可变场所的异常行为识别方法

张 军[1),2)] 刘志镜[1)]

[1)](西安电子科技大学计算机学院, 西安 710071) [2)](石家庄职业技术学院信息工程系, 石家庄 050081)

摘 要 在视觉分析中,人的同一动作在不同场景下会有截然不同的理解。为了判断在不同场景中行为是否为异常,在监控系统中使用双层词包模型来解决这个问题。把视频信息放在第 1 层包中,把场景动作文本词放在第 2 层包中。视频由一系列时空兴趣点组成的时空词典表示,动作性质由在指定场景下的动作文本词集合来确定。使用潜在语义分析概率模型(pLSA)不但能自动学习时空词的概率分布,找到与之对应的动作类别,也能在监督情况下学习在规定场景下运动文本词概率分布并区分出对应异常或正常行动结果。经过训练学习后,该算法可以识别新视频在相应场景下行为的异常或正常。

关键词 异常行为 兴趣点 双层词包 潜在含语义分析概率 可变场所

中图法分类号: TP391.41 **文献标识码**: A **文章编号**: 1006-8961(2009)10-2097-05

A Method of Abnormal Action Recognition in Variable Scenarios

ZHANG Jun[1),2)], LIU Zhi-jing[1)]

[1)](*School of Computer Science and Technology, Xidian University, Xi'an* 710071)

[2)](*Department of Information Engineering, Shijiazhuang Vocational Technology Institute, Shijiazhuang* 050081)

Abstract Different understanding results in different scenarios even for the same person to conduct visual analysis. In order to determine whether the behavior is abnormal in different scenarios, a double-layer bag-of-words model is proposed to solve the problem in surveillance system. The video information is processed in the first layer of Bag-of-Words, and the information of scenario-action text words is included in the second one. A video sequence is represented as a collection of spatial-temporal codebook by extracting space-time interest points. A behavior characteristic is represented as a collection of behavior text words in special scenarios. Probabilistic latent semantic analysis (pLSA) model is adopted to automatically learn the probability distributions of spatial-temporal words and the topics correspond to human action categories. PLSA also can learn the probability distributions of the motion text words in a scenario with supervisor and the topics correspond to anomalous or normal actions. The algorithm can categorize the human anomalous or normal action contained in the special occasion to a novel video sequence after being trained.

Keywords abnormal action, interest points, double-layer bag-of-words, probabilistic latent semantic analysis, variable scenarios

1 引 言

对视频内容的理解是目前机器视觉中最活跃的研究主题之一,其核心是利用图像处理、分析等技术从视频序列中检测、识别运动物体并对其行为进行描述与理解,在高级人机交互、智能监控及图像检索等方面具有广泛的应用前景和潜在的经济价值[1]。

基金项目:广东省教育部产学研结合项目(2006D90704017);河北省教育厅科研项目(Z2009127)

收稿日期:2009-06-19;**改回日期**:2009-07-18

第一作者简介:张 军(1972 ~),男。高级工程师,西安电子科技大学计算机应用博士研究生。研究方向视觉计算、监控系统。E-mail:zhang72jun@163.com

智能监控系统要用计算机协助人来完成监控任务，其中对运动目标的异常识别是监控中的重点和难点，国内外学者针对特定的场景提出了统计方法、物理参数的方法、时空运动的方法和模型的方法[2]。统计方法对噪声具有较强鲁棒性和运算量小的优点；物理参数的方法易于理解、不依赖于观测的角度，但依赖于被恢复场景的参数；时空运动的方法能表现出时间和空间的特性，具有计算复杂度低，使用简单，但易受噪声干扰；模型的方法具有难于从视频获取精确的模型和运算量大的问题。

目前监控系统只是针对特定场景内运动目标的检测或跟踪，一旦场景发生变化，原来的算法就失效，需要重新设计。本文结合行人走路的特点，采用具有统计特性和时空特性的时空兴趣点来表示物体的移动，使用 pLSA 来自动学习时空词的概率分布；考虑到场景的因素，提出了一种基于半监督的可变场景异常行为识别方法。

2 运动目标表示

物体的运动是时空信息的变化，常用的移动目标表示方法有基于边缘的静态特征、基于光流的动态特征和从局部获得的时空特征，基于时空兴趣点的方法能提供丰富的描述和表示。本文就采用时空兴趣点来表示目标的运动[3]。

2.1 特征表示

每段视频序列由抽取的时空兴趣点组成时空词。由于使用常规的兴趣点检测器所获取的特征信息太稀疏，为了能更好地表示复杂视频，采用可分线性过滤器[3]的方法。设摄像机是固定的或可以估计它的运动，视频中可分线性过滤器的响应函数：

$$R=(\boldsymbol{I}*\boldsymbol{g}*\boldsymbol{h}_{ev})^2+(\boldsymbol{I}*\boldsymbol{g}*\boldsymbol{h}_{od})^2 \qquad (1)$$

式中，$\boldsymbol{I}$ 为图像，$g(x,y;\sigma)$ 是 2 维高斯平滑核，适用于空间维度 (x,y)；$\boldsymbol{h}_{ev}$ 和 $\boldsymbol{h}_{od}$ 是应用在时间维度上的 1 维正交滤波器，$\boldsymbol{h}_{ev}=(t;\tau,\omega)=-\cos(2\pi t\omega)\times e^{-t^2/\tau^2}$，$\boldsymbol{h}_{od}=(t;\tau,\omega)=-\sin(2\pi t\omega)\times e^{-t^2/\tau^2}$。$\sigma$ 和 τ 这两个参数分别对应检测器的空间和时间尺度，为了简化计算设定 $\omega=4/\tau$；为了执行多尺度，必须在一系列空间和时间尺度执行检测器。

Dollár[3] 发现在任何区域中，空间分辨特性经过复杂的运动可以引起明显的反应，然而在经过单纯平移或无空间明显特性不会引起显著的反应，因此在响应函数最大值的区域提取时空兴趣点。图 1 是行人走路的兴趣点检测图。每个白色方块对应检测到的兴趣点。兴趣点区域大小由检测器尺度参数 σ 和 τ 确定。从图 1 可以看到兴趣点是位于发生明显变化的位置。通过检测兴趣点，产生视频序列的稀疏表示。从视频片段抽取到的兴趣点，构成了局部信息用于学习和目标分类。更为重要的是其局部模式有充分区分于其他对象的特征，并能提供合理的特征空间，从而建立良好的运动模式。此外这种做法不需做诸如背景减除等全局的特性预处理。

图 1 行走时兴趣点实例

Fig. 1 A case of walking interest point detection

要取得每个时空体的描述符，就需要计算出在 x，y 和 t 3 个方向的亮度梯度。在图像梯度计算之前在不同阶上平滑时空体。计算的梯度形成一个向量。向量的大小等于立方体像素点数乘以平滑阶数，再乘以梯度方向数。使用主成分分析把描述符投影到低维空间中。Dollár[3] 使用了不同的描述符，如规格像素值、梯度亮度和窗口光流。梯度亮度和窗口光流都能有效地描述运动信息[4]。

本文采用了基于图像梯度的描述符。这个的描述符既有空域尺度不变性，也有时域尺度不变性。更加复杂的描述符需要复杂的计算，因而使用代码本来处理尺度变化和摄像头的移动。

2.2 代码本形式

本文使用两个代码本。一个是视频代码本，另一种是文字代码本。所有视频序列词存储在视频代码本中，运动文本词包含在文本代码本中。

潜在的主题模式 pLSA[4] 适用于容量为 V 的词汇表，为了学习词汇表中的时空词，假定在训练数据集中，描述符集合对应所有检测到的时空兴趣点。利用 k-means 算法和欧氏距离作为聚类尺度来生成代码本[4]。每个聚类的中心定义一个时空词。这样每个检测到的兴趣点被划分到唯一的类成员，也就是时空词，因而一段视频就可以从代码表中找出一个时空词的集合与之对应。在文本字典中，潜在主题模型 pLSA 依赖于 M 个场景运动词。

3　利用双层词包识别异常行为

为了解决在不同场景中对行为是否为异常的判定,在监控系统中使用双层词包模型。

3.1　框架结构

图 2 把视频序列信息放在第 1 层包中,把行为的文本词放在第 2 层包中。视频序列由一系列时空兴趣点组合成时空信息词的集合来表示。行为特征是由在特定场景下的动作文本词的集合来表示。在分类识别过程中使用潜在语义分析概率模型来自动学习时空词的概率分布,并能找到主题词与相应的行动类别。同时也学习在规定场景下运动文本词概率分布并区分出相应的异常或正常行动结果。

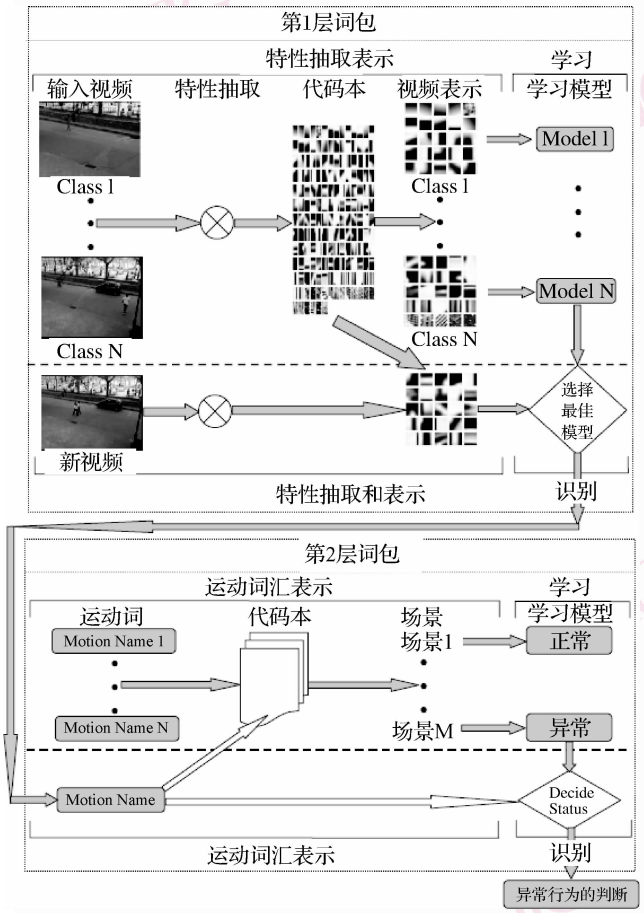

图 2　学习识别流程图

Fig. 2　Learning and recognizing flow chart

3.2　使用 pLSA 学习识别异常

LSA 是通过获取和表示词之间的关联意思来对长文本进行解读的一种理论方法。下面把 pLSA 用在文本和视频中。假设有 $N(j=1,\cdots,N)$ 视频序列包含词汇表 $V(1,\cdots,V)$ 个时空词。视频的集合可以表示为 $V\times N$ 的共生表 $\boldsymbol{N}$,其中 $m(w_i,d_j)$ 表示在视频 d_j 出现时空词 w_i 的数量,一个潜在的主题变量 z_k 与在视频 d_j 中出现时空词 w_i 相联系。每个主题对应的行动类别,如慢跑、下蹲等。联合概率 $P(w_i,d_j,z_k)$ 的图模型由图 3 水平部分表示。节点表示随机变量。带阴影的节点是观察变量,不带阴影的是隐含量。其中,d 代表视频序列,z 是行动类别和 w 是时空词。模型的参数使用最大期望值来进行无监督数的学习[4]。

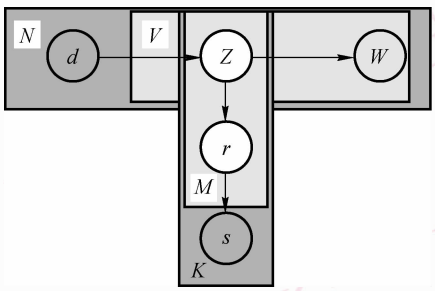

图 3　混合潜在含语义分析概率

Fig. 3　The hybrid probabilistic latent semantic analysis

$$P(d_j,w_i)=P(d_j)P(w_i\mid d_j) \tag{2}$$

观察值(d_j,w_i)假定是独立的,可以忽略主题词 z_k 来获取条件获取条件概率 $P(w_i\mid d_j)$。

$$P(w_i\mid d_j)=\sum_{k=1}^{K}P(z_k\mid d_j)P(w_i\mid z_k) \tag{3}$$

式中,$P(z_k\mid d_j)$ 的概率是主题词 z_k 发生在视频 d_j 中,$P(w_i\mid z_k)$ 是时空词 w_i 划分为动作 z_k 的概率。现在有 K 个主题词,也就是有 K 个动作。这个模型对每个视频序列作为 K 个行为种类的凸组合,即通过表面凸组合或行为概率 $P(w_i\mid z_k)$ 得到指定视频词分布 $P(w_i\mid d_j)$。使用权重特定因子 $P(z_k\mid d_j)$ 来表示特定的视频。每视频段作为混合行为类别建模,由混合柱状图组成特殊视频的柱状图对应运动种类。然后找到能适应模型直方图 $P(w_i\mid z_k)$ 和混合系数 $P(z_k\mid d_j)$ 来确定运动种类。为了确定出现在视频中时空词最大可能性概率模型,对目标函数使用最大期望值算法来得到参数的最大相似度[4]估计值:

$$\prod_{i=1}^{V}\prod_{j=1}^{N}P(w_i\mid d_j)^{m(w_i,d_j)} \tag{4}$$

从训练序列获得指定动作类别视频词分布 $P(w\mid z)$,给定一段新视频,这段未被理解的视频通过 $P(w\mid z)$ 投射到简单的平面。接着找出混合系数

$P(z_k \mid d_{\text{test}})$，它是位于经验分布 $\tilde{P}(w \mid d_{\text{test}})$ 和 $P(w \mid d_{\text{test}}) = \sum_{k=1}^{K} P(z_k \mid d_{\text{test}}) P(w \mid z_k)$ 之间的 KL 收敛[5]。使用最大期望值算法找到最终的答案。分类结果就是通过选择行为分类来最好地解释观察值：

$$Action\ Category = \arg\max_k P(z_k, d_{\text{test}}) \tag{5}$$

在文本词识别分类中也用 pLSA 方法对场景动作文本词进行相应的学习和分类，联合概率 $P(z_i, s_j, r_k)$ 的图模型由图 3 垂直方向表示。对应的变量如表 1 所示。

表 1 变量之间对应关系

Tab. 1 The relationship between variable and variable

	视频		文本	
	表示	说明	表示	说明
学习视频、文本数量	V	视频帧数	M	场景动作数
识别视频、文本数量	N	视频帧数	K	场景动作数
序列表示	d	新视频序列	s	新场景
视频、文本集合	w	视频集合	z	动作种类集合
主题词	z	动作种类	r	行为异常正常

最终结合具体场景的行为公式：

$$Result = \arg\max_m P\left(r_m, \frac{P(w_i \mid z_k) P(z_k \mid d_j)}{\sum_{l=1}^{K} P(w_i \mid z_l) P(z_l \mid d_j)}\right) \tag{6}$$

3.3 算法实验

为了测试所提算法的可行性，使用了 5 个场景在实时监控系统中进行测试，分别是会议室、街道、旁边的汽车、宾馆大堂和实验室。

在建立视频代码本时，采用 CASIA 行为分析数据库[6]，CASIA 行为分析数据库共有 1 446 条视频数据，是在室外环境下由在 3 个不同视角的摄像机拍摄而成，数据分为单人行为和多人交互行为，单人行为包括走、跑、弯腰走、跳、下蹲、晕倒、徘徊和砸车。本文采用室外摄像机的俯视方式，单人行为。因数据集中只包含单一行动视频序列，这将使我们能够在其他的场景中使用这些视频作为模板。使用上面介绍的方法提取时空兴趣点，检测参数 $\sigma = 1.5$，$\tau = 2$。每个时空片段由相关的向量和时空梯度描述。然后把描述符映射到低维空间。为了建立代码本，需要把所训练视频系列特征描述进行聚类。然而由于训练样本的特征数量庞大，因而只能使用学到代码本的一个子集作为特征集。最后使用 pLSA 来学习和识别行为分类。

为了测试视频代码本在识别中的有效性，在每一类取一组在学习中没使用到视频片段，图 4 给出了使用 3 900 代码本的混合矩阵。在训练时 pLSA 算法自动分配每个测试序列的一个行为类别。每一行的混合矩阵对应于行为分类，每一列对应到指定集。另外我们还测试识别率和代码本大小的关系，图 5 这表明识别精度依赖承认字典代码本的大小。

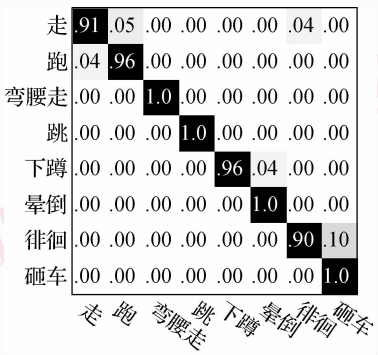

图 4 CASIA 识别混合矩阵

Fig. 4 Confusion matrix for the CASIA dataset

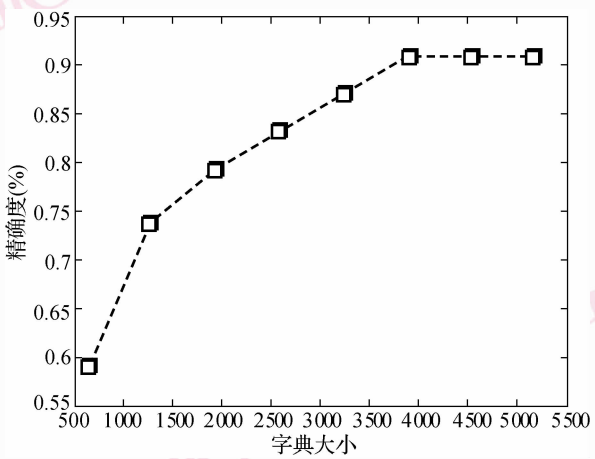

图 5 精确度与字典大小关系

Fig. 5 Classification accuracy vs. codebook size

4 结 论

在文本的学习和识别阶段，从监控中学习知识。从实时视频中，由专家统计的各种行为在特定场景发生的数量，并确定行为的正常或异常。这样可以得到一组发生在会议室、街道、路边、汽车边、宾馆大堂和实验室的动作行为词。建立好视频和文本代码

表后,用学习到的知识测试实时监控系统,在特定的场景中找到 CASIA 库中 8 种动作的正常或异常。其结果不但是能直接判断这 8 种动作,而且还可以间接确定由这 8 种动作组合的行为的正常性或异常性,如打斗、盗窃、踢车、撬锁等。从监控系统中的实测中可以得到的 86.2% 识别率。达到甚至超过了一些著名监控系统[7]的识别率。在一般场景(会议室,街道,宾馆大堂和实验室),很容易判断异常行动,但在某些需要智能的场合(汽车旁边),判断异常行为不是很准确。实验结果表明,使用本文方法能较好地解决不同场所下异常行为的判断问题。

参考文献(References)

1 Elgammal A, Duraiswami R, Harwood D, *et al*. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance [J]. Proceedings of the IEEE, 2002, **90**(7): 1153-1163

2 Hu Wei-ming, Tan Tie-niu, Wang Liang, *et al*. A survey on visual surveillance of object motion and behaviors[J]. IEEE Transactions on System, Man and Cybernetics, 2004, **34**(3): 334-352.

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

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

5 Hofmann T. Probabilistic latent semantic indexing [A]. In: Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval [C], California, USA, 1999, 50-57.

6 CBSR. Center for Biometrics and Security Research [EB/OL]. http://www.cbsr.ia.ac.cn/china/Action%20Databases%20CH.asp, 2008-04-13/2009-03-21.

7 Haritaoglu I, Harwood D, Davis L S. W4 real-time surveillance of people and their activities [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, **22**(8): 809-830.