人体动作行为识别研究综述*

李瑞峰 王亮亮 王 珂

(哈尔滨工业大学 机器人技术与系统国家重点实验室 哈尔滨 150080)

摘 要 人体动作行为识别因其在视频监控、虚拟现实、人机智能交互等领域的广泛应用而成为计算机视觉领域的研究热点.文中将人体动作行为识别问题归纳为计算机经过检测动作数据而获取并符号化动作信息,继而提取和理解动作特征以实现动作行为分类的过程,在此基础上,从运动目标检测、动作特征提取和动作特征理解 3 个方面对涉及到的技术进行回顾分析,对相关方法进行分类,并讨论相关难点和研究方向.

关键词 动作识别,运动目标检测,动作特征提取,动作特征理解
中图法分类号 TP 391.4

A Survey of Human Body Action Recognition

LI Rui-Feng, WANG Liang-Liang, WANG Ke
(*State Key Laboratory of Robotics and System*, *Harbin Institute of Technology*, *Harbin* 150080)

ABSTRACT

Action recognition has become a hotspot in the fields of video surveillance, virtual reality, human-computer interaction and others recently. In this paper, action recognition is comprehended as a process of detecting action data, called symbols of action message, and distinct actions based on action feature extraction and reception are further classified. On the basis, an overview of vision-based full-body action recognition techniques is presented within the domain of moving object detection, action feature extraction and action feature perception, and the corresponding methods are classified. Besides, the research trend of action recognition is discussed.

Key Words Action Recognition, Moving Object Detection, Action Feature Extraction, Action Feature Perception

1 引 言

人体动作行为识别是近年来计算机视觉领域的一个研究热点,其广泛应用于人机智能交互[1-3]、虚拟现实[4]和视频监控[5-6]等领域.尽管近年来国内外人体动作行为识别的研究取得了重要进展,但人体运动的高复杂性和多变化性使得识别的精确性和高效性并没有完全满足相关行业的实用要求.总体

* 国家自然科学基金项目(No. 61273336)、国家教育部博士点专项基金项目(No. 20122302120039)、中央高校基本科研业务费专项资金项目(No. HIT. NSRIF. 201172)资助
收稿日期:2013-05-17;修回日期:2013-06-13
作者简介 李瑞峰,男,1965 年生,教授,博士生导师,主要研究方向为智能机器人、工业机器人. E-mail: Lrf100@hit.edu.cn. 王亮亮(通讯作者),男,1987 年生,博士研究生,主要研究方向为机器视觉、模式识别. E-mail: yueyangmeng@gamil.com. 王珂,男,1979 年生,博士,讲师,主要研究方向为人工智能、机器视觉.

来说人体动作行为识别中的挑战来自以下两方面.

1) 空间复杂性. 不同光照、视角和背景等条件下会呈现不同的动作场景,而在不同的动作场景中相同的人体行为在姿态和特性上会产生差异. 即使在恒定的动作场景中,人体动作也会有较大的自由度,而且每种相同的动作在方向、角度、形状和尺寸方面有很大的差异性. 此外,人体自遮挡、部分遮挡、人体个体差异、多人物识别对象等问题都是动作识别复杂性在空间上的体现. 空间复杂性对人体动作行为识别结果的影响主要体现在精确性方面.

2) 时间差异性. 时间差异性是指人体动作发生的时间点不可预测,而且动作的持续间隔也不尽相同. 此外,动作在作用时间内也可能存在动作空白间隙. 时间差异性要求识别过程中能够辨别动作的起止时间,同时有效判断动作作用的有效时间和间隔,对动作在时域和时序范围内进行更加细致的分析,导致动作在不同速率、顺序和组合情况下都会存在差异. 时间差异性不仅对识别精确性产生影响,也会带来计算实时性和效率等影响识别高效性的问题.

人体动作行为的空间复杂性和时间差异性使得人体动作行为识别研究领域并没有统一有效的研究框架,相关技术也没有标准固定的分析分类方法. 针对基于视觉的全身人体运动行为识别研究,普遍通过 3 种方式进行分析和分类: 1) 将人体动作行为识别划分为几个相互联系的子过程,根据过程划分方式的不同对相应技术进行分析分类; 2) 典型问题方式,即选取人体动作行为识别中的部分典型问题作为对象,针对这些对象研究中涉及的方法进行分析分类; 3) 空间时域方式,即根据时域和空间上各项研究方法的差别对相应技术进行分析分类. 文献[7] ~ [12] 列出了人体动作行为识别一些综述性研究及其分析分类方法.

本文将人体动作行为识别问题归纳为计算机经过检测动作数据而获取并符号化动作信息,继而提取和理解动作特征从而实现动作行为分类的过程. 以此为基础,采用过程划分的研究方式,对运动目标检测、动作特征提取和动作特征理解 3 个过程进行研究.

人体动作行为识别由计算机在初始数据中检测出动作信息,即运动目标检测过程. 运动目标检测的目的是从静态图像中分割出人体前景,从包含动作信息的视频中分割出动作序列,从而得到容量更小但包含足够运动信息的数据,并以数学符号的形式表达出人体动作. 目前图像分割技术已较成熟,而且人体动作行为识别越来越注重识别的实时性,简单的静态图像形式的动作行为识别往往出现在基础实验性质的研究中,因此动作序列的分割是未来运动目标检测的研究方向. 但现有动作分割方法往往实现较复杂且分割效果不理想,许多研究通常省略运动目标检测这一步骤,直接利用只包含连续人体动作序列的视频作为动作特征提取对象,然后利用人体动作数据库进行方法的评估验证. 本文第二部分分析运动目标检测的研究现状,同时对一些流行的人体动作数据库进行介绍.

动作特征提取是为了进一步选取部分底层信息实现对人体动作的表征. 底层信息可以是经过运动目标检测得到的包含人体动作信息的数学符号形式的图像或视频,也可以是省略目标检测步骤而直接经过数学形式转换的动作序列. 动作特征提取的效果对人体动作行为识别有重要影响,本文第 3 节介绍了相关国内外的研究.

最后,在动作特征提取的基础上,在空间或时空领域完成动作特征理解,以通过数据的分析实现动作的分类. 动作特征理解可看成一个结合先验知识对数学符号进行训练和分类的过程,本文的第 4 节对其研究中的相关技术进行介绍.

2 运动目标检测

人体动作识别首先需要在初始数据中获得包含动作信息的感兴趣区域,这一预处理过程在人体动作识别中称为运动目标检测. 本节从两方面讨论运动目标检测技术: 动作图像的图像分割和动作视频的动作分割.

2.1 图像分割

图像分割是以预定义的一系列标准利用连续性和相似性上的差异将图像分割成若干子部分的过程; 当图像中预期的感兴趣区域已被分割出来后,分割过程应自行停止. 在人体动作行为识别过程中可通过图像分割得到动作图像中的人体区域,简化后续工作.

国内外普遍将图像分割方法分为基于区域的方法,基于边界的方法和数学形态学方法. 其中,基于区域的方法又分为阈值分割法,区域生长法和区域分裂合并法.

阈值分割法将像素点灰度值与定义的阈值进行比较而区别前景和背景,该方法直接,快速; 区域生长法以一个像素点为种子,通过与其邻域像素点的比较逐步获得新种子,进而得到种子合集的区域,该方法计算简单,对于较均匀的连通目标有较好的分

割效果,对噪声敏感; 区域分裂合并法将图像分割成一些区域,根据相似性检验标准将满足条件的合并,对复杂图像的分割效果较好,但算法较复杂,计算量大,还可能破坏区域的边界; 基于边界法检测灰度级或结构中不连续的地方得到边缘从而将图像分割,该方法对噪声敏感,只适合于噪声较小不太复杂的图像; 数学形态学法用具有一定形态的结构元素去量度和提取图像中的对应形状,该方法简单,易于实现,但对噪声敏感,适用于噪声较小的图像.

在人体动作识别研究对人体区域进行分割时,常见的方法可分为两种: 1) 直接利用或结合图像分割方法对人体区域进行分割,如 Tseng 等[13] 首先利用人脸的颜色特征检测并定位人脸,然后利用区域生长法对整个人体进行分割; 文献[14] 先对人体图像进行小波分析,进而得出分割边界,然后通过数学形态方法获得较精确的人体边界,最后通过梯度矢量流主动轮廓模型(GVF Snake) 获得最终的人体边界. 2) 先构建人体图像模型,然后在模型基础上结合图像分割方法通过人体区域的匹配完成人体图像分割,如 Zhao 和 Nevatia[15] 在贝叶斯框架下利用三维人体模型对图像中的人体区域进行分割,该三维人体模型根据头部、躯干和双腿的表征椭圆构成,通过求解对应的最优化参数即可得到人体轮廓边界; Gulshan 等[16] 利用 Kinect 和 OpenNI 库构建一个大容量的逐像素的人体数据库,根据此数据库利用 GrabCut 算法实现人体区域的分割; 文献[17] 则利用训练好的剪影图像作为模版,通过待分割图像中剪影的迭代匹配完成人体剪影的分割.

由于人体动作的空间复杂性,直接利用图像分割方法对人体区域进行分割的效果往往不如在人体图像模型基础上完成分割的效果好,但人体图像模型的构建复杂,实现的难度也较大,构建大容量的人体模型是人体区域分割的研究方向之一. 值得注意的是,图像分割方法效果的评估依赖于人的主观判断,分割效果的自评估方法越来越受到重视[18].

2.2　动作分割

人体行为是一些基本动作(走、跳、坐、跑等) 和相应的过渡动作在空间和时间上的组合. 动作分割的目的是将这些基本动作分离出来从而得到容量更小但包含足够运动信息的数据,并以数学符号的形式表达出人体动作. 总体来说,动作分割实现的难度较大,很多人体动作识别研究中往往直接利用人工分割好的动作序列进行动作的理解分类. 动作分割方法可分为局部分割和全局分割.

2.2.1　局部分割

局部分割法选取部分动作序列作为分析目标或提取局部特征以实现整个行为的分割. 边界检测法是一种常用而经典的局部动作分割方法,它认为人体行为边界是由加速度,速度或人体运动曲率的不连续性造成的,通过对局部的不连续性进行分析得出边界,就可实现动作分割. Ali 和 Aggarwal[19] 通过一个包含人体主要部位 3 个角度的特征向量来确定动作边界. Hanjalic 等[20] 利用逻辑故事单元(每个单元由一个或几个时序无关的事件表示) 来检测动作边界. Wang 和 Xiao[21] 利用一种连续线性动态系统根据光流信息对动作进行分割. Weinland 等[22] 利用多个相机中记录的一些动作示例进行无监督模式学习,并基于一种动作描述因子进行动作边界的划分. 此外,文献[23] ~ [25] 都是利用边界检测完成动作的分割. 局部分割法还可采用一个滑窗实现动作分割,滑窗与可能存在动作信息的区域逐个进行比较,在动作发生区域产生一个峰值. 如 Kim 等[26] 提出的一种基于滑窗利用阈值和权重的视频分割算法.

边界检测法简单实用,但不适用于包含多个动作的序列和运动场景的重构. 滑窗法比边界检测法需要更多的计算量,滑窗法不适合在训练阶段使用,对未知动作的估计也会有更大的不确定性,但它不需要假设动作边界,也可以更容易地与动作分类器融合.

2.2.2　全局分割

全局分割法将整个动作序列作为一个统计模型来处理. 目前应用较多的模型为隐马尔可夫模型(HMM). Zhai 和 Shah[27] 在各种视频中用马尔可夫链蒙特卡洛技术实现时序场景的分割. Niu 和 Abdel-Mottaleb[28] 提出基于 HMM 的动作分割方法,该方法用阈值自动有效地实现复杂动作分割. Shi 等[29] 则使用半马尔可夫模型实现动作分割.

全局动作分割法的效果相比于局部分割法更好,既不要对动作边界进行假设,计算量也不像滑窗法那样巨大,但它需要更多的训练数据,对动作数据库的要求更高. 不论是全局分割还是局部分割,由于人体动作的时间差异性和空间复杂性,人体动作分割的研究仍没有取得一个较为理想的效果,也仍是未来研究中的难点和热点.

2.3　人体动作数据库

为简化人体动作行为识别的预处理过程,便于直接对人体动作进行特征提取和理解,国内外研究中常采用一些人体动作数据库作为初始数据的采集样本,从而避免了运动目标检测过程所带来的精度

和效率问题. 此外,人体动作数据库也是识别方法在统一标准下各种不同性能的重要校验标准. 本节对目前流行的人体动作数据库进行介绍.

KTH 数据库由瑞典皇家理工学院的 Schldt 等[30] 在人体动作识别研究中提出. KTH 数据库包括 4 个场景下由 25 个不同的人执行的 6 类行为视频(慢走、小跑、跑、拳击、挥手、拍手),场景中背景相对静止,摄像机位置也相对固定,只有焦距的变化.

Weizmann 数据库是 Gorelick 等[31] 构建的一共包括 90 段视频的人体动作数据库,这些视频分别是由 9 个人执行 10 个不同的动作(弯腰、抬高、跳高、跳跃、跑、站在一边、小跳、走、挥手 1、挥手 2) 形成. 场景中背景、视角及摄像头都是静止的. 此外,该数据库提供标注好的前景轮廓视频的同时,也包含两个单独的视频序列,一个是不同视角下人体行走的视频,另一个为衣着和人物等方面有细微不同的行走动作序列.

UCF 体育动作数据库[32] 包含 13 个动作(跳水、高尔夫挥杆、击球、举重、骑马、赛跑、滑冰、击打棒球等) 的 150 段体育视频. 该数据库大部分运动类别在动作类型、人物表情、相机运动、视角、光照和背景灯方面有较大差异性.

INRIA XMAS 数据库是 Weinland 等[33] 提出,进而被广泛使用的用于多视角情形下的动作数据库. 它包含 14 种行为(看手表、绕胳膊、抓头、坐下、起立、转身、走、挥手、拳击、敲、指、捡、过头扔和从下方扔) 在 5 个视角获得. 室内方向和头顶一共安装 5 个摄像头,光照和背景基本不变.

KTH 数据库和 Weizmann 数据库约束了视角、人物等条件,是较为理想条件下构建的人体动作数据库. 针对此两种数据库,目前国内外研究的方法已达到较高的识别率. UCF 体育动作数据库约束条件相对较少,对人体动作识别研究的广泛性和实用性要求较高,是今后研究中将被更多使用的数据库,类似的数据库还有 UCF YouTube、UCF50、PASCAL 和 HMDB51 等. INRIA XMAS 数据库则是多视角这一研究热点的重要校验基石,类似的数据库还有 i3DPost Multi-view 和 MuHAVi 等.

3 动作特征提取

动作特征提取的目的是从动作底层数据中抽取部分特征信息对人体动作进行表征. 剪影、光流、梯度、时空特征和深度特征是目前人体动作识别中被普遍采用的表观特征.

3.1 剪影特征

由于剪影受颜色和纹理等不相关特征的影响较小,目前被广泛采用. 如 Ahmad 和 Lee[34] 提出一种称为剪影能量图像(Silhouette Energy Image, SEI) 的剪影时空表征方法,其中 SEI 由一系列图像的剪影构成,作为人体动作识别中运动和形状的表征. Li 等[35] 的研究认为运动物体的三维形状是在剪影重构形状(Shape-from-silhouette) 模型基础上产生的结果,基于此模型对人体动作进行表征. Cheung 等[36] 对 shape-from-silhouette 模型进行进一步研究,并将其扩展到多个动态连接物体的表征中.

剪影提取的效果对于基于剪影特征的人体动作识别方法有着关键影响,通常剪影可通过多种方式来表达. Agarwal 和 Triggs[37] 利用边界上规则分布的约 500 个点来表示剪影,然后用形状上下文将剪影编码为一个 60 维的特征分布. Kolev 等[38] 通过计算生成颜色信息最大概率的三维形状,构建一个关节剪影的概率方程. Howe[39] 提出一种用于单目姿势跟踪的剪影提取算法.

尽管剪影提取已取得较好效果,而且基于剪影的人体动作识别有诸多方面的优势,但人体自遮挡问题一直没有得到较好的解决.

3.2 光流特征

光流是运动识别中另一种重要的运动特征,是运动物体在观测成像面上的像素运动的瞬时速度,它建立于图像的变化仅仅来源于移动的假设基础之上. 光流的计算最初由 Horn 和 Schunck 于 1981 年提出,将二维速度场与灰度相联系. 光流计算的基本假设是图像模式中的点 (x,y) 在 t 时刻的灰度值为 $I(x,y,t)$,在较短的时间间隔内该值保持不变,即

$$\frac{\mathrm{d}I}{\mathrm{d}t}=0,$$

从而得到光流约束方程:

$$\frac{\partial I}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t}+\frac{\partial I}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}+\frac{\partial I}{\partial t}=0.$$

然后通过假设光流在整个图像上的变化平滑这一约束条件完成光流约束方程的求解. Chris 等[40] 提出图像运动的模糊神经系统分析方法,并将其应用到光流分析中; Marco[41] 利用遗传算法对光流进行计算. 此外,光流计算还可使用匹配法[42]、正则化方法[43] 和基于小波的方法[44] 等.

光流技术在人体动作识别领域有着广泛应用. Efros 等[45] 利用光流信息实现一定距离上人体动作的识别. Zhang[46] 根据光流场的二值图像构建目标模板,通过目标匹配的方法实现物体追

踪. Mahbub [47] 用光流场中特征点水平和垂直方向的平均差和标准差计算实现对运动及方向的检测. 基于光流技术的人体动作识别容易受光照和遮挡的影响,为克服这一缺陷,近年来开展许多相关研究. Denman 等 [48] 在进行人体追踪时对每个像素点的光流进行计算以得到更好的灵活性; Kinoshita 等 [49] 的人体追踪方法中使用的“一维光流”包含几个方向上的光流信息,也较好地实现了复杂背景下的人体跟踪. Ali 和 Shah [50] 在人体动作识别过程中则将光流信息转换为可更清晰地表征动作的运动特征,通过不同运动特征参数表示光流信息的各个方面和层面.

3.3　时空特征

时空特征主要指可以在动作场景中根据人体动作的特殊性对动作进行描述的算子. 近年来,其广泛应用于动作识别领域,主要包括时空立方体、时空兴趣点和时空上下文等.

时空兴趣点是一种容易提取和普遍采用的特征,它将人体动作信息以一些不关联的点的形式进行描述. 基于兴趣点的方法只需对提取的兴趣点进行分析就可实现人体动作的分析. 兴趣点检测常用的算法有 Harris 算法,Susan 算法和 SIFT 算法等. Yuan 等 [51] 利用三维 Harris 检测器对动作视频中特征点进行提取,在此基础上对动作进行时空表征,进而实现动作识别. Salmane 等 [52] 在光流中运用 Harris 角点检测算法得到更精确的光流特征. Zhang 和 Liu [53] 用量化的局部 SIFT 描述因子实现人体的表征. Sun 等则在综合二维、三维 SIFT 特征和整体特征的基础上构建统一的人体动作识别框架 [54] .

Biederman 等 [55] 对上下文特征进行定义,在此基础上,用于动作识别的上下文特征可分为: 场景上下文(“概率”) 、空间上下文(“位置”) 和尺度上下文(“尺寸”) 3 类. 其思想为构建动作场景中动作与人所处环境的相互关系,根据动作在场景中的关系信息(上下文特征) 综合地对动作进行识别,如文献[56] 利用“基于图像”,”基于人体”和”基于动作”的上下文特征构建动作识别的框架. 时空上下文是动作识别中另一种应用较为广泛的时空特征,文献[57] 利用马科夫逻辑网(Markov Logic Network, MLN) 对家庭环境中的上下文特征进行研究,并将其应用于人体动作识别中; Wu 等 [58] 则对厨房中动作与环境的上下文特征进行研究.

时空立方体特征是一种将提取到的时空兴趣点进一步映射到一个立方体上进行表征的一种技术. Hae 和 Milanfar [59-60] 用时空局部回归核(Space-Time Local Regression Kernels, 3D LSKs) 对图像中人体动作进行表征,通过时空立方体的匹配完成动作识别. 文献[61] 也通过坐标变换利用一个 LED 时空立方体将 Kinect 构建的三维人体关节模型进行表征.

基于兴趣点的动作识别因角点检测技术的发展而较易实现,但人体动作的复杂性使得该技术在实际应用中有较大难度; 基于上下文的动作识别效果取决于动作与动作场景关系的构建,避免了人体动作自身所带来的歧义; 基于时空立方体的动作识别依赖于其它特征的提取效果,但对三维人体识别这一研究热点有重要意义.

3.4　深度特征

深度信息能描述动作识别过程中人体的三维位置信息,为人体三维姿态识别提供方便,是目前研究热点. 深度信息可通过几个二维相机的组合或一个三维相机获取. 相机布局的误差和变化对多相机动作捕捉系统(Motion Capture, MoCap) 有较大影响,而三维相机价格昂贵且精度往往不能达到高要求. Khan 等 [62] 用多个未标定的二维相机获得距离信息,继而实现了人体跟踪. Khan 和 Shah [63] 发明了一种三维相机,它既不需要一套昂贵的设备,也可与照明设备同时受控. 2010 年 6,微软公司发布一款深度传感器——Kinect. Kinect 的软件开发工具包(Software Development Kit, SDK) 中提供人体三维姿态追踪的示例,在 1.2 ~ 3.5m 范围内实现较好的追踪效果. 文献[64] ~ [66] 等利用 Kinect 进行动作识别方面的研究.

近年来为了在人体动作识别中获得鲁棒性更强的特征表征,许多研究采用多特征融合的方式对特征进行描述. 融合的多特征既可更准确地表征运动,也可减少信息冗余,在精度和效率上有较大优势. 基于多特征融合的人体动作识别方法主要区别是选取的特征和融合策略的不同. 特征可任意选取颜色、光流、特征点和深度信息等进行组合,通过使用主成成分分析(Principal Component Analysis, PCA) [67] 或 Fisher 线性判别 (Fisher Linear Discriminant, FLD) [68] 等方法提取特征向量,然后根据融合策略进行融合. 融合策略可将所有的特征向量集合在一起,然后根据动态规划法 [69] ,神经网络法 [70] 或支持向量机 [71] 实现,也可使用一个特征向量与其他特征向量进行对比,根据之间的联系进行融合 [72-73] . 多特征融合技术在人体动作识别领域得到广泛应用,如 Paul 等 [74] 将颜色、纹理和剪影特征进行概率融合,实现人体动作多特征的表征. Serby 等 [75] 使用特征点、剪影、纹理和灰度特征的融合特征,Zhou 和

Aggarwal [76] 将颜色、剪影和深度信息进行融合. 文献[34]~[36]、[45]~[76]陈述了相关动作特征.

在人体动作识别过程中,选取单特征往往有其不可回避的弊端,如剪影特征的自遮挡问题及深度特征的复杂性问题等,因此多特征是未来的研究方向之一. 目前多特征的融合策略通常为几个特征向量的简单组合,因此多特征更巧妙的融合设计也是未来人体动作识别过程中的研究重点.

4 动作特征理解

在人体动作特征提取的基础上,动作特征理解可看作一个在空间或时空领域将提取到的人体运动特征与先验知识进行对比,通过数据的分析实现动作分类的过程. 人体动作识别的整个流程如图 1 所示. 由于人体动作识别的空间复杂性和时间差异性,动作特征理解在实现上有较大难度,通常有场景、动作类型、人物特性等方面的限制. 本文根据人体动作识别过程中特征理解时序上的差别,将特征理解中用到模型划分为人体模型和统计模型.

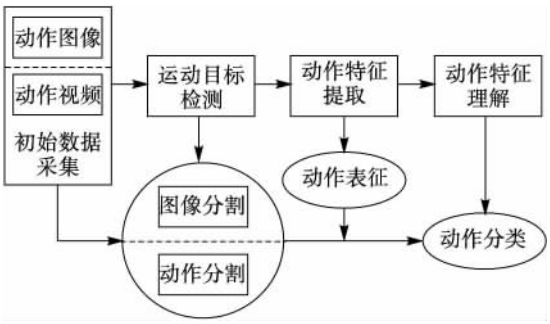


图 1 典型人体动作识别流程及分类图

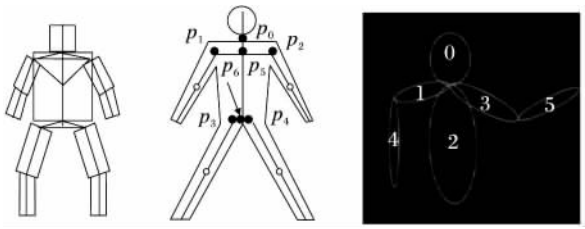
Fig. 1 A typical flow chart of action recognition

4.1 基于人体模型的特征理解

对于分类问题,最直观的方法就是将获得的动作表征与已有的动作模型直接进行比较,基于静态模型的理解方法基本思想是将图像序列转换为一组静态形状模式,然后在识别过程中和预先存储的运动样本相比较,根据相似度判别类型.

由于利用一个统一的人体模型表征任意个体,因此在人体特征提取效果较差时动作特征的理解精度较高,对个体差异有较强鲁棒性,但模型的构建复杂. 人体模型有二维模型和三维模型两种形式,二维模型通常通过二维形状表格来表征人体各部分,三维模型则利用与人体骨架模型对应的圆柱或圆锥几何模型.

二维模型相对简单,一般从图像中直接提取底层表观特征,据其估计人体二维模型参数. 通常将人体这种多关节体的投影划分成一系列在关节处相连接的子区域,每个子区域包含控制旋转平移和缩放的参数,通过调整参数对这些区域进行变形,使它们与二维模型相匹配,从而分割出人脸、躯干、四肢等不同区域. 二维模型在形状上有柱状(图 2(a))、椭圆形(图 2(c))和不规则形状等,文献[77]中人体动作行为监控系统使用的人体二维模型 BB6-HM 和文献[78]中人体跟踪系统构建的“外形连接学习模型”是不规则块状; Rohr [79] 使用的人体二维模型为七点人体模型(图 2(b)),该模型用 5 个 U 形形状带和一个人体躯干柱表达人体轮廓,用 7 个关键点来描述人体头部等部位的位置,并包含 4 个四肢处的关节点. 七点模型是一个柔性模型,可通过各组成部分和关节点的移动变形表达人体各种运动,但自遮挡问题会使该模型在估计人体轮廓时出现较大误差. Leungh 和 Yang [80] 在“First Sight”系统中改进了七点人体二维模型,这种人体模型通过图像估计出人体部分形状,通过这些形状来确定人体大致轮廓并确定部分关节点. 改进的七点模型降低了自遮挡问题带来的不确定性. 此外 Vinay [81] 提出一种圆块人体二维模型,该模型对各圆块的合并和分割鲁棒性较强,适用于人体追踪. Xu 等 [82] 通过曲率分析和椭圆拟合用多个连接的椭圆对人体上肢进行建模.



(a) 柱状模型 (b) 带状模型 (c) 椭圆形模型

(a) Block model (b) Ribbon model (c) Ellipse model

图 2 典型人体二维模型

Fig. 2 Some 2D human body models

人体二维模型在表达人体运动时不包含人体距离信息 [83] ,在自遮挡情况下表征人体动作时有很大的不确定性. 人体三维模型可充分利用人体运动学及人体三维形体属性等方面的先验知识,结合深度信息完成三维人体姿态数据的计算. 典型的人体三维模型是人体骨架模型. 三维人体骨架模型是通过

多个人体数据平均出来一个通用模型[84]，该模型定义了人体各个关节的连接关系、骨骼的长度及各个关节运动量的范围. 在定义人体骨架模型的基础上，将姿态定义为控制关节角度或位置的一组参数，而人的每个动作序列是一组姿态的集合. 此外锥台、椭圆柱、圆柱体等三维模型也是较为常见的人体三维模型组成部分.

Tong 等[85]首先在单目摄像头采集的人体图像 32 个连接骨骼处进行卷积曲面计算，构建一个可以随多项式、半径等参数的变化产生变形的人体骨骼模型; 然后通过分析卷积产生的曲面与图像中曲线的对应关系得到二维轮廓和三维姿势间的映射关系，实现人体二维动作图像在三维空间中的表达. Dekker[86]利用三维扫描仪获得人体距离信息样本，然后通过分析人体几何学与模型参数的关系构建人体三维模型，最后利用插值完成人体各部位形状和比例的重建. 此外，Andre 等[87]提出一种新的人体三维模型系统，该系统包含连接人体各个部位的动力学连接关系，表面皮肤网格和纹理图像. Kakadiaris 等[88]介绍一种人体部位鉴别策略(Human Body Part Identification Strategy, HBPIS)，首先单独构建一个人体站立的三维模型，然后通过调整不同部位的参数使其在不同视角变化中逐步优化该模型. Tatsuya 等[89]利用图像序列因式分解的方法得到三维运动参数来构建人体模型，但这种方法依赖于图像深度信息的连续性假设，不适用于户外等复杂场景下的情况.

三维人体模型虽然可以判断类似遮挡和碰撞问题，克服二维模型对人体遮挡和碰撞处理的不足，但深度信息的获得需要多视角图像或 3D 摄像机，模型的构建也更加复杂.

4.2　基于统计模型的特征理解

人体动作理解过程中基于统计模型的方法在一个时域中将人体序列的排列组合情况及各个时间点的动作进行相关的训练和分类. 模板匹配，动态规划法，动态时空规整法，状态空间法和状态统计法都是基于统计模型的特征理解方法.

4.2.1　模板匹配

动作识别中最直接简单的方法是模板匹配法，该方法事先对每一动作建立起特征数据样本模板，识别时只需按时间顺序将获取的待测动作特征数据与样本模板进行匹配，通过计算两者之间的相似度来判断是否属于样本动作. 特征理解中用于动作序列识别的典型模板有主动形状模型(Active Shape Models, ASM)、主动外观模型(Active Appearance Models, AAM)、运动历史图像(Motion History Images, MHI)、运动能量图像(Motion Energy Images, MEI)等.

AAM 既是一种物体形状描述技术[90]，也是一种用于解决图像中的目标搜寻的形状统计模型，而 ASM[91]的基本思想是选取一组训练样本，用一组特征点来描述样本的形状，然后对各样本的形状进行配准，对这些配准后的形状向量利用主分量分析方法进行统计建模得到物体形状的统计学描述. AAM 和 ASM 在人脸识别中被广泛应用，Kokkinos[92]和 Ma[93]分别将它们应用到人体运动识别中.

人体动作识别中更多地采用时序模板——MHI[94]和 MEI[95]来实现对特征的匹配. MEI 反映了人体动作所发生的区域及强度，MHI 则在一定程度上反映人体动作发生的时间及随时间变化情况. 由于 MHI 和 MEI 在实践中的鲁棒性等特点，它们在动作识别的研究中被广泛采用[96-98].

模板的建立直接影响着模板匹配法的匹配效果，需要一个稳定的大容量动作数据库作为基础，其计算量低，时间跨度的变化对其影响较大.

4.2.2　动态规划

动态规划算法的基本思想是多阶段最优化. 在匹配过程中，样本模板和待测模板无需考虑时间的对应关系，待测模板中每个时刻特征可与样本模板任意时刻特征进行匹配，搜索两个模板的最优匹配路径. 由于动态规划独立于时间序列和动作复杂性，近些年在动作识别领域逐渐得到应用. Hoai 等[99]利用动态规划方法对测试视频中的动作序列及训练库中的序列在表征模型的基础上进行相似度判断，以实现动作分割. Kang 等[100]利用动态规划方法提取物体中一些特征直线的模式，根据对直线的分组和匹配实现物体识别. Gherabi 等[101]使用动态规划方法进行形状匹配和模式识别.

由于动态规划算法不能提取典型的样本模板，每个动作都需建立多个样板模板，然后将待测模板与其逐个进行匹配，计算最小距离. 因此，计算量会随着训练样本数目的增加而增大，且易受噪声的影响，这是动态规划算法主要缺点.

4.2.3　动态时空规整

动态时空规整是在动态规划基础上发展起来的一种较好的非线性时间规整方法. 它将两个不同时间长度的动作特征模板，按照一定的时间规整曲线进行时间调整对应，使得两模板时间长度达到一致，然后再按时间顺序将待测模板的特征与样板模板中相邻的多个特征进行动态规划. Alajlan 等[102]利用动态时空规整算法搜索形状匹配中点与点之间的最

小代价值，快速完成物体形状匹配. Wang 和 Zheng[103] 进一步将动态时空规整应用到人体动作识别中.

动态时空规整的算法较好地解决人体运动在时间上的不确定性，具有概率简单，算法鲁棒的优点，即使测试序列模式与参考序列模式的时间尺度不能完全一致，只要时间次序约束存在，它仍能较好地完成测试序列和参考序列之间的模式匹配. 但该方法训练样本有限，随着样本数的增加效率会变得较低，因此该类方法只能处理部分动作的理解.

4.2.4 状态空间法

该方法将每个静态姿势定义为一个状态，每个状态间通过相互间的概率关系连接. 状态和状态之间的切换采用概率来描述，一个运动序列可看成一次这些状态或状态集合的遍历过程. 典型的状态空间模型有隐马尔可夫模型(Hidden Markov Models，HMM)，条件随机场(Conditional Random Field，CRFs)和动态贝叶斯网络(Dynamic Bayesian Network，DBN).

隐马尔可夫模型把一个总随机过程看成一系列状态的不断转移. HMM 认为模型的状态不能直接观测，能观测的是模型的观测量. 它可以看成是一个双重随机过程：1) 用于描述状态之间转移关系的基本的随机过程，采用 Markov 链形式；2) 用于描述状态和观测变量之间的统计对应关系的随机过程. 由于 HMM 模型能对任意分布建模，且对动态时间序列具有出色的处理能力，已成为人体运动识别中应用最广的模型. Yang 等[104] 对利用 HMM 将人体动作表征作为一个参数模型进行讨论. Yamato[105] 和 Hongeng[106] 研究基于 HMM 的人体动作识别方法. McCowan 等[107] 基于 HMM 在会议中实现一组动作的建模.

隐马尔可夫模型在处理问题时必须要给出一个严格的独立性假设. 但事实上，人体动作识别问题并不能满足这个假设；此外，使用隐马尔可夫模型只能使用有限的上下文特征，否则会带来数据稀疏等问题，导致识别精度的下降. 在文献[108]中，Mccallum 等提出一种基于最大熵原理的条件随机模型——最大熵马尔可夫模型(Maximam Entropy Markov Models，MEMM)以实现信息提取和分割，避免独立性假设.

Lafferty 等[109] 提出的条件随机场是一种综合最大熵模型和 HMM 优点的概率图模型. 同 HMM 类似，早期 CRFs 在语音识别和文字识别领域取得了巨大成功，近年来逐渐被应用到动作识别领域[110]. 如 Galleguillos 等[111] 用 CRFs 最大化人体特征标签，以实现分类. Taycher 等[112] 提出一种基于 CRFs 的人体跟踪方法. Natarajan 和 Nevatia[113] 用 CRFs 来表征多视角的合成动作. Wang 和 Ji[114] 提出一种融合上下文约束的动态条件随机场(Dynamic Conditional Random Field，DCRF)模型用来在图像序列中进行目标分割.

动态贝叶斯网络是模式识别领域中一种可同时获得变量间概率关系和时变规律的统计模型，它用一个变量的集合而不是一个变量来表示隐含状态，克服了 HMM 表达能力有限的缺陷. 近年来动态贝叶斯网络在人体动作识别领域受到越来越多的重视. 如 Pavlović 等[115] 提出一种基于 DBN 的开关线性动态系统(Switching Linear Dynamic System，SLDS)模型并将其应用到人体动作识别中. Du 等[116] 提出一种基于 DBN 的交互动作识别方法. 也可通过一个线性动态系统和 HMM 的融合建立一个混合状态 DBN 框架，对时变数据进行建模和分类[117].

状态空间法的高度模块化使其相比于基于动作模型的特征理解方法更适于复杂的特征理解，是目前的研究热点.

4.2.5 状态统计法

不同于状态空间法依赖于状态间的概率关系，状态统计法在全局状态下处理动作事件. 典型的技术有词袋模型、语法模型和支持向量机.

基于词袋模型的动作识别方法将动作表示为一些局部特征块，每个动作像句子一样表示为不同比例的单词，所有单词的集合组成一个词典. 通过统计单词出现的概率完成动作的理解过程. 近年来基于词袋模型的方法以其计算简单，对噪声、光照和局部遮挡良好的鲁棒性等特点得到了广泛关注[118].

Demirdjian 和 Wang[119] 定义一个时序特征词袋，通过将词袋特征作为特征向量用一个多级分类器实现动作中的识别. 类似地，Liu 和 Shah[120] 通过对视频中提取的三维特征点的量化构建时空特性的视频词袋. Marszalek 等[121] 则在词袋模型的框架上构建动作的视觉模型. Lazebnik 等[122] 首先将图像划分成一些部分，然后在传统词袋模型的基础上计算这些部分的特征直方图实现场景分类.

不同于词袋模型，语法模型将动作表征为一连串符号的同时，定义一系列的规则来连接这些动作符号，通过每个动作语法的不同实现动作特征的理解. 如 Lin 等[123] 提出用于视频监控的概率属性图语法模型，此模型会将场景中的一个事件理解为一些动作事件的集合，这些动作事件进而被划分为一系

列原子动作,从而实现时空中事件的语法表征,进而实现特征理解. 文献 [124] 同样将人体动作理解为一种语法结构,并在此基础上构建一套语法系统实现对人体动作的理解. 文献 [125] 和文献 [126] 也是基于语法模型完成人体动作的特征理解.

支持向量机(SVM)是适用于模式识别分类的有效方法,它是针对线性可分情况进行分析,对于线性不可分的情况,通过使用非线性映射算法将低维输入空间线性不可分的样本向高维特征空间转化使其线性可分,从而使得在高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能; 它基于结构风险最小化理论在特征空间中建构最优分割超平面,使得学习器得到全局最优化,并且在整个样本空间的期望风险以某个概率满足一定上界. Pontil 和 Verri [127] 使用 SVM 进行三维目标识别,该方法不进行姿态估计,在高维空间对图像上点进行处理. Schuldt 等 [128] 则通过 SVM 结合时空特征点的局部表征来实现动作识别. Wang 等 [129] 则在其研究中利用 SVM 对词袋模型中的重要特征进一步提取,并对图像进行分类. Simon 等 [130] 提出一种基于分割的支持向量机(kSeg-SVM)的方法,这种方法是空间词袋模型在时间上的一种延伸,训练则通过 SVM 框架完成.

在人体动作识别的研究中,针对不同的特征提取对象,特征理解方法的效果也不尽相同,而且特征理解方法的使用往往有一定的条件约束. 总体来说,基于人体模型的特征理解方法在一定范围内精确性较高,但模型的构建复杂,容量往往很难满足需求. 基于统计模型的方法适用性较好,但训练样本的选取和容量以及所带来的复杂度问题也是未来研究中的难点.

5　结 束 语

本文分析了基于视觉的人体动作识别研究现状,从运动目标检测、动作特征提取和动作特征理解三方面对相关技术进行分析. 通过以上工作,可看出人体动作行为识别的研究取得卓越的成果和进展. 但同时需要注意的是很多方法仍处于实验室研究阶段,现有系统在鲁棒性和实时性等方面距实际应用还有较大差距. 基于视觉的人体动作行为识别研究中的热点和方向主要内容如下.

1) 动作场景中噪声的规避. 如何降低人体运动场景中的光照、复杂背景、遮挡物等噪声对识别效果的影响一直是人体动作行为识别研究中的热点. 选取深度信息作为运动特征可以规避场景中的噪声,近年来在识别中被广泛采用并取得较好的识别效果,但其本身也有精度等方面的劣势. 一般情况下,研究中可通过限制光照条件等措施来约束对结果影响较大的噪声; 也可利用滤波等预处理方法对运动原始数据进行处理; 更自然有效的方法是采用多特征融合对运动进行表征,通过综合多特征的优势实现对单一噪声的规避.

2) 识别对象的多元化. 人体在衣着、身高、体态等方面存在诸多差异,同时识别对象的多元化也体现在不同且未知的人物数量方面. 尽管本文重点是单一对象的动作识别,但涉及到的文献中很多都有多人动作识别方面的尝试,对于衣着、身高、体态等方面有差异的多人动作识别研究是近年的研究热点. 同时,识别对象本身的自遮挡问题一直是研究中的难点,而且动作姿态的三维识别也越来越成为研究的主要方向.

3) 视角变化的研究. 相机在不同视角下会呈现不同的状态,人物和场景的大小、方向和形状都会发生变化. 然而多视角和多相机的研究在动作识别中往往有不可比拟的优势,因此如何克服视角变化的影响也会是动作识别未来的研究方向. 在多视角人体动作识别中,可通过视角变换对人体动作进行二维建模,通过模型中相同点在不同位置的匹配和分析实现不同视角下人体动作特征的表征(如文献 [131]); 也可通过人体姿态三维重构技术实现人体相关节点在三维空间的映射,利用人体三维模型中的姿态来克服二维空间中视角变化所产生的差异(如文献 [132] ~ [133]).

4) 动作视频自分割. 目前人体动作行为识别研究中用到的动作视频都是经过处理的仅包含特定研究对象的动作序列. 在识别过程中首先完成对包含任意动作信息的分割是人体动作识别能够进入实际应用的重要过程. 动作分割方法的研究近年来逐渐受到人们的重视,一个稳定的大容量动作数据库对于动作的分割及整个识别过程至关重要.

人体动作行为识别在视频监控、虚拟现实、人机智能交互等领域有广泛的应用前景. 本文对近年该领域一些研究进行分析和总结,探讨了其中存在的问题和发展方向. 由于动作识别问题的广泛性,许多重要的技术在本文中并没有完全涉及. 值得注意的是,人脸识别和手势识别与人体动作行为识别的研究有许多相似之处.

人体动作行为识别是一个综合性课题,模式分类的数学方法、图像传感技术及人体运动学研究中

的创新理论在动作识别领域的应用是推动其发展的重要动力.

参 考 文 献

[1] Mokhber A, Achard C, Milgram M. Recognition of Human Behavior by Space-Time Silhouette Characterization. Pattern Recognition Letters, 2008, 29(1): 81-89

[2] Polat E, Yeasin M, Sharma R. Robust Tracking of Human Body Parts for Collaborative Human Computer Interaction. Computer Vision and Image Understanding, 2003, 89(1): 44-69

[3] Kjellström H, Romero J, Kragić D. Visual Object-Action Recognition: Inferring Object Affordances from Human Demonstration. Computer Vision and Image Understanding, 2011, 115(1): 81-90

[4] Suma E A, Krum D M, Lange B, *et al*. Adapting User Interfaces for Gestural Interaction with the Flexible Action and Articulated Skeleton Toolkit. Computers & Graphics, 2012, 37(3): 193-201

[5] Ayers D, Shah M. Monitoring Human Behavior from Video Taken in an Office Environment. Image and Vision Computing, 2001, 19(12): 833-846

[6] López M T, Fernández-Caballero A, Fernández M A, *et al*. Visual Surveillance by Dynamic Visual Attention Method. Pattern Recognition, 2006, 39(11): 2194-2211

[7] Wang Liang, Hu Weiming, Tan Tieniu. A Survey of Visual Analysis of Human Motion. Chinese Journel of Computers, 2002, 25(3): 225-237 (in Chinese)
(王 亮,胡卫明,谭铁牛. 人运动的视觉分析综述. 计算机学报, 2002, 25(3): 225-237)

[8] Aggarwal J K, Park S. Human Motion: Modeling and Recognition of Actions and Interactions // Proc of the 2nd International Symposium on 3D Data Processing, Visualization and Transmission. Thessaloniki, Greece, 2004: 640-647

[9] Moeslund T B, Hilton A, Krüger V. A Survey of Advances in Vision-Based Human Motion Capture and Analysis. Computer Vision and Image Understanding, 2006, 104(2/3): 90-126

[10] Ling Zhigang, Zhao Chunhui, Liang Yan, *et al*. Survey on Vision-Based Human Action Understanding. Application Research of Computers, 2008, 25(9): 2570-2578 (in Chinese)
(凌志刚,赵春晖,梁 彦,等. 基于视觉的人行为理解综述. 计算机应用研究, 2008, 25(9): 2570-2578)

[11] Poppe R. A Survey on Vision-Based Human Action Recognition. Image and Vision Computing, 2010, 28(6): 976-990

[12] Weinland D, Ronfard R, Boyer E. A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition. Computer Vision and Image Understanding, 2011, 115(2): 224-241

[13] Tseng H C, Shyu J J, Chang J Y, *et al*. Exploiting Automatic Image Segmentation to Human Detection and Depth Estimation // Proc of the IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Processing. Paris, France, 2011: 19-25

[14] Cheng Jinyong, Liu Yihui. Human Body Image Segmentation Based on Wavelet Analysis and Active Contour Models // Proc of the International Conference on Wavelet Analysis and Pattern Recognition. Beijing, China, 2007: 265-269

[15] Zhao Tao, Nevatia R. Bayesian Human Segmentation in Crowded Situations // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA, 2003, II: 459-466

[16] Gulshan V, Lempitsky V, Zisserman A. Humanising GrabCut: Learning to Segment Humans Using the Kinect // Proc of the IEEE International Conference on Computer Vision Workshops. Barcelona, Spain, 2011: 1127-1133

[17] Ando H, Fujiyoshi H. Human-Area Segmentation by Selecting Similar Silhouette Images Based on Weak-Classifier Response // Proc of the 20th International Conference on Pattern Recognition. Istanbul, Turkey, 2010: 3444-3447

[18] Raut S, Raghuvanshi M, Dharaskar R, *et al*. Image Segmentation -A State-of-Art Survey for Prediction // Proc of the International Conference on Advanced Computer Control. Singapore, Singapore, 2009: 420-424

[19] Ali A, Aggarwal J K. Segmentation and Recognition of Continuous Human Activity // Proc of the IEEE Workshop on Detection and Recognition of Events in Video. Vancouver, Canada, 2001: 28-35

[20] Hanjalic A, Lagendijk R L, Biemond J. Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems. IEEE Trans on Circuits and Systems for Video Technology, 1999, 9(4): 580-588

[21] Wang Jinjun, Xiao Jing. Human Behavior Segmentation and Recognition Using Continuous Linear Dynamic Systems // Proc of the IEEE Workshop on Application of Computer Vision. Tampa, USA, 2013: 61-67

[22] Weinland D, Ronfard R, Boyer E. Automatic Discovery of Action Taxonomies from Multiple Views // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006, II: 1639- 1645

[23] Davis J W, Bobick A F. The Representation and Recognition of Human Movement Using Temporal Templates // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Minneapolis, USA, 1997: 928-934

[24] Rubin J M, Richards W A. Boundaries of Visual Motion. Technical Report, AIM-835. Cambridge, USA: Massachusetts Institute of Technology, 1985

[25] Rui Yong, Anandan P. Segmenting Visual Actions Based on Spatio-Temporal Motion Patterns // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head Island, USA, 2000, I: 111-118

[26] Kim W H, Jeong T I, Kim J N. Video Segmentation Algorithm Using Threshold and Weighting Based on Moving Sliding Window // Proc of the 11th International Conference on Advanced Communication Technology. Pyeongchang County, Repulic of Korea, 2009: 1781-1784

[27] Zhai Y, Shah M. Video Scene Segmentation Using Markov Chain Monte Carlo. IEEE Trans on Multimedia, 2006, 8(4): 686-697

[28] Niu Feng, Abdel-Mottaleb M. HMM-Based Segmentation and Recognition of Human Activities from Video Sequences // Proc of the International Conference on Multimedia and Expo. Amsterdam,

Holland, 2005: 804-807

[29] Qinfeng Shi, Li Wang, Li Cheng, *et al*. Discriminative Human Action Segmentation and Recognition Using Semi-Markov Model [EB/OL]. [2013-3-17]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4587557

[30] Schuldt C, Laptev I, Caputo B. Recognizing Human Actions: A Local SVM Approach // Proc of the 17th International Conference on Pattern Recognition. Cambridge, USA, 2004, III: 32-36

[31] Blank M, Gorelick L, Shechtman E, *et al*. Actions as Space-Time Shapes // Proc of the 10th IEEE International Conference on Computer Vision. Beijing, China, 2005, II: 1395-1402

[32] Rodriguez M D, Ahmed J, Shah M. Action MACH: A Spatiotemporal Maximum Average Correlation Height Filter for Action Recognition [EB/OL]. [2013-3-17]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4587727

[33] Weinland D, Ronfard R, Boyer E. Free Viewpoint Action Recognition Using Motion History Volumes. Computer Vision and Image Understanding, 2006, 104(2/3): 249-257

[34] Ahmad M, Lee S W. Variable Silhouette Energy Image Representations for Recognizing Human Actions. Image and Vision Computing, 2010, 28(5): 814-824

[35] Li Guan, Franco J S, Pollefeys M. 3D Occlusion Inference from Silhouette Cues [EB/OL]. [2013-3-17]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4270170

[36] Cheung K M G, Baker S, Kanade T. Shape-from-Silhouette of Articulated Objects and Its Use for Human Body Kinematics Estimation and Motion Capture // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Madison, USA, 2003, I: 77-84

[37] Agarwal A, Triggs B. Recovering 3D Human Pose from Monocular Images. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(1): 44-58

[38] Kolev K, Brox T, Cremers D. Fast Joint Estimation of Silhouettes and Dense 3D Geometry from Multiple Images. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34(3): 493-505

[39] Howe N R. Silhouette Lookup for Monocular 3D Pose Tracking. Image and Vision Computing, 2007, 25(3): 331-341

[40] Chris D G, Peter D, Christopher P M, *et al*. A Compact Optical Flow Cell for Nurse in Aqueous Halide Determination. Measurement Science and Technology, 1999, 10(4): N34-N37

[41] Tagliasacchi M. A Genetic Algorithm for Optical Flow Estimation. Image and Vision Computing, 2007, 25(2): 141-147

[42] Sun Changming. Fast Optical Flow Using 3D Shortest Path Techniques. Image and Vision Computing, 2002, 20(13/14): 981-991

[43] Francomano E, Tortorici A, Calderone V. Regularization of Optical Flow with *M*-Band Wavelet Transform. Computers & Mathematics with Applications, 2003, 45(1/2/3): 437-452

[44] Chen Lifen, Liao H M, Lin Jachen. Wavelet-Based Optical Flow Estimation. IEEE Trans on Circuits and Systems for Video Technology, 2002, 12(1): 1-12

[45] Efros A A, Berg A C, Mori G, *et al*. Recognizing Action at A Distance // Procc of the 9th IEEE International Conference on Computer Vision. Nice, France, 2003, II: 726-733

[46] Zhang Haiyan. Multiple Moving Objects Detection and Tracking Based on Optical Flow in Polar-Log Images // Proc of the International Conference on Machine Learning and Cybernetics. Qingdao, China, 2010: 1577-1582

[47] Mahbub U, Imtiaz H, Rahman Ahad M A. An Optical Flow Based Approach for Action Recognition // Proc of the 14th International Conference on Computer and Information Technology. Dhaka, Bangladesh, 2011: 646-651

[48] Denman S, Fookes C, Sridharan S. Improved Simultaneous Computation of Motion Detection and Optical Flow for Object Tracking // Proc of the Conference on Digital Image Computing: Techniques and Applications. Melbourne, Australia, 2009: 175-182

[49] Kinoshita K, Enokidani M, Izumida M, *et al*. Tracking of a Moving Object Using One-Dimensional Optical Flow with a Rotating Observer // Proc of the 9th International Conference on Control, Automation, Robotics and Vision. Singapore, Singapore, 2006: 1-6

[50] Ali S, Shah M. Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(2): 288-303

[51] Yuan Chunfeng, Li Xi, Hu Weiming, *et al*. 3D R Transform on Spatio-Temporal Interest Points for Action Recognition // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 724-730

[52] Salmane H, Ruichek Y, Khoudour L. Object Tracking Using Harris Corner Points Based Optical Flow Propagation and Kalman filter // Proc of the 14th International IEEE Conference on Intelligent Transportation Systems. Washington, USA, 2011: 67-73

[53] Zhang Zhuo, Liu Jia. Recognizing Human Action and Identity Based on Affine-SIFT // Proc of the IEEE Symposium on Electrical & Electronics Engineering. Kuala Lumpur, Malaysia, 2012: 216-219

[54] Sun Xinghua, Chen Mingyu, Hauptmann A. Action Recognition via Local Descriptors and Holistic Features // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 58-65

[55] Biederman I. Perceiving Real-World Scenes. Science, 1992, 177(4043): 77-80

[56] Moore D J, Essa I A, Hayes M H. Exploiting Human Actions and Object Context for Recognition Tasks // Proc of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece, 1999, I: 80-86

[57] Wu Chen, Aghajan H. Using Context with Statistical Relational Models: Object Recognition from Observing User Activity in Home Environment // Proc of the Workshop on Use of Context in Vision Process. Boston, USA, 2009: 22-27

[58] Wu Jianxin, Osuntogun A, Choudhury T, *et al*. A Scalable Approach to Activity Recognition Based on Object Use // Proc of the 11th IEEE International Conference on Computer Vision. Rio De Janeiro, Brazil, 2007: 1-8

[59] Seo H J, Milanfar P. Detection of Human Actions from a Single

Example // Proc of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan, 2009: 1965-1970

[60] Seo H J, Milanfar P. Action Recognition from One Example. IEEE Trans on Pattern Analysia and Machine Intelligence, 2011, 33(5): 867-882

[61] Zheng Xiao, Fu Mengyin, Yang Yi, *et al*. 3D Human Postures Recognition Using Kinect // Proc of the 4th International Conference on Intelligent Human-Machine Systems and Cybernetics. Nanchang, China, 2012: 344-347

[62] Khan S, Javed O, Rasheed Z, *et al*. Human Tracking in Multiple Cameras // Proc of the 8th IEEE International Conference on Computer Vision. Vancouver, Canada, 2001, I: 331-336

[63] Khan S, Shah M. Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapping Fields of View. IEEE Trans on Pattern Analysis and Machine Intelligence, 2003, 25(10): 1355-1360

[64] Alcoverro M, Lopez-Mendez A, Pardas M, *et al*. Connected Operators on 3D Data for Human Body Analysis // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 9-14

[65] Lu Xia, Chen C C, Aggarwal J K. Human Detection Using Depth Information by Kinect // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Colorado Springs, USA, 2011: 15-22

[66] Smisek J, Jancosek M, Pajdla T. 3D with Kinect // Proc of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 1154-1160

[67] de Castro L N, Von Zuben F J. Learning and Optimization Using the Clonal Selection Principle. IEEE Trans on Evolutionary Computation, 2002, 6(3): 239-251

[68] Mika S, Ratsch G, Weston J, *et al*. Fisher Discriminant Analysis with Kernels // Proc of the IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing IX. Madison, USA, 1999: 41-48

[69] Zhang Xinhua. An Information Model and Method of Feature Fusion // Proc of the International Conference on Signal Process. Dalian, China, 1998: 1389-1392

[70] Battiti R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. IEEE Trans on Neural Network, 1994, 5(4): 537-550

[71] Shi Yan, Zhang Tianxu. Feature Analysis: Support Vector Machines Approaches // Proc of the SPIE Conference on Image Extraction, Segmentation and Recognition. Wuhan, China, 2001: 245-251

[72] Liu Chengjun, Wechsler H. A Shape and Texture Based Enhanced Fisher Classifier for Face Recognition. IEEE Trans on Image Processing, 2001, 10(4): 598-608

[73] Yang Jian, Yang Jingyu, Zhang D, *et al*. Feature Fusion: Parallel Strategy vs. Serial Strategy. Pattern Recognition, 2003, 36(6): 1369-1381

[74] Brasnett P, Mihaylova L, Bull D, *et al*. Sequential Monte Carlo Tracking by Fusing Multiple Cues in Video Sequences. Image and Vision Computing, 2007, 25(8): 1217-1227

[75] Serby D, Meier E K, Van Gool L. Probabilistic Object Tracking Using Multiple Features // Proc of the 17th International Conference on Pattern Recognition. Cambridge, UK, 2004, II: 184-187

[76] Zhou Quming, Aggarwal J K. Object Tracking in an Outdoor Environment Using Fusion of Features and Cameras. Image and Vision Computing, 2006, 24(11): 1244-1255

[77] Folgado E, Rincón M, Carmona E J, *et al*. A Block-Based Model for Monitoring of Human Activity. Neurocomputing, 2011, 74(8): 1283-1289

[78] Thome N, Merad D, Miguet S. Learning Articulated Appearance Models for Tracking Humans: A Spectral Graph Matching Approach. Signal Processing: Image Communication, 2008, 23(10): 769-787

[79] Rohr K. Towards Model-Based Recognition of Human Movements in Image Sequences. CVGIP: Image Understanding, 1994, 59(1): 94-115

[80] Leung M K, Yang Y H. First Sight: A Human Body Outline Labeling System. IEEE Trans on Pattern Analysis and Machine Intelligence, 1995, 17(4): 359-377

[81] Sharma V. A Blob Representation for Tracking Robust to Merging and Fragmentation // Proc of the IEEE Workshop on Applications of Computer Vision. Breckenridge, USA, 2012: 161-168

[82] Da Xu R Y, Kemp M. Multiple Curvature Based Approach to Human Upper Body Parts Detection with Connected Ellipse Model Fine-Tuning // Proc of the 16th International Conference on Image Processing. Cairo, Egypt, 2009: 2577-2580

[83] Wren C R, Azarbayejani A, Darrell T, *et al*. Pfinder: Real-Time Tracking of the Human Body. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(7): 780-785

[84] Sato T, Kanbara M, Yokoya N, *et al*. Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-Baseline Stereo Using a Hand-Held Video Camera. International Journal of Computer Vision, 2002, 47(1/2/3): 119-129

[85] Minglei Tong, Yuncai Liu, Huang T S. 3D Human Model and Joint Parameter Estimation from Monocular Image. Pattern Recognition Letters, 2007, 28(7): 797-805

[86] Dekker L D. 3D Human Body Modeling from Range Data. Ph. D Dissertation. London, UK: University of London, 2000

[87] Gagalowicz A, Quah C K. 3D Model-Based Marker-Less Human Motion Tracking in Cluttered Environment // Proc of the IEEE 12th International Conference on Computer Vision Workshops. Kyoto, Japan, 2009: 1042-1049

[88] Kakadiaris I A, Metaxas D. 3D Human Body Model Acquisition from Multiple Views // Proc of the 15th International Conference on Computer Vision. Cambridge, USA, 1995: 618-623

[89] Osawa T, Wu Xiaojun, Wakabayashi K, *et al*. Human Tracking by Particle Filtering Using Full 3D Model of Both Target and Environment // Proc of the 18th International Conference on Pattern Recognition. Hong Kong, China, 2006, II: 25-28

[90] Lanitis A, Taylor C J, Cootes T F. Automatic Interpretation and Coding of Face Images Using Flexible Models. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(7): 743-756

[91] Cootes T F, Edwards G J, Taylor C J. Active Appearance Models. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001,

23(6): 681-685

[92] Kokkinos I, Maragos P. Synergy between Object Recognition and Image Segmentation Using the Expectation-Maximization Algorithm. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(8): 1486-1501

[93] Ma Jia, Ren Fuji. Detect and Track the Dynamic Deformation Human Body with the Active Shape Model Modified by Motion Vectors // Proc of the International Conference on Cloud Computing and Intelligence Systems. Beijing, China, 2011: 587-591

[94] Bobick A F, Wilson A D. A State-Based Approach to the Representation and Recognition of Gesture. IEEE Trans on Pattern Analysis and Machine Intelligence, 1997, 19(12): 1325-1337

[95] Bobick A F, Davis J W. The Recognition of Human Movement Using Temporal Templates. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(3): 257-267

[96] Meng Hongying, Pears N, Bailey C. A Human Action Recognition System for Embedded Computer Vision Application [EB/OL]. [2013-3-17]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4270418

[97] Naiel M A, Abdelwahab M M, El-Saban M. Multi-View Human Action Recognition System Employing 2DPCA // Proc of the IEEE Workshop on Applications of Computer Vision. Kona, USA, 2011: 270-275

[98] Tian Yingli, Cao Liangliang, Liu Zicheng, *et al*. Hierarchical Filtered Motion for Action Recognition in Crowded Videos. IEEE Trans on Systems, Man, and Cybernetics-Part C: Applications and Reviews, 2012, 42(3): 313-323

[99] Hoai M, Lan Zhenzhong, de la Torre F. Joint Segmentation and Classification of Human Actions in Video // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2011: 3265-3272

[100] Kang D J, Ha J E, Kweon I S. Fast Object Recognition Using Dynamic Programming from Combination of Salient Line Groups. Pattern Recognition, 2003, 36(1): 79-90

[101] Gherabi N, Gherabi A, Bahaj M. A New Algorithm for Shape Matching and Pattern Recognition Using Dynamic Programming // Proc of the International Conference on Multimedia Computing and Systems. Ouarzazate, Morocco, 2011: 1-6

[102] Alajlan N, El Rube I, Kamel M S, *et al*. Shape Retrieval Using Triangle-Area Representation and Dynamic Space Warping. Pattern Recognition, 2007, 40(7): 1911-1920

[103] Wang Jing, Zheng Huicheng. View-Robust Action Recognition Based on Temporal Self-similarities and Dynamic Time Warping // Proc of the IEEE Conference on Computer Science and Automation Engineering. Zhangjiajie, China, 2012: 498-502

[104] Yang Jie, Xu Yangsheng, Chen C S. Hidden Markov Model Approach to Skill Learning and Its Application to Telerobotics. IEEE Trans on Robotics and Automation, 1994, 10(5): 621-631

[105] Yamato J, Ohya J, Ishii K. Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Champagne, USA, 1992: 379-385

[106] Hongeng S, Nevada R, Bremond F. Video-Based Event Recognition: Activity Representation and Probabilistic Recognition Methods. Computer Vision and Image Understanding, 2004, 96(2): 129-162

[107] McCowan L, Gatica-Perez D, Bengio S, *et al*. Automatic Analysis of Multimodal Group Actions in Meetings. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(3): 305-317

[108] Mccallum A, Freitag D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation // Proc of the 17th International Conference on Machine Learning. Stanford, USA, 2000: 591-598

[109] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data // Proc of the 18th International Conference on Machine Learning. Montreal, Canada, 2001: 282-289

[110] Sminchisescu C, Kanaujia A, Li Zhiguo, *et al*. Conditional Models for Contextual Human Motion Recognition // Proc of the 10th IEEE International Conference on Computer Vision. Beijing, China, 2005, II: 1808-1815

[111] Galleguillos C, Rabinovich A, Belongie S. Object Categorization Using Co-occurrence, Location and Appearance [EB/OL]. [2013-3-19]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4587799

[112] Taycher L, Demirdjian D, Darrell T, *et al*. Conditional Random People: Tracking Humans with CRFs and Grid Filters // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 222-229

[113] Natarajan P, Nevatia R. View and Scale Invariant Action Recognition Using Multiview Shape-Flow Models [EB/OL]. [2013-3-17]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4587716

[114] Wang Yang, Ji Qiang. A Dynamic Conditional Random Field Model for Object Segmentation in Image Sequences // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005, I: 264-270

[115] Pavlovic V, Rehg J M, Tat-Jen Cham, *et al*. A Dynamic Bayesian Network Approach to Figure Tracking Using Learned Dynamic Models // Proc of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece, 1999, I: 94-101

[116] Youtian Du, Chen Feng, Xu Wenli, *et al*. Recognizing Interaction Activities Using Dynamic Bayesian Network // Proc of the 18th International Conference on Pattern Recognition. Hong Kong, China, 2006, I: 618-621

[117] Pavlovic V, Frey B J, Huang T S. Time-Series Classification Using Mixed-State Dynamic Bayesian Networks // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Fort Collins, USA, 1999: 2609-2615

[118] Niebles J C, Li Feifei. A Hierarchical Model of Shape and Appearance for Human Action Classification [EB/OL]. [2013-3-17]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4270157

[119] Demirdjian D, Wang S. Recognition of Temporal Events Using Multiscale Bags of Features // Proc of the IEEE Workshop on Computational Intelligence for Visual Intelligence. Nashville,

USA, 2009: 8-13

[120] Liu Jingen, Shah M. Learning Human Actions via Information Maximization [EB/OL]. [2013-3-17]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4587723

[121] Marszalek M, Laptev I, Schmid C. Actions in Context // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 2929-2936

[122] Lazebnik S, Schmid C, Ponce J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006, II: 2169-2178

[123] Lin Liang, Gong Haifeng, Li Li, *et al*. Semantic Event Representation and Recognition Using Syntactic Attribute Graph Grammar. Pattern Recognition Letters, 2009, 30(2): 180-186

[124] Summers-Stay D, Teo C L, Yang Yezhou, *et al*. Using a Minimal Action Grammar for Activity Understanding in the Real World // Proc of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura, Portugal, 2012: 4104-4111

[125] Ryoo M S, Aggarwal J K. Recognition of Composite Human Activities through Context-Free Grammar Based Representation // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006: 1709-1718

[126] Fogassi L, Ferrari P F, Gesierich B, *et al*. Parietal Lobe: From Action Organization to Intention Understanding. Science, 2005, 308(5722): 662-667

[127] Pontil M, Verri A. Support Vector Machines for 3D Object Recognition. IEEE Trans on Pattern Analysis and Machine Intelligence, 1998, 20(6): 637-646

[128] Schuldt C, Laptev I, Caputo B. Recognizing Human Actions: A Local SVM Approach // Proc of the 17th International Conference on Pattern Recognition. Cambridge, UK, 2004, III: 32-36

[129] Wang Mengyue, Zhang Changlin, Song Yan. An Improved Multiple Instance Learning Algorithm for Object Extraction // Proc of the Chinese Conference on Pattern Recognition. Chongqing, China, 2010: 1-5

[130] Simon T, Nguyen M H, Cohn J F, *et al*. Action Unit Detection with Segment-Based SVMs // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 2737-2744

[131] Liu Jingen, Shah M, Kuipers B, *et al*. Cross-View Action Recognition via View Knowledge Transfer // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3209-3216

[132] Gong Dian, Medioni G. Dynamic Manifold Warping for View Invariant Action Recognition // Proc of the IEEE Conference on Computer Vision. Barcelona, Spain, 2011: 571-578

[133] Weinland D, Özuysal M, Fua P. Making Action Recognition Robust to Occlusions and Viewpoint Changes // Proc of the 11th European Conference on Computer Vision. Heraklion, Greece, 2010: 635-648