第 18 卷 第 2 期
2013 年 2 月

中国图象图形学报
Journal of Image and Graphics

Vol. 18, No. 2
Feb., 2013

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2013)02-141-11
论文引用格式: 章毓晋. 时空行为理解[J]. 中国图象图形学报,2013,18(2):141-151.

时空行为理解

章毓晋
清华大学电子工程系,北京 100084

摘 要: 利用视觉信息了解世界是人类视觉和计算机视觉的共同目标。充分利用客观的时空信息,对场景中感兴趣目标的行为进行理解是近年计算机视觉的一个前沿研究内容。本文对该领域的基本情况、主要概念、研究焦点、典型技术、发展情况给予介绍,以期引起相关研究人员的关注,共同参与相关工作,推动计算机视觉的进展。

关键词: 时空技术,动作基元,动作,活动,事件,行为,图像理解

Understanding spatial-temporal behaviors

Zhang Yujin
Department of Electronic Engineering, Tsinghua University, Beijing 100084, *China*

Abstract: The common goals of human vision and computer vision are to understand the world via visual information. Effective utilization of all spatial-temporal information and understanding the behaviors of interesting objects in a scene is a current focus in computer vision research. In this paper, we provide a general introduction for the basic situation, main concepts, research issues, typical techniques and developments of this area, hope to draw the attention of research community, to jointly work in it and to push forward the front of computer vision.

Key words: spatial-temporal techniques, action primitives, action, activity, event, behavior, image understanding

0 引 言

人类视觉过程可看作是一个从感觉(感受到的是对 3D 世界之 2D 投影得到的图像)到知觉(由 2D 图像认知 3D 世界内容和含义)的复杂过程[1]。视觉的最终目的从狭义上说是要对场景做出对观察者有意义的解释和描述,从广义上讲,还有基于这些解释和描述并根据周围环境和观察者的意愿制定出行为规划。计算机视觉是指用计算机实现人的视觉功能,希望能根据感知到的图像对实际的目标和场景做出有意义的判断[2]。这实际上也就是图像理解的目标。

计算机视觉和图像理解的一个重要工作就是通过对场景获得的图像进行加工从而解释场景、指导行动。为此,需要判断场景中有哪些景物,它们随时间如何改变其在空间的位置、姿态、速度、关系等。简言之,要在时空中把握景物的动作,确定动作的目的,并进而理解它们所传递的语义信息。

基于图像/视频的自动目标行为理解是一个很有挑战的研究问题。它包括获取客观的信息(采集图像序列),对相关的视觉信息进行加工,分析(表达和描述)信息内容,以及在此基础上对图像/视频的信息进行解释以实现学习和识别行为。

上述工作的跨度很大,其中动作检测和识别近期得到很多关注和研究,也取得了明显的进展。相对来说,高抽象层次的行为识别与描述(与语义和智能相关)研究还不多,许多概念的定义还不很明确,许多技术还在不停地发展更新中。

收稿日期: 2012-09-15; 修回日期: 2012-11-21
基金项目: 国家自然科学基金项目(61171118);教育部高等学校博士学科点专项科研基金项目(SRFDP-20110002110057)
第一作者简介: 章毓晋(1954—),男,教授,1989 年于比利时列日大学获应用科学专业博士学位,主要研究方向为图像工程。E-mail: zhang-yj@tsinghua.edu.cn

1 时空技术

时空技术是面向时空行为理解的技术，是一个相对较新的研究领域，目前的工作正在不同的层次展开，下面是一些概括情况。

1.1 新的领域

笔者撰写的图像工程综述系列从对1995年的文献统计开始至今已进行了17年[3]。在图像工程综述系列进入第二个十年时（对2005年的文献统计开始），随着图像工程研究和应用新热点的出现，在图像理解大类中增加了1个新的小类——C5：时空技术（3D运动分析、姿态检测、对象跟踪、行为判断和理解）[4]。该小类强调的是综合利用图像/视频中所具有的各种信息以对场景及其中目标的动态情况做出相应的解释。

过去这7年里综述系列收集的C5小类文献数量共有76篇，它们在各年的分布如图1中直方图所示，图1中还给出了用3阶多项式对各年文献数量进行逼近得到的变化趋势。总地来说，这还是一个相对较新的领域，研究成果还不算多，发展趋势也不太稳定。

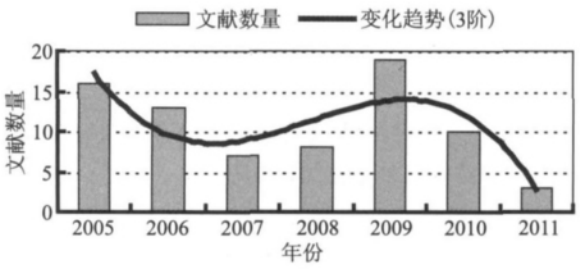

图1 时空技术文献数量

Fig. 1 The numbers of papers in spatial-temporal techniques

1.2 多个层次

目前时空技术研究的主要对象是运动着的人或物，以及场景中景物（特别是人）的变化。根据其表达和描述的抽象程度从下到上可分为多个层次：

1）动作基元（action primitives）：用来构建动作的原子单元，一般对应场景中短暂的运动信息。

2）动作（action）：由主体/发起者的一系列动作基元构成的有具体意义的集合体（有序组合），一般动作代表简单的常由一个人进行的运动模式，且常仅持续秒的量级。人体动作的结果常导致人体姿态的改变。

3）活动（activity）：为完成某个工作或达到某个目标而由主体/发起者执行的一系列动作的组合（主要强调逻辑组合）。活动是相对大尺度的运动，一般依赖于环境和交互人。活动常代表由多个人进行的序列（可能交互的）复杂动作，且常持续较长的时段。

4）事件（event）：指在特定时间段和特定空间位置发生的某种活动。通常其中的动作由多个主体/发起者执行（群体活动）。对特定事件的检测常与异常活动有关，这方面的综述可见文献[5]。

5）行为（behavior）：主体/发起者主要指人或动物，强调主体/发起者受思想支配而在特定环境/上下境中改变动作、持续活动和描述事件等。

下面以乒乓球运动为例，给出各个层次的一些典型示例（图2）。运动员的移步、挥拍等都可看做典型的动作基元。运动员完成一个发球（包括抛球、挥臂、抖腕、击球等）或回球（包括移步、伸臂、翻腕、抽球等）都是典型的动作，但一个运动员走到挡板边把球拣回来则常看做一个活动。而两个运动员来回击球以赢得分数也是典型的活动场面。运动队之间的比赛等一般作为一个事件来看待，比赛后颁奖也是典型的事件。运动员赢球后握拳自我激励虽然可看做一个动作，但更多的时候被看做运动员的一个行为。当运动员打出漂亮的对攻后，观众的鼓掌、呐喊、欢呼等也都归于观众的行为。

需要指出，在许多研究中对后3个层次的概念常不严格区分地使用。例如，将活动称为事件，此时一般指一些异常的活动（如两人发生争执，老人走路

图2 乒乓球比赛中的几个画面

Fig. 2 Some frames from a table tennis game

跌倒等); 将活动称为行为,此时更强调活动的含义(举止)、性质(如行窃的动作或翻墙入室的活动称为偷盗行为)。另外,它们在一定程度上也有密切联系,如有许多人类的活动直接与它们的行为相关[6]。在下面的讨论中,除特别强调,将主要用(广义的)活动来统一代表后 3 个层次。

2　时空兴趣点

场景的变化源于景物的运动,特别是加速运动。视频图像局部结构的加速运动对应场景中加速运动的景物,它们处在图像中有非常规运动数值的位置,这些位置(兴趣点)的信息对理解场景很有帮助。

在时空场景中,对兴趣点的检测有从空间向时空扩展的趋势[7]。

2.1　空间兴趣点的检测

在图像空间里,可以使用线性尺度空间表达 L^{sp} 来对图像 f 建模,f^{sp}: $\mathbf{R}^2\to\mathbf{R}$。

$$L^{sp}(x,y;\sigma_l^2)=g^{sp}(x,y;\sigma_l^2)\otimes f^{sp}(x,y) \tag{1}$$

即将 f^{sp} 与具有方差 σ_l^2 的高斯核 g 卷积

$$g^{sp}(x,y;\sigma_l^2)=\frac{1}{2\pi\sigma_l^2}\exp\left[-(x^2+y^2)/2\sigma_l^2\right] \tag{2}$$

典型的 Harris 兴趣点检测器的思路是确定 f^{sp} 在水平和垂直两个方向均有明显变化的空间位置。对给定的观察尺度 σ_l^2,这些点可借助在方差为 σ_l^2 的高斯窗中求和得到的二阶矩的矩阵来计算,即

$$\boldsymbol{\mu}^{sp}(\cdot;\sigma_l^2,\sigma_i^2)=g^{sp}(\cdot;\sigma_i^2)\otimes\{[\nabla L(\cdot;\sigma_l^2)][\nabla L(\cdot;\sigma_l^2)]^{T}\}=g^{sp}(\cdot;\sigma_i^2)\otimes\begin{bmatrix}(L_x^{sp})^2 & L_x^{sp}L_y^{sp}\\ L_x^{sp}L_y^{sp} & (L_y^{sp})^2\end{bmatrix} \tag{3}$$

式中,L_x^{sp} 和 L_y^{sp} 是在局部尺度 σ_l^2 根据 $L_x^{sp}=\partial_x[g^{sp}(\cdot;\sigma_l^2)\otimes f^{sp}(\cdot)]$ 和 $L_y^{sp}=\partial_y[g^{sp}(\cdot;\sigma_l^2)\otimes f^{sp}(\cdot)]$ 算得的高斯微分。这个二阶矩描述符可看做一幅 2D 图像在一个点的局部邻域的朝向分布协方差矩阵。所以 $\boldsymbol{\mu}^{sp}$ 的本征值 λ_1 和 λ_2($\lambda_1\leqslant\lambda_2$)构成 f^{sp} 沿两个图像方向变化的描述符。如果 λ_1 和 λ_2 的值都很大,则表明有一个感兴趣点。为检测这样的点,可以检测角点函数的正极大值

$$H^{sp}=\det(\boldsymbol{\mu}^{sp})-k\cdot\mathrm{tr}^2(\boldsymbol{\mu}^{sp})=\lambda_1\lambda_2-k(\lambda_1+\lambda_2)^2 \tag{4}$$

2.2　时空兴趣点的检测

将在空间的兴趣点检测扩展到时空中,即去检测在局部时空体[8]中图像值沿时和空都有大变化的位置。具有这样性质的点将对应在时间上具有特定位置的空间兴趣点,其处在具有非常数值运动的时空邻域内。检测时空兴趣点是一种提取底层运动特征的方法,不需要背景建模。这里可先将给定的视频与一个 3D 高斯核在不同的时空尺度进行卷积。然后在尺度空间表达的每一层都计算时空梯度,将它们在各个点的邻域结合起来以得到对时空二阶矩矩阵的稳定估计。从矩阵中就可提取出局部的特征。

为对时空图像序列建模,可使用函数 f: $\mathbf{R}^2\times\mathbf{R}\to\mathbf{R}$ 并通过将 f 与各向非同性高斯核(不相关的空间方差 σ_l^2 和时间方差 τ_l^2)卷积构建它的线性尺度空间表达 L: $\mathbf{R}^2\times\mathbf{R}\times\mathbf{R}_+^2\to\mathbf{R}$,即

$$L(\cdot;\sigma_l^2,\tau_l^2)=g(\cdot;\sigma_l^2,\tau_l^2)\otimes f(\cdot) \tag{5}$$

式中,时空分离的高斯核为

$$g(x,y,t;\sigma_l^2,\tau_l^2)=\frac{\exp\left[-\frac{x^2+y^2}{2\sigma_l^2}-\frac{t^2}{2\tau_l^2}\right]}{\sqrt{(2\pi)^3\sigma_l^4\tau_l^2}} \tag{6}$$

对时间域使用一个分离的尺度参数是非常关键的,因为时间和空间范围的事件一般是独立的。另外,使用兴趣点算子检测出来的事件同时依赖于空间和时间的观察尺度,所以对尺度参数 σ_l^2 和 τ_l^2 需要分别对待。

类似于在空间域,考虑一个时-空域二阶矩的矩阵,它是一个 3×3 的矩阵,包括用高斯权函数 $g(\cdot;\sigma_i^2,\tau_i^2)$ 卷积的一阶空间和时间微分:

$$\boldsymbol{\mu}=g(\cdot;\sigma_i^2,\tau_i^2)\otimes\begin{bmatrix}L_x^2 & L_xL_y & L_xL_t\\ L_xL_y & L_y^2 & L_yL_t\\ L_xL_t & L_yL_t & L_t^2\end{bmatrix} \tag{7}$$

其中,将积分尺度 σ_i^2 和 τ_i^2 与局部尺度 σ_l^2 和 τ_l^2 根据 $\sigma_i^2=s\sigma_l^2$ 和 $\tau_i^2=s\tau_l^2$ 联系起来。一阶微分定义为

$$\begin{aligned}L_x(\cdot;\sigma_l^2,\tau_l^2)&=\partial_x(g\otimes f)\\ L_y(\cdot;\sigma_l^2,\tau_l^2)&=\partial_y(g\otimes f)\\ L_t(\cdot;\sigma_l^2,\tau_l^2)&=\partial_t(g\otimes f)\end{aligned} \tag{8}$$

为检测感兴趣点,在 f 中搜索具有 $\boldsymbol{\mu}$ 的显著本征值 λ_1、λ_2、λ_3 的区域。这可将定义在空间的 Harris 角点检测函数,即式(4)通过结合 $\boldsymbol{\mu}$ 的行列式和秩扩展到时空域

$$H=\det(\boldsymbol{\mu})-k\cdot\mathrm{tr}^3(\boldsymbol{\mu})=\lambda_1\lambda_2\lambda_3-k(\lambda_1+\lambda_2+\lambda_3)^3 \tag{9}$$

3 动态轨迹学习和分析

动态轨迹学习和分析[9]试图通过对场景中各个运动目标位置和变化结果的理解和刻画来提供对监控场景状态的把握。图3所示为对视频进行动态轨迹学习和分析的流程框图,首先对目标进行检测(如在车上对行人检测[10])并跟踪,接着用所获得的轨迹自动地构建场景模型,最后用该模型描述监控的状况和提供对活动的标注。

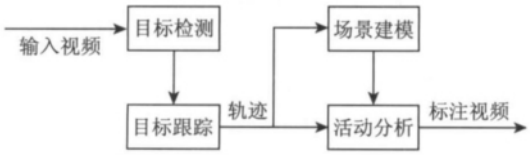

图3 动态轨迹学习和分析的流程框图

Fig. 3 Flowchart for dynamic trajectory learning and analysis

在场景建模中,先将有事件发生的图像区域定义为兴趣点(POI),然后在接下来的学习步骤中定义活动路径(AP),该路径刻画目标是如何在感兴趣点间运动/游历的。这样构建的模型可称为POI/AP模型。

在POI/AP学习中的主要工作包括:

1)活动学习(activity learning):通过比较轨迹来进行,尽管轨迹长度可能不同,关键是要保持对相似性的直观认识。

2)适应(adaption):研究管理POI/AP模型的技术。这些技术要能在线地适应增加新活动、除去不再继续的活动,并验证模型。

3)特征选择(feature selection):确定对特定任务正确的动力学表达层次。例如,仅使用空间信息就可确定汽车走哪条道,但要检测事故还常需要速度信息。

3.1 自动场景建模

借助动态轨迹对场景的自动建模包括以下3个要点[11]:

1)目标跟踪

对目标的跟踪需要在每一帧中对各个可观察到的目标进行身份维护。例如,在T帧视频中被跟踪的目标会生成一系列可推断出来的跟踪状态,即

$$S_T=\{s_1,s_2,\cdots,s_T\} \tag{10}$$

式中,$s_t,t=1,2,\cdots,T$可描述诸如位置、速度、外观、形状等目标特性。这些轨迹信息构成进一步分析的基石。通过认真分析这些信息就可识别和理解活动。

2)兴趣点检测

场景建模的第1个任务就是找出图像中的感兴趣区域。在指示跟踪目标的地形图中,这些区域对应图中的结点。常考虑的两种结点包括入/出区域和停止区域。以教师走进教室走上讲台为例,前者对应教室门而后者对应讲台。

入/出区域是目标进入或离开视场或被跟踪目标出现或消失的位置。这些区域常可借助2D高斯混合模型建模,$Z\sim\sum_{i=1}^{W}w_iN(\mu_i,\sigma_i)$,其中有$W$个分量。这可用EM算法来解。进入点数据包括在第1个跟踪状态所确定的位置,而离去点数据包括在最后1个跟踪状态所确定的位置。它们可用一个密度准则来区分,在状态i的混合密度定义为

$$d_i=\frac{w_i}{\pi\sqrt{|\sigma_i|}}>L_d \tag{11}$$

它测量高斯混合的紧凑程度。其中,阈值

$$L_d=\frac{w}{\pi\sqrt{|C|}} \tag{12}$$

指示信号聚类的平均密度。这里,$0<w<1$是用户定义的权重,C是在区域数据集中所有点的协方差矩阵。紧凑的混合指示正确的区域而宽松的混合指示由于跟踪中断而导致的跟踪噪声。

停止区域源于场景地标点,即目标在一段时期内趋于固定的位置。这些停止区域可用两种不同的方法来确定:(1)在该区域中被跟踪点的速度低于某个事先确定的很低的阈值;(2)所有被跟踪点至少在某个时间段保持在一个有限的距离环中。通过定义一个半径和一个时间常数,第(2)种方法可保证目标确实保持在特定的范围,而第(1)种方法仍有可能包括运动很慢的目标。对活动分析,除了要确定位置也需要把握在每个停止区域所花去的时间。

3)活动路径学习

要理解行为,需要确定出活动路径。可以使用POI从训练集中滤除虚警或跟踪中断的噪声,只保留在进入区域后开始并在终止区域前结束的轨迹。经过停止区域的跟踪轨迹分成分别对应进入区域的和离开区域的两段,一个活动要定义在目标开始运动和结束运动的两个感兴趣点之间。

3.2 学习路径

由于路径刻画了目标运动的情况,一个原始的

轨迹可表示成动态测量的序列。例如,常用的轨迹表达就是一个运动序列

$$\boldsymbol{G}_T = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_T\} \tag{13}$$

式中,运动矢量

$$\boldsymbol{g}_t = [x^t, y^t, v_x^t, v_y^t, a_x^t, a_y^t]^{\mathrm{T}} \tag{14}$$

表示从跟踪中获得的目标在时刻 t 的动态参数,即位置 $[x, y]^{\mathrm{T}}$,速度 $[v_x, v_y]^{\mathrm{T}}$ 和加速度 $[a_x, a_y]^{\mathrm{T}}$。

仅使用轨迹就可能以无监督的方式学习 AP,其基本流程如图 4 所示。

图 4　轨迹学习步骤

Fig. 4　Steps in trajectory learning

主要包括 3 个步骤:

1) 轨迹预处理

路径学习研究中的大部分工作都是要获得适合聚类的轨迹。当进行跟踪时主要困难来源于时间变化的特性,这导致轨迹长度的不一致。需要采取步骤以保障在不同尺寸的输入之间可以进行有意义的比较。另外,轨迹表达在聚类中应直观地保持原始轨迹的相似性。

轨迹预处理主要包括两个内容:

(1) 归一化　归一化的目的是保证所有轨迹有相同的长度 T_a。两种简单的技术是填零和扩展。填零就是在较短轨迹的后面增加一些等于零的项。扩展是将在原轨迹最后时刻的部分延伸扩展到需要的长度。它们都有可能把轨迹空间扩得非常大。除了检查训练集以确定轨迹的长度 T_a 外,也可利用先验知识进行重采样和平滑。重采样结合插值可保证所有轨迹有相同的长度 T_a。平滑可用来消除噪声,平滑后的轨迹也可插值和采样到固定的长度。

(2) 降维　降维将轨迹映射到新的低维空间,从而可以使用更鲁棒的聚类。这可通过假设一个轨迹模型并确定能最好地描述该模型的参数来实现。常用技术包括矢量量化,多项式拟合,多分辨率分解,隐马尔可夫模型,子空间方法,频谱方法和核方法等。

2) 轨迹聚类

聚类是在没有标记的数据中确定结构的常用机器学习技术。在观察场景时,收集运动轨迹并将其结合进类似的类别中。为了产生有意义的聚类,轨迹聚类过程要考虑 3 个问题: (1) 定义一个距离(对应相似性) 测度; (2) 聚类更新的策略; (3) 聚类验证。

3) 路径建模

轨迹聚类后,可根据所得到的路径建立图(graph) 模型以进行有效的推理。路径模型是对聚类的紧凑表达。可以用两种方式对路径建模。第 1 种方式考虑完整的路径,从端点到端点的路径不仅有平均的中心线,两边还有包络指示路径范围,沿路径可能有一些中间状态给出测量顺序(如图 5(a)); 第 2 种方式将路径分解为一些子路径,或者说路径表示成子路径的树,预测路径的概率从当前结点指向叶结点(如图 5(b))。

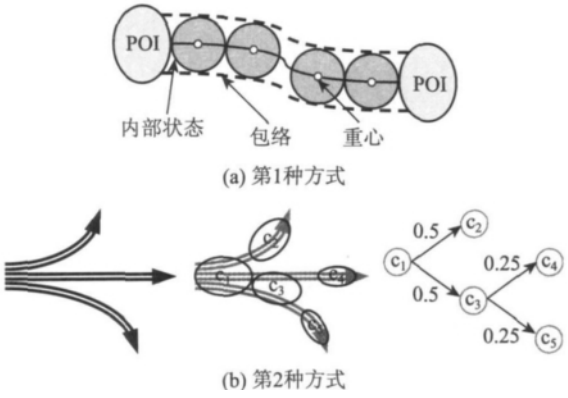

图 5　两种对路径建模的方式

Fig. 5　Two methods for path modeling

3.3　自动活动分析

一旦建立了场景模型,就可以对目标的行为和活动进行分析。例如,监控视频的一个基本功能就是对感兴趣事件进行验证。一般来说,只有在特定环境下才可定义是否感兴趣。例如,停车管理系统会关注是否还有空位可以停车,而在智能会议室系统中关心的是人员之间的交流。除了仅仅识别特定的行为,所有非典型的事件也需要检查。通过对一个场景进行长时的观察,系统可以进行一系列的活动分析,从而学习到哪些是感兴趣的事件。

一些典型的活动分析事例如下:

1) 虚拟篱笆(virtual fencing),任何监控系统都有一个监控范围,在该范围的边界上设立哨兵就可对范围内发生的事件进行预警。这相当于在监控范围周围建立了虚拟篱笆,一旦有入侵就触发分析,如控制高分辨率的云台摄像机(PTZ) 获取入侵处的细节,开始对入侵数量的统计等。

2) 速度分析(speed profiling),虚拟篱笆只利用了位置信息,借助跟踪技术还可获得动态信息实现基于速度的预警,如车辆超速或路面堵塞。

3) 路径分类(path classification),速度分析只利用了当前跟踪的数据,实际中还可利用由历史运动模式获得的活动路径(AP)。新出现目标的行为可借助最大后验(MAP)路径来描述,即

$$L^* = \arg\max_k p(l_k \mid G) = \arg\max_k p(G, l_k)\, p(l_k) \quad (15)$$

这可帮助确定哪个活动路径能最好地解释新的数据。因为先验路径分布 $p(l_k)$ 可用训练集来估计,所以问题就简化为用 HMM 来进行最大似然估计。

4) 异常检测(abnormality detection),异常事件的检测常是监控系统的重要任务。因为活动路径能指示典型的活动,所以如果一个新的轨迹与已有的不符就可发现异常。

5) 在线活动分析(online activity analysis),能够在线地分析、识别、评价活动比使用整个轨迹来描述运动更重要。一个实时的系统要能够根据尚不完整的数据快速地对正在发生的行为进行推理(常基于图模型),其工作步骤包括:路径预测;异常跟踪。

6) 目标交互刻画(object interaction characterization),更高层次的分析期望能进一步描述目标间的交互。与异常事件类似,严格地定义目标交互也很困难。在不同的环境下,不同的目标间有不同类型的交互。以汽车碰撞为例,每辆汽车有其空间尺寸,可看做其个人空间。汽车在行驶时,其个人空间在汽车周围要增加一个最小安全距离(最小安全区),所以时空个人空间会随运动而改变,速度越快,最小安全距离增加越多(尤其在行驶方向上)。

最后需要指出,对简单的活动,仅靠目标位置和速度就能进行分析,但对更复杂的活动则可能还需要更多的测量,如加入剖面的弯曲度以判别古怪的行走。为提供对活动和行为的更全面覆盖,常需要使用多摄像机网络。活动轨迹还可来源于由互相连接的部件而构成的目标(如人体),这里活动需要相对于一组轨迹来定义。

4 动作分类和识别

基于视觉的人体动作识别是对图像序列(视频)用动作(类)标号进行标记的过程。在对观察到的图像或视频获得表达的基础上,可将人体动作识别变成一个分类的问题。

4.1 动作分类

对动作的分类可采用多种形式的技术[12]:

1) 直接分类

在直接分类的方法中,并不对时间域以特别关注。这些方法将观察序列中所有帧的信息都加到单个表达中或对各个帧分别进行动作的识别和分类。

在很多情况下,图像的表达是高维的。这导致匹配计算量非常大。另外,表达中也可能包括噪声等特征。所以,为分类需要在低维空间获得紧凑、鲁棒的特征表达。降维既可采用线性的方法也可采用非线性的方法。例如,PCA 是一种典型的线性方法,而局部线性嵌入是一种典型的非线性方法。

直接分类所用的分类器也可不同。鉴别型分类器关注如何区分不同的类别,而不是模型化各个类别,典型的如 SVM。在自举框架下,用一系列弱分类器(每个仅使用 1D 表达)来构建一个强分类器。除 AdaBoost 外,LPBoost 可以获得稀疏的系数且能很快收敛。

2) 时间状态模型

生成模型学习观察和动作间的联合分布,对每个动作类建模(考虑所有变化)。最典型的是隐马尔可夫模型,其中的隐状态对应动作进行的各个步骤。隐状态对状态转移概率和观察概率进行建模。

鉴别模型学习在观察条件下动作类别的概率。它们并不对类别建模但关注类间的差别。这种模型对区分相关的动作比较有利。条件随机场是一种典型的鉴别模型,

3) 动作检测

基于动作检测的方法并不显式地对图像中目标表达建模,也不对动作建模。它将观察序列与编号的视频序列联系起来,以直接检测(已定义的)动作。例如,可将视频片段描述成在不同时间尺度上编码的词袋,每个词都对应一个局部片(patch)的梯度朝向。具有低时间变化的片可以忽略掉,这样表达将主要集中于运动区域。

对人体动作的表达和描述的主要方法可分为两类:(1)基于表观的方法:直接利用对图像的前景、轮廓、光流等的描述;(2)基于人体模型的方法:利用人体模型表达行为人的结构特征,如将动作用人体关节点序列来描述。不管采用那类方法,实现对人体的检测以及对人体重要部分(如头部、手、脚

等) 的检测和跟踪都会起重要的作用。

4.2　动作识别

动作及活动的表达和识别是一个相对有历史但还不太成熟的领域[13]。采用的方法依赖于研究者的目的。在场景解释中,表达可独立于导致活动产生的目标(如人或车) ; 而在监控应用中,一般关注人的活动和人之间的交互。在整体(holistic) 的方法中,全局的信息要优于部件的信息,这在要确定人的性别时比较适用。而对简单的动作如走或跑,也可考虑使用局部的方法,其中更关注细节动作或动作基元。

4.2.1　整体识别

整体识别强调对整个人体目标或单个人体的各个部分进行识别。例如,可基于整个身体的结构和整个身体的动态信息来识别人的行走、行走的步态等。这里绝大多数方法基于人体的剪影或轮廓而不太区分身体的各个部分。例如,有一种基于人体的身份识别技术使用了人的剪影并对其轮廓进行均匀采样,然后对分解的轮廓用 PCA 处理。为计算时-空相关性,可在本征空间比较各个轨迹。另一方面,利用动态信息除可辨识身份外也可确定人正在做什么工作。基于身体部件的识别则通过身体部件的位置和动态信息来对动作进行识别。

4.2.2　姿态建模

对人体动作的识别与对人体姿态的估计密切相连。人体姿态可分为动作姿态和体位姿态,前者对应人在某一个时刻的动作行为,后者对应人体在 3D 空间的朝向。

对人体姿态的表达和计算方法主要可分为 3 种: 1) 基于表观的方法; 2) 基于人体模型的方法; 3) 基于 3D 重构的方法。

可以基于时空兴趣点(参见第 2 节) 来对姿态进行建模。如果基于时空 Harris 角点的检测,得到的时空兴趣点多处于运动突变的区域。这样的点数量较少,属于稀疏型,容易丢失视频中重要的运动信息,导致检测失效。所以还可借助运动强度提取稠密型的时空兴趣点,以充分捕获运动产生的变化。提取出时空兴趣点后,对每个点先建立描述符,然后再对每个姿态建模。

近期在对自然场景中的姿态估计方面有一个趋势是为克服在无结构场景中用单视图进行跟踪中的问题多采用在单帧图中进行姿态检测。例如,基于鲁棒的部件检测并对部件进行概率组合已能在复杂的电影中获得对 2D 姿态的较好估计。

4.2.3　活动重建

动作导致姿态的改变,如果将人体的每个静止姿态定义为一个状态,那么借助状态空间法(也称概率网络法) ,将状态之间通过转移概率来切换,则一个活动序列的构建可通过在对应姿态的状态之间进行一次遍历而得到。

基于对姿态的估计,从视频自动重建人体活动方面也已有了明显进展。原始的基于模型的分析-合成方案借助多视角视频采集从而有效地对姿态空间进行搜索。当前的许多方法更注重获取整体的身体运动而不很强调精确地构建细节。

单视图人体活动重建也基于统计采样技术有了很多进展。目前比较关注的是利用学习得到的模型来约束基于活动的重建。研究表明使用强有力的先验模型对单视图中跟踪特定活动很有帮助。

4.2.4　交互活动

交互活动是比较复杂的活动。可以分为两类: 1) 人与环境的交互,如人开车,拿一本书; 2) 人际交互,常指两人(也可多人) 的交流活动或联系行为,它是将单人的(原子) 活动结合起来而得到的。对单人活动可借助概率图模型来描述。概率图模型是对连续动态特征序列建模的有力工具,有比较成熟的理论基础。它的缺点是其模型的拓扑结构依赖于活动本身的结构信息,所以对复杂的交互活动需要大量的训练数据以学习图模型的拓扑结构。为了将单人活动结合起来,可以使用统计关系学习的方法。它是一种将关系/逻辑表示、概率推理、机器学习和数据挖掘等进行综合以获取关系数据似然模型的机器学习方法。

4.2.5　群体活动

量变引起质变,参与活动目标数量的大幅增加,带来了新的问题和新的研究。例如,群体目标运动分析主要以人流、交通流以及自然界的密集生物群体为对象,研究群体目标运动的表达与描述方法,分析群体目标的运动特征以及边界约束对群体目标运动的影响。此时,对特殊个体的独特行为的把握有所减弱,更关注的是把个体抽象而对整个集合活动的描述。例如有的研究借鉴宏观运动学理论,探索粒子流的运动规律,建立粒子流的运动理论。在此基础上,对群体目标活动中的聚合、消散、分化、合并等动态演变现象进行语义分析,以期解释整个场景的动向和态势。

4.2.6 场景解释

与对场景中目标的识别不同,场景解释主要考虑整幅图像而不去验证特定的目标或人。实际使用的许多方法仅考虑摄像机拍到的结果,从中通过观察目标运动而不一定确定目标的身份来学习和识别活动。这种策略在目标足够小、可表示成2D空间中一个点时是比较有效的。

5 活动和行为建模

一个通用的动作/活动识别系统包括从一个图像序列到高层解释的若干工作步骤[14]:

1) 获取输入视频或序列图像;

2) 提取精练的底层图像特征;

3) 从底层特征到中层动作描述;

4) 从基本的动作出发进行高层语义解释。

一般实用的活动识别系统是分层的。底层包括前景-背景分割模块,跟踪模块和目标检测模块等。中层主要是动作识别模块。高层最重要的是推理引擎,它将活动的语义根据较低层的动作基元进行编码,并根据学习的模型进行整体的理解。

如第1节中指出的,从抽象程度来看,活动的层次要高于动作。如果从技术的角度来看,对动作和活动的建模和识别常采用不同的技术,而且有从简单到复杂的特点。目前许多常用的动作和活动的建模和识别技术可如图6给以分类[14]。

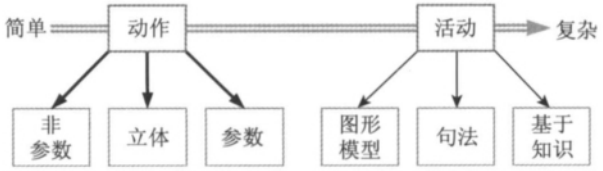

图6 动作和活动建模识别技术的分类图

Fig.6 Classification of approaches for action and activity recognition

5.1 动作建模

对动作建模的方法主要可分为3类:非参数建模、立体建模、参数时序建模。非参数的方法从视频的每帧中提取一组特征,将这些特征与存储的模板匹配。立体的方法并不逐帧提取特征,而是将视频看做像素强度的3D立体并将标准的图像特征(如尺度空间极值,空域滤波器响应)扩展到3D。参数时序方法对运动的时间动态建模,从训练集中估计对一组动作的特定参数。

5.1.1 非参数建模方法

常见的非参数建模方法有如下几种:

1) 2D模板

这类方法包括如下步骤:先进行运动检测,然后在场景中跟踪目标。跟踪后,建立一个包含目标的裁剪序列。尺度的改变可借助归一化目标尺寸来补偿。对给定的动作计算一个周期性的指标(index),如果周期性很强,就进行动作识别。为进行识别,利用对周期的估计将周期序列分割成独立的周期。将平均周期分解为若干个时间上的片段并对各个片段中的每个空间点计算基于流的特征。把每个片段中的流特征平均到单个帧中。一个活动周期中的平均流帧就构成每个动作组的模板。

2) 3D目标模型

3D目标模型是对时空目标建立的模型,典型的如广义圆柱体模型,2D轮廓叠加模型等。在2D轮廓叠加模型中包含了目标的运动和形状信息,据此可提取目标表面的几何特征,如峰、坑、谷、脊等。如果把2D轮廓换成背景中的块(blob),就得到二值时空体。

3) 流形学习方法

很多动作识别都涉及到高维空间的数据,特征空间随维数变得按指数形式稀疏,为构建有效的模型需要大量的样本。利用学习数据所在的流形可确定数据的固有维数,该固有维数的自由度较小,可帮助在低维空间设计有效的模型。

5.1.2 立体建模方法

常见的立体建模方法有以下几种:

1) 时空滤波

时空滤波是对空间滤波的推广,采用一组时空滤波器对视频体数据滤波。根据滤波器组的响应进一步推出特定的特征。

2) 基于部件的方法

一个视频体可看做许多局部部件的集合体,各个部件有特殊的运动模式。一种典型的方法是使用第2节的时空兴趣点。除了使用Harris兴趣点检测器,也可对从训练集中提取的时空梯度进行聚类。另外,还可使用词袋模型来表示动作,其中词袋模型可通过提取时空兴趣点并对特征聚类得到。

在大多数基于部件的方法中,对部件的检测常基于线性操作,如滤波、时空梯度等,所以描述符对表观变化、噪声、遮挡等比较敏感。但另一方面,由于本质上的局部性,这些方法对非稳态背景比较

鲁棒。

3) 子体匹配

子体匹配(sub-volume matching) 指在视频和模板中的子体间的匹配。例如,可借助从时空运动相关的角度将动作与模板匹配。这种方法与基于部件方法的主要区别是它并不需要从尺度空间的极值点提取动作描述符而是检查两个局部时空块(patch) 间的相似度(通过比较两个块间的运动) 。不过,对整个视频体都进行相关计算会很耗时。解决此问题的一种方法是将目标检测中很成功的快速 Haar 特征(盒特征) 推广到 3D。

子体匹配的优点是对噪声和遮挡比较鲁棒,如果结合光流特征,则也对表观变化比较鲁棒。子体匹配的缺点是易受背景改变的影响。

4) 基于张量的方法

张量是对矩阵在多维空间的推广。一个 3D 时空体可以自然地看做一个有 3 个独立维的张量。例如,人的动作,人的身份和关节的轨迹可看做一个张量的 3 个独立维。通过将总的数据张量分解为主导模式(类似 PCA 的推广) ,就可以提取对应人的动作和身份(执行动作的人) 的标志。当然,也可直接将张量的 3D 取为时空域的 3D,即(x, y, t) 。

基于张量的方法提供了一种整体匹配视频的直接方法,不需要考虑前几种方法所用的中层表达。另外,其他种类的特征(如光流、时空滤波器响应等) 也很容易通过增加张量维数而结合进来。

5.1.3　参数建模方法

前面两种建模方法仅比较适合简单的动作,下面介绍的建模方法更适合跨越时域的复杂动作,如芭蕾舞视频中的舞步,乐器演奏家用复杂的手势演奏等。

1) 隐马尔可夫模型

隐马尔可夫模型是状态空间的一种典型模型,它对时间序列数据的建模很有效,有很好的推广性和鉴别性,适用于需要递推概率估计的工作。在构建离散隐马尔可夫模型的过程中,将状态空间看做一些离散点的有限集合。随时间的演化模型化为一系列从一个状态转换到另一个状态的概率步骤。隐马尔可夫模型最早用于识别动作是用于识别网球击打(shot) 动作,如反手击球,反手截击,正手击球,正手截击,扣杀等。其中,将一系列减除背景的图像模型化为对应特定类的隐马尔可夫模型。隐马尔可夫模型也可用于对随时间变化动作(如步态) 的建模。

2) 线性动态系统

线性动态系统比隐马尔可夫模型更一般化,它并不限制状态空间是有限符号的集合而可以是 $\mathbf{R}^k$ 空间中的连续值,其中 k 是状态空间的维数。最简单的线性动态系统是一阶时不变高斯-马尔可夫过程,可表示为

$$\boldsymbol{x}(t) = A\boldsymbol{x}(t-1) + \boldsymbol{w}(t) \quad \boldsymbol{w} \sim N(0, \boldsymbol{P}) \tag{16}$$

$$\boldsymbol{y}(t) = C\boldsymbol{x}(t) + \boldsymbol{v}(t) \quad \boldsymbol{v} \sim N(0, \boldsymbol{Q}) \tag{17}$$

式中,$\boldsymbol{x} \in \mathbf{R}^d$ 是 d 维状态空间,$\boldsymbol{y} \in \mathbf{R}^n$ 是 n 维观察矢量,$d \ll n$,$\boldsymbol{w}$ 和 $\boldsymbol{v}$ 分别是过程和观察噪声,它们都是高斯分布的,均值为零,协方差矩阵分别为 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$。线性动态系统可看作对具有高斯观察模型的隐马尔可夫模型在连续状态空间的推广,更适合于处理高维时间序列数据,但仍不太适合用于非稳态的动作。

3) 非线性动态系统

考虑下面一系列动作: 一个人先弯腰捡起一个物品,然后走向一个桌子并将物品放在桌上,最后坐在一把椅子上。这里面有一系列短的步骤,每个步骤都可用线性动态系统建模。整个过程可看做在不同的线性动态系统之间的转换。最一般的时变线性动态系统形式为

$$\boldsymbol{x}(t) = A(t)\boldsymbol{x}(t-1) + \boldsymbol{w}(t) \quad \boldsymbol{w} \sim N(0, \boldsymbol{P}) \tag{18}$$

$$\boldsymbol{y}(t) = C(t)\boldsymbol{x}(t) + \boldsymbol{v}(t) \quad \boldsymbol{v} \sim N(0, \boldsymbol{Q}) \tag{19}$$

对比式(16) (17) ,这里 A 和 C 都可随时间变化。

5.2　活动建模和识别

活动相比于动作,不仅持续时间长,而且大多数人们所关注的活动应用,如监控和基于内容的索引,都包括多个动作人。他们的活动不仅互相作用而且也与上下文实体互相影响。为对复杂的场景建模,需对复杂行为的本征结构和语义进行高层次的表达和推理。

5.2.1　图形模型

仅简单介绍两种常见的图形模型。

1) 信念网络

贝叶斯网络就是一种简单的信念网络。它先将一组随机变量编码为局部条件概率密度,再对它们间的复杂条件依赖性进行编码。动态信念网络(也称动态贝叶斯网络) 是对简单贝叶斯网络通过结合随机变量间的时间依赖性而得到的一种推广。对比只能编码一个隐变量的传统隐马尔可夫模型,动态信念网络可以对若干随机变量间的复杂条件依赖关系进行编码。

2) 皮特里网

皮特里网是一种描述条件和事件之间联系的数

学工具。它特别适合模型化和可视化如排序,并发,同步和资源共享等行为。皮特里网是包含两种结点——位置和过渡的双边图,其中位置指实体的状态而过渡指实体状态的改变。

5.2.2 合成方法

合成方法主要借助语法概念和规则来实现。

1) 语法

语法利用一组产生式规则描述处理的结构。类似于语言模型中的语法,产生式规则指出如何从词(活动基元)构建句子(活动),以及如何识别句子(视频)满足给定语法(活动模型)的规则。早期对视觉活动进行识别的语法用于识别将物体拆解的工作,此时语法中还没有概率模型。其后得到应用的是上下文自由语法,它被用来对人体运动和多人交互进行建模和识别。

2) 随机语法

用于检测低层基元的算法本质上是概率算法。所以,随机上下文自由语法对上下文自由语法进行了概率扩展,更适合用于将实际的视觉模型结合起来。随机上下文自由语法可用于对活动(其结构假设已知)的语义进行建模。

在很多情况下,常需要将一些附加的属性或特征与事件基元相关联。例如,事件基元发生的准确位置对描述一个事件很可能是很重要的,但这有可能没有事先记录在事件基元集合中。在这些情况下,属性语法比传统语法就有更强的描述能力。概率属性语法已用于在监控中处理多代理的活动。

5.2.3 基于知识和逻辑的方法

知识和逻辑有密切的联系。

1) 基于逻辑的方法

基于逻辑的方法依靠严格的逻辑规则来描述一般意义上的领域知识以描述活动。逻辑规则对描述用户输入的领域知识或使用直观且用户可读的形式来表示高层推理结果很有用。

2) 本体论的方法

在使用前述方法的大多数实际配置中,符号活动的定义都是以经验方式来构建的。如语法的规则或一组逻辑的规则都是手工指定的。尽管经验构建设计较快且在多数情况下效果很好,但推广性较差,仅限于所设计的特定情况。所以,还需要对活动定义的集中表达或独立于算法的活动本体。本体可以标准化对活动的定义,允许对特定的配准进行移植,使不同的系统增强互操作性,以及方便地复制和比较系统性能。典型的实际例子包括分析护理室中的社会交往,对会议视频进行分类,对银行交互行动的设置等。

另外,国际上从2003年开始举办视频事件竞赛工作会议(video event challenge workshop)以整合各种能力构建一个基于通用知识的领域本体。会议已定义了6个视频监控的领域:1)周边和内部的安全;2)铁路交叉的监控;3)可视银行监控;4)可视地铁监控;5)仓库安全;6)机场停机坪安全。会议还指导了两种形式语言的制定,一种是视频事件表达语言(VERL),它帮助完成基于简单的子事件实现复杂事件的本体表达;另一种是视频事件标记语言(VEML),它用来对视频中的VERL事件进行标注。

6 结 语

本文对时空行为理解这一个近年计算机视觉的前沿研究领域的基本情况、主要概念、研究焦点、典型技术、发展情况等给予了概括的介绍。从研究进展和趋势来看,在时空技术的5个层次中,较高层的研究更有挑战性。而从研究内容所涉及的范围来看,除去计算机视觉本身技术以外,医学、心理学、行为学、认知学、社会学等学科也将在其中发挥积极的作用。

参考文献(References)

[1] Kong B. Comparison between human vision and computer vision [J]. Nature Magazine, 2002, 24(1): 51-55. [孔斌. 人类视觉与计算机视觉的比较 [J]. 自然杂志, 2002, 24(1): 51-55.]

[2] Shapiro L, Stockman G. Computer Vision [M]. New Jersey, USA: Prentice Hall, 2001.

[3] Zhang Y J. Image engineering in China: 2011 [J]. Journal of Image and Graphics, 2012, 17(5): 603-612. [章毓晋. 中国图像工程: 2011 [J]. 中国图象图形学报, 2012, 17(5): 603-612.]

[4] Zhang Y J. Image engineering in China: 2005 [J]. Journal of Image and Graphics, 2006, 11(5): 601-623. [章毓晋. 中国图像工程: 2005 [J]. 中国图象图形学报, 2006, 11(5): 601-623.]

[5] Popoola O P, Wang K. Video-based abnormal human behavior recognition: a review [J]. IEEE-SMC-C: 2012, 99(99): 1-14.

[6] Khair N M, Yaacob S, Hariharan M, et al. A study of human emotional: Review [C]//Proceedings of 2012 International Conference on Biomedical Engineering. Penang, Malaysia: IEEE,

2012: 393-399.

[7] Laptev I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64(2/3): 107-123.

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

[9] Morris B T, Trivedi M M. A survey of vision-based trajectory learning and analysis for surveillance [J]. IEEE-CSVT, 2008, 18(8): 1114-1127.

[10] Jia H X, Zhang Y J. A survey of computer vision based pedestrian detection for driver assistance system [J]. Acta Automatica Sinica, 2007, 33(1): 84-90. [贾慧星, 章毓晋. 车辆辅助驾驶系统中基于计算机视觉的行人检测研究综述 [J]. 自动化学报, 2007, 33(1): 84-90.]

[11] Makris D, Ellis T. Learning semantic scene models from observing activity in visual surveillance [J]. IEEE-SMC-B, 2005, 35(3): 397-408.

[12] Poppe R. A survey on vision-based human action recognition [J]. Image and Vision Computing, 2010, 28: 976-990.

[13] 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: 90-126.

[14] Turaga P, Chellappa R, Subrahmanian V S, et al. Machine recognition of human activities: a survey [J]. IEEE-CSVT, 2008, 18(11): 1473-1488.