利用级联 SVM 的人体检测方法

李同治 丁晓青 王生进
(清华大学电子工程系,智能技术与系统国家重点实验室,北京 100084)

摘 要 从图像中检测出人体是计算机视觉应用中的关键步骤。通过一个由简到繁的级联线性 SVM 分类器将级联拒绝的机制与梯度方向直方图特征相结合,实现了一个准确和快速的人体检测器,整个检测器由级联的线性 SVM 分类器组成。实验结果表明,在保持 Dalal 算法检测准确性的同时,大幅的提高了检测速度,每秒平均可以处理 12 帧左右的 320×240 的图像。

关键词 行人检测 风险敏感 SVM 分类器 由简到繁的检测器
中图法分类号:TP391.4 **文献标识码**:A **文章编号**:1006-8961(2008)03-0566-05

Human Detection with a Coarse-to-fine Cascade Linear SVM

LI Tong-zhi, DING Xiao-qing, WANG Sheng-jin
(*State Key Laboratory of Intelligent Technology and Systems, Department of Electronic Engineering, Tsinghua University, Beijing* 100084)

Abstract Finding human in images is critical for several applications in computer vision. We combine the cascade-of-rejection approach with the Histograms of Oriented Gradient (HOG) to form a fast and precise human detector. The detector consists of coarse-to-fine linear SVM classifiers. Our experiments show that our method can process average 12 frames with 320 × 240 image per second, while maintaining the comparable accuracy to Dalal's method.

Keywords human detection, cost sensitive SVM, coarse-to-fine detector

1 引 言

人体检测在计算机视觉中有许多重要的应用,例如视频监控、智能汽车及智能交通、机器人和高级人机交互等。然而,由于自身姿态的变化、衣服的多样性和光照等因素的影响,人体的外观变化非常大,导致人体检测是一个非常困难的问题。

近年的文献中出现了大量的人体检测方法。这些方法根据实现的途径可以分为两类。第 1 类是基于局部模型的方法,采取从部件到整体的途径。例如,Papageorgiou 等人将人体分为脸、左臂、右臂和腿 4 个部件,然后分别训练了 4 个部件检测器,最后根据部件之间的几何约束来检测整个人体[1]。Mikolajczyk 等人利用方向-位置联合直方图特征建立了一个基于人脸,头部和肩部等部件的上肢检测器[2]。Ronfard 等人用"图解结构"来描述人体的各个部件之间的关系[3],利用一阶和二阶梯度特征描述各个部件的外观特性,然后用支持向量机(support vector machine, SVM)分类器构造了一个人体检测器,类似的工作还有文献[4]、[5]。Leibe 等人利用关键点提取、Hough 投票和 Chamfer 距离模板匹配的方法[6],建立了一个自底向上和自上向下相结合的人体检测方法。第 2 类方法是基于单一检测窗口的方法,在尺度和位置空间应用分类器判断所有的子图像是否为目标。Papageorgiou 等人提出了利用修改过的 Haar 特征和多项式 SVM 作为分类器的人体检测器[7]。Viola 等人利用 Haar-like 特征和运动特征结合级联的 Adaboosting 分类器,构造了一个快速的视频行人检测器[8]。

基金项目:国家高技术研究发展计划 863 项目(2006AA01Z115);国家自然科学基金项目(60472002)
收稿日期:2007-07-31;**改回日期**:2007-11-29

第一作者简介:李同治(1980~),男。2003 年获得西安电子科技大学学士学位。现为清华大学电子工程系博士研究生。主要研究方向为模式识别和计算机视觉。E-mail:ltz03@mails.tsinghua.edu.cn

最近,Dalal 等人提出了一种性能优异的单一窗口人体检测方法[9]。他们的方法利用小块上的梯度方向直方图(histograms of oriented gradient,HOG)来描述图像,实验结果证明,该描述方法结合 SVM 分类器可以有效地区分出人体和非人体。但是,其缺点是速度慢,每秒钟只能处理 1 帧 320×240 的图像,并且扫描的窗口非常稀疏(两个方向的扫描步长都是 8 个像素),这与实际应用还有很大的距离。

本文提出的方法对文献[9]的方法进行了加速。采取了类似文献[8]提出的级联拒绝机制,用由简到繁多级分类器实现整个检测过程。用相对简单的特征排除掉大量的非人体图像后,采用更加精细的特征作进一步的验证。同时,采用了直方图积分图对特征提取进行了加速。实验结果表明,该方法在保持文献[9]算法准确性的同时,大大提高了检测的速度。

2　梯度方向直方图及 Dalal 的算法

方向直方图特征在计算机视觉领域已经应用了很长时间,但是,直到 Lowe 提出的应用于图像匹配的 SIFT(scale invariant feature transform)特征的出现[10],才得到了比较成熟的应用,Lowe 利用局部梯度方向直方图描述图像块,然后据此匹配具有尺度不变性的特征点。类似的特征还有形状上下文(Shape Context)[11]和边缘方向直方图(edge orientation histograms,EOH)[12]等。Dalal 提出的 HOG 就是类似 SIFT 的图像描述方法。

Dalal 提出的 HOG 与 Lowe 的 SIFT 描述方法之间的区别在于后者是基于关键点检测,是一种稀疏的描述方法。而 HOG 是将图像均匀地分成相邻的小块(cell),然后在所有的小块内统计梯度方向直方图,用这些直方图来描述图像,是一种非稀疏的描述方法。

HOG 特征的计算过程[13]如下:

第 1 步,计算两个方向的梯度,采用简单的[-1,0,1]模板计算每个位置的梯度幅值和方向;

第 2 步,图像按空间位置均匀的分成相邻的小块,称为“cell”,在 cell 内按照设定好的方向量化间隔统计梯度方向直方图,应用梯度的幅值进行投票,然后相邻的 cell(2×2)组成一个大块,称为“block”,相邻的 block 之间相互重叠;

第 3 步,在 block 内采用二范数(L2)归一化直方图来消除光照的影响。一个检测窗口内的所有 block 内的归一化直方图组成最后的特征向量,最后应用线性 SVM 分类器进行判断。

HOG 描述方法有以下优点:HOG 表示的是边缘(梯度)的结构特征,因此,可以描述局部的形状信息;位置和方向空间的量化,在一定程度上可以抑制平移和旋转带来的影响;同时采取在局部区域应用 L2 归一化,可以部分抵消光照带来的影响。实验证明,HOG 用于行人检测性能要优于 Haar,PCA-SIFT 和 Shape context 等方法[9]。另外,该方法在 VOC2006 测试中人体检测上也取得了最好的结果。但是,正如前面已经提到,该方法最大的缺点就是速度慢。为了提高速度,首先在特征提取阶段,引入积分直方图[14]进行加速。这样计算任意一个小块(cell)内的直方图时,每个量化间隔只需 4 次读内存的操作和加减运算。

3　由简到繁的级联 SVM 分类器

基于统计方法的目标检测的基本流程是按照一定比例对图像进行放缩,然后在放缩后的多个图像中用一定尺寸的窗口进行穷举搜索和判别,最后将所有尺度下的检测结果进行融合。以 320×240 像素的图像为例,如果按照 1.2 的尺度步长缩小图像,并以 64×128 大小的窗口和 2 个像素空间步长搜索人体,共需判断 12 000 余个窗口,并且大部分的窗口都是背景图像。面对如此巨大的计算量和搜索空间,为了达到实时的检测速度,分类器必须采用由简到繁的分层级联结构,由前面若干层相对简单的分类器快速排除掉绝大部分的非目标窗口,后面的复杂分类器再进一步排除与目标相似度大的非目标窗口。

线性 SVM 分类器运算简单、推广能力好,并且在 Dalal 的工作中已经证明结合 HOG 特征可以很好的区分人体和非人体[9],因此,分层级联分类器中,每层都采用线性 SVM 分类器,通过逐级增加特征的精度来实现由简到繁,逐级地滤除非人体窗口。检测时由于大部分子窗口都是非目标,是目标的窗口只占很小的比例,把人体窗口判断为非人体(False Negative)的风险要远大于把非人体窗口判断为人体(False Positive)的风险,为此采取了风险敏感的线性 SVM 分类器。

3.1　风险敏感 SVM

SVM 理论主要是由 Vapnik 在 90 年代中期提出的一种新的统计模式分类方法[15]。它以结构化风险最小化(structural risk minimization,SRM)作为优

化准则,在理论上保证了分类器较好的推广性能。线性 SVM 的判别函数为

$$y=\mathrm{sgn}(\boldsymbol{\omega}\cdot x+b) \tag{1}$$

令 $\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i)\mid 1\leqslant i\leqslant N\}$,表示训练集,其中 $y_i\in\{-1,+1\}$ 为训练数据的标签,$\boldsymbol{x}_i$ 为特征向量,则常规的线性 SVM 优化目标为求 $\boldsymbol{\omega}$ 和 b,使其满足

$$\min\left\{\frac{1}{2}\|\boldsymbol{\omega}\|^2+c\sum_i\xi_i\right\} \tag{2}$$

约束条件为

$$y_i(\boldsymbol{\omega}\cdot\boldsymbol{x}_i+b)\geqslant 1-\xi_i \tag{3}$$

$$\xi_i\geqslant 0,\forall i \tag{4}$$

式中,$c>0$ 为惩罚因子,ξ_i 为松弛因子。

前面已经提到,检测问题中,两类的风险是不对称的。采用级联 SVM 分类器,则要求每级在保证足够高检测率的前提下最小虚警率,所以,漏警的风险要大于虚警的风险,不能等同的对待两类错误。在常规 SVM 分类器设计中,并没有区分来自两类的不同错误,而是达到分类间隔最大和总分类错误最小的一个折中,这样必须通过调整阈值才能达到分类器在高风险类别上的满意效果。但是,阈值调整后的分类器已经不是最优意义下的分类器,其性能不能达到最优的分类效果。为此,文中采取了两类风险不对称的风险敏感 SVM 分类器(cost sensitive SVM,CS-SVM)的设计方法[16],克服常规 SVM 分类器设计中存在的不足,将其应用于复杂背景下人体检测这个典型的两类风险不对等的分类问题。而 CS-SVM 的优化目标由式(2)变为

$$\min\left\{\frac{1}{2}\|\boldsymbol{\omega}\|^2+C_p\sum_{i|y_i=1}\xi_i+C_n\sum_{i|y_i=-1}\xi_i\right\} \tag{5}$$

约束条件不变,式中 C_p,C_n 分别为正负样本的风险,一般取 $C_P>C_n$。

3.2 由简到繁的级联 SVM 分类器

采用逐级增加特征精度的方法构造级联分类器。从 HOG 特征的计算过程可以看出,影响特征向量维数的主要参数就是 cell 的大小,cell 的面积越大,特征的维数越低。当然,不是 cell 越小,检测性能就越好,根据文献[9]算法,当检测窗口为 64×128 时,8×8 大小的 cell 是最优的,如果继续减小导致类内变化增大,检测性能反而下降。设定 cell 大小从 32×64 到 8×8。前面几层采用较粗糙的特征(cell 面积大),特征维数低,排除掉相对容易排除的窗口,后面用较精细的特征来区分和行人相似的窗口。具体的训练流程如下:

(1) 给定输入 每级的特征参数(cell 的大小 $w_i\times h_i$ 和两类的权重 C_{pi},C_{ni}),每级的最低检测率 $d_{\min}$,以及级数 l 和虚警率要求 $f_{\min}$。正样本集 $\boldsymbol{P}$ 与负样本集 $\boldsymbol{N}$。

(2) 初始化 $i=0$,$d_i=1$,$f_i=1$,$\hat{f}_i=1$,其中 d_i,f_i 分别为第 i 级的检测率,虚警率,$\hat{f}_i$ 为级联数目为 i 时,整个级联分类器的虚警率。

(3) $i=i+1$,根据当前级的参数 $w_i\times h_i$ 以及 C_{pi},C_{ni} 训练线性 SVM 分类器,调整阈值使检测率 d_i 满足 $d_i\geqslant d_{\min}$,在训练样本上测试第 i 级分类器的虚警率 f_i。

(4) 计算当前整个级联分类器的虚警率 $\hat{f}_i=\hat{f}_{i-1}\times f_i$,如果 $\hat{f}_i<\hat{f}_{\min}$ 或者 $i\geqslant l$,训练结束;

否则,到第(5)步。

(5) 清空负样本集,扫描背景图像,将识别错误的子图像中进行聚类,样本数目最多的一类作为负样本集合,回到第(3)步,继续训练。

在第(5)步中,由于前面几级负样本数目过多,由于受到计算机能力的限制,按照一定的概率均匀采样出部分样本使用。一般情况下,正样本分布相对集中,而负样本则分布比较分散,为此,将负样本聚类减少类其内变化,使每层分类器针对相对集中的负样本,提高单层分类器的性能,通过多个线性 SVM 的组合实现非线性分类面。

4 实验结果及其分析

实验用的数据是 INRIA (the french national institote for reasearch incomputer science and control) Person Dataset[9]。训练数据包括 1 208 个人和对应的镜像图像,总共 2 416 张正样本训练图像,图 1 中是一些训练样本示意图。1 218 张没有人的图像作为负样本。测试数据包括 566 个人对应的 1 132 张正样本图像和 453 张背景照片的负样本。该数据集的特点是衣服、姿态、光照和视角等变化大,是一个难度非常高的数据库。采用的评测办法同文献[9]中一样,即采用 Miss Rate VS. FPPW (false positive per window)曲线来比较。

本文设计了一个 10 级分类器,每级的参数采用如表 1 所示的设置,即采用逐渐减小 cell 的大小来增加特征的维数,实现由简到繁。特征数目指每级用到的大块(block)的个数,其中每块用 36 维的梯度方向直方图来描述,实验中采取了二范数对 block 内的直方图进行归一化处理。

图 1　INRIA 中的训练样本

Fig. 1　Some sample images from INRIA dataset

表 1　级联 SVM 各级用到的特征数目

Tab. 1　Feature number of each stage

级序号	1	2	3	4	5	6	7	8	9	10
特征数目	3	6	10	10	21	21	36	36	55	105

图 2 表示的是拒绝率与级联层数的关系。可以看出,仅仅利用前 4 层已经拒绝掉了 90% 以上的负样本,前 4 层总共用到了 29 个 block。拒绝掉一个负样本,平均需要计算 13.6 个 block,而 Dalal 的方法需要计算 105 个 block。因此,平均计算量是单一线性 SVM 的 1/8。同时,在提取特征时候采用积分直方图,又降低了特征提取的运算负担。对于 320 ×240 的图像,采取缩放图像的方法在尺度空间搜索,检测4 000个窗口平均需要时间 100ms,而Dalal 的方法需要 1s,本文提出的级联方法速度是单一 SVM 的 10 倍。

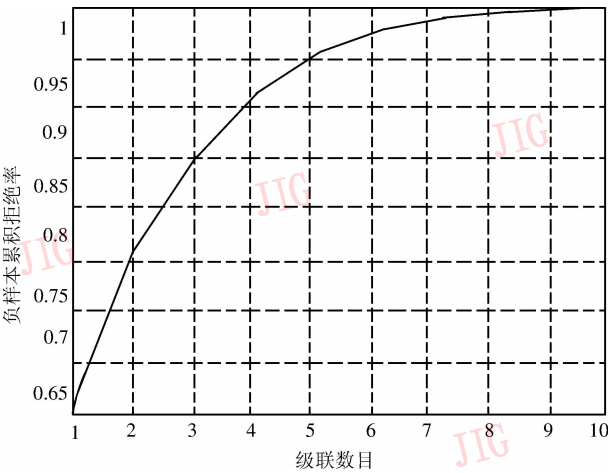

图 2　级联分类器累积拒绝比例示意图

Fig. 2　The accumulated rejection rate over the cascade levels

图 3 中是对数尺度下的 DET(detection error tradeoff)曲线,即 Miss Rate(1-Recall 或 FalseNeg/(TruePos + FalseNeg))对 FPPW(FalsePos/(TrueNeg + FalsePos))曲线。首先逐级增加级联数目得到前面的数据,然后通过调整最后一级的阈值,得到 FPPW小于 10^{-4}时的数据。单一的线性 SVM 结果,是按照文献[9]中的方法和参数实现的结果。在 FPPW 低于 10^{-3}时,级联 SVM 性能要差于单一的 SVM,这是由于在级联的前面几层,在较大的尺度上提取的 HOG 特征,鉴别能力有限;当 FPPW 高于 10^{-3}时,两者的性能非常接近,级联的检测率比于单一的 SVM 低 1%。而在实际应用中,FPPW 高于 10^{-4}才有意义,此时级联方法的检测率非常接近单一的线性 SVM,同时由于检测速度快,可以通过增加检测窗口数目来提高检测率。

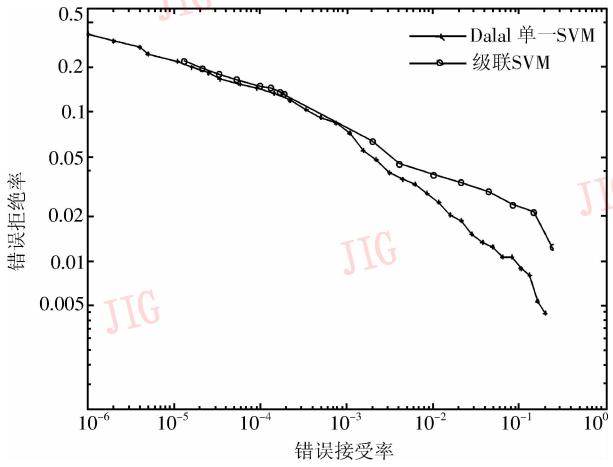

图 3　级联 SVM 与单一 SVM 的性能比较

Fig. 3　Comparing Dalal algorithm and our method

图 4 和图 5 是一些检测结果,采取了类似于人脸检测中的后处理合并方法[17]。其中图 5 是本文收集的数据,与训练样本的相关性差。

5　总　结

通过一个由简到繁的级联线性 SVM 分类器将级联拒绝机制与梯度方向直方图特征相结合,实现了一个接近实时应用的人体检测系统。实验结果表明,系统的性能在保持已有算法检测准确性的同时,加快了检测速度,接近于实时应用。但是,由于前面几层的特征比较粗糙,造成初始阶段漏检率较大,影响了系统的性能。下一步工作,准备通过特征选择方法来挑选出少数的具有鉴别性的特征供前面几级使用,进一步提高系统的性能。

图 4 在 INRIA Person Dataset 测试图像上检测结果

Fig. 4 Results on INRIA Person Dataset

图 5 本文收集数据上的检测结果

Fig. 5 Results on dataset

参考文献(References)

1 Mohan A, Papageorgiou C, Poggio T. Example-Based object detection in images by components [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, **23**(4): 349 ~ 361.

2 Mikolajczyk K, Schmid C, Zisserman A. Human detection based on a probabilistic assembly of robust part detectors [A]. In: Proceedings of 8th European Conference on Computer Vision [C]. Prague, Czech, 2004: 69 ~ 82.

3 Ronfard R, Schmid C, Triggs B. Learning to parse pictures of people [A]. In: Proceedings of 6th European Conference on Computer Vision [C]. Copenhagen, Denmark, 2002: 700 ~ 714.

4 Felzenszwalb P F, Huttenlocher D P. Pictorial stuctures for object recognition [J]. International Journal of Computer Vision, 2005, **61**(1): 55 ~ 79.

5 Ioffe S, Forsyth D A. Probabilistic methods for finding people [J]. International Journal of Computer Vision, 2001, **43**(1): 45 ~ 68.

6 Leibe B, Leonardis A, Schiele B. Pedstrain detection in crowed scenes [A]. In: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition [C]. San Diego, CA, USA, 2005: 878 ~ 885.

7 Papageorgiou C, Poggio T. A trainable system for object detection [J]. International Journal of Computer Vision, 2000, **38**(1): 15 ~ 33.

8 Viola P, Jones M. Detecting pedestrians using patterns of motion and appearance [A]. In: Proceedings of the IEEE International Conference on Computer Vision [C]. Nice France, 2003: 734 ~ 741.

9 Dalal N, Triggs B. Histograms of oriented gradients for human detection [A]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Beijing, China, 2005: 886 ~ 893.

10 Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, **60**(2): 91 ~ 110.

11 Belongie S, Malik J. Matching shapes [A]. In: Proceedings of the 2001 IEEE International Conference on Computer Vision [C]. Vancouver, Canada, 2001: 454 ~ 461.

12 Levi K, Weiss Y. Learning object detection from a small number of examples: the importance of good features [A]. In: Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition [C]. Washington, DC, USA, 2004: 53 ~ 60.

13 Dalal N. Finding people in images and videos [D]. France: the French National Institute for Research in Computer Science and Control, (INRIA), 2006.

14 Porikli F. Integral histograms: a fast way to extract histograms in Cartesian spaces [A]. In: Proceedings of 2005 IEEE Conference on Computer Vision and Pattern Recognition [C]. San Diego, CA, USA, 2005: 829 ~ 836.

15 Vapnik V N. The nature of statistical learning theory [M]. New York: Spring Press, 1995.

16 Ma Y, Ding X. Face detection based on cost-sensitive support vector machines [A]. In: Proceedings of First Internatimal workshop on Pattern Recognition with Support Vector Machines [C]. Niagara Faus, Canada, 2002: 260 ~ 267.

17 Viola P, Jones M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2004, **57**(2): 137 ~ 154.