文章编号:1673-9469(2013)02-0098-05 doi:10.3969/j.issn.1673-9469.2013.02.025

基于DBAN的监控视频数据挖掘

姜振凤,迟庆云
(枣庄学院 信息科学与工程学院,山东 枣庄 277160)

摘要:视频挖掘主要涉及三个层次的工作:视频数据预处理,视频特征数据提取及视频模式发现与表示。针对监控视频数据,以人体姿势识别和行为理解为挖掘任务开展视频数据挖掘研究。提出了一个带有二维身体部位表示法的动态贝叶斯动作网(DBAN)基本框架来提高人体姿势定位的准确性及行为识别的精度,并通过实验证明了该方法的有效性。

关键词:视频挖掘; 监控视频; 动态贝叶斯动作网
中图分类号:TP391.4 **文献标识码**:A

Data mining of surveillance video based upon DBAN

JIANG Zhen-feng, CHI Qing-yun
(College of Information Science and Engineering, Zaozhuang University, Shandong Zaozhuang 277160, China)

Abstract: Video mining principally involves 3 levels of operation, namely, video data preprocessing, video feature data extraction, the discovery and representation of video mining pattern. Aiming at the above manipulations, the current study probe into the researches of video data mining by taking the recognition of body posture and the comprehension of action as the mining tasks. The paper puts forward a DBAN basic framework with the notation of planar body part so as to enhance the accuracy of body posture positioning and the precision of its action recognition, which has been given testimony to its availability through one empirical research.

Key words: video mining; surveillance video; DBAN

视频挖掘是一种以挖掘目的为指导的视频内容分析过程,将视频挖掘应用到监控视频上,可以使计算机自动从海量监控视频中提取出视频内容特征、语义信息及一些视频模式[1],这样就提高了监控视频的职能应用。监控视频中对人的行为理解与描述是近年来被广泛关注的研究热点。目前,国内外已经有大量研究工作来对人体行为或事件进行建模和检测,比较流行的方法包括使用支持向量机(SVM)、隐马尔可夫模型(HMMs)、动态贝叶斯网络(DBNs)及条件随机场模型(CRFs)等方法[2]。其中,动态贝叶斯网络能在有限时间内,将变量之间的因果关系用联合概率关系的形式表示出来,被认为非常适合对视频中人体运动这种既具有特征相关性又具有时序相关性的复杂特征进行建模[3]。

本文在对DBN研究的基础上,将人体动作描绘成一系列简单的动作基元(action primitives),并映射到动态贝叶斯网络,构建了一种基于动态贝叶斯动作网(Dynamic Bayesian Action Network,即DBAN)的模型框架,主要用于视频中对人体位置跟踪及行为识别的同步处理,实现监控视频中对人体行为信息的高层次语义挖掘[4]。

1 监控视频数据预处理

进行视频挖掘的第一个步骤就是对监控系统采集的视频图像进行预处理,得到高质量的视频对象(如:像素块、视频帧、视频段、场景等)[5]。本文针对静止背景下单个摄像机捕捉的监控视频数据开展研究,主要考虑对原始监控视频图像进行滤噪预处理。

收稿日期:2012-11-29
基金项目:枣庄学院青年科研项目(项目号:2011QN41)
作者简介:姜振凤(1981-),女,辽宁辽阳人,硕士,讲师,从事多媒体数据挖掘。

由于受到图像采集设备自身及环境等因素的干扰,监控视频序列在获取和传输时常常会受到各种各样的噪声干扰。为了更好地获得监控视频挖掘中相关语义信息,增强有关信息的可检测性和最大限度地简化数据,提高特征抽取、图像分割、匹配和识别的可靠性,需要对采集的视频图像进行滤噪处理[6]。当前大多数视频监控系统都属于背景始终静止不变,只有前景目标运动这类情况。在摄像机固定、环境没有变化的前提下,背景图像上的像素点可以用一个高斯分布描述,针对这一特点,本文参考文献[6]中的高斯滤波去噪和邻域去噪相结合的方法对图像进行滤噪预处理。

2 人体行为特征数据提取及建模

视频挖掘可分为两个层次,一是直接基于特征的底层视频语义信息挖掘,二是以特定挖掘任务为指导的高级模式知识挖掘。挖掘的数据实际上是视频对象的特征(物理特征、运动特征、特征之间的关系特征)及其语义信息描述[5]。

2.1 行为描述

目前,对人体行为描述的方法可分为两类,一类是基于表观的方法,它直接由图像的前景、轮廓、光流等描述行为;另一类是基于人体模型的方法,即利用人体模型获取行为者的结构特征,用人体关节序列描述人体行为。Johansson 通过试验证明人体关节点模型包含了可用于行为识别的丰富信息[7]。本文使用一种基于 3D 关节点序列的人体行为描述法,主要思想是将一个复合动作分解为一系列简单的基本动作,称为动作基元(primitive action)[8]。由于在一个基元中,各个身体部位都是基于单个坐标轴旋转的,每个基元可以简单的定义为由身体各部位旋转构成的结合体。例如,我们把行走看作一个复合动作,它包括四个基元:左腿向前→右腿交叉左腿→右腿向前→左腿交叉右腿。在行走过程中,大腿的转动与臀部紧密结合,小腿的转动与膝盖紧密结合。图 1[4] 显示了一个有关蹲伏动作的动作基元行为描述示例。

这种描述法可以从二维姿势和动作边界中获取,也可以利用动作捕捉器(MoCAP)技术获得,但鉴于 MoCAP 数据的获取需要耗费大量的时间且硬件要求较高,本文参考文献[9] 使用二维标注的方式从视频中进行模型学习。由于在人体姿势的角度测量表征中,每一个关键姿态都标示了一种不连续性,为了得到每个复合动作的表示方法,我们首先手动的选取每个动作的关键姿势,然后获取每个关键姿势的 3D 模型。在这一阶段,每个复合动作本质上是一组带有时间间隔的 3D 关键姿势序列。对于每个连续的关键姿势对,我们将基元定义成由一个关键姿势到另一个关键姿势的时间步长变换,也就是说基元在一个持续时间里描述了行为者从一个状态转换成另一个新的状态。

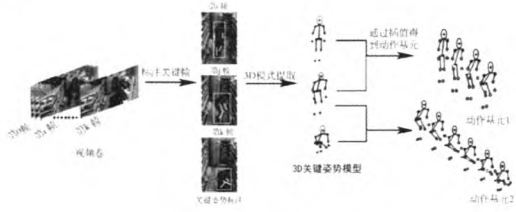

图1 蹲伏动作模型示例
(包含3个关键姿势和两个动作基元)

Fig.1 Illustration for crouch action model
(Containing 3 key poses and 2primitives)

2.2 基于 DBAN 的行为建模

本文参考文献[4],以参数化的函数方式给出动作模型,并将其嵌入到 DBN 中,构建基于 DBAN 的人体动作模型(如图 2 虚线部分所示)。在 DBAN 模型中,顶层的节点与复合动作(CA)相对应,如行走,蹲伏等;第二层对应动作基元(PA);第三层对应人体姿势(P)。其中,与每个基元相关联的持续时间节点 D_t 可用于捕捉从动作基元开始所持续的时间,即帧数。因此,在时间 t 的状态 S_t 是由元组(CA_t, PA_t, D_t, P_t)表示的。与文献[6]不同的是,本文使用 2D 身体部位模型来表示观测的 3D 投影姿势(如图 2 中的第四个层次),这部分内容将在第四部分作详细介绍。需要说明的是,为了清晰起见,前景观测节点在图中未显示。

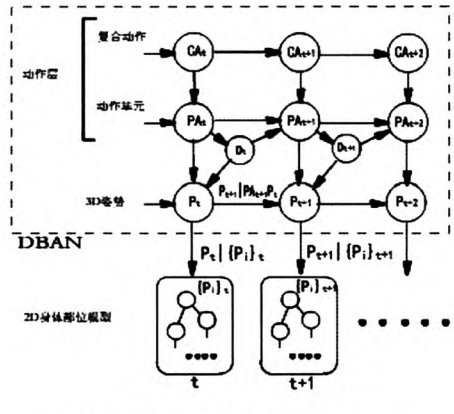

图2 基于DBAN的人体动作模型

Fig.2 Action model based on DBAN

对于长度为 T 的观测序列，其最优状态序列 $S^*_{[1:T]}$ 可通过公式(1)进行计算。其中，$\Phi_i(s_{t-1}, s_t, I_t)$ 是观测及转换势值；ω_i 是权重向量，它模拟了势函数的相对重要性。需要说明的是，该公式参考了文献[8]的隐马尔可夫模型的多变量状态表示法。

$$S^*_{[1:T]} = arg\ max \sum_{t=1}^{T}\sum_{f}\omega_f\phi_f(s_{t-1}, s_t, I_t) \quad (1)$$

观测及转换势值：由于动作基元的转换是通过正符号对数函数中的事件持续时间进行建模，使得停留在相同动作基元 PA_t 的概率减少到平均持续时间 $\mu(PA_t)$，且增加了转换为新基元的概率。因此，公式中的姿势转换势值可使用在训练中学习得到的人体关节位移的平均值及方差的正态分布建模。对于某一状态观测势值的获取，本文参考文献[4]，使用前景重叠及差异图像匹配的方法从视频中提取特征数据对观测势值进行描述。不同之处在于我们通过投影 3D 姿势获得一个 2D 身体部位模型，这个 2D 部位模型能够有效地进行局部搜索并更准确地与观测结果相对应。

相对权重向量：在 DBAN 中，特征权重估算用公式表示为在整个训练集 T 中似然误差对数函数值的最小化，该问题可通过 *Voted Perceptron* 算法[8]有效的计算获得。为了避免在所有帧中的姿势标注，文献[4]介绍了一种改进的 *Latent State Voted Perceptron* 算法来处理缺失数据。该算法对每个训练序列计算带有当前权重向量最可能的状态序列。如果这个估算复合行为不正确，真实信息状态序列从没有动作预测步骤的标注动作序列中进行估算，观测结果和真实信息状态序列之间的特征误差在整个训练集中收集并被用于更新权重向量。

3 视频模式发现

为了更准确的对投影的 3D 姿势进行定位，本文通过使用一个 2D 身体部位模型来对观测姿势进行校准，从而提高识别精度。另外，为了弥补 2D 部位校准所需的额外时间，我们提出了一个基于动作熵的方案来确定每一帧中所保存的样本数目，并结合 DBAN 实现对人体行为及姿势跟踪的同步处理。

3.1 构建 2D 身体部位模型

本文中所使用的身体部位模型与文献[10]中的 Pictorial Structures 类似。模型包括 10 个节点，每个节点对应一个身体部位，包括头、躯干、上臂、下臂、大腿和小腿。依据捕获的身体部位之间的运动学关系，将每个身体部位相关联的一对关节点进行拟合构建一个平面矩形模型，并通过各个部位的边缘进行连接。

对于给定视角的 3D 姿势，我们通过对其进行投影来估算人体关节的二维方位及身体部位的相对深度序列。接下来，我们来确定哪些部分是基于深度序列可视地，哪些部分是在部位矩形之间成对重叠地。在我们的实验中，我们认为百分比能见度低于 50% 视为被遮挡。此外，当 3D 姿势投影到 2D 模式时，一些身体部位因太小而无法进行观测时，这些部位也是无法用于定位的。图 3 显示了对于给定视角的 3D 姿势构建 2D 部位模型的流程图。其中图(d)中与右上臂相对应的节点由于遮挡不可见，使用了一个虚线边界进行标记。

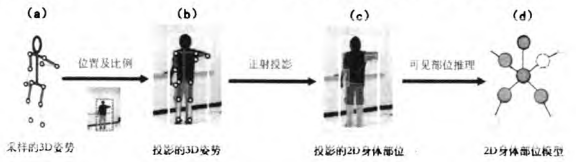

图3 从3D姿势中估算的2D身体部位模型

Fig. 3 2D body part models estimated from the 3D pose

3.2 3D 姿势探测与定位

在构建身体部位模型的基础上，使用基于模板匹配的方法来探测图像中的人体部位。用一个椭圆形的模板来为头部建模，躯干用定向的矩形来建模，每只手臂用一对按透视法缩短的线来建模。我们用模型中每一点的强度和方向梯度来定义一个身体部位(假设为 p)的似然度评分 $\Phi(p)$。

$$\Phi_{edge}(p) = \sum_{p_i \in p} d_{mag}(I(p_i)) \times d_{ori}(p_i I(p_i)) \quad (2)$$

其中，$d_{mag}(I(p))$ 是从图像 p 点到最近边缘像素的近似欧几里得距离。它可以通过在边缘似然图上使用广义距离变换进行有效的计算[11]。$d_{ori}(p, I(p))$ 是方向似然度，它是一个在模型 p 点的法向量和与图像 $I(p)$ 上的对应点之间的“点积”。由于方向信息经常有很多“噪声”数据，我们通过量化八向方位的方法来对法向量进行粗略估计。

3.3 人体姿势跟踪及行为识别

本文应用带有 2D 部位模型的 DBAN 来实现监控视频数据中对人体姿势跟踪及行为识别的同步处理。

初始化:我们从动作集中的所有复合动作中抽取样本姿势进行状态分布初始化。为了保证视角不变,针对每个来自于复合动作模型中的样本姿势,所有可能视角都要考虑。

样本状态预测:对于每一个状态 S_t,我们以单位时间步长来增加状态的持续时间。给定当前动作(CA_t,PA_t) 及新的持续时间,然后我们对下一个动作状态(CA_{t+1}, PA_{t+1})进行取样。如果发生基元变换,将持续时间设置为 0,标志开始一个新的基元。接下来,我们从姿势"转换势值" Φp(P_t, PA_{t+1}, P_{t+1})进行抽样来选择下一个姿势 P_{t+1}。其中,转换势值代表与基元 PA_{t+1} 相对应的函数 $f_{pA_{t+1}}(p,p',a)$ 中参数 α 的分布。

抽样状态与观察结果的拟合:首先应用一个步行监测器(pedestrian detector)来发现视频中的行为人,然后我们应用一个组合形状和前景影像跟踪器来对每一帧中的行为人进行跟踪定位[12]。利用跟踪器得到位置和测量信息,还可以对动作模型中的 3D 姿势进行校准。给定调整后的 3D 姿势,我们通过正射投影将该姿势构造成 2D 部位模型。为了实现姿势 P 的准确定位,首先通过预期的部位构形(configuration)及其周围的一个小的临近域为每个身体部位 p_i 应用一个边缘模板,然后我们使用消息传递的方法[8]在获取的身体部位分布之间执行运动学制约(kinematic constraints)。对于一个给定的观测特征映射,整个姿势 P 的后验似然度通过公式 3 得出。

$$F(P_t,X)=\sum_{i\in V}\Phi_i(x_i|p_i)+\sum_{ij\in E}\psi_{ij}(p_i,p_j) \qquad (3)$$

式中,V - 所有身体部位集合;E - 运动学连接的"部位对"集合;x_i - 部位 i 的似然映射($x_i\in X$)。

校正后的最优 2D 姿势是通过后验似然度 F(P_t, X)取最大值得到的。此外,对于遮挡姿势的处理,文献[10]考虑的是像素级的遮挡限制,与之相比,我们只须考虑那些几乎全部可见(可见度超过 80%)的部位,这使得我们的推理更为有效,姿势的定位也更为准确。

状态样本的选择:人体姿势匹配通过从动作模型中进行姿势抽样并将模型同观测图像进行拟合来实现的。这一过程需要将所有的动作模型与观测序列进行匹配,从而找到最佳匹配的方法来推断行为标签。由于维护所有可能的状态序列是不可能也是没有必要的,文献[4]使用贪心策略来保存具有最高评分的前 N 个状态序列,但这种贪心选择步长的设定有一定局限性,可能导致抽样样本数量过少,或影响算法的效率和准确性。针对以上问题,本文设置了一个在每一帧中进行维护的最小样本数 $N_{\min}$ 及动作类中所允许的最大样本数 $N_{\max}$。通过在当前有效行为中计算熵的方式,定义了一个不确定性度量值[12],使得所有可能的动作样本都将得到很好的体现。

要计算当前有效动作的熵,我们计算动作类的似然向量 $v=\{v_{CA}\}$,这里的 v_{CA} 是在属于动作类别 CA 的当前帧中所有状态的最高似然度评分。给定了动作类的似然度向量 v,然后我们在当前帧 t 定义大小为 Nt 的目标样本集 $N_t\propto(\sum_{CA}v_{CA}\log(v_{CA}))\times N_{\max}$。

当不确定性较高时,保留的样本数量也相对较大,使得来自于多个动作类的样本得以呈现且可以避免样本数量的缺失;当不确定性较低时,样本都可能属于相同或少量的动作类,因此只要少量的样本就足够进行准确的推理,且保存较少的样本也能够加速推理过程。

4 实验

本文采用单视角数据库 Weizmann[13] 数据集对方法进行验证。该数据集包含了 Bend(弯腰)、Jack(原地跳步)、Jump(跳步前行)、P_jump(原地纵跳)、Run、Side(横向步行)、Skip(单腿跳行)、walk、wave_s(单手挥舞)和 wave_d(双手挥舞)共 10 个动作类别,由 9 个行为者在单一静态背景下完成的 93 个动作视频。每个视频又包含了行为者多个连续实例的演示。在实验中,我们将实例进行自动分段处理,并根据行为者将实例动作序列分割为训练集合测试集,也就是说动作模型在行为者动作序列集的一个子集上进行训练,其他部分进行测试。每个动作模型由视频语义标注获得。动作特征权重是随机进行初始化的,在训练集上达到最高的那个权重将被用于测试。

在实验设置方面,由于参考了文献[4]中算法,使得本文中的学习动作模型的训练要求较低。我们按照 1:7 的训练和测试比开展实验。在推理过程中,我们为每个动作设置的最小样本数 Nmin 为 3,在任意帧的最大样本数 $N_{\max}$ 为 15。根据基于熵的样本设置选择,每一帧实际的样本数在 3 到 15 之间变化,且每帧所保存的平均样本数为 7。为了评估本文所提出的方法的性能,我们计算了动作分类的精确度及姿势估算误差,并得出了算法在整个数据集进行动作识别的混淆矩阵。图 4

显示了按1:7训练测试比的混淆矩阵。从实验结果上看,由于部位探测器的精确度问题使得Jump类动作识别度不高,其他动作(包括含有遮挡姿势的动作如:walk类)都得到了较为准确的识别。

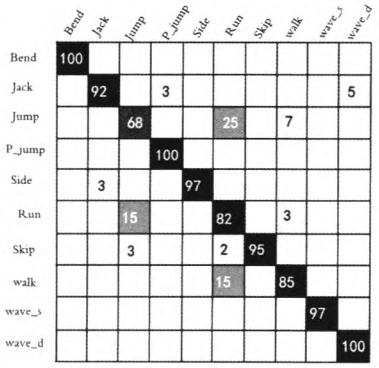

图4 关于数据集的混淆矩阵

Fig. 4 Confusion matrix on dataset

5 结束语

通过二维部位模型对每个身体部位执行局部调整和定位搜索,实现了更为精确的人体姿势跟踪,提高了监控视频对人体姿势识别及行为理解的同步处理效率。本文中涉及的算法还可以用于监控视频中异常行为模式的发现,如通过训练建立正常行为模式或直接训练异常模式来实现监督的和非监督的异常行为检测过程。

参考文献:

[1] 罗青山. 面向视频挖掘的视觉内容分析 [D]. 上海 : 上海交通大学,2009.

[2] V SHET, S N PRASAD, A ELGAMMAL. Multi - cue exemplar - based nonparametric model for gesture recognition[C]. ICVGIP, India: Gwalior, 2004(4):21 - 22.

[3] P NATARAJAN , R NEVATIA. View and scale invariant action recognition using multiview shape - flow models [C]. Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008) , USA: Anchorage, 2008(6):24 - 26.

[4] P NATARAJAN, V K SINGH, R NEVATIA . Learning 3D action models from a few 2D videos for view invariant action recognition[C]. Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2010) , USA :San Francisco, 2010(6):13 - 18.

[5] 代科学, 李 强 , 李国辉. 视频挖掘研究进展 [J]. 计算机科学,2010, 37 (10): 11 - 15.

[6] 赵海勇,刘志镜,张 浩. 基于模板匹配的人体日常行为识别[J]. 湖南大学学报,2011,38(2):88 - 92.

[7] 谷军霞, 晓 青,王生进. 基于人体行为3D模型的2D行为识别[J]. 自动化学报,2010,36(1):46 - 53.

[8] M COLLINS. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms[C]. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), USA:Philadelphia,2002(7): 6 - 7.

[9] F LV , R NEVATIA. Single view human action recognition using key pose matching and viterbi path searching [C]. Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), USA :Minneapolis , 2007(6): 18 - 23.

[10] P F FELZENSZWALB, D P HUTTENLOCHER. Pictorial structures for object recognition[J]. International Journal on Computer Vision . 2005, 61(1):55 - 79.

[11] L SIGAL, M J BLACK. Measure locally, reason globally: Occlusion - sensitive articulated pose Estimation [C]. Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), USA: New York, 2006(6):17 - 22 .

[12] VIVEK KUMAR SINGH, RAM NEVATIA. Human action recognition using a dynamic bayesian action network with 2D part models [C]. ICVGIP, India: Chennai, 2010(12):12 - 15.

[13] L GORELICK, M BLANK, E. SHECHTMAN, et al. Actions as space - time shapes [J]. T - PAMI 2007, 29(12):2247 - 2253.

(责任编辑 刘存英)