多摄像机监控中基于贝叶斯因果网的人物角色识别

明安龙 马华东 傅慧源
(北京邮电大学智能通信软件与多媒体北京市重点实验室 北京 100876)

摘 要 很多传统视觉监控的研究工作集中于行人跟踪、行为和事件检测、步态或人脸识别等,然而角色识别却研究较少.针对多摄像机监控中角色识别的应用问题,该文作者提出了一种基于贝叶斯因果网的角色识别方法.该方法不仅用到了通常的一些人物视觉特征,而且还考虑了时间特征、空间统计特征和一些其它特征.作者将这些特征向量的概率分布参数化,特征向量成员之间的因果关系通过有向无环图的方式来表达,然后通过提取的特征来计算概率以识别人物角色.实验的结果证明了方法的有效性.

关键词 监控;多摄像头;角色识别;贝叶斯因果网
中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.02378

Bayes Causal Network Based Method for Role Identification in Multi-Camera Surveillance

MING An-Long MA Hua-Dong FU Hui-Yuan
(*Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing* 100876)

Abstract Many works of conventional surveillance are focused on people tracking, behavior or event detection, gait or face based recognition, etc. However, role identification is also very important but usually paid less attention. In video surveillance, video analysis and video data mining, it is better for us to treat detected or tracked people with different strategies considering their roles. This paper proposes a multi-camera system to identify people with specific roles using a causal network. Not only visual features but also spatio-temporal features, and some object specific features are involved in the new system. Multiple cameras benefit locating the position of moving objects and overcoming occlusions. Experimental results demonstrate the effectiveness of the method.

Keywords surveillance; multiple cameras; role identification; Bayes causal network

1 引 言

我们将一个人"角色"理解为其在视觉监控场景中伴随着某种责任的表现形式,如医生和病人、售货员和顾客等.传统的监控活动一般是跟踪行人和识别可疑的行为,如 DOTS[1]研究了新颖的监控用户接口,这些接口主要是基于行人跟踪技术.然而,一个人的行为是否可疑,不仅取决于行为本身,还要考虑这个人的角色.举例而言,对于一个大楼工作人员而言拿走一台笔记本电脑或进入一个控制室是正常的事情,但是对一个大楼访问者而言这些行为却非常可疑.因此,角色识别在视频监控中有比较重要的意义.不过在过去的工作中,角色识别讨论得比较少,比较接近的工作是 Gao 等[2]和 Hung 等[3]提出的会议主持人检测.

收稿日期:2010-08-22.本课题得到国家"八六三"高技术研究发展计划项目基金(2009AA01Z305)、国家自然科学基金(60833009,60903072)资助.**明安龙**,男,1979 年生,讲师,主要研究方向为多媒体系统与网络、计算机视觉. E-mail: minganlong@bupt.edu.cn. **马华东**,男,1964 年生,教授,博士生导师,主要研究领域为多媒体系统与网络、传感器网络、网格计算、形式化技术等. **傅慧源**,男,1986 年生,博士研究生,主要研究方向为计算机视觉.

某人的角色通常被视为高层信息而隐藏在多媒体数据中.高层信息一般不直接在数据中表明,但是这些信息有可能通过一个自动化过程被提取出来[4].角色识别将不同的实体与不同的角色一一对应起来.在监控视频、视频分析和视频挖掘中,识别或辨别一个人物所扮演的角色,从某种意义上讲是一个语义信息提取的过程.这些角色可能是室内的清洁员、大楼保安员或商场售货员等.我们可以总结出这些角色的一些共有的特点,而这些特点一般是这些角色的工作性质所决定的:

(1)他们通常穿着职业相关的工作服,如大楼保安员一般穿着保安专用服装、佩戴通话机或手持警棍;

(2)他们工作的场所通常也是视频监控关注的场景,即他们和其他人一样混杂在监控视频中;

(3)他们的活动区域通常是相对固定的,如售货员通常位于柜台附近,清洁员有自己负责的清洁区域,而大楼保安员有自己固定的巡视区域;

(4)他们出现在被监控的场所中时间长度和频率是有规律的,如保安巡视的时间比较有规律;清洁员一般在早上、中午、晚上人少的时候做清洁等.

角色识别在视觉监控、视频分析和视频数据挖掘中非常重要.在视觉监控中,那些特殊角色的人往往被认为是拥有较高或者较低安全系数的实体.从实时监测中减少系统负载的角度出发,我们可以忽略这些人或者相反地仅仅集中注意力在这些人身上.从视觉监控的语义提取角度,一些行为可能对普通人来讲是禁止的(如购物者进入工作车间),但是这些行为对特殊角色的人来说却是允许的,如果不考虑角色的话可能会产生错误的警报.同样的,在视频分析和视频数据挖掘中很有必要排除那些特殊角色的人,因为不排除的话往往得出的统计结果也会是错误的,如某场所的人员计数、大楼保安员或清洁员等很容易被重复计数,而实际上他们不是我们所关注的过客.

视频中的角色识别不同于传统的目标识别、目标分类和事件检测.视频中的事件检测是从视频中检索"故事",属于新的或是事先未定义的事件识别而不是一个人的角色识别.换言之,视频中事件检测是一个无监督的学习任务(没有标定的训练样本),而视频中的角色识别则是一个推理的过程.目前,基于视频的事件检测方法一般是基于预先定制的语法规则或者语义描述,通过一些变量的概率分析来达到分类或者识别的目的.

此外,角色识别和传统的目标识别的区别或差异可总结如下:

(1)传统的目标识别或分类是基于观察特征的,而角色识别则是基于一个推理过程;

(2)角色识别相对于传统的目标识别而言属于更高层的语义信息,这意味着目标识别的结果可能被用于角色识别;

(3)在传统的目标识别中,经常使用的仅仅是相对底层的特征,如视觉特征、音频特征等;而角色识别不仅考虑底层特征,而且包括其它高层特征;

(4)传统的目标识别中,特征向量的各个成员变量经常是独立统计的;而角色识别中所使用特征向量的各个成员变量则往往是统计相关的,我们必须考虑成员变量之间的因果关系.

在本文中,我们提出了一种使用贝叶斯因果网来进行多摄像机监控环境下的人物特殊角色识别的方法.多视角将有利于移动目标的定位和克服遮挡.本文方法的特点如下:

(1)综合视觉特征、时空特征和其它特征用于角色识别;

(2)特征向量中成员变量之间的因果关系通过基于贝叶斯因果网的有向无环图的形式表示;

(3)视觉特征和一些统计特征从存在重叠视域的多摄像机监控的视频中获取以获得更大场景的表示.

2　相关工作

角色识别试图从多媒体数据记录中提取高层语义信息,因为人物角色通常并没有直接在数据中表示出来而需要一个推理的过程才能得到.相似的工作还包括 Barzilay 等[5]提出了一个基于词汇特殊性的方法来进行电台广播中的角色识别,这种方法使用的是音频特征;Javed 等[6]描述了考虑视觉通道的技术来划分电视节目片段中主持人和嘉宾的一种方法(视觉通道指的是一种图像序列端到端的传输方式),但是这种方法主要依赖于传输过程中的某些信息而不是对视频场景的内容进行分析来得到角色识别的结果;Vinciarelli[4]提出了一种电台广播新闻的参与者中说话人角色识别的方法,例如主持人和嘉宾,该方法使用了伯努利分布来进行角色的建模和识别;Hung 等[3]提出了一个系统架构来检测一个多人会议中起主导作用的那个人,在这个架构中使用了不同的音频和视频线索,Hung 等独立地提取每种测量值,然后选择那个总分最高的人作为整个会场的主持人.总体而言,角色识别是一件具有挑战性的事情,它需要在内容理解的基础上进行分析

和判断.按照所采用特征种类的不同,角色识别的相关工作大致可被划分为下面3类:

(1)基于文本的方法.首先将相关文件转换成文本,然后基于文本内容进行检索.这类方法相对简单,文献[5]就属于这种方法.然而,通常包含在视频中的内容比较文本而言要丰富得多,主观上和客观上都很难转换成文本的描述.

(2)基于组合底层特征的方法.这类方法单独考虑不同的底层特征并将这些特征组合起来,文献[3]就是属于这类方法.然而在这类方法中,不同特征与待识别角色之间、特征之间的相互关系被忽略了.

(3)基于高层特征的方法.这类方法试图提取若干高层信息作为特征来进行角色识别,如身份、情绪状态、有说服力的内容和各种不同类型的事件等.文献[4,6]等就是属于这一类方法.然而这类方法存在一些不确定性,一是特征的层次越高,当特征提取错误时对结果的负面影响越大;二是如何将这些高层特征有效合理地组合起来.

在本文提出的方法中,不仅是底层的视觉特征,还有时空统计特征以及其它的高层特征都被用于角色识别.在某些情况下我们可以安全地直接假设特征之间是否相关.我们用有向无环图来表达这种因果依赖关系.当一个分类器使用的各种特征之间的依赖关系是不可知的时候,我们会认为它们之间的条件是独立的.本文给出一个贝叶斯因果网的模型样例和相应的条件概率,然后通过已知的证据就可以对角色的结果进行推理.

3 方法概述

角色识别问题有一些重要的特点:(1)它是一个多角色分类问题;(2)我们可以利用的事实是有限的;(3)即使这些有限的事实在某些时候也可能部分缺失,缺失的部分对识别没有贡献.我们尝试选择一个模型将所有这些有限的事实融合在一起.这个模型不仅仅可从目前已知的有限事实中形成一个最好的角色识别结果,而且当部分事实缺失的时候依然可以知道怎么去推断.

3.1 角色特征

为了实现角色识别,需要知道一些有相互依赖关系的特征,这些特征有的是从视频中提取得到的,有的是从先验知识得到的.为了从视频中区分角色,必须基于一些事实和证据.通过观察我们发现:

(1)活动时间.通常监控场所中特殊角色的人活动时间是非常有规律的,如主要集中在上班时间;

(2)热点区域.特殊角色的人活动比较有规律,如经常在某些固定的地方停留、徘徊;

(3)频率规律.特殊角色的人通常在一段时间内反复出现于一些区域;

(4)衣着特征.特殊角色的人一般身穿工作服;

(5)随身物品.特殊角色的人身边通常有一些特殊物品,如大楼保安员经常会携带对讲机或者警棍.

3.2 贝叶斯因果网的选用

本文选择贝叶斯因果网作为执行角色识别的模型.因果网是一系列随机变量的联合概率分布的表示,它由两个部分组成:(1)一个有向无环图,其中节点表示随机变量而边缘线表示变量之间的依赖关系;(2)每个变量的条件概率分布.相较于其它方法,如非单调逻辑[7]、D-S证据理论[8]和模糊逻辑[9],因果网在解决推理问题方面更具优势.

它是严格的数学语言,有完全的语义表达能力,适合计算机处理[10];

它直观易懂,可以很方便地扩展到多角色分类;

提供了一个推理模型,当某些事实缺失的时候,相关缺失证据对于被查询变量的影响会减弱;

此外,Laskey等[11]曾经提出使用一个因果网来解决战斗机身份识别中固有的不确定性、复杂性和短时间限制的问题,但该方法主要利用的是雷达信息.

3.3 本文的解决方案

本文角色识别解决方案的框架图如图1所示.该方案逻辑上主要分为两个部分:特征提取部分和

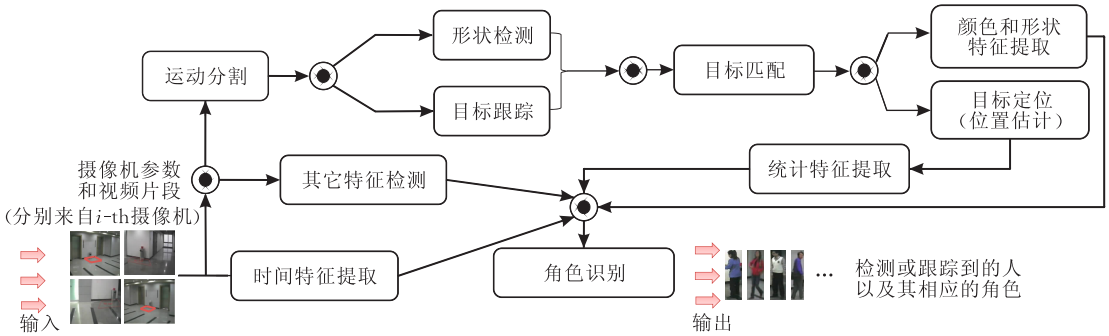


图1 角色识别解决方案的框架图

角色识别部分.在特征提取阶段,我们提取时间特征、统计特征、颜色形状特征以及其它特征.所有这些特征由贝叶斯因果网进行融合,输出的结果是检测出的或跟踪到人以及他们的角色.

4　特征提取与分析

4.1　时间特征

通常,特殊角色的人会有一个相对稳定的工作时间表,这与具体的工作环境有关.在不应该出现的时候特殊角色的人却出现在场景中,对于监控而言是需要关注的.如我们通过观察发现一个大楼清洁员每天大体的活动规律如表 1 所示.

表 1　一个清洁员的日工作时间表

时间区间	活动
07:00 — 08:00	经常
08:00 — 18:00	有时
18:00 — 19:00	经常
19:00 — 08:00	极少

如果将表 1 所示的工作时间表视为先验知识,我们可以依据古典概率理论来进行时间特征概率分布的初始化.

4.2　统计特征

统计特征所依据的一个事实是特殊角色的人拥有相对固定的活动区域.根据这一现象,我们在监控视频中划定一些“热区”,如本文在垃圾桶附近划定“热区”,然后计算监控目标在监控场景逗留期间进入“热区”的次数和停留时间.特殊角色的人因为职责所限,与“热区”相关的这个统计值会很高.而要提取统计特征,我们需要对场景进行俯视图建模、目标跟踪、目标定位及统计分析.

(1)俯视图建模.本文选取某教学楼 9 层的大厅作为实验场景.图 2 以俯视图的形式展示了摄像机在场景中的部署情况.本文按图 2 所示建立场景的世界坐标系(Z 轴正方向向上).“热点”区域设置在垃圾桶附近,有了俯视图模型后,我们就可以通过目标跟踪和目标定位来确定运动目标是否在“热区”之中,最后确定统计特征.

(2)目标跟踪.目标跟踪采用基于隐马尔可夫模型的粒子滤波算法,并在跟踪的过程中考虑跳帧的负面影响,利用一个全局的运动检测器来修正跟踪;不同视角的目标匹配算法采用基于区域 SIFT 的目标匹配算法,具体算法描述参见文献[12].

通过目标跟踪算法,我们得到的输出是一个人体的矩形区域.这个矩形区域在目标匹配和目标定位中将作为输入.目标跟踪将人体的矩形区域作为输出结果,不妨选取人体矩形框的下边中点为接触点,也就是用于计算的特征点.当然,这样选取的特征点肯定引入了误差,图像分析所勾勒出的人体范围矩形框下底边中心并不是精确的人体与地面接触

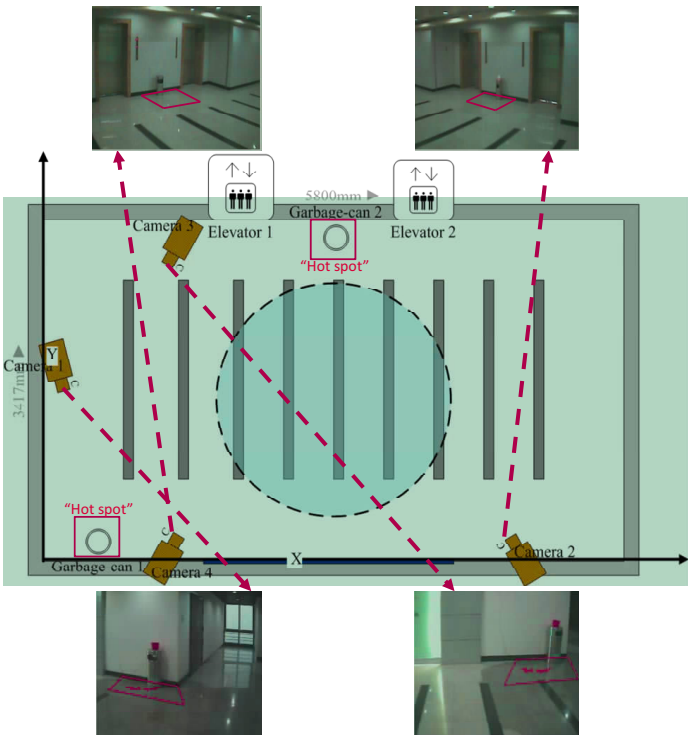


图 2　实验场景的俯视图建模

范围的中心点，特别是当人处于移动状态下时，误差还会更大，我们在下文的目标定位中需要充分考虑到这个误差.

(3)目标定位. 由于大多数情况下人体目标是位于场景中地面上的，从计算机视觉的世界坐标系来看，俯视图下的目标定位只需关心 X 和 Y 轴坐标，Z 轴坐标在需要三维重建的时候才是必须的. 如果将人体看作一条垂直于地面的直线，定位时关心的 X 和 Y 轴坐标就是这条直线与地面的交点坐标 $(X,Y,0)$. 此时目标定位演变成估算人体与地面接触的点，即人脚的位置. 在实际场景中，人与地面接触的范围远不止一个点，所以需要确定在这个接触范围内选取一个具有代表性的点来进行计算.

现在介绍计算特征点 p，也就是人与地面的接触点的方法. 根据立体视觉原理，用点 p 的图像点坐标 (u_p,v_p)，可以得到从摄像机光心 C 引出的射线 Cp. 双目视觉的定位需要从另一幅图像中找出 p 的对应点 p'，根据另一台摄像机的光心 C' 也引出一条射线 $C'p'$，联立方程求两条射线的交点以得到 p 的世界坐标. 在这里，由于事先引入了一个约束条件：p 点必在 $Z=0$ 平面上，所以计算 p 点世界坐标的几何意义就演变成了求 Cp 这条射线与 $Z=0$ 平面的交点. 具体实现上，摄像机内外参数估计我们使用 Zhang 等[13] 提出的方法，该方法的代码已经作为 Intel公司 OpenCv 开源视觉处理库的一部分，其中涉及到的标定物是一个黑白相间的国际象棋棋盘，我们使用的是大的棋盘模型(1.05m×1.05m). 这样我们就可以得到 4 个摄像机分别对应的平移视点间的 4 个单应矩阵. 最后，我们使用 Sankaranarayanan 等[14] 提出的方法对视角重合的两两摄像机之间的目标位置的估计结果进行融合，以得到目标定位结果.

(4)统计分析. 为了进行运动目标的角色识别，我们对该运动目标在监控场景的“进入”状态到“离开”状态这个时间区间的空间位置分布进行统计. 我们约定：一个从前面的监控视频跟踪到当前视频的目标称为 Tracked Object，简称 TO；而从当前视频开始分割出来的目标称为 Detected Object，简称 DO. 则一个 TO 的统计特征(Statistical Location Feature，SLF)的计算如算法 1 所示.

算法 1. 统计特征提取.

输入：TO，DO，TO 相应的 *SLF* //TO，DO 为矩形区域

输出：TO 的 *SLF* //*SLF* 为整数

Begin

1. if (只存在一个 TO，并且这个 TO 对应到一个 DO)
2. if (TO 的空间位置在“热区”中)
3. Then $SLF \leftarrow SLF+1$, return *SLF*.
4. endif
5. else if(存在多个 TO 对应到一个 DO)
//说明发生群聚或遮挡
6. if(DO 的空间位置在“热区”中)
7. Then $SLF \leftarrow SLF+1$, return *SLF*.
8. endif
9. else if(存在一个 DO 没有任何一个 TO 与之对应)
//说明有新的目标加入
10. Then 设这个 DO 为新的 TO，新 To 的 *SLF* 置为 0.
11. endif
12. return *SLF*.

End.

在实际的系统中每秒会采集至少 10 帧以上的数据，逐帧的处理没有必要而且无法满足实时性的需求. 实际处理中可采用输入一个时间间隔，每个时间间隔中抽样一帧进行统计.

4.3 颜色形状特征

特殊角色的人可能包括大楼清洁员、大楼保安员、售货员、送货员等. 这些人的一个共同特点是穿着统一的职业服装. 基于这种先验知识，我们可以考虑从视频中提取一些底层视觉特征来帮助我们进行推理和判断. 首先，我们使用一个简单的分类器去区分人和其它的移动目标. 其次，如果一个移动目标被确定识别为人，我们会给出一个 *YCbCr* 颜色空间的一个统计分析，这个分析的结果可以作为底层颜色特征. 我们使用一个垂直方向的投影直方图[15] 来区分任何其它移动的目标，前提是一个人的形状可以由其投影直方图表示. 设 $I(x,y)$ 是一个检测出来的运动区域的二值图，x 和 y 分别是代表水平和垂直方向的坐标. 设 H 和 W 分别是这个分割出来的运动区域的高和宽，Y 轴的垂直投影直方图是相同水平坐标的所有像素个数总计，这样我们可以得到 S：一个垂直投影直方图的分布.

$$S=\frac{\sum_{x=1}^{W-1}|h(x+1)-h(x)|}{\sum_{x=1}^{W}h(x)} \tag{1}$$

其中 h 是垂直投影直方图，其结果是由前景中的像素投影到垂直方向得到.

$$h(x)=\sum_{y=1}^{H}I(x,y) \tag{2}$$

其中 $1\leq x\leq W$，$1\leq y\leq H$. 通常来讲一个孤立的人的垂直投影从外表上看要比其它非人目标的垂直投影要显得陡峭，因此其分布 S 的值也要大于其它非人目标. 同时一个非人的移动目标的前景像素的边

缘往往要比一群人的前景像素的边缘平滑,因此一群人的分布 S 的值也要大于其它非人目标.基于以上分析,我们可以确定一个阈值来区分人和非人目标.

在一个被跟踪的矩形区域内,我们可以给出一个 $YCbCr$ 颜色空间的统计.如:我们知道一个大楼清洁员通常穿着一件蓝色的制服,我们首先将图像从 RGB 颜色空间转化至 $YCbCr$ 颜色空间,给定一个阈值向量(Cb_l,Cb_h,Cr_l,Cr_h)、一个跟踪的矩形区域 Rv_i,则我们可以得到颜色特征值 CS:

$$CS=|\{p|p\in Rv_i, Cb_l\leq \text{“}Cb \text{ value of } p\text{”}\leq Cb_h] \text{ and } Cr_l\leq \text{“}Cr \text{ value of } p\text{”}\leq Cr_h\}| \quad (3)$$

其中,p 为像素点,CS 可理解为形状为人体的区域中服装区域主体颜色的统计.

4.4　补充特征

除了上述的几个特征以外,识别特殊角色的人还有其它一些“蛛丝马迹”可以参考.如一个大楼清洁员一般会推着一个用于盛放垃圾的手推车,或者拿着扫帚、拖把等物品.本文中,我们使用一个前景提取并结合支撑向量机的算法来检测手推车.

4.5　一个因果网实例

给定标定的节点 T,S,R,C,A(T 代表时间段特征,S 代表统计特征,C 代表颜色形状等特征,A 代表其它特征,R 代表人物角色).每个节点表示一个系统成员,其取值是离散的(每个系统成员的变量以它相应的小写字符来表示).设 e(与 T,S,C,A 相应)代表从多摄像机视觉监控视频中提取出来的证据.那么问题是:我们如何设计一个贝叶斯因果网模型,该模型可以求解条件概率 $P(r|e)$?

设条件概率 $P(r|e)$ 表示在证据 e 的情况下人物为某种角色的概率.每个节点可能的取值是离散的,如 R 可以取 $r_1=$“清洁员”,$r_2=$“保安”,$r_3=$“行人”.在这些离散的状态,$P(r|e)$ 可能有连续取值.在贝叶斯因果网中,每个链接是有方向的并连接两个节点,链接的方向代表了一个节点对另一个节点的因果关系.

我们考虑通过以下的证据来判别人物角色的状态:(1)时间特征.什么时间特殊角色的人最有可能出现?(2)统计特征.什么地方特殊角色的人最有可能反复出现?(3)颜色和形状特征.是人还是物?人的衣着的颜色?(4)其它特征.周围有什么东西?

这些特征之间的关系我们可以根据知识来推断,如时间特征和统计特征不存在着因果联系,而角色类型和颜色特征之间存在因果联系.上述证据中,知道(1)、(2)和(3),则可推断人物角色;而若人物为大楼清洁员的话,必有证据(3)、(4).所以人物角色依赖于(1)、(2)和(3),证据(4)依赖于人物角色.这些特征之间可认为是条件独立的.

根据以上分析,我们设计了一个贝叶斯因果网模型实例(图 3).该模型拥有 5 个节点,分别用 5 个大写英文字符表示,每个节点相关的状态用相应的小写字符表示,如节点 T 拥有状态$\{t_1,t_2,\cdots\}$,这些状态总体简单表示为 $\boldsymbol{t}$;节点S拥有状态$\{s_1,s_2,\cdots\}$,这些状态总体简单表示为 $\boldsymbol{s}$.简单概率可表示为$P(\boldsymbol{t})$、$P(\boldsymbol{s})$、$P(\boldsymbol{c})$和 $P(\boldsymbol{a})$,条件概率可表示为 $P(\boldsymbol{r}|\boldsymbol{t})$、$P(\boldsymbol{r}|\boldsymbol{s})$、$P(\boldsymbol{c}|\boldsymbol{r})$和 $P(\boldsymbol{a}|\boldsymbol{r})$.其中,$P(\boldsymbol{t})$等可以表示一个成员为 $P(t_i)$的概率值向量,$P(\boldsymbol{r}|\boldsymbol{t})$可被描述为一个以 $P(r_i|t_j)$为元素的概率值矩阵.

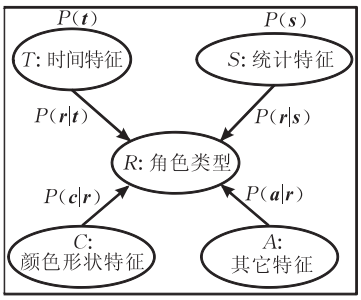


图 3　一个贝叶斯因果网模型

设 $e_X,e_Y,\cdots$等表示从多摄像机监控视频中提取出来的 $X,Y,\cdots$等节点的证据.这样问题就可以转化为求解 $P(\boldsymbol{r}|e_{T,S,C,A})$.

从图 3 可知:

$$P(\boldsymbol{r}|e_{T,S,C,A})\propto P(\boldsymbol{r}|e_{T,S})P(e_{C,A}|\boldsymbol{r}) \quad (4)$$

而我们有

$$P(e_{C,A}|\boldsymbol{r})=P(e_C,e_A|\boldsymbol{r})=P(e_C|\boldsymbol{r})P(e_A|\boldsymbol{r}) \quad (5)$$

但是 $P(\boldsymbol{r}|e_{T,S})$则要复杂得多:

$$\begin{aligned}P(\boldsymbol{r}|e_{T,S})&=P(\boldsymbol{r}|e_T,e_S)\\&=\sum P(\boldsymbol{r}|t_1,t_2,\cdots,s_1,s_2,\cdots)\cdot\\&\quad P(t_1,t_2,\cdots,s_1,s_2,\cdots|e_T,e_S)\end{aligned} \quad (6)$$

其中 $P(t_1,t_2,\cdots,s_1,s_2,\cdots|e_T,e_S)$可进一步求解:

$$P(t_1,t_2,\cdots,s_1,s_2,\cdots|e_T,e_S)=\prod_{i=1}^{|T|}P(t_i|e_T)\prod_{i=1}^{|S|}P(s_i|e_S) \quad (7)$$

根据上述的式(4)~(7),我们就可以求出 $P(\boldsymbol{r}|e_{T,S,C,A})$,其结果最后还需要归一化.

5　实验结果与分析

本文给出的实验是一个大楼清洁员的角色识别.下面从大楼的楼道多摄像机监控的具体应用背

景出发，描述一下角色识别的任务和需求：因为北京高校老师和学生防盗意识不强，经常发生高校内笔记本电脑或者其它物品失窃事件. 因此在大楼室内智能视觉监控中，如何实时地发现可疑人物和事件，或者事后从监控录像中挖掘出可疑人物和事件，是一项很大的挑战，这将大大地节约人力和时间成本. 要达到这样的目标，必须首先区分可疑人物和安全人物. 众所周知，楼道清洁员虽然在楼道中形迹可疑（在某些时段频繁出现、可能拿走一些物品、徘徊在实验室门口等），但是我们在监控和事后挖掘中需要把他们排除掉.

实验使用 4 个 D-Link 的 DSC-5300 型无线网络摄像头参与目标定位. 实验中采用标定块的方法[13]对 4 个摄像头内参数进行了分别标定. 实验中所使用的视频分别由 4 个摄像机分别从不同的视角拍摄，时间跨度约为 5 天. 然而这些视频中的很大部分没有人出现在场景中，我们随机选取了 44 个视频片段（总共约 113 人次出现），图 4 显示了一些片段的截图）来进行网络学习和角色识别，并且将识别的结果与手动标注的真实的结果进行比较. 这 4 段视频分成 11 个组（＃1～＃11），每组 4 段视频，分别对应 4 个摄像机拍摄的内容.

图 4 实验所用视频素材

5.1 特征状态

上文中我们已经构建了一个贝叶斯因果网，下面给出模型中节点状态的设置. 节点 R 表示角色的种类，本文中仅进行大楼清洁员的识别，所以 r_1 代表大楼清洁员，r_2 代表行人. 节点 T 表示活动时间，$t1$：[07：00，08：00]，$t2$：[08：00，18：00]，$t3$：[18：00—19：00]，$t4$：[19：00，07：00].

节点 S 表示统计特征，s_1：$SLF \geq \theta$，s_2：$SLF < \theta$. 节点 S 如表 2 所示.

表 2 节点 *S*

	条件	r_1	r_2
s_1	$SLF \geq \theta$	$P(r_1 \mid s_1)$	$P(r_1 \mid s_1)$
s_2	$SLF < \theta$	$P(r_1 \mid s_2)$	$P(r_2 \mid s_2)$

节点 C 颜色形状特征，c_1：穿蓝色衣服的人，c_2：穿其它颜色衣服的人. 节点 C 如表 3 所示.

表 3 节点 *C*

	c_1	c_2
条件	穿蓝色衣服的人	穿其它颜色衣服的人
r_1	$P(c_1 \mid r_1)$	$P(c_2 \mid r_1)$
r_2	$P(c_1 \mid r_2)$	$P(c_2 \mid r_2)$

节点 A 表示其它特征，a_1：场景中至少有一个手推车且离人物较近，a_2：场景中至少有一个手推车且离人物较远，a_3：场景中没有手推车. 节点 A 如表 4 所示.

表 4 节点 *A*

	a_1	a_2	a_3
条件	存在手推车（近）	存在手推车（远）	不存在手推车
r_1	$P(a_1 \mid r_1)$	$P(a_2 \mid r_1)$	$P(a_3 \mid r_1)$
r_2	$P(a_1 \mid r_2)$	$P(a_2 \mid r_2)$	$P(a_3 \mid r_2)$

表中“远”和“近”的定义如下：

“近”：某人处于视角 1 或 3（分别对应摄像机 1 或 3），同时手推车也位于视角 1 或 3；某人处于视角 2 或 4，同时手推车也位于视角 2 或 4；

“远”：某人处于视角 2 或 4，但手推车位于视角 1 或 3；某人处于视角 1 或 3，但手推车也位于视角 2 或 4.

具体参数估计的结果将在下面讨论.

5.2 网络参数估计

因果网在给定一些变量值的情况下是最为有效的，而其网络参数估计是得到所有变量的概率分布，如各种证据. 然而在本文的这种情况下，我们首先需要得到变量的条件概率表. 一般来说，条件概率的分布可以从数据分析中估计得到，当然某种情况下也可以通过问题性质的分析得到. 这里我们的任务是在一个有条件的参数限制下（实验中 $\theta=8$）进行参数估计，并且因果网的网络结构已知. 为了完成这个任务，我们使用最大似然估计（因为本文是小样本训练）.

考虑一个由 n 个变量 $X=\{X_1, X_2, \cdots, X_n\}$ 组成

的因果网$\mathcal{N}$.不失一般性,设其中节点X_i共有r_i个取值,其父节点$parent(X_i)$的取值共有q_i个组合.若X_i无父节点,则$q_i=1$.那么网络参数为$\vartheta_{i,j,k}=P(X_i=k)|parent(X_i)=j)$,其中$i$的取值范围是$1\sim n$,而对于一个固定的$i,j,k$的取值范围分别是从$1\sim q_i$和$1\sim r_i$.用$\boldsymbol{\vartheta}$表示所有$\vartheta_{i,j,k}$组成的向量.

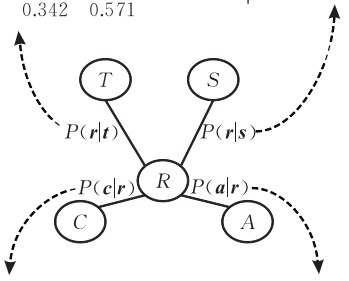
设$\mathcal{N}=(\mathcal{S},\vartheta_{\mathcal{N}})$表示一个因果网,其中$\mathcal{S}$为网络结构.用$P_{\mathcal{N}}(X)$表示$\mathcal{N}$的联合概率分布.在$\mathcal{N}$中,把参数替换成最大似然估计$\vartheta^*$,得到另一个贝叶斯网$\mathcal{N}=(\mathcal{S},\vartheta^*)$,将其联合概率分布记为$P^*(X)$.设$\overline{P}(X)$是基于数据组$\mathcal{D}$的经验分布,即对$X$的任一取值$x$,$\overline{P}(X)=\dfrac{\mathcal{D}\text{中满足 }X=x\text{ 的样本数目}}{\text{样本量 }m}$,则依据文献[16],当$\overline{P}(parent(X_i)=j)>0$时有

$$\vartheta_{i,j,k}=P^*(X_i=k|parent(X_i)=j)=\overline{P}(X_i=k|parent(X_i)=j) \tag{8}$$

由公式(8),我们就可以通过小样本学习得到网络参数.实验中我们使用8个组(#4～#11,共有54人次)来进行参数估计.参数估计的结果如图5所示.

	r_1	r_2	r_3
t_1	0.846	0.613	0.164
t_2	0.111	0.125	0.889
t_3	0.714	0.872	0.286
t_4	0.429	0.342	0.571

	r_1	r_2	r_3
s_1	0.765	0.665	0.245
s_2	0.162	0.217	0.838



	c_1	c_2
r_1	1.000	0.000
r_2	0.000	1.000
r_3	0.375	0.625

	a_1	a_2	a_3
r_1	0.50	0.364	0.136
r_2	0.35	0.545	0.105
r_3	0.25	0.281	0.469

图5　使用最大似然估计的网络参数估计结果

5.3　比较与分析

前文中我们设置了参数,并根据参数推导出了每个角色的概率,参数的取值可以通过有监督的学习得到.下面比较我们的方法与使用时间特征、统计特征、颜色形状特征或者其它特征进行角色识别的识别率.实验中使用t_1、t_2、t_3 3个时段的3份样本视频(#1,#2,#3)作为测试集,t_4时段大楼监控场景中人很少,通常很长时间都见不到人出现,所以没有考虑.

从图6、图7比较的结果可知我们的方法得到了最高的召回率和准确率,这说明我们算法的性能相对比较突出.此外,本文的方法即使某些特征的信息缺失(按缺省值),也能得出一个角色识别的结果,避免了此种情况下不能识别的问题.从图6知我们方法的召回率未超过80%,这是因为约有17.4%人次没有被系统检测到或者跟踪到.这是由于目标检测和跟踪精度的问题所致.

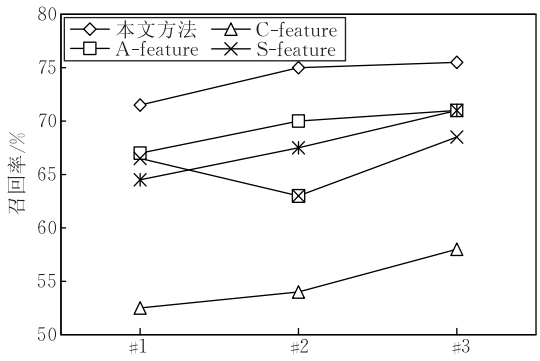


图6　召回率的比较

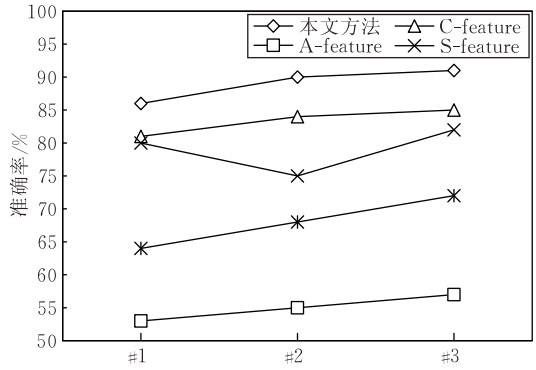


图7　准确率的比较

6　总结与将来的工作

本文提出了一种使用贝叶斯因果网来进行多摄像机监控环境下的人物角色识别的方法.多视角将有利于移动目标的定位和克服遮挡.而在因果网的框架下,本文的方法可适用多角色识别.然而,特征提取的误差会对角色识别的结果造成负面的影响,怎样定量的分析这种影响、并且进一步的克服或减小这种影响还需要深入的研究.此外,当角色的种类增加时,如何构建可靠的因果网模型也是我们将来的工作.

参　考　文　献

[1] Girgensohn Andreas et al. DOTS: Support for effective video surveillance//Proceedings of the ACM International Conference on Multimedia. Augsburg, Germany, 2007: 423-432

[2] Gao Xinbo, Tang Xiaoou. Unsupervised video shot segmentation and model-free anchorperson detection for news video story parsing. IEEE Transactions on Circuits and Systems for Video Technology, 2002, 12(9): 765-776

[3] Hung H et al. Using audio and video features to classify the most dominant person in a group meeting//Proceedings of the ACM International Conference on Multimedia. Augsburg, Germany, 2007: 835-838

[4] Vinciarelli A. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. IEEE Transactions on Multimedia, 2007, 9(6): 1215-1226

[5] Barzilay R et al. The rules behind the roles: Identifying speaker roles in radio broadcasts//Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence. Austin, 2000: 679-684

[6] Javed D, Rasheed Z, Shah M. A framework for segmentation of talk and game shows//Proceedings of the International Conference on Computer Vision. Canada, 2001: 532-537

[7] Donini F M et al. Nonmonotonic reasoning. Artificial Intelligence Review, 1990, 4: 163-210

[8] Shafer G. A Mathematical Theory of Evidence. Princeton, NJ: Princeton University Press, 1976

[9] Zadeh L A. Fuzzy sets. Information and Control, 1965, 8(3): 338-353

[10] Pearl J. Reasoning with belief functions: An analysis of compatibility. The International Journal of Approximate Reasoning, 1990, 4(5/6): 363-389

[11] Laskey K, Laskey G. Combat identification with Bayesian networks//Proceedings of the 7th Annual Command and Control, Research and Technology Symposium. 2002: 1-13

[12] Ming An-Long, Ma Hua-Dong. Region-SIFT descriptor based correspondence between multiple cameras. Chinese Journal of Computers, 2008, 31(4): 650-661(in Chinese)
(明安龙,马华东. 多摄像机之间基于区域 SIFT 描述子的目标匹配. 计算机学报, 2008, 31(4): 650-661)

[13] Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11): 1330-1334

[14] Sankaranarayanan A C, Chellappa R. Optimal multi-view fusion of object locations//Proceedings of the IEEE Workshop on Motion and Video computing. Copper Mountain, Co, 2008: 1-8

[15] Zhao T, Nevatia R, Lv F. Segmentation and tracking of multiple humans complex situations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(9): 1208-1221

[16] Zhang N L, Guo H. Introduction to Bayesian Networks. Beijing: Science Press, 2006

MING An-Long, born in 1979, lecturer. His research interests include multimedia system and network, computer vision.

MA Hua-Dong, born in 1964, professor, Ph. D supervisor. His research interests include multimedia systems and networking, grid computing, sensor networks and formal method.

FU Hui-Yuan, born in 1986, Ph. D. candidate. His research interests focus on computer vision.

Background

This work is supported by the National Natural Science Foundation of China under grant Nos. 60833009 and 60903072; the National High Technology Research and Development Program (863 Program) of China under grant No. 2009AA01Z305.

In many places, multiple surveillance cameras have been installed and it is possible to have constant human monitoring of multi-view video streams. Multi-view surveillance systems exploit multiple video streams to enhance observation capabilities in detecting objects in the scene. Multiple cameras provide a wider coverage of the scene and redundant data that help solve occlusions and improve accuracy. In most cases, the attention of a surveillance system is focused on people moving in the scene, aiming to track people, extract all the possible visual information and identify suspicious behaviors. However, we should refer to not only the behavior itself but also the role of people. For instance, somebody's wandering on an artificial meadow in a public space is very suspicious but normal when his role is a city gardener. Role identification is also important but paid less attention in conventional video surveillance. Role is usually viewed as high level information hidden in data. Many approaches are designed to extract high level information from videos. Role identification is the subtask of information extraction for dealing with the assignment of specific roles to various entities and it requires a inferring process. To identify a person's role, to some extent, is a question of retrieving semantic information. It needs to be pointed out that the term "role" is a subjective concept to some extent. The goal of object classification is to identify the type of an individual entity, e. g. , a person or a vehicle. Generally, object classification can be attempted by using shape and color information. In the above problem, we need more visual information of detail and spatiotemporal information. This has led to difficulties to a live vision system. This paper presents problems of building a vision system for role identification at a hall environment. It involves the challenges faced by a role identification system, which is assumed to have a representation of people tracking, consistent labeling, data acquisition and fusion. The objectives here are to give consistent labeling of people and recognize there different roles.