2442-2449

[22] Dollar P, Rabaud V, Cottrell G, Belongie S. Behavior recognition via sparse spatio-temporal features//Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing, China, 2005: 65-72

[23] Le Q V, Zou W Y, Yeung S Y, Ng A Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3361-3368

[24] Wang Jiang, Chen Zhuo-Yuan, Wu Ying. Action recognition with multiscale spatio-temporal contexts//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3185-3192

[25] Wu Xin-Xiao, Xu Dong, Duan Li-Xin, Luo Jie-Bo. Action recognition using context and appearance distribution features//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 489-496

[26] Liu Jingen, Benjamin K, Silvio S. Recognizing human actions by attributes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3337-3344

[27] Yao Bang-Peng, Jiang Xiao-Ye, Khosla A, et al. Human action recognition by learning bases of action attributes and parts//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 1331-1338

[28] Maji S, Bourdev L, Malik J. Action recognition from a distributed representation of pose and appearance//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3177-3184

[29] Cheng Zhong-Wei, Qin Lei, Huang Qing-Ming, et al. Group activity recognition by gaussian processes estimation// Proceedings of the International Conference on Pattern Recognition. Istanbul, Turkey, 2010: 3228-3231

[30] Wang Heng, Kläser A, Schmid C, Liu Cheng-Lin. Action recognition by dense trajectories//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3169-3176

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

[32] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition//Proceedings of the ACM Multimedia. Augsburg, Bavaria, Germany, 2007: 357-360

[33] Hu Qiong, Qin Lei, Huang Qing-Ming, et al. Action recognition using spatial-temporal context//Proceedings of the International Conference on Pattern Recognition. Istanbul, Turkey, 2010: 1521-1524

[34] Berg T, Berg A, Shih J. Automatic attribute discovery and characterization from noisy Web data//Proceedings of the European Conference on Computer Vision. Heraklion, Crete, Greece, 2010: 663-676

[35] Farhadi A, Endres I, Hoiem D, Forsyth D. Describing objects by their attributes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009: 1778-1785

[36] Marszałek M, Laptev I, Schmid C. Actions in context// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009: 2929-2936

[37] Jiang Zhuo-Lin, Lin Zhe, Davis L S. Recognizing human actions by learning and matching shape-motion prototype trees. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3): 533-547

[38] Jiang Zhuo-Lin, Lin Zhe, Davis L S. A tree-based approach to integrated action localization, recognition and segmentation// Proceedings of the 3rd Workshop on Human Motion. Heraklion, Crete, Greece, 2010: 114-127

[39] Ryoo M S, Aggarwal J K. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities//Proceedings of the IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1593-1600

[40] Ji Rong-Rong, Yao Hong-Xun, Sun Xiao-Shuai. Actor-independent action search using spatiotemporal vocabulary with appearance hashing. Pattern Recognition, 2011, 44(3): 624-638

[41] Wang Xiao-Gang, Ma Xiao-Xu, Grimson E. Unsupervised activity perception by hierarchical Bayesian models//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 2007: 1-8

[42] Liu Jin-Gen, Shah M. Learning human actions via information maximization//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

[43] Niebles J C, Wang H, Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision, 2008, 79(3): 299-318

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

[45] Shi Qinfeng, Cheng Li, Wang Li, Smola A. Human action segmentation and recognition using discriminative semi-Markov models. International Journal of Computer Vision, 2011, 93(1): 22-32

[46] Weinland D, Boyer E, Ronfard R. Action recognition from arbitrary views using 3D exemplars//Proceedings of the International Conference on Computer Vision, Rio de Janeiro, 2007:1:7

[47] Savarese S, DelPozo A, Niebles J C, Li Fei-Fei. Spatial-temporal correlations for unsupervised action classification// Proceedings of the IEEE Workshop on Motion and Video Computing. Copper Mountain, Colorado, USA, 2008: 1-8

[48] Schuldt C, Laptev I, Caputo B. Recognizing human actions: A local SVM approach//Proceedings of the International Conference on Pattern Recognition. Cambridge, England, UK. 2004: 32-36

[49] Natarajan P, Nevatia R. Online, real-time tracking and recognition of human actions//Proceedings of the IEEE Motion and Video Computing, Copper Mountain. Colorado, USA, 2008: 1-8

[50] Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

[51] Lafferty John D, Andrew M C, Pereira Fernando C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the Eighteenth International Conference on Machine Learning. Williamstown, MA, USA, 2001: 1-8

[52] Natarajan P, Nevatia R. View and scale invariant action recognition using multiview shape-flow models//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

[53] Konstantinos B, Stefanos Z, Louis M, Maja P. Infinite hidden conditional random fields for human behavior analysis. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(1): 170-177

[54] Chris E, Syed Z M, Marshall F T, et al. Exploring the trade-off between accuracy and observational latency in action recognition. International Journal of Computer Vision, 2013, 101(3): 420-436

[55] Nevatia R, Zhao T, Hongeng S. Hierarchical language-based representation of events in video streams//Proceedings of the IEEE Workshop on Event Mining. Madison, Wisconsin, USA, 2003: 39

[56] Ryoo M S, Aggarwal J K. Stochastic representation and recognition of high-level group activities//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 183-200

[57] Si Zhangzhang, Pei Mingtao, Yao Benjamin, Zhu Song-Chun. Unsupervised learning of event AND-OR grammar and semantics from video//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 41-48

[58] Liu Jin-Gen, Luo Jie-Bo, Shah M. Recognizing realistic actions from videos "in the Wild"//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA, 2009: 1996-2003

[59] Nga D H, Yanai K. Automatic construction of an action video shot database using Web videos//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 527-534

[60] Reddy K K, Shah M. Recognizing 50 human action categories of Web videos. Machine Vision and Applications, 2012, 5(9): 1-11

[61] Yan Ke, Sukthankar R, Hebert M. Event detection in crowded videos//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2007: 1-8

[62] Rodriguez M D, Ahmed J, Shah M. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

[63] Jiang Kai-Hua, Hu Zhi-Peng, Chen Zhong-Wei, et al. Pairwise event detection in surveillance video//Proceedings of the TRECVID. Boston, MA, USA, 2010: 1-8

[64] Ellis A, Ferryman J. PETS2010 and PETS2009 evaluation of results using individual ground truthed single views// Proceedings of the 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, 2010. Boston, MA,USA, 2010: 135-142

[65] Oh S, Hoogs A, Perera A, et al. A large-scale benchmark dataset for event recognition in surveillance video//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3153-3160

[66] Sullivan J, Carlsson S. Recognizing and tracking human action//Proceedings of the European Conference on Computer Vision. Copenhagen, Denmark, 2002: 629-644

[67] Zhang Zhang, Tan Tieniu, Huang Kaiqi. An extended grammar system for learning and recognizing complex visual events. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(2): 240-255

[68] Li Hong-Song, Li Da. Some advances in human motion analysis. Pattern Recognition and Artificial Intelligence, 2009, 22(1): 70-78(in Chinese)
(黎洪松, 李达. 人体运动分析研究的若干新进展. 模式识别与人工智能, 2009, 22(1): 70-78)

[69] Xie Yue-Lei, Chang Hong, Li Zhe, et al. A unified framework for locating and recognizing human actions//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 25-32

[70] Zhang Yan-Hao, Qin Lei, Yao Hong-Xun, Huang Qing-Ming. Abnormal crowd behavior detection based on social attribute-aware force model//Proceedings of the International Conference on Image Processing. Orlando, FL, USA, 2012: 2689-2692

HU Qiong, born in 1986, Ph. D. candidate. Her research interests include computer vision and patter recognition.

QIN Lei, born in 1977, Ph. D., associate professor. His research interests include computer vision and pattern recognition.

HUANG Qing-Ming, born in 1965, Ph. D., professor, Ph. D. supervisor. His research interests include multimedia analysis, image processing, computer vision and pattern recognition.

Background

This work was supported in part by National Basic Research Program (973 Program) of China under Grant No. 2009CB320906, in part by National Natural Science Foundation of China under Grant Nos. 61025011, 61133003, 61035001 and 61003165, in part by Beijing Natural Science Foundation under Grant No. 4111003.

Action recognition has been a hot spot issue in computer vision field with broad applications in multiple areas in the past a few years. Due to background clutter, camera motion, occlusion, object scale and illumination condition changes, how to extract "good" features and acquire robust feature descriptions are crucial to human action recognition. In the literatures, existing features can be divided into four categories, namely, (1) color, shape, texture based static features, (2) optical flow, trajectory based dynamic features, (3) space-time volume based spatial-temporal features, and (4) some description based features. All these features capture different aspects of a dynamic process, i. e., a certain type of human action. However, none of the aforementioned features escape suffering from some constraints or limitations. Therefore, it is still a long way to go in the direction of robust feature extraction and representation.

Another equally important problem in the human action recognition task is action modeling and classification. According to this survey, there mainly are three different approaches involved in this stage. The simplest one is static/dynamic template matching, the basic idea of which is natural and easy to understand, but the performance degrades heavily in realistic scenarios. Another widely used approach is statistics based methods, which try to learn the intrinsic model of each human action type based on large numbers of training data. Consequently, the model and computation complexity increases exponentially with the enrichment of training data, such as Flickr, YouTube, etc. The third type comes from the syntactic analysis and parsing techniques. Some researches propose the event recognition language (ERL) to describe a human action or event, and then use some grammar analysis to recognize events. This type of method starts from a high-level semantic view, and the accuracy depends on the low-level processing such as segmentation, detection and tracking.

To foster the advancement of this field, plenty of benchmark datasets and international competitions are published each year. Though no universal consensus has been reached on the definition and concepts of this problem, people are trying to advance the research towards this direction.