DOI:10.3969/j.issn.1673-4785.201212064
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130515.0939.010.html

视频序列的人体运动描述方法综述

孙倩茹[1,2],王文敏[1],刘宏[1,2]
(1.北京大学深圳研究生院 深圳物联网智能感知技术工程实验室,广东 深圳 518055; 2.北京大学 机器感知与智能教育部重点实验室,北京 100871)

摘 要:视频中的人体运动分析是计算机视觉领域的重要课题,同时也是近年来备受关注的前沿研究方向之一.在明确实际视频中存在的若干种难点,如人体遮挡、视频模糊、拍摄视角变化等基础上,从经典的人体运动特征提取、特征选择以及特征融合 3 个方面,对基于视频序列的人体运动描述方法和研究现状进行了概述,归纳出人体运动描述算法的研究难点,并分析了人体运动分析的技术发展趋势.指出了利用不同特征间存在的互补性质探求高性能特征选择和特征融合机制是人体运动描述技术发展的必然趋势,从处理简单实验场景视频向挑战高难度实际场景视频的转化是运动视频分析未来发展的方向.

关键词:视频序列;人体运动描述;特征提取;特征选择;特征融合
中图分类号: TP391 **文献标志码**:A **文章编号**:1673-4785(2013)03-0189-10

中文引用格式:孙倩茹,王文敏,刘宏. 视频序列的人体运动描述方法综述[J]. 智能系统学报, 2013, 8(3): 189-198.
英文引用格式:**SUN Qianru, WANG Wenmin, LIU Hong. Study of human action representation in video sequences[J]. CAAI Transactions on Intelligent Systems**, 2013, 8(3): 189-198.

Study of human action representation in video sequences

SUN Qianru[1,2], WANG Wenmin[1], LIU Hong[1,2]
(1. Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School of Peking University, Shenzhen 518055, China; 2. Key Laboratory for Machine Perception (Ministry of Education), Peking University, Beijing 100871, China)

Abstract: Recently analysis of human actions in videos has become an important issue in the field of computer vision. Much attention has been paid to this frontier research. In this paper, we first explicitly defines several existing difficulties in real-world videos, such as body occlusion, video fuzzy, shooting angle change and then conducts a survey based on the popular methods and present situation research studies on human action representation. Next, we focus attention on three aspects of feature extraction, feature selection and feature fusion, and then summarize the research difficulties in algorithms of action description, and analyze the technical development trend of human action analysis. It was pointed out that the inevitable trend of human action representation technology is to explore high-performance feature selection and feature merging mechanism by making use of the complementary mechanism among different features, and the development trend of motion video analysis in the future is to change from processing simple experimental scene videos to the challenge of real-world scene videos with high difficulties.

Keywords: video sequences; human action representation; feature extraction; feature selection; feature fusion

运动信息是视频的重要特性之一.近年来,摄像机等视频录制设备价格的不断降低,计算机性能的不断提高,使得视频相关领域的应用需求不断增加.视频中的人体运动分析是指通过计算机视觉和模式识别的各类技术手段,实现对视频序列中存在特定人体运动的智能化表示和标记.它在较多计算机视频应用领域有着广泛且重要的研究应用,因此对人体运动描述和识别算法的研究就成为了计算机视觉领域备受关注的前沿研究课题.近几年,国际上一些权威期刊如 IJCV (International Journal of Computer Vision)、CVIU (Computer Vision and Image Under-

收稿日期:2012-12-31. **网络出版日期**:2013-05-15.
基金项目:国家自然科学基金资助项目(60875050,60675025);国家"863"计划资助项目(2006AA04Z247);深圳市科学和技术创新委员会资助项目(JC201005280682A, JCYJ20120614152234873, CXC201104210010A).
通信作者:孙倩茹. E-mail:qianrusun@sz.pku.edu.cn.

standing)、PAMI(IEEE Transactions on Pattern Analysis and Machine Intelligence)、IVC(Image and Vision Computing),以及重要的国际学术会议,如 ICCV(International Conference on Computer Vision)、CVPR(IEEE Conference on Computer Vision and Pattern Recognition)、ECCV(European Conference on Computer Vision)等,已将基于视频信息的人体运动分析研究作为其主体内容之一[1].目前,在低噪声环境下获取的视频中进行运动检测和识别已经可以达到较高的识别效率,但是针对实际环境中的视频,人体运动描述和识别仍然面临很多难题.

1 人体运动描述的研究难点

由于人体运动识别需要将视频中包含的人体运动进行准确地描述和正确地分类,因此这是一项极富挑战性的研究工作.另外,当此类方法应用到实际视频中时,由于视频中存在的种种现象,如人体遮挡、视频模糊、拍摄视角变化等,所需要解决的问题就变得更加复杂.为了避免研究这些复杂情形,很多研究方法都集中在对视频质量和运动发生环境严格受限的理想数据库的实验上.而且,研究者为了得到鲁棒的运动描述特征,对视频中的人体运动进行了前提性的假设,如假设已经实现了鲁棒的人体跟踪,排除轻微的相机晃动和图像模糊以及对观察视角进行了若干个简单的划分.这些都从根本上限制了方法本身在实际视频中的应用.

为了解决这些问题,首先需要对问题本身进行分析,然后对识别过程中出现的各类难点问题进行总结.

(a) 人体遮挡 (b) 相机晃动 (c) 严重阴影

(d) 类内多样性 (e) 类间相似性

图 1 人体运动分析中的客观难点举例

Fig.1 Examples of problems in human action analysis

为了便于后面的论述,本文先给出几个术语.

1)类内多样性(intra-class variations).它指的是相同的运动存在不同的个体和视角.人体运动者处于不同的年龄阶段,会拥有不同的外表,同时运动速度和时空变化程度都有较大的差异.例如图 1(d)所示的 2 个骑单车运动,它们的不同之处就在于实验者的着装以及拍摄视角,这就导致了在特征获取阶段得到的特征存在着很大的差异性,最直观的轮廓特征就几乎完全不同.为了解决类内多样性,需要探求一种抓住运动本质的鲁棒运动特征.

2)类间相似性(inter-class similarity).它指的是不同的运动看上去有很大的相似性,这与类内多样性是相对的一种困难情况.例如图 1(e)中显示的 2 张灰度图像,2 个人好像都是在跑步,但是结合原多帧图像序列可以判断第 1 幅子图是在跑步而第 2 幅子图是在单腿向前跳.在视频中跑和跳出现了多帧极其类似的情况,这就给区分这 2 个运动带来了极大的模糊性.并且,当分类的运动种类增多时,这种类间相似造成的模糊性也会随之增大,进而导致识别率降低,这就要求继续研究高区分度的人体运动描述特征和模型.

3)人体遮挡(body occlusion).实际场景中的人体经常会被场景中的其他人或物体遮挡住部分或者全部的身体,有时还会因为视角的问题产生自遮挡的问题.这类问题严重影响了运动特征的有效提取和描述过程.此时,识别算法获取的特征是不完整的,甚至会误导识别结果,降低识别率.例如图 1(a)中的交互行为“拳击”,当摄像头角度固定时,2 个人拳击的过程中会不停地挪动,遮挡是很常见的,一旦遮挡发生就会造成子特征或者整体轮廓类特征混乱,对识别的进行会造成严重影响.另外,当全遮挡发生的时候,根本无法完成目标定位或者运动物体的定位,这是显而易见的实际难题.

4)视角转变(view point variation).当摄像机的视角发生大的转变时,所观察到的运动在计算机看来就有可能完全不同.例如,图 1(d)中的自行车运动,侧面得到的特征和背面有很大的不同,如轮廓、姿势等.当然,远近视角会造成尺度的变化,这也是需要在特征选取过程中考虑的因素.

5)相机运动(camera motion).相机运动是造成运动序列变化的另一种根本性因素,不合理的相机运动设置会造成严重的运动扭曲,其中就包括相机抖动的情况.相机在运动过程中会造成运动视角的转变以及背景的更新,因此固定相机和移动相机所拍摄的同一运动过程就会显现出不同的状态.一般会采用预处理的方式对相机移动造成的影响进行运动补偿,但是当视频中包含快速的背景变化或者难以进行轨迹参数化的相机运动时,预处理是完全无效的.

6)动态背景(dynamic background).实际场景中经常包含同时运动的多个人和物体,因此,背景是不断变化的.当存在这种变化时,运动识别主要面临的问题是会出现局部或者全身遮挡,导致目标定位和识别变得复杂和困难,背景减除也变得困难,运动特

征提取会因为严重的背景噪声而变得效率低下.

7)其他环境因素(environmental conditions).录像设置和场景选择也是影响运动分析的重要因素.例如,室外场景中存在的阴影、光照变化以及人群拥挤都会严重影响人体运动识别结果.

2　经典的人体运动描述方法

通常来说,不同的运动具有不同的计算复杂度,运动表示方法的不同会直接影响后续识别的效率.Bobick[2]将人的运动分为 3 类:动作(movement)、行动(action)和行为(activity),这 3 类运动分别处于 3 个不同复杂度的层次上.动作是运动的基元,是最基本的运动,是形成其他复杂、高级运动的基础.一般来说人体动作在执行过程中会持续较短的时间,其识别方法一般可以采用几何或概率统计的方法[3].一般来说,运动的表示与应用场合有紧密的关系,对于不同的情况通常会选择不同的运动表示方法.譬如,在对一个大的场景下进行较远距离的监控时,只需要提取运动目标的轨迹就可以满足需要,而在识别近距离人体动作时,对人的肢体进行 2D 或者 3D 建模则会起到更好的效果[3].

动作描述是指给定一段包含人体运动的视频,需要建立起视频(观察)到高维特征空间一种合理的映射,用特征或者特征的组合形式来表述这段运动.参考在第 1 节中提到的人体运动识别的难点,主要解决途径就是探求具备高类间区分度且对类内元素具有很好的"聚类"作用的特征描述方法.好的运动描述方法可以使人体运动识别系统实现高识别效率,因此近些年来,运动的特征描述成为运动识别的重点研究之一.

根据近些年的相关研究成果,关于运动的特征描述方法主要可以分为四大类[2]:1)基于时空形状模板(spatio-temporal shape template)的运动描述方法;2)基于光流(optical flow)特征的运动描述方法;3)基于运动轨迹(trajectories)的运动描述方法;4)基于兴趣点(interest points)的运动描述方法.

2.1　基于时空形状模板的运动描述方法

时空形状模板是一种较早的用于运动识别的方法.该方法在训练过程中通过对视频序列中检测到的人体形状建立起一组与特定运动相对应的人体形状序列.运动识别的过程实际就是模板匹配的过程,获得训练样本的高质量轮廓模板是这类方法的前提.因此,这类方法要求以高精度的人体轮廓分割(一般会使用背景减除)为前提,因此当出现复杂背景情形时,如相机晃动、人体阴影、人体遮挡或者多个运动目标,此类方法的识别率会变得比较低,甚至会完全失效.

Bobick 和 Davis[2] 于 2001 年提出使用人体轮廓模板训练分类器的方法,他们采集单一视角的轮廓并对其进行聚类以提取可计算的特征向量.首先利用连续的轮廓形状建立一种运动能量图(MEI),用于表征运动发生的位置信息;再建立一种运动历史图(MHI),用于表征轮廓序列的灰度值变化情况;最后通过对这 2 种特征图像的参数化描述建立运动特征向量(如图 2).Blank[4] 和 Yilmaz[5] 先后提出了结合运动信息的人体 3D 体积模型,通过获取轮廓序列计算 3D 体积的特征值(如体积大小、时空角点位置等)来计算运动描述向量.除了利用单纯的轮廓和体积信息之外,Wang[6] 为了探索人体轮廓的运动流形轨迹的内在结构,于 2007 年提出采用 LPP[7](R 变换)对提取的人体轮廓序列进行轨迹分析,他们在多个具有挑战性的数据库上对这种方法的鲁棒性进行了实验验证,均取得了较好的识别效果.

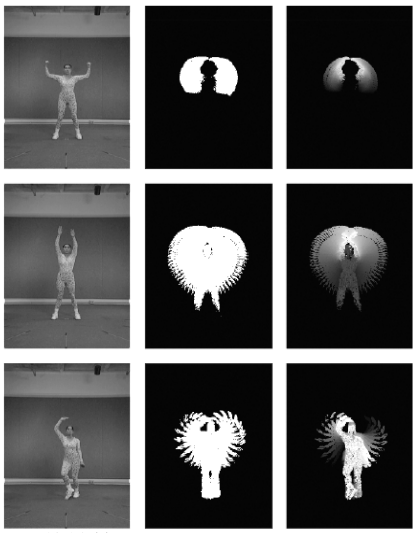


图 2　人体运动描述子:MEI 和 MHI

Fig.2　Human action descriptors: MEI and MHI

后续的很多文献致力于获取对视角和尺度变化较为鲁棒的轮廓描述方法,但是在处理实际视频中出现的遮挡、拥挤、大视角大尺度视频变化等问题时,基于时空形状模板的运动描述方法难以满足识别要求.

2.2　基于光流特征的运动描述方法

基于光流特征的运动描述方法是将人体运动联合背景变化作为一个变化的整体,然后通过获取主运动区域来定位人体运动.光流法不需要预先获取图像背景,而且计算结果仅仅依靠连续帧的相对运动,不受复杂背景的影响,因而在基于对象的运动估计、运动检测和跟踪等领域都有广阔的应用前景[8].光流的基本计算以 2 帧图像亮度恒定为前提,用泰勒级数一阶展开,使得光流计算受限于 2 帧图像间

的运动不能大于 1 个像素,因此只有当相邻 2 帧间的运动不大于 1 个像素时,标准光流算法才比较可靠.Efros[9] 于 2003 年率先在这个领域有所突破,他先是利用跟踪算法锁定远距离的人体,然后再对跟踪目标框内的视频流进行光流变化的检测和统计,同时将光流变化划分为上下左右 4 个独立通道,获取邻接跟踪通道中光流的时间相关数据作为最终的运动描述子.如图 3 所示,其中图 3(a)为原始图像,图 3(b)是光流图像,图 3(c)是分离出来的 x 和 y 方向上的光流分量,图 3(d)是半波整流产生的 4 个独立通道分量,图 3(e)是得到的模糊运动通道.这种方法虽然没有利用轮廓分割,但是需要鲁棒的人体跟踪,因此目标的全遮挡和尺度变化对检测性能的影响很大.其后,Fathi[10] 提出了结合底层和中层特征对视频中的运动区域进行整体描述,其中中层特征就是采用上述光流方法.

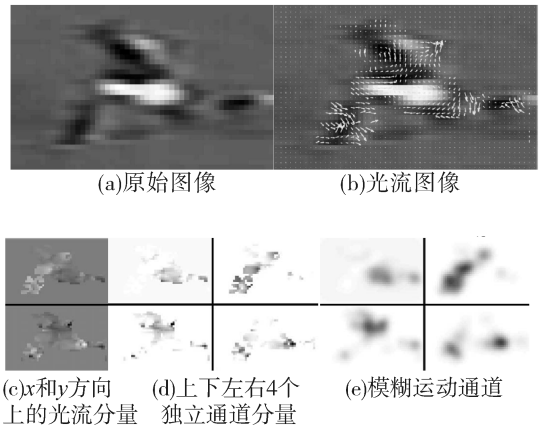


图 3 光流及其光流描述子

Fig.3 Optical flow and its corresponding description

另一类光流描述方法是将光流算子作为局部特征,通过经典统计方法训练分类器.Danafar[11] 将跟踪通道内的局部光流按照横向和纵向划分为 2 个独立通道进行特征直方图统计,将统计结果作为运动描述算子,再利用 SVM 分类器实现最终的运动分类和识别.实验表明该局部特征统计方法对环境噪声和视角变化有良好的鲁棒性.

2.3 基于运动轨迹的运动描述方法

运动轨迹模型是利用时空描述子记录行动物体的运动轨迹或肢体运动轨迹,通过轨迹特征来表述相应的运动.其优点是可以记录关于目标运动的整体发生时间特性,增强运动间的区分度;缺点在于它通常需要将 3D 空间的轨迹映射到 2D 再进行数学化描述,这就造成了视角模糊性,进而增加了运动种类间的模糊性.为了尽量减少轨迹的模糊性,很多学者通过增加与运动轨迹相联系的描述信息,如局部特征、身体轮廓等,建立起三维综合信息,实现最终的人体运动描述.

Rao 和 Shah[12] 于 2001 年研究了人手运动过程中轨迹的视角不变性,他们的方法是通过计算轨迹的时空曲率实现的,其中运动轨迹是通过肤色跟踪器记录人手动作执行过程得到的.Sheikh[13] 于 2005 年提出了使用人体的多个肢点记录轨迹来表示某一个人体的全身运动情况,其描述空间是包含了时间轴的 4D 空间.这一方法的明显缺点是无法对长时间视角变化的运动轨迹进行清晰有效的表示.为了探索这一问题的解决方案,John[14] 于 2010 年提出了一种多视角情况下跟踪铰链式人体运动的框架,该方法在文中设定的一系列限制条件下取得了较好的实验效果.Ali[15] 则另辟蹊径,提出了获取关节轨迹的混沌不变量作为不同运动之间可用于相互区分的特征,该特征在 2 个经典的人体运动数据库上取得了较好的运动分类效果.图 4 中描绘了人体跑步运动的 3 种轨迹曲线[15].

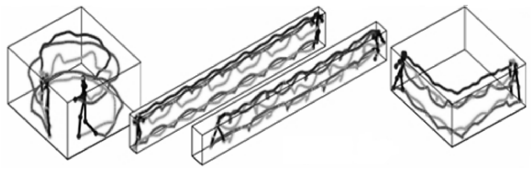


图 4 跑步运动的 3 种运动轨迹

Fig.4 Running trajectories toward three orientations

针对视频中存在的运动识别难点,Wang[16] 和 Raptis[17] 分别于 2011 年和 2012 年提出利用视频中局部特征的稠密运动轨迹进行运动识别和分析,这是目前针对实际拍摄中相机抖动的 2 种性能较好的方法.其缺点是计算和分析的代价较大,检测到的局部特征维度较高,因此对计算机的计算能力要求高.同时,所记录运动轨迹的稠密性决定了这类方法对计算机的存储能力要求也很高.

2.4 基于兴趣点的运动描述方法

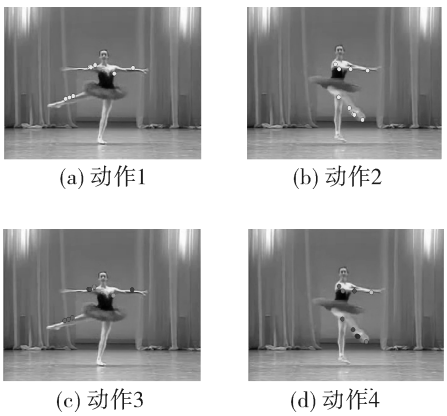
基于兴趣点的人体运动描述模型是建立在 2 个关键步骤上的:兴趣点的检测和兴趣点周围局部区域的描述.在现有的各种运动表示方法和模型中,基于兴趣点的运动描述是研究者们最热衷的一类方法.与前面提到的人体模型表示方法和轨迹表示方法比较,兴趣点模型最大的优点就是不需要跟踪移动人体或对其进行任何轮廓轨迹的建模,并且兴趣点是对显著区域的稀疏采样,因此其存储和计算代价较小.这类模型的缺点是无法解决动态背景干扰问题.下面分 2 个部分论述这类模型的主要框架和实现方法:兴趣点检测和局部区域描述.

2.4.1　兴趣点检测

兴趣点是指当运动发生时在视频中检测到的运动显著位置的集合.对于不具备连续性的运动来说,很多基于模板匹配的方法会失效,此时兴趣点检测显得尤为重要.更重要的是,兴趣点检测不需要考虑视角变换和运动事件周期的变化.近些年,在图像识别领域出现了很多兴趣点定义和检测的方法,比较著名的是 Harris[18] 于 1988 年提出的图像角点检测,2003 年 Laptev 等[19]将 Harris 角点检测拓展到三维视频数据的显著区域定位上,提出了 3D 时空兴趣点的检测方法.2D 兴趣点完全忽略视频数据中的时域变化信息,而 3D 兴趣点周围局部区域内的三维灰度数据无论是在时空域还是在时域上都包含了比较丰富的像素变化信息,所以它们普遍具有很强的特征描述能力并且应用广泛.这种检测角点的缺点在于处理比较平滑少纹理的视频数据时,检测足够多的有效显著区域是比较困难的.为了解决这一显著点过于稀疏的问题,Dollar[20] 在 2005 年提出了一种基于周期性运动的兴趣点检测方法(图 5 所示),他利用 2D 空间高斯平滑核函数 g 和 1 对 1D 时间高斯核函数 h_{ev}、h_{od}(式(1)和(2))构造了三维响应函数 R(式(3)).所要获取的兴趣点个数是通过手动调整 2 个核函数的尺度参数来进行设定的.这个兴趣点检测器虽然应用广泛;但是检测器本身还是存在一些缺点,比如运动物体边缘较为平滑(和背景区分度很小)即灰度值变化较小时,检测足够的显著区域也是比较困难,另外,这种检测是在单一尺度(固定尺度参数)下的.

$$\begin{cases} h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2}, & (1) \\ h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)\mathrm{e}^{-t^2/\tau^2}; & (2) \end{cases}$$

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2. \quad (3)$$



(a) 动作1　(b) 动作2　(c) 动作3　(d) 动作4

图 5　芭蕾舞运动中的兴趣点检测和相应的分类标记

Fig.5　Interest points of Ballet motions and their corresponding labeled results

2006 年,Oikonomopoulos[21] 提出了一种改进的兴趣点检测器,他利用光流信息来降低相机运动或背景连续变化所带来的背景干扰问题.具体算法是将光流场中的热力熵信息与 Dollar 方法中的灰度梯度值相结合之后对视频中的显著区域进行检测.该方法使用的时空尺度是利用测试得到的最佳尺度.为了增强兴趣点的描述有效性,该方法还采用了聚类的手段进行特征选择,排除掉了一些低显著度的兴趣点.

总的来说,这类显著区域检测器主要利用的是图像变化信息,因此其检测性能还是很好的.但是这类检测器最大的缺陷是主要适用于静止相机拍摄的包含运动信息的视频.为此,2007 年 Wong 和 Cipolla[22] 提出将空间域检测和时间域检测分开进行的思想,这样就可以在空间上做合理化的背景减除来提取主要运动区域,进而适应移动摄相机的情况.

2.4.2　局部区域描述

近年来,在对兴趣点周围局部区域的描述,即局部特征的获取上,相关研究人员花费了很大的精力.Schuldt[23] 是这方面工作的先驱,他先是利用 Laptev 的检测器[19] 检测兴趣点,再对兴趣点周围的立体区域提取灰度值变化的标准化差分算子作为局部特征,最后经过聚类算法计算所有获取特征的统计直方图.这种方法可以避免摄像头移动带来的干扰,但是对于相似度较高的运动(如跑步和单腿向前跳)识别效果比较差.

较先提出兴趣点检测算子的 Dollar[20] 同时提出了局部特征描述的方法.他在文章中对 3 种不同的描述子进行了测试:像素级别的归一化描述子、亮度梯度描述子和基于光流统计的描述子.其中利用亮度梯度描述子的分类器达到了最好的运动识别效果.该方法还利用 PCA 来降低特征维度,提高了计算和存储的效率.

为了将检测到的显著区域较好地表示出来,以达到较高的特征区分度,Scovanner[24] 在 2007 年提出了改进的 3D-SIFT 算子(图 6 所示),利用 3 个维度的高斯差分结果计算局部灰度特征,这是一种时间域上扩展的 SIFT 方法[25].此外,还有很多特征融合的描述子建立方法[26-27] 都取得了较高的人体运动识别率.

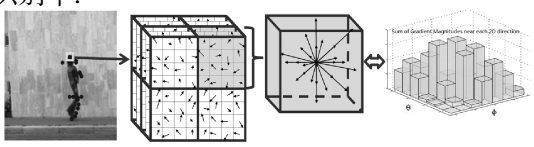


图 6　3D-SIFT 描述算子的提取和建立过程

Fig.6　The extraction process of a 3D-SIFT descriptor

为了更深层次地探索稀疏兴趣点所夹带的空间运动信息,Sun 等[28] 于 2012 年提出将点与点之间的距离信息(NGLD 相关图,如图 7 所示)与传统局部特征的 Bag-of-Words 模型[24] 相结合.在相同的数据库上进行验证时,此方法取得了更加鲁棒的人体运动识别效果.

(a) WEIZMANN数据库

(b) UCF运动数据库

图 7 数据库举例及其对应的 NGLD 相关图

Fig.7 Dataset samples and their corresponding NGLD correlograms

轨迹描述模型和兴趣点描述模型最大的问题在于它们记录的运动信息量较为稀疏,容易被实际视频中出现的噪声轨迹和噪声点干扰,导致识别率低,甚至完全失效.鉴于实际视频中运动背景和前景噪声干扰常常是不可避免的,剔除检测过程中得到的冗余特征就变得尤为重要.在这方面,较为常用的方法就是进行特征选择.

3 特征选择方法

利用上述描述方法对视频中的运动信息进行特征提取之后,所得到的底层描述特征中常常包含冗余和噪声成分,此时,特征选择和特征降维就显得尤为关键.进行特征选择的主要目的是:1)避免特征冗余导致的数据溢出;2)生成更加快速有效的运动表示模型;3)将特征数据所包含的内在结构信息尽量多地提取出来.综上,特征选择的目的就是选择尽量少的特征子群,达到尽量高的运动识别效率.

Cover[29] 和 Jain[30] 分别在 1974 年和 2000 年的文献中证明,将相互独立的优质特征进行联合所得到的描述算子,相比较于单特征并不一定能产生更好的识别效果.Guyon 等[31] 在 2003 年的文献中表明,相关性较高的特征之间往往会存在互补特性,去除冗余特征并不是简单地剔除相关性高的特征.由此可见,高相关性特征选择和分析对实现合理高效的运动描述是非常重要的.

对视频中的人体运动识别问题来说,特征选择方法可以总结为两大类:滤纸法和包装器法.滤纸法是一个预处理过程,通过计算特征的相关性和目标运动的区分能力实现特征选择.而包装器方法则通过评估所得到分类器的分类性能来进行特征选择.两者主要的区别就在于包装器方法的实现借助于目标分类器,而滤纸法则不需要.因此,相对于包装器方法,滤纸法计算速度较快,选择效率较低.

3.1 滤纸法

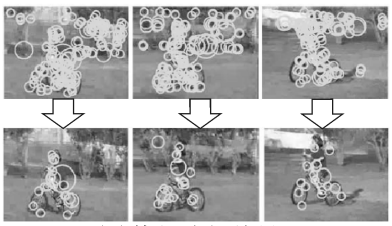
滤纸法的思路就是考虑每个独立的特征,对其与目标运动类的相关度进行评估排名,选择前 k 个特征即可.其中比较经典的特征评估算法如 1992 年 Kira 和 Rendell 提出的救援算法[32],该算法通过计算不同特征对不同运动类别的线性从属值的大小进行特征排序.Zaffalon 和 Hutter[33] 于 2002 年提出了另外一种基于互信息量度的非线性从属关系计算方法.其中互信息量度是指一个独立特征对一个特别运动类别的区分度大小.例如,一个特征被所有的运动类所共享,那么这个特征的区分能力就是很低的;相反,如果一个特征仅属于一个运动类别,那么它的区分度是最高的.

Peng[34] 在 2005 年提出了一种较为复杂的特征选择框架,称为“最小冗余-最大相关性”(MRMR).它将所选择特征类和运动类间的互信息最大化,而将特征与特征间的相关性最小化.实验证明,该方法对分类问题里的特征选择是有效的.

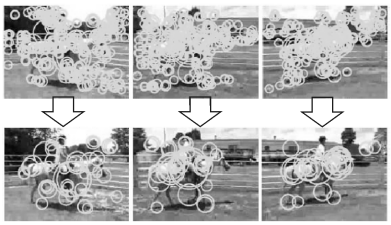
3.2 包装器法

包装器方法是利用机器学习方法实现的.大多数的情况该方法利用一定数量的测试样本得到区分度较高的特征子集,但由于每个特征都会占用一个测试回合,因此比较耗时.此外,实际视频中包含了高度的类间相似和类内变化,所以普通的特征选择方法的效率较低.

Liu[35] 提出了一种非常规的 PageRank(PR)方法来选择高区分度的描述特征.该方法通过建立有向图来评估每个特征与其他特征的匹配度,将匹配度最高的特征视为最具区分度的特征.图 8 是使用 PageRank 方法得到的特征选择结果.特征选择结果表明,PR 方法剔除了由于背景干扰和相机抖动得到的大部分噪声点.但是此方法不考虑动态背景,即默认得到的特征都表示的是前景物体,因此不适用于多物体的运动特征分析.针对这一问题,Gilbert 等[26] 在 2009 年提出了另一种完全不同的思路,他们注重对特定运动类别中出现的局部特征的记录和统计,不涉及任何前景轮廓信息,因此更加适合于分析非时空尺度视频中获得的人体运动特征.



(a)特征选择结果1



(b)特征选择结果2

图 8　使用 PageRank 方法得到的特征选择结果

Fig.8　Feature selection using PageRank

上述 2 种特征选择方法利用了特征在分类时表现的不同类间区分能力,其中滤纸法在不进行任何训练的情况下,力求得到可以实现最优二分类的独立特征.也就是说,该方法在选择特征的时候不考虑类间公共特征.与此不同的是包装器法利用同一特征在不同运动类别上表现出来的特性,通过训练分类器根据类间独立特征和类间共享特征对运动分类的影响进行评估,进而评价特征的优劣,实现最终的特征选择.

4　特征融合方法

特征融合是指为了得到更好的运动表示模型和更高的运动识别率,将来自不同获取渠道的特征进行合理的信息融合.一般来说,只有对具有互补作用的特征进行融合时才会提高运动识别率.另外,由于不同特征在维度、尺度和可行性上都有区别,所以直接融合反而会带来性能的降低.由此可见,合理有效的融合算法是极为重要的.现有的特征融合方法一般可以划分为 2 种:特征层面的融合和决策层面的融合.

4.1　特征层面的融合方法

所谓特征层面的融合,是指对不同特征的特征空间进行合并,最终使用一个融合之后的特征对运动进行表示.这是最常见的一种融合思路.Lin 等[36]指出可以将运动特征和形状特征进行加权融合,其中各项的最优权值利用部分训练样本的交叉验证计算得到.Schindler 等[37]将 3 类局部特征(时空梯度、光流、SIFT)进行了融合,在统一框架的实验中提升了 4.5%的人体运动识别率.

这类方法比较直观简单,但是特征直方图的融合要求严格正确的归一化操作.如果特征之间存在严重的维度差异,就必须调整特征空间的维度,否则特征融合之后的性能反而会变得很差.

4.2　决策层面的融合方法

决策层面的融合是指首先使用多种特征分别训练分类器,然后将得到的几个分类决策进行判决得到最终的分类或者识别结果.显然,这种方法是以独立特征的识别结果为前提的.

表 1 是典型人体运动特征建立、特征选择和特征融合等方法在典型人体运动数据库[4,19]上的识别率的比较.

表 1　典型的运动描述方法的识别率比较

Table 1　Recognition rates of traditional action representation methods　%

人体描述方法	KTH 数据库[19]	WEIZMANN 数据库[4]
Sun[28]	—	100.00
Sun[38]	94.00	97.80
Lin[36]	93.43	—
Wang[39]	92.51	100.00
Liu[35]	92.30	—
Ikizler[40]	94.00	—
Fathi[10]	90.50	100.00
Zhang[41]	91.33	92.89
Kläser[42]	91.40	84.30
Niebles[27]	83.30	90.00
Liu[43]	94.16	—
Zhao[44]	91.17	—
Gilbert[45]	89.92	—
Savarese[46]	86.83	—
Nowozin[47]	84.72	—
Dollar[20]	81.17	85.20

5　总结与展望

视频中的人体运动描述作为一个新的研究领域,在实际应用上存在着很多问题,在今后的若干年中仍会是一个研究热点.如下几个方面已经成为未来的发展趋势.

1)从视频中获取足够显著的人体运动特征.基于视频的人体运动特征提取是人体运动识别领域的重要研究内容.由于人体是非刚性结构,且运动过程中存在遮挡等问题,使得基于视频的人体运动捕获非常困难;然而目前国内外的研究成果对诸多问题进行简化,且大多只能对标准数据库中的运动进行有效的捕获.从实际场景的运动视频中捕获、提取出显著的人体运动信息始终是推动整个运动识别领域发展的重要研究方向.

2)定义和提取运动序列中的最小运动基元.动作识别中没有明确定义的基元,即基本动作单元.大多数的运动描述中并没有涉及识别前的动作分割,然而运动的划分是很多实际视频处理过程中必须考虑的重要问题.反过来讲,运动序列是由一系列特征的阶段构成的,如果能够将其与人体运动特点相结合对其合理地细化,用最小基元作为识别单位,对提高鲁棒性有很大帮助.这里需要研究的课题是:如何定义最小基元以及如何有效地提取最小基元的表示形式.

3)对现有的运动描述模型进行深层次改进.主要考虑运动描述方法在视频处理过程中对大信息量的承载,这就导致对特征降维方法的迫切需求.另外就是利用不同特征间存在的互补性质探求合理的特征选择和特征融合机制,例如,局部兴趣点和运动轨迹的融合特征、光流算子和场景模型描述子的结合.更深层次的研究,可以就已有的特征描述算法提出合理的高层次特征建立框架[48-50],或寻求更高区分度的特征实现更加高效的动作描述.

4)将已有算法与实际应用平台相结合.人体运动分析的主要应用场景有家庭环境、公共场所、危险环境和其他一些特定场景[51].当具体的算法应用到这些实际场景中时,运动描述与运动识别方法在系统上的实际效果就成为主要的测评标准.当前人机之间的通信仅局限于几个特定的姿势,这个局限是人的姿势结构不易理解造成的,而且跟踪多人的系统由于摄像机的分辨率、计算机处理能力和视角的影响而不能准确地估计身体姿势.为了完成优化尺度和广域的分析,可以寻求准确实时的多摄像机的信息融合方法,以便机器更好地理解人的肢体运动或者行为[52].

本文主要对运动识别方法中的特征描述模型进行了研究.归纳起来,利用视频信息进行运动描述的发展趋势主要是:从2D图像(空间)信息向3D视频(时空)信息转化;从单一特征表述到多特征融合的方法转化;从整体特征向局部子特征的方向转化;从处理简单运动视频向高复杂度视频的方向转化;从受限制的实验环境到实际应用场景中的转化.

参考文献:

[1]王亮,胡卫明,谭铁牛.人运动的视觉分析综述[J].计算机学报, 2002, 25(3): 225-237.
WANG Liang, HU Weiming, TAN Tieniu. A survey of visual analysis of human motion[J]. Chinese Journal of Computers, 2002, 25(3): 225-237.

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

[3]傅晓英.基于半连接HMM模型的人体行为识别研究与实现[D].北京:北京交通大学, 2009: 19.
FU Xiaoying. Study and implementation of human action recognition based on semi-connected HMM[D]. Beijing: Beijing Jiaotong University, 2009: 19.

[4]BLANK M, GORELICK L, SHECHTMAN E, et al. Actions as space-time shapes[C]//International Conference on Computer Vision. Beijing, China, 2005: 1395-1402.

[5]YILMAZ A, SHAH M. Actions sketch: a novel action representation[C]//IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 984-989.

[6]WANG L, SUTER D. Learning and matching of dynamic shape manifolds for human action recognition[J]. IEEE Transactions on Image Processing, 2007, 16(6): 1646-1661.

[7]HE Xiaofei, NIYOGI P. Locality preserving projections[M]//THRUN S, SAUL L K, SCHOLKOPF B. Advances in Neural Information Processing Systems. Cambridge, USA: The MIT Press, 2003.

[8]冯波.基于光流计算的典型行为识别算法研究[D].西安:西北工业大学, 2006: 25-26.

[9]EFROS A A, BERG A C, BERG E C, et al. Recognizing action at a distance[C]//International Conference on Computer Vision. Nice, France, 2003: 726-733.

[10]FATHI A, MORI G. Action recognition by learning mid-level motion features[C]//IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8.

[11]DANAFAR S, GHEISSARI N. Action recognition for surveillance applications using optic flow and SVM[C]//Asian Conference on Computer Vision. Tokyo, Japan, 2007: 457-466.

[12]RAO C, SHAH M. View-invariance in action recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Kauai, USA, 2001: 316-322.

[13]SHEIKH Y, SHEIKH M, SHAH M. Exploring the space of a human action[C]//International Conference on Computer Vision. Beijing, China, 2005: 144-149.

[14]JOHN V, TRUCCO E, MCKENNA J. Markerless human motion capture using charting and manifold constrained particle swarm optimisation[C]//British Machine Vision Conference (Workshops). Aberystwyth, UK, 2010: 1-11.

[15]ALI S, BASHARAT A, SHAH M. Chaotic invariants for human action recognition[C]//International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1-8.

[16] WANG H, KLASER A, SCHMID C, et al. Action recognition by dense trajectories[C]//IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3169-3176.

[17] RAPTIS M, KOKKINOS I, SOATTO S. Discovering discriminative action parts from mid-level video representations[C]//IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 1242-1249.

[18] HARRIS C, STEPHENS M. A combined corner and edge detector[C]//Proceedings of the Fourth Alvey Vision Conference. Manchester, UK, 1988: 147-151.

[19] LAPTEV I, LINDEBERG T. Space-time interest points [C]//International Conference on Computer Vision. Nice, France, 2003: 432-439.

[20] DOLLAR P, RABAUD V, COTTRELL G, et al. Behavior recognition via sparse spatiotemporal features[C]//IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Beijing, China, 2005: 65-72.

[21] OIKONOMOPOULOS A, PATRAS I, PANTIC M. Spatiotemporal salient points for visual recognition of human actions[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2006, 36(3): 710-719.

[22] WONG S, CIPOLLA R. Extracting spatiotemporal interest points using global information[C]//International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1-8.

[23] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]//International Conference on Pattern Recognition. Cambridge, UK, 2004: 32-36.

[24] SCOVANNER P, ALI S, SHAH M. A 3-dimensional SIFT descriptor and its application to action recognition[C]// Proceedings of the 15th International Conference on Multimedia. Augsburg, Germany, 2007: 357-360.

[25] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[26] GILBERT A, ILLINGWORTH J, BOWDEN R. Fast realistic multi-action recognition using mined dense spatio-temporal features[C]//International Conference on Computer Vision. Kyoto, Japan, 2009: 925-931.

[27] NIEBLES J, WANG H, FEI-FEI L. Unsupervised learning of human action categories using spatial-temporal words [J]. International Journal of Computer Vision, 2008, 79 (3): 299-318.

[28] SUN Qianru, LIU Hong. Action disambiguation analysis using normalized Google-like distance correlogram [C]// 11th Asian Conference on Computer Vision. Daejeon, Korea, 2012: 425-437.

[29] COVER T M. The best two independent measurements are not the two best[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1974, 4(1): 116-117.

[30] JAIN A K, DUIN R P W, MAO J. Statistical pattern recognition: a review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37.

[31] GUYON I, ELISSEEF F. An introduction to variable and feature selection [J]. Journal of Machine Learning Research, 2003, 3: 1157-1182.

[32] KIRA K, RENDELL L A. The feature selection problem: traditional methods and a new algorithm[C]//Proceedings of the Tenth National Conference on Artificial Intelligence. San Jose, USA, 1992: 129-134.

[33] ZAFFALON M, HUTTER M. Robust feature selection by mutual information distributions[C]//International Conference on Uncertainty in Artificial Intelligence. Edmonton, Canada, 2002: 577-584.

[34] PENG H, LONG F, DING C. Feature selection based on mutual information: criteria of maxdependency, max-relevance, and min-redundancy [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[35] LIU J, LUO J, SHAH M. Recognizing realistic actions from videos "in the wild" [C]//IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1996-2003.

[36] LIN Z, JIANG Z, DAVID L S. Recognizing actions by shape-motion prototype trees [C]//International Conference on Computer Vision. Kyoto, Japan, 2009: 444-451.

[37] SCHINDLER G, ZITNICK L, BROWN M. Internet video category recognition[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Anchorage, USA, 2008: 1-7.

[38] SUN X, CHEN M, HAUPTMANN A. Action recognition via local descriptors and holistic features[C]//IEEE Conference on Computer Vision and Pattern Recognition. Kyoto, Japan, 2009: 58-65.

[39] WANG H, ULLAH M M, KLASER A, et al. Evaluation of local spatiotemporal features for action recognition[C]// British Machine Vision Conference. London, UK, 2009: 124.1-124.11.

[40] IKIZLER N, CINBIS R G, DUYGULU P. Human action recognition with line and flow histograms[C]//International Conference on Pattern Recognition. Tampa, USA, 2008: 1-4.

[41] ZHANG Z, HU Y, CHAN S, et al. Motion context: a new representation for human action recognition[C]//European Conference on Computer Vision. Marseille, France, 2008:

817-829.

[42] KLASER A, MARSZALEK M, SCHMID C. A spatio-temporal descriptor based on 3D-gradients[C]//British Machine Vision Conference. Leeds, UK, 2008: 995-1004.

[43] LIU J, SHAH M. Learning human actions via information maximization[C]//IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA, 2008: 1-8.

[44] ZHAO Z P, ELGAMMAL A M. Information theoretic key frame selection for action recognition[C]//British Machine Vision Conference. Leeds, UK, 2008: 109.1-109.10.

[45] GILBERT A, ILLINGWORTH J, BOWDEN R. Scale invariant action recognition using compound features mined from dense spatio-temporal corners[C]//European Conference on Computer Vision. Marseille, France, 2008: 222-233.

[46] SAVARESE S, POZO A D, NIEBLES J, et al. Spatial-temporal correlations for unsupervised action classification [C]//IEEE Workshop on Motion and Video Computing. Copper Mountain, USA, 2008: 1-8.

[47] NOWOZIN S, BAKIR G H, TSUDA K. Discriminative subsequence mining for action classification[C]//International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1-8.

[48] 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.

[49] CUI Peng, WANG Fei, SUN Lifeng, et al. A matrix-based approach to unsupervised human action categorization[J]. IEEE Transactions on Multimedia, 2012, 14(1): 102-110.

[50] SADANAND S, CORSO J J. Action bank: a high-level representation of activity in video[C]//IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 1234-1241.

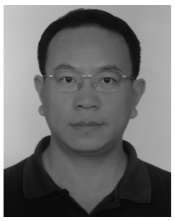
[51] BALLAN L, BERTINI M, DEL BIMBO A, et al. Effective codebooks for human action representation and classification in unconstrained videos[J]. IEEE Transactions on Multimedia, 2012, 14(4): 1234-1245.

[52] 楼晓燕.智能监控系统若干关键技术的研究和实现[D].杭州:浙江工业大学, 2005: 8.
LOU Xiaoyan. Research and implementation of some key technique on intelligent monitoring system[D]. Hangzhou: Zhejiang University of Technology, 2005: 8.

作者简介:

孙倩茹,女,1987 年生,博士研究生,主要研究方向为计算机视觉、模式识别.

王文敏,男,教授,主要研究方向为 Web 技术、嵌入式软件系统、智能终端技术.主持省、市重大科技专项 6 项,作为主要研发人员参与国家科技支撑计划项目 1 项、省部产学研结合项目 1 项,提交发明专利申请 1 项(第一发明人),取得软件著作权 1 项(第一著作权人).获得国家(首批)青年自然科学基金、省部级科技奖 3 项、市级科技奖 1 项、部级鉴定 1 项.发表学术论文 30 余篇.



刘宏,男,1967 年生,教授,博士生导师,中国人工智能学会副秘书长.主要研究方向为计算机视听觉、智能机器人.先后承担国家自然科学基金 7 项、国家“863”计划、“973”计划项目 5 项. 获国家航天科技进步奖. 发表学术论文 130 余篇,其中被 SCI、EI 检索 90 余篇.