动作识别与行为理解综述

徐光祐 曹媛媛
(清华大学计算机科学与技术系 普适计算教育部重点实验室,北京 100084)

摘 要 随着“以人为中心计算”的兴起和生活中不断涌现的新应用,动作识别和行为理解逐渐成为计算机视觉领域的研究热点。主要从视觉处理的角度分析了动作识别和行为理解的研究现状,从行为的定义、运动特征提取和动作表示以及行为理解的推理方法 3 个方面对目前的工作做了分析和比较,并且指出了目前这些工作面临的难题和今后的研究方向。

关键词 以人为中心计算 动作识别 行为理解
中图法分类号: TP391 **文献标识码**: A **文章编号**: 1006-8961(2009)02-0189-07

Action Recognition and Activity Understanding: A Review

XU Guang-you, CAO Yuan-yuan
(*Key Laboratory of Pervasive Computing, Ministry of Education, Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract As the “Human-centered computing” is getting more popular and novel applications are evolving, action recognition and activity understanding are attracting researchers in the field of computer vision. In this paper, we review the state-of-the-art work on action and activity analysis with focus on three parts: Definition of activity, low-level motion features extraction and action representation, and reasoning method for activity understanding. Furthermore, open problems for future research and potential directions are discussed.

Keywords human-centered computing, action recognition, activity understanding

1 引 言

计算正渗透和影响到人们生活的各个方面,根据传感器数据来识别和理解人的动作和行为就成为未来“以人为中心的计算”中的关键[1]。其中基于视觉的动作识别和行为理解尤为重要。因为在人之间的人际交互过程中,视觉是最重要的信息。可以帮助人们迅速获得一些关键特征和事实,如对方的表情、手势、体态和关注点等,这些视觉线索综合起来反映了对方的态度,潜在意图和情绪等信息。未来人机交互和监控中,机器要感知人的意图很大程度上就需要依靠视觉系统。此外,视觉传感器体积小、被动性和非接触式的特点,使得视觉传感器和视觉信息系统具备了无所不在的前提。近年来,在对计算机视觉提出的层出不穷的新要求中,行为理解是一个具有挑战性的新课题,在诸如智能家居,老年人看护,智能会议室等应用中都起着至关重要的作用。它要解决的问题是根据来自传感器(摄像机)的原始图像(包括图像序列)数据,通过视觉信息的处理和分析,识别人体的动作,并在上下境信息的指导下,理解人体动作的目的、所传递的语义信息。行为理解作为近几年开始兴起的研究,正在逐渐获得越来越多的关注。

基金项目:国家自然科学基金项目(60673189)
收稿日期:2008-11-28;**改回日期**:2008-12-03

第一作者简介:徐光祐(1940~),男。教授,博士生导师。IEEE 高级会员,CCF 会员。主要研究领域为计算机视觉,人机交互,普适计算技术。E-mail:xgy-dcs@mail.tsinghua.edu.cn

人体检测、定位以及人体的重要部分(头部,手等)的检测,识别和跟踪是人体行为理解的基础,在解决这些基本问题的基础上,更重要也更困难的问题就是动作识别和行为理解。对动作识别的研究可以追溯到 20 世纪九十年代。2002 年,相关研究的论文数量经历了一个飞跃式的增长,这些研究大多假设是在结构化环境和孤立动作条件下。所谓的结构化环境就是预先设定和可人为控制的环境,例如,用于计算机输入的手势识别,这时视觉数据采集的光照、视角、距离等因素都是固定或已知的。这就为改善视觉处理的脆弱性提供了有利条件。但与此同时,也带来了很大的局限性。例如,基于生理特征的生物特征识别,目前的方法只适合于愿意在规定环境下给予系统配合的合作对象。与此相对,能在自然环境下,基于行为特征的生物特征识别就更为困难,但它具有容易被对象接受,或不易被察觉的优点。对于各种目的的视觉监控来说,能工作在自然环境下,至关重要。例如,为帮助老人延长独立生活或改善生活质量的视觉监控和提示,都需要能在老人生活的日常环境提供相应的服务。近年来,对日常生活和工作中动作和行为的理解正成为热点。这是所谓的“日常活动”(ADL)的分析和理解。由于人们在日常生活环境中的动作是自然和连续的,而且与环境有密切的联系,因此,给识别和理解带来一系列具有挑战性的难题:(1)分布式视觉信息处理方法和系统。通过多摄像机信息的融合来克服由于视角、距离、遮挡、光照等多种环境因素带来的干扰和不确定性是有效但也是富有挑战性的课题;(2)自然连续动作和行为的分割及多层次模型。人类的日常活动和行为是人体自然和连续的动作,其中包含了多种类型的运动和动作:无意识的人体移动,为了操作物体的动作,以及为了进行相互交流,例如打招呼的动作和姿态。此外复杂的人类活动和行为是由一系列动作或操作组成的。系统必在一个多层次的模型指导下对人体动作进行分割和分类。而分割和分类又需要有来自对动作和行为理解的高层模型指导;(3)基于上下境的行为理解。对动作和行为的理解需要了解当时视觉环境以及应用的情境。这也就是所谓的要具有“觉察上下境”或基于上下境的视觉处理和分析方法。因为相同的动作在不同的情境下传递不同语义。上下境的指导作用体现在以下 2 方面:①在现实的视场中可能需要处理的视觉对象和任务非常多,而计算资源是有限的。此外还有实时处理的要求。这时必须根据上下境来确定视觉处理的关注点;②在上下境的指导下对动作传递的语义进行推理。

行为理解的研究包含着从底层到高层的多层处理。底层处理中的人体检测和跟踪、动作识别、手势识别和轨迹分析等已经有了较多的研究和综述。而高层的处理方法,如对行为的建模、上下境在行为推理中的指导等研究还在起步阶段。

2 行为理解的研究现状

如引言中所述,行为理解包含了从底层处理到高层推理的全过程,涉及底层运动特征的提取和表示、行为的表示方法,以及高层行为语义的推理模型。下面的综述首先从行为的定义开始,然后讨论特征提取和动作表示,最后分析常见的几种行为推理模型。

2.1 行为表示的模型

目前对于行为的表示还没有一个通用的模型,大部分的研究都是针对特定的应用采用某种行为表示模型,最常见的是分层结构模型,而各个层次表示的内容取决于应用的需要。人体的行为就其目的而言可大致分为:(1)与环境交互。例如对物体的操作;Moeslund 等人提出了 action/motor primitives, actions, and activities 的分层模型[2]。在 Park 等人提出的驾驶员动作的表示模型中[3],底层为身体某个部位的运动,如头转动,躯干前移,伸胳膊等。中间层是由底层各部位的运动组合而成的一个复杂动作。最高层为人与周围设备的交互动作,即驾驶员对汽车部件的操作动作,如向左转动方向盘。(2)人际交互。Aggarwal 等人在 2 人交互的分析中[4],把交互行为分为 3 个层次。最高层是交互行为;中间层为单个人体的动作;最底层是对身体某个部分运动的检测和识别。群体交互,例如会议室场景更是需要多层次的表示[5]。关于行为的分层表示方法还可参考[2-4,6-10]。其中特别需要注意的是 Gonzàlez 等人在动作-行为的层次表示中增加了情境[7]。情境可认为是最高层的上下境,它用于解决行为理解的歧义问题。比如挥手这个动作在“足球赛”和“地铁站”这两种情境中显然是有不同的含义。

综上所述,分层模型已经成为研究者们公认的一种行为的表示方法,只是在不同的研究背景和任

务下,层次的数量和每个层次的定义各不相同。得到较多认可的表示模型大致包括如下几个层次:运动,动作,活动或操作,行为。这些层次大致是按照时间的尺度来进行分割的。但这样的分层方法在复杂的情况下,有时显得无能为力。更为实用的是按照任务过程进行分解。例如,老人在厨房中的做饭活动,它可分为取食品、处理食材、烹饪、上菜等过程。其中每个步骤,又可进一步分解,例如,烹饪又可按菜谱分为若干步骤。这样的分解是应用导向的。作为一个表示模型除了定义各层表示的含义以外还需要定义它们之间的关系和运行机制。Crowley 等人提出了情境网络的运行框架[11]。Dai 等人提出了一个基于多层次“上下境-事件”的模型[5],认为行为的层次结构中,上层的行为就是下层动作定义的环境,所以就称为上下境。它定义了什么是下层中发生的有意义的动作,即事件。相邻层次之间的“上下境-事件”关系可递归地延伸到所有的层次。所以,这个模型具有通用性。

2.2　运动特征的提取和动作表示

视觉或者其他底层运动特征的提取和表示是进行高层行为理解的推理所必需的基础工作。较早开始的对动作行为分析的工作很多是采用主动传感器来获得人体某个部位的运动信息[12-16]。这类工作主要是通过人体的四肢或躯干佩戴的各种传感器来获取该部位的运动特征,然后分析动作行为,由于当前以人为中心的计算强调用户感觉自然,嵌入式的传感器破坏了用户的感受,给用户的行动造成不便,因此,目前越来越多的研究开始转向用摄像机这种非嵌入式被动的传感器获取人体的动作特征。

基于视觉的动作表示按特征的性质大致可以分为两类,一类是 3 维特征,另一类是 2 维图像特征。3 维特征本身具有视角不变性,适用于分布式视觉系统下的动作体态表示。Campbell 等人提出了基于立体视觉数据的 3 维手势识别系统[17]。Jin 等人建立了基于 3 维模型的动作识别系统[18]。3 维模型通常参数多,训练复杂,计算量大。如果是基于立体视觉的原理还可能要遇到匹配中的对应性困难。相比之下,基于 2 维图像特征的表示计算相对简单,适用于视角相对固定的情况。下面具体介绍一些基于 2 维特征的动作表示。

Liu 等人只对坐、站、躺几个日常生活中最基本的动作做了分析[19]。计算了前景区域每个像素的距离投影

$$DP = \left(\sum_{i=1}^{M} (H^i - H_c)^2, \sum_{i=1}^{M} (V^i - V_c)^2 \right) \quad (1)$$

式中,H^i 和 V^i 表示前景像素在水平和竖直方向上的坐标,H_c 和 V_c 表示前景中点的坐标,M 是前景像素点的个数。每一个动作都用距离投影的高斯分布来表示。这种特征抽取方法是视角相关的,文中使用了与人体朝向成 90°的固定视角。这个视角上最容易抽取出区分度大的人体形状特征。

Niebles 等人把每个动作的一系列视频帧都看作是一组特征描述词的集合[20],特征描述词通过提取时空兴趣点得到。定义响应函数如下:

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

式中,$g(x,y,\sigma)$ 是 2 维高斯平滑核函数,应用在空间维度上,h_{ev}和 h_{od}都是 1 维 Gabor 滤波器,分别定义为 $h_{ev}(t;\tau,\omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$和 $h_{od}(t;\tau,\omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$并运用在时间维度上。一般情况下,复杂动作发生的区域会产生较大的响应,局部响应最大的点作为兴趣点。并用梯度或者光流来描述。

Park 等人用多高斯混合模型表示人体 5 个主要部分(头、脸、胳膊、躯干和下身)的颜色分布[3],并用椭圆拟合,Kalman 滤波器随时对参数进行更新。动态贝叶斯网络被用来检测动作和姿态,驾驶员行为被用一个表达式表示,表达式组成如下:{agent-motion-target},其中 agent 表示动作实施者,如头、手等;motion 表示动作;target 表示驾驶室的操作仪器。

Chung 等人用水平和竖直方向上的一对投影来表示当前的体态[21];Robertson 等人采用了基于光流的动作描述子来描述动作[22],继而与样本集中样本逐个匹配来识别动作类型;Turaga 等人也是提取光流作为每一帧中动作的特征[23];Ryoo 等人用人体外框的长、宽和中心点的坐标被作为特征[24]。Wang 等人在办公室异常行为识别的研究中[25]对提取出的人体区域采用 R 变换[26],提取动作形状,R 变换具有尺寸和旋转不变性,可以应对人离摄像机距离不同造成的尺寸变化。

以上这些工作都是在固定视角下用 2 维运动特征表示动作。这时可在有利的视角下观测动作,但也限制了对象的活动范围,使它难以适应实际应用的环境。现实生活中,观测对象活动范围较广,位置

变化大,导致视角多变;同时由于生活环境中的家具等也会对人体造成遮挡。因此,需要分布式视觉系统的支持,通过多摄像机信息的融合来克服由于视角多变,活动范围广以及遮挡带来的各种问题。这是富有挑战性的难题。

基于人体特征例如人头或四肢的运动特征将可简化信息融合和动作分析。Kim 等人是在分布式环境下检测人体的躺、站、坐等简单动作[26],通过自适应的背景相减得到前景区域,然后用椭圆拟合和 omega 曲线头肩部检测算法检测头部,在任何时刻,所有的摄像机都会进行全部的底层处理,得到人体的高度,人体位置,头部位置,人体长宽比和手部的位置,一个专门的模块将负责从每个摄像机处理的结果进行人的匹配,并选出没有遮挡的处理结果作为行为理解的观测向量。再如 Park 等人在分布式视觉系统下研究两人交互的行为[13],文中考虑到了视角对动作特征抽取带来的影响,因此,首先讨论了摄像机选择的问题。他根据不同摄像机得到的前景区域的离散度选择最佳视角,可以理解为选择像平面中两人距离最大的视角,因此,避免了遮挡问题。将分割出的人体区域在水平方向投影,计算得到人体的中轴,然后人体被按照一个指定参数分割为头,上身和下身 3 个部分。用 HSV 颜色空间表示每个像素点,用混合高斯模型表示身体的 3 个部分。可以同时分析上身和腿部的动作。而在不同视角下检测人体特征本身也是一个困难的问题,这是这种方法需要付出的代价。

除了上述由于成像环境限制造成的困难以外,现实生活中的很多动作,例如厨房中的烹饪操作,很细微,难以单独依靠视觉来检测和识别。而动作所使用的工具或接触的物体将可提供关于动作明确的线索。因此,有学者提出了根据使用的物品来协助识别对象动作。如果知道装面包的容器被使用了,这往往比识别到人伸手拿东西这个动作蕴含更多的语义。Wu 等人将水壶、电话、果汁等 33 个物品贴上电子标签(RFID)[27],并在用户的手腕上带上接收器。当用户使用某个物品时,接收器就会接收到该物品上电子标签发出的 ID 信号。通过对使用物品的分析能够识别出烧水,打电话,喝果汁等 16 种行为。Wang 等人也类似地充分利用了关于“所使用物品”的“常识”对行为理解的指导意义[28],通过在物品上贴 RFID,手腕上带接收器来获得物品使用信息。

另外有一些工作[29-31]认为人的轨迹甚至人在某个功能物体(如冰箱、沙发等)附近停留的时间可以用来解释人的行为,这样的假设就完全避开了复杂困难的动作分析以及传感器对人体造成的不便,在这类工作中,环境上下境信息和场景知识受到极大重视,成为进行行为理解推理所依赖的重要线索。

2.3 行为理解的推理方法

行为理解的推理中广泛采用了基于图模型的推理方法,如隐马尔科夫模型(HMM),动态贝叶斯网络(DBN),条件随机场(CRF)等;也有的研究采用其他的推理方法,如文献[14]使用基于规则的决策树来对一系列表示动作及对象的三元表达式进行分类。文献[32]、[33]采用模板匹配的方法,将检测到的运动特征与训练好的样本逐个匹配,匹配的结果即为对行为识别的结果。文献[31]使用了有限状态自动机,每个状态表示当前人体的位置,来对人的轨迹进行分类,识别异常事件。

在目前的行为分析领域中,HMM 是较常使用的一种推理模型[34-35]。HMM 是一种有效的时变信号处理方法,它隐含了对时间的校正,并提供了学习机制和识别能力。根据不同应用环境下行为的特性,很多研究对 HMM 进行了适应性扩展,比如 Hierarchical HMM, Coupled HMMs[36], Parameterized-HMMs[37]等。大部分的模型采用了分层的结构来对应行为的分层特性。文献[38]在群体交互动作识别中采用两层 HMM 模型,下层模型对群体中的个体进行动作识别,识别结果作为上层群体行为识别模型的观测。文献[39]也采用了分层的模型分析行为,由 3 层在不同时间粒度上依次增加的 HMM 组成。HMM 虽然是对时间序列建模的一种简单而有效的模型,但是当行为变得复杂或者在长时间尺度上存在相关性,就不满足马尔可夫假设[27],同时考虑了行为的分层结构和状态的持续时间,提出了 S-HSMM(switching hidden semi-markov model),是 HSMM 模型的两层扩展,底层表示了自动作及其持续时间,高层表示了一系列由底层自动作组成的行为。文中给出的实验结果证明了比 HSMM 和 HMM 对行为具有更强的模型表示能力。

也有研究将 DBN 引入到行为理解中[40-42]。由于 HMM 在一个时间片段上只有一个隐藏节点和一个观测节点,在一个时刻需要将所有的特征压缩到一个节点中,那么所需要的训练样本将是巨大的(相当于联合概率密度函数);而 DBN 在一个时间

段上是任意结构的贝叶斯网络,可以包含有多个因果关系的节点,即用条件概率来形成联合概率,训练相对要简单,也给模型的设计提供了更大的灵活性,能够更准确地表达状态之间以及状态和观测之间真实的关系,但是设计起来要比 HMM 复杂。文献[43]对 DBN 和分层的 HMM 做了详细的比较并且给出了模型选择和表示时需要考虑的几个因素:(1)可用于训练和测试的数据;(2)变量被观测到的可能性;(3)数据之间的内在关系;(4)应用的复杂度。

也有些研究放弃了产生式模型而采用区分式模型来分析行为。文献[44]首先采用了 CRF 模型用于行为识别,考虑到 HMM 最大的缺点就是输出独立性假设,导致不能考虑上下境的特征,限制了特征的选择。但是实际情况是,行为的当前状态往往与一个长的时间尺度上的观测存在相互的依赖,并且观测之间很可能不是相互独立的。CRF 不需要对观测进行建模,因此,避免了独立性假设,并且可以满足状态与观测之间在长时间尺度上的交互。结合产生式模型和区分式模型的优势对行为理解进行推理将成为未来的研究方向。

3　结　语

以上对动作识别和行为理解的现状做了简要的综述,但就建立能在复杂的现实世界中提供有效服务的计算机视觉系统而言,还缺少了两个关键的部分,这就是:(1)如何从复杂的现场背景下快速,可靠地检测和识别人体(物体)。物体在现实世界中的位置和光照情况多种多样甚至还有遮挡,但人类还是能在混乱的场景中快速地检测和识别各种物体。这是目前的机器视觉远未达到的能力。视觉认知,计算机视觉和认知神经科学的文献中有很多证据说明上下境信息极大地影响搜索和识别物体的效率[45-46]。上下境的影响是多层次的,其中包括:语义层(例如,桌子与椅子经常出现在同一图像中);空间构造层(例如,键盘一般是在显示器的下方);姿态层(例如,椅子通常是朝向桌子;汽车一般是沿着道路方向停靠)等。研究还证明空间上下境可为场景预测中可能发生的动作提供有用的线索[47]。总之,基于上下境的视觉关注机制是解决上述困难的关键;(2) 上下境指导下的行为理解。生活中人体动作的语义不仅取决于本身的状态而且取决于场景中其他人和物体的当前和历史的状态,也就是取决于上下境。相同的动作在不同的上下境中代表着不同的语义,在会议这样的群体交互场景下尤为突出[5]。例如,“举手”的动作,在“大会报告”的场景下,表示“希望提问”;在“会议表决”时表示决定的取向等。以上两个关键问题都涉及如何在视觉计算感知和利用中上下境信息。这也就是当前所谓的基于上下境的视觉和觉察上下境的视觉方法。从视觉处理的策略来说,目前大多数都是采用自底向上的,从局部到整体的方式,而基于上下境的视觉处理是采用自顶向下,从整体到局部的方式。这在一定程度上反映了人类视觉系统的处理方式。因此,这是重要的值得注意的研究方向。

参考文献(References)

1　Alejandro J, Daniel G P, Nicu S, *et al*. Human-centered computing: toward a human revolution [J]. Computer, 2007, **40**(5):30-34.

2　Moeslund T B, Hilton A, Krüger V. A survey of advances in vision-based human motion capture and analysis [J]. Computer Vision and Image Understanding, 2006, **104**(3): 90-126.

3　Park S, Trivedi M. Driver activity analysis for intelligent vehicles: issues and development framework [A]. In: Proceedings of IEEE Intelligent Vehicles Symposium [C], Las Vegas, Nevada, USA, 2005:644-649.

4　Aggarwal J K, Park S. Human motion: modeling and recognition of actions and interactions[A]. In: Proceedings of Second International Symposium on 3D Data Processing, Visualization and Transmission [C], Thessaloniki, Greece, 2004: 640-647.

5　Dai Peng, Tao Lin-mi, Xu Guang-you. Audio-visual fused online context analysis toward smart meeting room[A]. In: Proceedings of International Conference on Ubiquitous Intelligence and Computing [C], Hong Kong, China, 2007: 11-13.

6　Bobick A, Movement, activity, and action: the role of knowledge in the perception of motion[A]. Philosophical Transactions of the Royal Society of London, 1997, **352**(1358):1257-1265.

7　Gonzàlez J, Varona J, Roca F X, *et al*. A Spaces: action spaces for recognition and synthesis of human actions[A]. In: Proceedings of International Workshop on Articulated Motion and Deformable Objects [C], Palma de Mallorca, Spain, 2002: 21-23.

8　Jenkins O C, Mataric M. Deriving action and behavior primitives from human motion capture data [A]. In: Proceedings of International Conference on Robotics and Automation[C], Washington DC, USA, 2002: 2551-2556.

9　Nagel H H. From image sequences towards conceptual descriptions [J]. Image and Vision Computing, 1988, **6**(2): 59-74.

10　Mori T, Kamisuwa Y, Mizoguchi H, *et al*. Action recognition system based on human finder and human tracker [A]. In: Proceedings of the 1997 IEEE/RSJ International Conference on Intelligent Robots

and Systems[C], Beijing, China, 1997:1334-1341.

11 Crowley J L, Coutaz J. Context aware observation of human activity, multimedia and expo [A]. In: Proceedings of IEEE International Conference on ICME '02 [C], Lausanne, Switzerland, 2002: 909-912.

12 Park S, Kautz H. Hierarchical recognition of activities in daily living using multi-scale, multi-perspective vision and RFID [A]. In: Proceedings of 4th International Conference on Intelligent Environments[C], Seattle, WA, USA, 2008: 1-4.

13 Ward J A, Lukowicz P, Troster G, *et al*. Activity recognition of assembly tasks using body-worn microphones and accelerometers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, **28**(10): 1553-1567.

14 Yin J, Yang Q, Pan J J. Sensor-based abnormal human-activity detection [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, **20**(8): 1082-1090.

15 Yang S I, Cho S B. Recognizing human activities from accelerometer and physiological sensors [A]. In: Proceedings of IEEE International Conference on Multi-sensor Fusion and Integration for Intelligent Systems[C], Seoul, Korea, 2008: 100-105.

16 Purwar A, Jeong D U, Chung W Y. Activity monitoring from real-time tri-axial accelerometer data using Sensor network [A]. In: Proceedings of International Conference on Control, Automation and Systems[C], Seoul, Korea, 2007: 2402-2406.

17 Campbell L W, Becker D A, Azarbayejani A, *et al*. Invariant features for 3D gesture recognition [A]. In: Proceedings of International Conference on Automatic Face and Gesture Recognition [C], Killington, Vermont, USA, 1996: 157-162.

18 Jin N, Mokhtarian F. Image-based shape model for view-invariant human motion recognition[A]. In: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance[C], London, UK, 2007: 336-341.

19 Liu C D, Chuug P C, Chung Y N. Human home behavior interpretation from video streams [A]. In: Proceedings of the 2004 IEEE International Conference on Networking, Sensing & Control [C], Taipei, Taiwan, China, 2004: 192-197.

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

21 Chung P C, Liu C D. A daily behavior enabled hidden Markov model for human behavior understanding [J]. Pattern Recognition, 2008, **41**(5): 1572-1580.

22 Robertson N, Reid I. Behavior understanding in video: a combined method [A]. In: Proceedings of IEEE International Conference on Computer Vision[C], Beijing, China, 2005: 808-815.

23 Turaga P K, Veeraraghavan A, Chellappa R. From videos to verbs: mining videos for activities using a cascade of dynamical systems [A]. In: Proceedings of Conference on Computer Vision and Pattern Recognition[C], Minneapolis, Minnesota, USA, 2007: 1-8.

24 Tabbone S, Wendling L, Salmon J P. A new shape descriptor defined on the Radon transform [J]. Computer Vision and Image Understanding, 2006, **102**(1-2): 42-51.

25 Wang Y, Huang K, Tan T N. Abnormal activity recognition in office based on R transform [A]. In: Proceedings of IEEE Conference on Image Processing[C], San Antonio, TX, USA, 2007: I-341-344.

26 Kim K, Medioni G G. Distributed visual processing for a home visual sensor network[A]. In: Proceedings of IEEE Workshop on Applications of Computer Vision[C], Copper Mountain, Colorado, USA, 2008: 1-6.

27 Wu J X, Osuntogun A, Choudhury T, *et al*. A scalable approach to activity recognition based on object use [A]. In: Proceedings of IEEE International Conference on Computer Vision[C], Beijing, China, 2007: 1-8.

28 Wang S, Pentney W, Choudhury T. Common Sense based joint training of human activity recognizers [A]. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence[C], Hyderabad, India, 2007: 2237-2242.

29 Duong T V, Bui H H, Phung D Q, *et al*. Activity recognition and abnormality detection with the switching hidden semi-Markov model [A]. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition [C], San Diego, CA, USA, 2005: 838-845.

30 Nguyen N T, Phung D Q, Venkatesh S. Learning and detecting activities from movement trajectories using the hierarchical hidden markov model [A]. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition[C], San Diego, CA, USA, 2005: 955-960.

31 Mahajan D, Kwatra N, Jain S, *et al*. A framework for activity recognition and detection of unusual activities [A]. In: Proceedings of Indian Conference on Computer Vision, Graphics, Image Processing [C], Kolkata, India, 2004: 37-42.

32 Doll'ar P, Rabaud V, Cottrell G, *et al*. Behavior recognition via sparse spatio-temporal features [A]. In: Proceedings of 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance[C], Beijing, China, 2005: 65-72.

33 Liu C D, Chuug P C, Chung Y N. Human home behavior interpretation from video streams [A]. In: Proceedings of IEEE International Conference on Networking, Sensing & Control[C], Taipei, China, 2004: 192-197.

34 Xu G, Ma Y F, Zhang H J, *et al*. Motion based event recognition using HMM [A]. In: Proceedings of IEEE International Conference on Pattern Recognition[C], Quebec, Canada, 2002: 831-834.

35 Sun X D, Chen C W, Manjunath B S. Probabilistic motion parameter models for human activity recognition [A]. In: Proceedings of International Conference on Pattern Recognition [C], Quebec, Canada, 2002: 443-446.

36 Brand M, Oliver N, Pentland A. Coupled hidden Markov models for complex action recognition [A]. In: Proceedings of International Conference on Computer Vision and Pattern Recognition[C], Puerto Rico, 1997: 994-999.

37 Wilson A, Bobick A. Recognition and interpretation of parametric

gesture [A]. In: Proceedings of International Conference on Computer Vision[C], Bombay, India,1998: 329-336.

38 Zhang D, Gatica-Perez D, Bengio S, *et al*. Modeling individual group actions in meetings: a two-layer HMM framework[A]. In: Proceedings of IEEE CVPR Workshop on Detection and Recognition of Events in Video[C], Washington, DC, USA,2004: 117-125.

39 Olivier N, Horovitz E, Garg A. Layered representations for human activity recognition [A]. In: Proceedings of IEEE International Conference on Multimodal Interfaces [C], Pittsburgh, PA, USA, 2002: 3-8.

40 Luo Y, Wu T D, Hwang J N. Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks [J]. Computer Vision and Image Understanding, 2003, **92**(2-3): 196-216.

41 Du Y T, Chen F, Xu W L, *et al*. Recognizing interaction activities using dynamic Bayesian network [A]. In: Proceedings of International Conference on Pattern Recognition [C], New York, USA, 2006: 618-621.

42 Buxton H, Gong S G. Advanced visual surveillance using Bayesian networks [A]. In: Proceedings of International Conference on Computer Vision[C], Boston, MA,USA, 1995:111-123.

43 Oliver N, Horvitz E. A comparison of HMMs and dynamic Bayesian networks for recognizing office activities [A]. In: Proceedings of 10th International Conference on User Modeling [C], Edinburgh, UK, 2005: 199-209.

44 Sminchisescu C, Kanaujia A, Metaxas D. Conditional models for contextual human motion recognition [J]. Computer Vision and Image Understanding, 2006,**104**(2-3): 210-220.

45 Oliva1 A, Torralba A. The role of context in object recognition[J]. Trends in Cognitive Sciences, 2007,**11**(12): 520-527.

46 Torralb A. Contextual priming for object detection [J]. International Journal of Computer Vision, 2003, **53**(2): 169-191.

47 Zibetti E, Tijus C. Perceiving action from static images: The role of spatial context [J]. Lecture Notes in Computer Science,2003, 2680: 397-410.