基于部位检测和子结构组合的行人检测方法

胡 斌 王生进 丁晓青

(智能技术与系统国家重点实验室 清华信息科学与技术国家实验室 清华大学电子工程系 北京 100084)

摘 要 提出了一种基于部位检测和子结构组合的、可用于辅助驾驶或视频监控系统中行人检测的方法。首先使用头部分类器在整幅图像中检测,得到感兴趣区域;然后在每个感兴趣区域内使用头部、躯干、腿部以及左臂和右臂5个人体部位检测器分别检测并使用基于子结构的检测组合方法对部位检测结果进行组合,以得到最终结果。在不同数据库上的实验结果表明,本方法可以有效地用于移动或静止摄像机所拍摄的视频图像中的多姿态及部分遮挡的行人检测。

关键词 视频图像,行人检测,部位检测器,子结构组合

中图法分类号 TP24 **文献标识码** A

Pedestrian Detection Method Based on Part Detector and Substructure Assemble

HU Bin WANG Sheng-jin DING Xiao-qing

(State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract We presented a pedestrian detection method based on part detector and substructure assemble which can be used for driving assistant system or video surveillance system. Firstly we used the head classifier to search the whole image in order to get the ROI(Region of Interesting). Then in each ROI five part detectors including head, torso, leg, left arm and right arm were used individually. Finally we verified the detection result through part detector assemble method based on substructure. Experiments on different database show that our method has high performance in detecting pedestrians with numerous poses and partial occlusion in cluttered background.

Keywords Video images, Pedestrian detection, Part detector, Substructure assemble

1 引言

从静态图像或视频序列中有效、可靠地检测行人是一项很有意义的研究工作,在图像复原与检索、视频监控、辅助驾驶等很多领域都有着迫切的应用需求。但要在现实环境中很好地检测出行人,也面临着相当多的困难:各种不同的姿态、外形、衣着使不同场景中的行人具有很大的差异;行人可能身处的室内环境、城市室外环境、自然环境之间的背景变化也是巨大而普遍的;目标的颜色和光照变化也会给检测带来很大困难;最后,行人与背景以及行人之间的遮挡也是一个普遍存在而且较难解决的问题。因此,如何设计一种能够有效解决上述问题的鲁棒的行人检测算法,就成为计算机视觉相关领域的重点研究内容。图1中给出了一些不同现实场景中包含行人的图像。从图中可以看出,由于环境的不同造成了行人图像之间存在着很大的差异。

早期的行人检测多以直接检测行人图像所具有的相关特征为主,如运动特征[1,2]和图像特征(竖直边缘[3]、局部区域的熵[4]和纹理[5]等),其它还有基于步态分析的检测[6]和基于形状模板的检测[7]。但由于前面提到的种种困难,该类方法得到的检测器通常不具有良好的鲁棒性,很难用于实际的系统检测。在Viola和Jones成功地将基于Adaboost的机器学习方法用于人脸检测[8]并扩展到行人检测之后[9],基于机器学习(主要以Adaboost方法为主)的行人检测方法已经逐渐成为了目前该领域最常用的方法之一。本文的算法就是以Adaboost为基础建立起来的。

图1 现实场景行人图像示例

基于Adaboost算法的行人检测研究中,一个主要方向就

到稿日期:2008-12-11 返修日期:2009-03-24 本文受863国家重点基金项目(2006AA01Z115),973国家重点基金项目(2007CB311004)资助。

胡 斌(1976—),男,助理研究员,主要研究方向为图像处理等,E-mail:binhu@tsinghua.edu.cn;王生进 男,教授,主要研究方向为图像处理等;丁晓青 女,教授,主要研究方向为图像处理等。

是如何创建和选择构成最终检测器的大量底层弱分类器特征。目前比较成功的几种弱分类器特征包括 Haar 小波特征[9]、HOG 特征[10]、Edgelet 特征[11]和 Shapelet 特征[12]，其中小波特征最为简单并容易实现。但由于行人图像复杂多变，单独用小波特征来做整体检测时，很难同时满足检测率和误检率的需要；而后几种方法虽然在性能上都比小波特征有了较大的提高，但也都存在计算复杂、运行时间长的缺点，很难用于辅助驾驶系统之类的实时检测。由于本文的算法同时考虑在视频监控和辅助驾驶系统中的应用，希望尽可能满足实时性的要求，因此采用了最简单的 Haar 小波特征来构成检测器。

目前的行人检测方法主要分为基于整体的检测和基于部位的检测两大类。在待检测图像较简单、图像中的行人遮挡不严重的情况下，基于整体的检测方法通常会有不错的结果，如 Papageorgiou 等的 SVM 检测器[13]、Gavrila 的边缘模板检测器[7]和 Wu 等的 Markov 检测器[14]等。但由于整体检测器需要整个人体的绝大部分部位信息，因此当遮挡出现较多时，该类方法的性能就会出现明显的下降。而在现实场景中，背景物体对人的遮挡或人之间的遮挡是会经常出现的，这就给基于整体检测的算法造成了很多困难，因此基于部位检测的方法逐渐成为目前研究的热点。该类方法首先分别检测事先定义的组成人体的各个部位，然后通过一定的准则将部位检测的结果进行组合，得到最终的检测器。

在基于部位的行人检测方法中，Mohan 等[15]将人体划分为 4 个部分：头肩、腿、左臂、右臂，使用 Harr 小波特征学习 SVM 的检测器。Shashua 等[5]将人体划分为 9 个区域，对每个区域基于方向直方图特征训练各自的分类器。Mikolajczyk 等[16]将人体划分为 7 个部分：正面头脸、侧面头脸、正面头肩、背面头肩、侧面头肩、腿部，对每一个部位，通过类 SIFT 的方向特征学习分类器。但这些方法中对部位的分割和组合多是以先验知识为主，缺乏理论分析，而且对最后的组合检测器能够容忍的丢失部位情况没有一个明确的概念，影响了算法的推广性。文献[17]基于理论分析提出了一种新的模型选择策略，用于学习最优的检测器组合，来得到最小的误检率，同时满足设计需求中对丢失部位的容忍能力。虽然其算法是针对人脸检测提出的，但稍加改进就可以很好地用于行人检测中。

2 部位分割及样本获取

基于部位组合的检测方法，要解决的首要问题就是部位的划分准则，即在什么部位划分以及划分为几个部位。对于人体而言，比较合理和直观的划分位置就是按照头、四肢以及躯干来分离部位。而对于部位的数量，过少的话，无法体现部位检测的目的和优势；过多的话，其组合算法将会趋于复杂而且会增加整体检测的时间。由于本文采用的是基于机器学习的部位检测方法，因此综合考虑检测性能和检测时间并通过大量的实验验证，最后将整个人体划分为 5 个部位进行检测和组合：头部、左臂、右臂、躯干和腿部。下面进行具体介绍。

为了使算法的结果具有通用性并便于同其它的算法进行比较，采用了目前行人检测系统中最常使用的 MIT 行人数据库[18]和 INRIA 数据库[19]，作为产生部位数据的原始数据样本。分割位置及示例图像如表 1 及图 2 所示（左为 MIT 样本，右为 INRIA 样本）。

表 1 部位检测器位置大小

部位	位置(x,y,width,height)	
	MIT 数据库	INRIA 数据库
头部	(17,11,30,30)	(15,8,30,30)
躯干	(17,28,30,45)	(15,28,30,45)
腿部	(17,60,30,60)	(15,52,30,60)
左臂	(12,34,12,40)	
右臂	(40,34,12,40)	

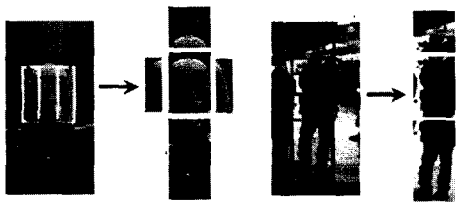

图 2 部位检测器分割图

对 MIT 数据库而言，由于所有行人的位置和姿态都比较规范，因此可以直接统一用于部位分割。现将该数据库中所有 1848 个样本（原始 924 个样本加左右镜像后的样本）中的前 1348 个用于训练，后 500 个用于测试。INRIA 数据库中的样本情况则比较复杂，包含有很多姿态变化较大的行人图像，如果直接统一用于分割，会产生大量错误的样本。因此首先将 INRIA 数据库的训练集样本进行了分类，共得到规范姿态图像（与 MIT 数据库中的基本类似）1181 幅，复杂背景图像 995 幅以及大变化姿态图像 236 幅，而只使用规范姿态的 1181 幅生成部位样本。但就是在这些规范姿态的图像中，大部分行人的左右臂也很难单独分割出来，因此对 INRIA 数据库只分为头部、躯干和腿部 3 个部位。这样，在正样本方面共得到 2529 幅头部、躯干和腿部的训练图像以及 1348 幅左臂和右臂的训练图像。另外，使用从网络搜集得到的 8000 幅不含行人的图像作为负样本图像。

3 感兴趣区域产生

文献[17]主要是针对人脸进行检测。由于人脸图片中人脸所占比例一般比较大而且对实时性的要求并不是很高，因此其算法是使用所有的部位检测器在整个待检测图像中分别进行全图检测，然后将结果进行组合。对行人检测任务而言，这样的时间消耗是很难接受的，因此本文用比较经典的感兴趣区域假设和目标确认的两步过程进行检测。

首先使用基于头部的检测器在全图进行粗检测，如图 3 所示。蓝色检测框便为头部检测结果，通过调整头部检测器的阈值使其检测率基本达到 100%，虽然这样不可避免地产生了大量的误检，但还是会排除掉很大一部分图像空间，从而大大加快目标确认阶段的检测速度。然后按比例扩展检测到的窗口，作为感兴趣区域，如图中绿色框所示。可以看出，在感兴趣区域中行人都被检测到了。当然，由于参数设置的问题导致感兴趣区域中有很多是误检。下一步就是在每个感兴趣区域中采用部位组合方法进行确认，去除掉误检的感兴趣区域，从而得到最终的结果。整个检测的流程如图 4 所示。

图 3　感兴趣区域产生示意图

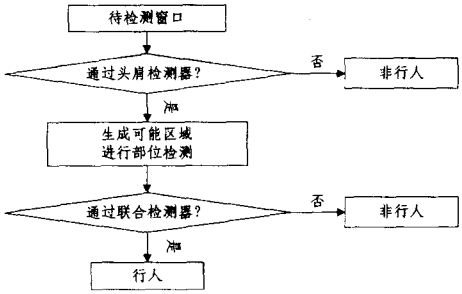

图 4　整体检测流程图

4　部位检测器组合

对基于部位检测的方法而言，由于一般情况下各个部位检测器的鉴别性较差，因此基于部位的检测器通常会产生大量的误检。另外，由于遮挡和/或部分检测器的不完备，可能造成部位的丢失。这些情况都对现有的基于部位的检测器提出了挑战。为了处理误检，通常需要考虑部位内部的几何约束。而如果约束是在所有的部位上(或绝大部分部位上)定义的，则很少数量部位(甚至只有一个)的丢失都会造成约束无效。这样，基于该种约束的检测器在容忍部位丢失方面的性能就不会很好。为了解决该问题，采用了文献[17]提出的子结构概念来描述几何约束。一个子结构就是指几个部位检测器(本文采用 3 个)及其相对大小和位置关系的组合。基于子结构的部位检测器整体组合模型如图 5 所示。

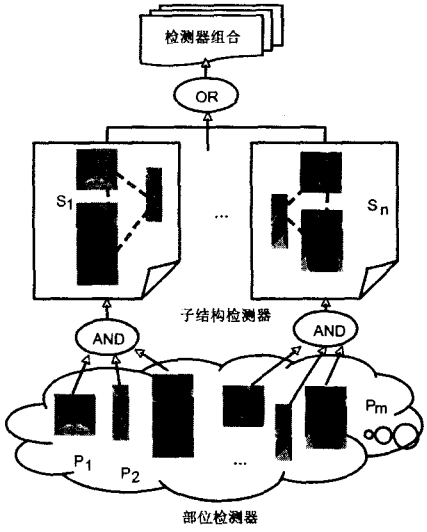

图 5　部位检测器组合结构图

假定目标 o 包含了 m 个部位 $\{p_1, p_2, \cdots, p_m\}$，每个部位 p_i 都和一个部位检测器相关联。而一个子结构检测器是由一个部位检测器集和这些部位之间的约束构成的。只有当包含的所有部位都检测到，而且部位之间的约束满足时，才会认为该子结构检测器被检测到。子结构检测器的一条关键的原则是它不允许部位丢失，也就是说所有相关的部位都必须被检测到才能构成一个有效的子结构检测，这条原则可以在检测中减少误检。最终的检测器组合包含了一个子结构检测器集。另一条主要原则是，只要有一个子结构被检测到，检测器整体组合就会给出一个正确的检测。这条原则可以增加检测到目标的可能性。

4.1　子结构检测器

假定对部位 p_i 得到了 k_i 个检测，称之为部位候选。每个候选目标可以用一个标记状态 $L=\{l_1, l_2, \cdots, l_n\}$ 来表示。这里的 $l_i \in \{1, 2, \cdots, k_i\}$ 表示部位 p_i 的候选。如果 p_i 丢失，则 $l_i=0$。

每个子结构 $S_j=\{P_j, h_j(L_j)\}$ 有两部分，$P_j \subseteq O$ 包含了与该子结构相关的所有部位，L_j 是标记状态 L 在 P_j 上的投影，$h_j(L_j)$ 是决定候选集 L_j 和训练数据拟合程度的决策函数。在本文的行人检测系统中，选择部位检测器的三元组构成子结构，其决策函数如下：

$$h_j(L_j)=\begin{cases}0, & \exists i, p_i \in P_j, l_i=0 \\ \log H_j(L)-\lambda_j, & \text{otherwise}\end{cases}$$

其中，λ_j 是预先定义的阈值，该阈值的选取要求是在训练集上能有非常高的检测率(>99%)。H_j 是从正样本训练数据中计算得到的似然项，定义为三元组子结构中所包含的 3 个部位的内角分布、两两距离之比以及两两尺度比的线性组合，如下式所示：

$$H_j=\alpha_1 G(\theta_1, \theta_2)+\alpha_2 G\left(\frac{S_1}{S_2}, \frac{S_2}{S_3}\right)+\alpha_3 G\left(\frac{L_{12}}{L_{23}}, \frac{L_{23}}{L_{31}}\right)$$

其中，θ, S, L 分别代表部位检测器之间的内角、尺度和距离，$\alpha_1+\alpha_2+\alpha_3=1$。$G$ 代表一个 2D 高斯分布 $G(x)=(x-\mu)^T\Sigma^{-1}(x-\mu)$，其中 μ 为均值向量，Σ 为协方差矩阵。此高斯分布的参数可以从训练样本中学习得到例如对由头部、左臂和右臂 3 个部位组成的子结构，通过训练样本得到的参数如表 2 所列。其它子结构与其类似，在此就不一一列出。在实际检测中，仅当决策函数 $h_j(L_j)$ 的值大于 0 时，该子结构才是有效的。与文献[17]的决策函数相比，去掉了在负样本数据上对结果影响很小的似然项，以简化计算过程。

表 2　训练得到的高斯分布参数

	均值向量	协方差矩阵
内角	(1.157 1.122)	$\begin{pmatrix}0.134 & -0.091\\ -0.091 & 0.129\end{pmatrix}$
尺度比	(0.699 1.183)	$\begin{pmatrix}0.074 & -0.075\\ -0.075 & 0.261\end{pmatrix}$
距离比	(0.918 1.232)	$\begin{pmatrix}0.165 & -0.049\\ -0.049 & 0.125\end{pmatrix}$

4.2　选取最优子结构组合

为了得到最优的结果，采用了覆盖集的概念来选取子结构。最主要的一项选取原则是能够最大程度地容忍部位的丢失。如图 6 所示，对一个具有 5 个部位的检测目标，图 6(a)使用了 4 个二元组的子结构。可以看出，当部位 A 丢失时，所

有的子结构均将检测失败，从而造成目标丢失，因此这个子结构组合方式不合理。同样，图 6(b)使用了 2 个三元组的子结构，也存在部位 A 丢失目标便无法检测的缺点，所以其组合方式同样不合理。图 6(c)给出了一种 3 个三元组的子结构组合方式，可以看出在任何一个部位丢失的情况下，总是还至少存在一个有效的子结构，从而该目标就能够被检测到。该组合方式便是一个覆盖集。

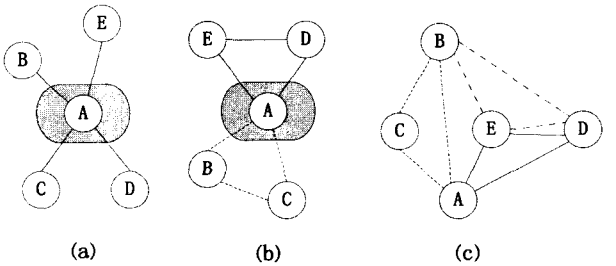

图 6　覆盖集解释图

具体来说，对一个具有 n 个部位的目标，假定有一个包含 m 个部位的子结构集，对任意 t 个部位，在该子结构集中都至少存在一个子集，该子集中的所有部位都包含在这 t 个部位中，则这个子结构集就被称为一个(n,t,m)的覆盖集。有了该覆盖集，则在实际检测中，只要检测到该目标的任意 t 个部位(说明有 $n-t$ 个部位漏检)，该目标都可以被检测到。

在本文的行人检测系统中，要求当上述 5 个人体部位中的任意一个部位没有检测到时，通过其它的部位检测仍然可以准确地检测到行人，因此至少需要 3 个子结构来构成检测器集合($T(5,4,3)=3$)[17]，这远远少于三元组子结构的全部数目($C_5^3=10$)。

为了从 10 个子结构中选取 3 个最优且能构成覆盖集的子结构，采用了一种自底向上顺序选取的办法。下面给出一般情况下该算法的描述。假设总共有 n 个部位检测器，每个子结构都是部位检测器的三元组，因此共有 $m=C_n^3$ 个可能的子结构；又假设要在检测到任意 t 个部位的时候，总有至少一个子结构是有效的。选取的整个算法过程如下：

输入：部位检测器数目 n，子结构集合 $T=\{S_1,S_2,\cdots,S_m\}$，子结构 S_j 的误检率 f_j^S，最少检测到的部位数目 t

输出：最优的检测器集合 $\varepsilon_{opt}\subseteq T$

1)将子结构按照误检率从小到大排序；

2)选取误检率最小的 3 个子结构；

3)检查这 3 个子结构是否构成一个覆盖集，若是，则转到 4)；若不是，则替换其中一个子结构，使得误检率的增量最小，返回 3)；

4)输出最优检测器集合 。

实际检测时，对于每个子结构，其所包含的每个部位检测器都可能会有多个检测结果。如何得到好的组合结果，可以看作是一个组合优化问题。我们采用遗传算法来进行组合选择，首先随机选取各个部位检测器的检测结果，并将其决策函数的输出值作为遗传算法的适应度。经过一定轮数的迭代，可以找到接近于最优的部位检测器组合，并得到它的决策函数值。如果决策函数值大于预先设定的该子结构的阈值，则认为该子结构是有效的。如果至少有一个子结构是有效的，则检测器集合将该样本标记为正例。与文献[17]相比，该组合选择方法在运行速度和选择结果上都有所改善。

5　实验结果

对于前面得到的 5 个部位的训练数据，使用扩展的 Haar 小波特征[20]和经典的层叠式分类器[8]分别进行训练，得到各自均为 15 层的 5 个部位检测器。在测试集上的 ROC 曲线如图 7 所示。将本文算法的结果同普通的 Haar 全身检测器和最新的 Shapelet 检测器进行了比较，从图 7 中可以看出我们的系统的性能有了明显的改善。

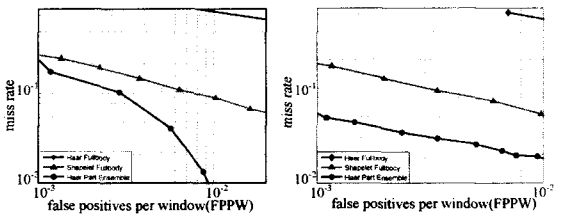

图 7　不同数据集上 ROC 曲线比较图

图 8 给出了我们的算法在不同实际环境图像中的部分检测结果，其中第一行和第二行是 INRIA 数据库中的测试图像，第三行为 CAVIAR 数据库中的测试图像，第四行为实验室自己拍摄的室外行人图像。在每一行中，第一列为产生了感兴趣区域的图像，蓝色框为头部检测器的检测情况，绿色框为根据头部检测结果按比例扩大后产生的感兴趣区域；第二列为使用部位组合检测后满足条件的区域，可以看出各个部位的检测情况；第三列为我们算法的最终检测结果；为了比较，我们在第四列图中给出了基于目前最新的 Shapelet 特征的行人全身检测器的检测结果。可以看出，当出现遮挡或背景较复杂时，基于全身的检测器就会失效，而基于部位的检测器仍可以很好地工作。总之，在各种不同的复杂背景以及存在遮挡的情况下，我们的系统都能够很好地检测出行人，体现了本文算法的鲁棒性。

在 P4 2.8G，2G 内存的机器上，检测一幅 320×240 大小的图片，本文算法在未经过任何后期优化情况下的检测时间低于 500ms，远远小于 Shapelet 算法的检测时间(10s 左右)。可见本文的算法不仅在检测性能上有了大幅度提高，在检测时间上也大大加快。

结束语　本文提出了一种基于部位组合的行人检测方法，通过将行人分解为不同的部位进行检测，来解决实际场景中存在的遮挡问题；通过基于子结构的组合算法来最优化各个部位的总体组合性能；通过感兴趣区域假设和目标确认的由粗到精的两步检测过程来进一步提高检测速度。在不同数据集上的检测结果表明，我们的方法可以有效地用于各种不同场景下的行人检测任务。

当然，由于考虑实时性的要求，本文只采用了最简单的类 Harr 小波特征作为 Adaboost 分类器的底层特征，这在一定程度上影响了系统最后的性能。如果实际任务中对实时性的要求不高，其它一些具有更好性能的特征也完全可以在感兴趣区域假设阶段和/或目标确认阶段应用到的算法框架中，来获得更好的检测结果，这也是我们下一步要继续研究的工作。

（自上到下第 1、2 行为 INRIA 测试图像，第 3 行为 CAVIAR 测试图像，第 4 行为自拍测试图像。从左至右第 1 列为感兴趣区域产生图像，第 2 列为部位检测器结果，第 3 列为最终结果，第 4 列为 Shapelet 特征检测结果）

图 8　现实场景行人检测结果图

参 考 文 献

[1] Ramprasad P, Randal N. Detection and recognition of periodic, nonrigid motion[J]. International Journal of Computer Vision, 1997, 23(3): 261-282

[2] Stein G P, Mano O, Shashua A. A robust method for computing vehicle ego-motion[C]// Proceedings of IEEE Intelligent Vehicles Symposium. 2000: 362-368

[3] Broggi A, Bertozzi M, Fascioli A, et al. Shape-based pedestrian detection[C]// Proceedings of IEEE Intelligent Vehicles Symposium. 2000: 215-220

[4] Curio C, Edelbrunner J, Kalinke T, et al. Walking pedestrian recognition[J]. IEEE Transactions on Intelligent Transportation System, 2000, 1(3): 155-163

[5] Shashua A, Gdalyahu Y, Hayun G. Pedestrian detection for driving assistance systems: single-frame classification and system level performance[C]// Proceedings of IEEE Intelligent Vehicles Symposium. 2004: 1-6

[6] Wohler C, Kressel U, Anlauf J K. Pedestrian recognition by classification of image sequences global approaches vs. local spatio-temporal processing[C]// Proceedings of IEEE International Conference on Pattern Recognition. 2000: 540-544

[7] Gavrila D M. Pedestrian detection from a moving vehicle[C]// Proceedings of European Conference on Computer Vision (ECCV). 2000: 37-49

[8] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2001, 1: 511-518

[9] Viola P, Jones M, Snow D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision (IJCV), 2005, 63(2): 153-161

[10] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2005, 2: 886-893

[11] Wu Bo, Nevatia R. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors[C]// Proceedings of IEEE International Conference on Computer Vision (ICCV). 2005, 1: 90-97

[12] Sabzmeydani P, Mori G. Detecting pedestrians by learning shapelet features[C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2007

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

[14] Wu Ying, Yu Ting, Hua Gang. A statistical field model for pedestrian detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2005, 1: 1023-1030

[15] Mohan A, Papageorgiou C, Poggio T. Example-based object detection in images by components[J]. IEEE Transactions on Pattern Analysis And Machine Intelligence (PAMI), 2001, 23(4): 349-361

[16] Mikolajczyk K, Schmid C, Zisserman A. Human detection based on a probabilistic assembly of robust part detectors[C]// Proceedings of European Conference on Computer Vision (ECCV). 2004, 1: 69-81

[17] Dai Shengyang, Yang Ming, Wu Ying, et al. Detector ensemble [C]// Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2007

[18] http://cbcl.mit.edu/projects/cbcl/software-datasets/PeopleData1Readme.html

[19] http://pascal.inrialpes.fr/data/human/

[20] Lienhart R, Maydt J. An Extended Set of Haar-like Features for Rapid Object Detection[C]// Proceedings of IEEE International Conference on Image Processing. 2002, 1: 900-903