一种联合文本和图像信息的行人检测方法

周炫余[1,2],刘 娟[1,2],卢 笑[3],邵 鹏[1,2],罗 飞[1,2]
(1. 武汉大学软件国家重点实验室,湖北武汉 430072;2. 武汉大学计算机学院,湖北武汉 430072;
3. 湖南大学电气与信息工程学院,湖南长沙 410082)

摘 要: 针对纯视觉行人检测方法存在的误检、漏检率高,遮挡目标以及小尺度目标检测精度低等问题,提出一种联合文本和图像信息的行人检测方法. 该方法首先利用图像分析的方法初步获取图像目标的候选框,其次通过文本分析的方法获取文本中有关图像目标的实体表达,并提出一种基于马尔科夫随机场的模型用于推断图像候选框与文本实体表达之间的共指关系(Coreference Relation),以此达到联合图像和文本信息以辅助机器视觉提高交通场景下行人检测精度的目的. 在增加了图像文本描述的加州理工大学行人检测数据集上进行的测评结果表明,该方法不仅可以在图像信息的基础上联合文本信息提高交通场景中的行人检测精度,也能在文本信息的基础上联合图像信息提高文本中的指代消解(Anaphora Resolution)精度.

关键词: 行人检测; 马尔科夫随机场; 文本-图像信息联合; 共指关系; 指代消解
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)01-0140-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.01.020

A Method for Pedestrian Detection by Combining Textual and Visual Information

ZHOU Xuan-yu[1,2], LIU Juan[1,2], LU Xiao[3], SHAO Peng[1,2], LUO Fei[1,2]
(1. *State Key Lab of Software Engineering, Wuhan University, Wuhan, Hubei* 430072, *China*;
2. *Computer School, Wuhan University, Wahan, Hubei* 430072, *China*;
3. *College of Electrical and Information Engineering, Hunan University, Changsha, Hunan* 410082, *China*)

Abstract: Existing vision-based pedestrian detection methods encounter many flaws, such as high false and miss detection rates, low detection accuracy on partial occluded and small scale objects, etc. In this paper, we propose a pedestrian detection method combining textual and visual information together. First, we use a vision-based method to initially localize the candidate visual objects. Second, we analyze the text information to get the text mentions corresponding to the visual objects. Finally, we propose a Markov random field-based model to infer the coreference relations between the candidate visual objects and textual mentions, so that the visual and textual information can be fused efficiently. The experimental results on the Caltech pedestrian detection benchmark enriched with textual description information have shown that the proposed method can not only improve the pedestrian detection accuracy by combining textual information with visual information, but also outperform the baseline anaphora resolution model by combining visual information with textual information.

Key words: pedestrian detection; Markov random field; text and image information combination; coreference relation; anaphora resolution

1 引言

行人检测是智能车辆的重要研究内容,也是计算机视觉的研究热点[1]. 传统的纯视觉的行人检测方法是基于滑动窗口的检测机制[2],并通过挖掘有强大描述能力的特征[3]、设计强判别能力的分类器[4]、以及对多视图、多姿态检测问题的研究[5],以提高该类方法的检测精度和速度. 然而,在复杂的城市道路环境下,车辆和行人通常一起出现并存在相互遮挡的现象,智能车辆无法对当前的交通环境做出正确的感知;此外,算法对远距离(低

收稿日期:2015-12-10;修回日期:2016-05-26;责任编辑:李勇锋
基金项目:国家自然科学基金(No. 61272274);国家自然科学基金青年项目(No. 61402340);湖北省自然科学基金(No. 2014CFB194)

分辨率、小尺度)目标检测精度低,以致车辆无法及时准确的检测出行人.针对上述两个问题,研究者们提出了相应的解决方法[6,7].但这些纯视觉的方法只是为了解决某一具体检测问题,并没有同时考虑交通场景中所有难点[8],如何在原有视觉信息的基础上融入新的信息以提高检测精度成为了当前研究的难点.随着语音识别和人机交互水平的提高,通过语言辅助机器视觉系统控制智能车辆通过复杂的城市道路环境成为了可能.当有行人被车辆或者被其它物体部分遮挡导致视觉处理方法没有成功检测出行人时,乘客可以通过简单的语言描述提示智能车辆,例如:"前方有两辆汽车,其中深色汽车的右方有两个行人,他们有可能横穿马路".简单的语言描述可以为机器视觉提供丰富的信息[9],并辅助智能车辆更好的感知当前道路环境.

文本信息与图像信息正确、有效联合的关键步骤是找出文本中所有的实体表达(Mention)和图像中的目标实体(Entity)之间的共指关系.而正确找出文本中所有的目标实体需要对文本中所有的实体表达进行指代消解(Anaphora resolution).在过去的十年中,对联合文本和图像信息的研究主要集中在图像检索[10]和自然语言生成[11,12],也有少量研究者通过自然语言帮助机器理解视觉场景[13]和图像语义划分[9].

针对纯视觉行人检测算法的不足,引入交通场景的文本描述以增强智能车辆对场景的理解.为了实现这一目的,提出了一种联合文本和图像信息的马尔科夫随机场模型用于提高交通场景的行人检测精度.该方法首先通过 HOG-SVM[3]算法初步获得交通场景中所有可能存在行人的候选框,其次通过马尔科夫随机场模型结合文本信息和图像信息进一步对候选框进行推理,判断其是否为行人.方法在增加了图像描述文本的 Caltech 行人检测数据库[14]上进行了广泛的测评,结果表明,所提出的方法性能比其它几类经典的纯视觉行人检测方法有明显的提高.同时该模型能有效地利用图像目标实体和文本中实体表达的共指关系,反向推导出文本中各实体表达之间的指代关系,在所标注描述文本上的测评表明,反馈信息能有效的提高指代消解模型的精度.

2　联合文本和图像信息的行人检测方法

为了有效的联合文本和图像信息,将找出文本中的实体表达与图像中的目标实体的共指关系的问题视为马尔科夫随机场模型(MRF)的推理问题.该模型分为文本信息层(ψ_{text})和图像信息层(ψ_{rgb}),模型中变量关系如图 1 所示.黑色方框表示图像信息,其中 φ_c 表示单个图像目标的一元势函数,$\varphi_{c,c}$ 表示相邻图像目标的二元势函数;蓝色方框表示文本信息,其中 $\varphi_{a,a}$ 表示文本中相邻两个实体表达之间的势函数;橙色方框表示文本实体表达与图像目标实体的共指信息,其中 $\varphi_{a,c}$ 表示文本实体和图像目标共指的势函数.

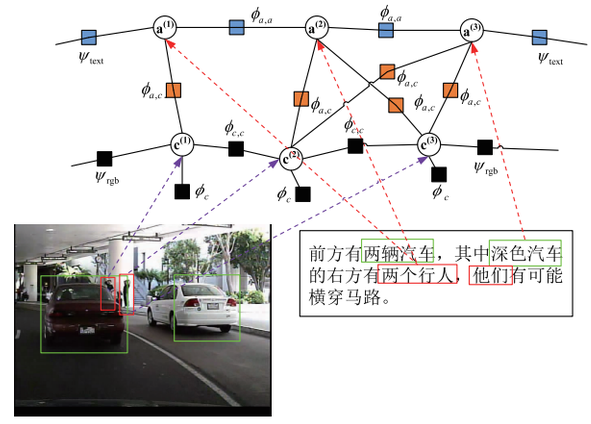


图1　马尔科夫网络结构及模型变量关系图

2.1　文本分析

2.1.1　文本预处理

文本预处理主要包括分词、词性标注、命名实体识别、句法树分析.为了保证预处理的模块的精度以及数据流的格式的一致性,文本预处理各基础自然语言预处理技术均由斯坦福自然语言处理小组提供*.

2.1.2　实体表达识别

本文将交通场景中的文本实体表达分为名词短语和代词两种表现形式.

名词短语:在场景描述中存在多个行人和车辆,单纯的名词不能有效的提供文本属性帮助文本实体表达和图像中的目标实体对齐,因此需要引入名词短语.抽取所有的命名实体以及句法分析树上的 NN、NP、NR 等节点做为候选名词短语.借鉴文献[15]中的待消解项识别模块,得到最终的名词短语.通过句法分析树以及句法分析树的相互依赖关系可以获得修饰名词的各类属性特征,例如:正前方、左侧、右侧(方位属性);两个行人(单复数属性);深色的和浅色的(颜色属性);车辆和行人(生命属性)等.

代词:由于每一帧图像都采用了多个句子进行描述,为了语言的简洁和通顺,在描述图像的文本中使用了一定数量的代词.将代词(或者为名词短语)和前文中所出现的名词短语关联起来的过程,叫做指代消解.本文基于 Soon 的所提出的思想[16]实现了一种最为经典的中文指代消解模型,通过该模型可以得到每对实体表达之间的指代概率.

2.2　图像分析

在交通场景中,行人往往和车辆同时出现,且它们

---

* http://nlp.stanford.edu/software/

在场景中存在某种位置关系. 车辆具有刚体形状,其视觉检测较行人检测更为简单. 本文中的马尔科夫随机场模型对一系列可能存在行人或车辆的目标窗口进行推理判断. 为了得到这些候选目标窗口,首先利用线性分类器 SVM 和经典的基于梯度的特征 HOG[3],分别训练行人和车辆分类器. 然后利用训练好的行人和车辆检测分类器采用滑动窗口策略对交通场景 RGB 图进行扫描,同时为了降低漏检率,以较低的分类阈值对目标和背景进行区分,得到图像的候选框.

2.3 文本实体表达和图像目标实体的共指模型

2.3.1 问题的设定与模型的建立

模型首先确定文本中出现的实体表达(名词短语或者代词)与图像候选框的共指关系,同时对候选框的类别(行人、车辆及背景)进行推导判断. 模型的变量描述如下,用 I 表示图像信息,T 表示图像的文本信息,图像中的每个候选框用随机变量 $c_i(i=1,\cdots,C)$ 表示,且 $c_i\in\{0,1,2\}$ 表示第 i 个候选框的类别,其中 $c_i=0$ 表示此候选框为误检(背景),$c_i=1$ 表示行人,$c_i=2$ 表示车辆. T 中的每个实体表达用随机变量 $a_j(j=1,\cdots,K)$ 表示,且 a_j 的值表示该实体表达与图像中的哪一个候选的检测框对应,即 $a_j\in\{0,1,\cdots,C\}$,当 $a_j=0$ 时表示该实体表达所表示的实体在图像中没有出现. 当实体表达表示复数时,为了避免文本中同一实体表达指向图像中的多个候选框,生成与实体表达数量相同的随机变量 a. 至此,确定 a_j 的取值就确定了文本信息中的实体表达与图像中的目标实体的共指关系. 文中提出的 MRF 模型各变量关系如图 1 所示,其数学模型描述如式(1)所示:

$$p(\boldsymbol{a},\boldsymbol{c})=\frac{1}{Z}\exp(-E(T,I,\boldsymbol{w})) \tag{1}$$

其中 $E(T,I,\boldsymbol{w})$ 是在图像信息 I 和文本信息 T 下的能量函数,$\boldsymbol{w}$ 是 MRF 模型的参数,Z 为式(1)的规范化因子(normalization factor)具体由式(2)所示:

$$Z=\sum_{(a,c)}\exp(-E(T,I,\boldsymbol{w})) \tag{2}$$

通过最小化能量函数 $E(T,I,\boldsymbol{w})$,实现最大化马尔科夫随机场联合概率 $p(\boldsymbol{a},\boldsymbol{c})$,能量函数的定义如式(3)所示:

$$E(T,I,\boldsymbol{w})=-\left(\begin{array}{l}\sum\limits_{i=1}^{C}\boldsymbol{w}_c\varphi_c(c_i)+\sum\limits_{(i,i')\in B_I}\boldsymbol{w}_I\varphi_{c,c}(c_i,c_{i'})\\+\sum\limits_{j=1}^{K}\boldsymbol{w}_a\varphi_{a,c}(a_j)+\sum\limits_{(j,j')\in B_T}\boldsymbol{w}_T\varphi_{a,a}(a_j,a_{j'})\end{array}\right) \tag{3}$$

其中 $\boldsymbol{w}=[\boldsymbol{w}_c;\boldsymbol{w}_I;\boldsymbol{w}_a;\boldsymbol{w}_T]^{\mathrm{T}}$ 是模型的参数,$\varphi_c(c_i)$ 是图像信息所确定的第 i 个候选框的一元势函数,$\varphi_{c,c}(c_i,c_{i'})$ 是由图像上下文关系定义的检测框之间的二元势函数,$\varphi_{a,c}(a_j)$ 是确定文本实体表达与图像检测框之间共指关系的势函数,$\varphi_{a,a}(a_j,a_{j'})$ 是确定文本信息中各实体表达之间指代关系的二元势函数.

2.3.2 模型势函数的定义

图像信息势函数定义:图像候选框的一元势函数特征向量定义为 $\boldsymbol{x}_I=[s_p,s_v,w,h,cw_r,cw_h]^{\mathrm{T}}$,其中 s_p 是行人检测分类器 f_p 在该候选框的得分;s_v 是车辆检测分类器 f_v 在该候选框的得分;w、h 分别代表检测框本身相对于图像的宽和高;cw_r、cw_h 代表候选框中心点相对图像的位置. 由于行人检测框和车辆检测框大小不一致,因此对同一候选框计算不同分类器在该候选框上的得分将采用以下策略:若当前候选框 c_i 为车辆检测分类器的候选框时,以候选框的中心点为中心点虚拟一个行人检测框 vc_i,此时 s_v 和 s_p 的取值分别是车辆和行人检测分类器 f_v 和 f_p 在 c_i 和 vc_i 上的得分;若当前候选框 c_i 为行人检测分类器的候选框时,则相应的以候选框的中心点虚拟一个车辆检测框 vc_i. 利用带五阶多项式核的支持向量回归(SVR)模型,建立关于行人、车辆及背景的回归模型,并利用 Platt 等[17]提出的方法将该模型的输出值转换为概率值作为各候选框的势函数的值,具体如式(4)所示:

$$\varphi_c(c_i=c)=\mathrm{Score}_{\mathrm{SVR}}(c_i=c) \tag{4}$$

图 1 所示的马尔科夫模型中的图像信息层中,相邻检测框可能是行人候选框、车辆候选框或误检目标的候选框,因此他们之间的二元势函数是用于描述真实目标候选框(包括车辆和行人)与误检目标候选框之间的相似性或差异性. 为了更好的凸显图像候选框之间的相似性或者差异性,定义每个候选框的二元势函数特征向量为 $\boldsymbol{x}_{II'}=[h^i,s_p^i,s_v^i]^{\mathrm{T}}$,其中 h^i 为第 i 个候选框的中心点位置的纵坐标值,s_p^i 是行人检测分类器 f_p 在第 i 个候选框的得分,s_v^i 是车辆检测分类器 f_v 在第 i 个候选框的得分. 根据上述向量定义两个相邻的候选框 $(c_i,c_{i'})$ 之间的势函数为 $\varphi_{c,c}(c_i,c_{i'})$,具体势函数如式(5)所示:

$$\varphi_{c,c}(c_i,c_{i'})=\sigma(1/\|\boldsymbol{x}_i-\boldsymbol{x}_{i'}\|^2) \tag{5}$$

其中 $\sigma(x)=1/(1+\exp(-1.5x))$ 是逻辑函数.

文本-图像信息共指势函数定义:交通场景的文本描述提供了车辆的颜色和在图像中的方位信息. 势函数 $\varphi_{a,c}(a_j)$ 表示文本信息中第 j 个文本实体表达与图像中第 i 个图像候选框之间指向同一目标实体的可能性. 其中 x_c,x_{geo},x_{ori} 分别表示检测框的颜色、几何、候选框相对于图像的方位特征;s_p 为行人检测分类器 f_p 在该候选框的得分;s_v 为车辆检测分类器 f_v 在该候选框的得分;t_c,t_{ori},t_{life} 分别为文本信息中提取的颜色、方位和生命属性特征. 将候选框的 RGB 颜色信息转换到 HSV 空间后提取 H 分量的 256 维直方图特征作为图像的颜色

属性,对应于文本信息中深色和浅色两种颜色属性;图像几何特征包括检测框的宽、高、宽高比;检测框方位是指候选框在图像中的位置信息,对应于文本信息中的正前方、左侧、右侧三种常用的方位属性. 在文本分析中提取每一个实体表达的颜色特征、方位特征、生命属性特征做为文本特征集合. 当实体表达为代词时,通过文本分析中所提及的指代消解模型得到代词所指代的名词短语,代词的特征属性与其存在指代关系的名词短语相同. 将文本-图像信息特征向量 x_{TI} 训练带 RBF 核的 SVM 分类器,并转换为概率值作为势函数 $\varphi_{a,c}(a_j)$ 的值,具体势函数如式(6)所示:

$$\varphi_{a,c}(a_j = i \mid I, T) = \mathrm{Score}(a_j = i) \tag{6}$$

文本信息势函数定义:二元势函数 $\varphi_{a,a}$ 表示文本中每对实体表达 $(a_j, a_{j'})$ 之间的存在指代关系的势函数,根据 Soon 等[16]提出实体表达对模型(Mention-pair model)的思想在已标注的图像描述文本训练集上训练出一个基于 SVM 指代消解分类模型,通过该模型判断待消解项(Anaphoricity)之间是否存在指代关系. 将分类模型的输出值转换成概率值作为相邻实体表达之间的势函数的值. 当名词短语或者代词是复数时,通过 $\varphi_{a,a}$ 对指向同一候选框的不同随机变量 a 进行惩罚,具体势函数如式(7)所示:

$$\varphi_{a,a}(a_j, a_{j'} \mid T) = \begin{cases} 0, \text{if } a_j = a_{j'} \text{ and } a_j, a_{j'} \text{ plural} \\ \mathrm{Score}_{\mathrm{SVM}}(a_j, a_{j'}), \text{otherwise} \end{cases} \tag{7}$$

2.4 模型学习及推断

根据结构化学习(structured learning)方法的理论[18],对模型公式(1)中各势函数参数的学习可转化为最大边界(max-margin)问题进行求解,即在给定所有训练样本 $T^{(n)}, I^{(n)}$ 及其标注 $(\boldsymbol{a}, \boldsymbol{c})^{(n)}, n = 1, \cdots, N$,对于所有的 $(\boldsymbol{a}, \boldsymbol{c}) \neq (\boldsymbol{a}, \boldsymbol{c})^{(n)}$,式(8)中的能量不等式成立:

$$E((\boldsymbol{a}, \boldsymbol{c})^{(n)}, T^{(n)}, I^{(n)}, \boldsymbol{w}) \leqslant E((\boldsymbol{a}, \boldsymbol{c}), T^{(n)}, I^{(n)}, \boldsymbol{w}) \tag{8}$$

所有真值情况下的能量 $E((\boldsymbol{a}, \boldsymbol{c})^{(n)}, T^{(n)}, I^{(n)}, \boldsymbol{w})$ 小于其他非真值情况下的能量 $E((\boldsymbol{a}, \boldsymbol{c}), T^{(n)}, I^{(n)}, \boldsymbol{w})$ 一个较大的间隔 $\Delta((\boldsymbol{a}, \boldsymbol{c}), (\boldsymbol{a}, \boldsymbol{c})^{(n)})$,因而得到一系列约束条件集合,具体如式(9)所示:

$$E((\boldsymbol{a}, \boldsymbol{c})^{(n)}, T^{(n)}, I^{(n)}, \boldsymbol{w}) \leqslant E((\boldsymbol{a}, \boldsymbol{c}), T^{(n)}, I^{(n)}, \boldsymbol{w}) - \Delta((\boldsymbol{a}, \boldsymbol{c}), (\boldsymbol{a}, \boldsymbol{c})^{(n)}) \tag{9}$$

由于式(9)对于每一个样本的标注 $(\boldsymbol{a}, \boldsymbol{c})^{(n)}$ 存在指数多个约束条件,类似于切平面算法(cutting plane algorithm)[19],可以通过求解式(10)寻求最违背的约束条件:

$$(\hat{\boldsymbol{a}}, \hat{\boldsymbol{c}})^{(n)} = \arg\min_{(a,c)} E((\boldsymbol{a}, \boldsymbol{c}), I^{(n)}, T^{(n)}, \boldsymbol{w}) - \Delta((\boldsymbol{a}, \boldsymbol{c}), (\boldsymbol{a}, \boldsymbol{c})^{(n)}) \tag{10}$$

这使得对参数 $\boldsymbol{w}$ 的求解可转化为最小化以下目标函数的求解问题,具体如式(11)所示:

$$\min_{w} \frac{\gamma}{2} \|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} l^n(\boldsymbol{w}) \tag{11}$$

$l^n(\boldsymbol{w}) \equiv E((\boldsymbol{a}, \boldsymbol{c}), T^{(n)}, I^{(n)}, \boldsymbol{w}) - E((\boldsymbol{a}, \boldsymbol{c})^{(n)}, T^{(n)}, I^{(n)}, \boldsymbol{w})$,$\gamma$ 是参数复杂度与损失函数之间的平衡因子. 综上所述,需要对损失函数 $\Delta((\boldsymbol{a}, \boldsymbol{c}), (\boldsymbol{a}, \boldsymbol{c})^{(n)})$ 进行明确定义以衡量假设值与真值的对应情况. 对于 c 的类别,定义 0-1 损失函数,此外对目标检测结果的定义为 0-1 损失函数,具体如式(12)所示:

$$\Delta_a(a_j, \hat{a}_j) = \begin{cases} 1, & \text{if IOU}(a_j, \hat{a}_j) \leqslant 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

其中 $\mathrm{IOU}(a_j, \hat{a}_j)$(Intersection-over-Union)为检测结果与真值的交集和并集的商,由此可见,损失函数对检测结果与真值覆盖率小于 50% 的结果进行惩罚.

对于复数形式的实体表达所产生的随机变量 a,只要所共指的候选框为真实标注值的任意一个,其损失函数为 0,否则为 1. 此外,定义复数实体表达对所产生的相邻随机变量间的损失函数为:同时指向同一候选框的不同随机变量,定义其损失函数为 1,否则为 0,具体如式(13)所示:

$$\Delta_{\mathrm{plural}}(a_j, a_k) = \begin{cases} 1, & \text{if } a_j = a_k \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

由于模型的推断属于 NP 难问题,利用 Schwing 等[21]提出的逼近算法进行求解,该算法将线性规划的整数约束松弛为非负约束,并利用对偶分解实现并行计算,同时保证了算法的收敛性能. 此外,由于不需要对图的结构和势函数进行任何约束,该算法具有很强的实用性.

3 实验结果与分析

3.1 实验设置

文章在 Caltech 行人检测数据集中抽取 1200 张图片进行文本描述,其中 700 张用于训练本文所提出的模型,其余 500 张为模型的测试集. 采用 Dollár 等[14]提出的行人检测标注方式对数据集中的车辆进行标注. 利用上述标注信息分别训练行人和车辆的 SVM + HOG 检测分类器. 每张图片的语言描述均由人工完成. 所有的文本处理均由文本分析中所述的自然语言处理技术完成,并对所有处理步骤进行人工检查.

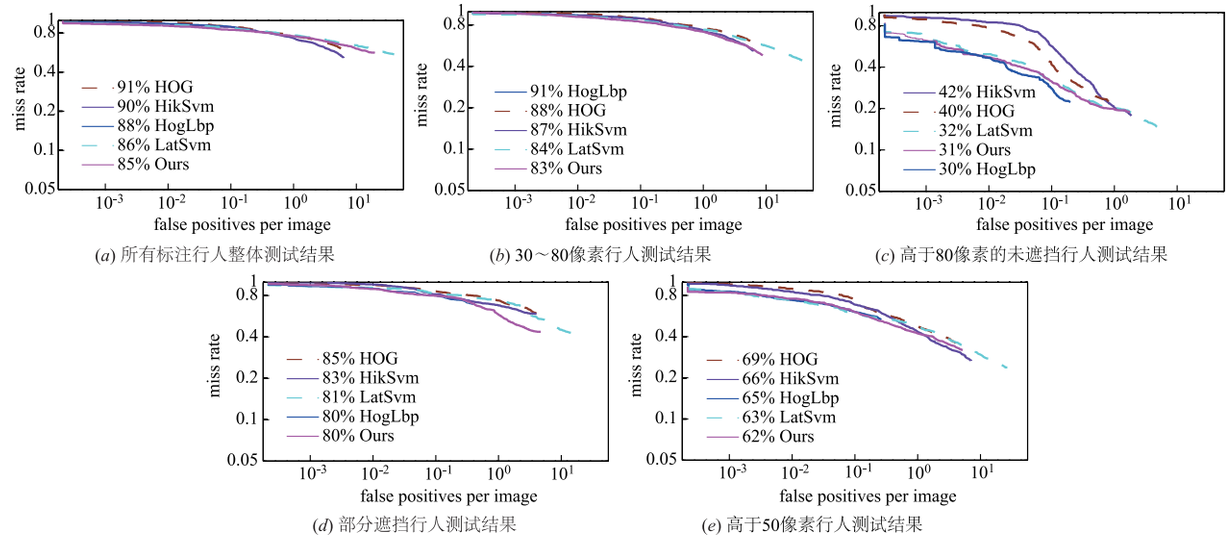
本文从行人检测和中文指代消解两方面对算法进行测评. 行人和车辆检测采用 Dollár 等[14]提出的全帧图像测评方法,绘制不同检测置信度阈值下 FPPI 曲线,比较文献[14]中提到的四类经典纯视觉行人检测算法,包括 HOG[3]、LatSVM[20]、HOG-LBP[7]、HikSVM[21]的检测效果. 指代消解采用 MUC[22]、BCUBED[23]、CEAFE[24]三种指代消解测评方法对所提及的模型和 Soon 等[16]提出实体表达对模型进行测评和比较.

3.2 实验结果与分析

所提出的模型和几种经典的基于图像的行人检测算法,在扩充文本描述的 Caltech 行人检测数据集上进行测评和比较. 图 2 通过 FPPI 曲线显示了几种算法在不同尺度和部分遮挡情况下的行人检测结果,该曲线反映了漏检率(miss rate)和每帧图像误检率(false positives per image)之间的关系(对数-对数). 由于检测分类器的阈值设定影响漏检率和每帧图像误检率,当分类阈值越高,漏检率越高但误检率降低,反之则亦然,因此 FPPI 曲线给出了检测算法在不同阈值下的性能. FPPI 曲线下降的越快,表示行人检测算法在测试集上测试效果越好. 此外,采用由 Dollár 等[14]定义的对数-平均漏检率(log-average miss rate)指标对不同检测算法的性能进行量化比较,这一指标是通过对均匀分布在对数空间的 $10^{-2}-10^{0}$ 范围内的对应九个不同 FPPI 数值(分别为 0.0100,0.0178,0.0562,0.1000,0.1778,0.3162,0.5623,1.000)的漏检率取均值所得.

图 2(a)显示了各算法在所有行人目标上的检测结果;图 2(b)显示了各算法对处于中距离尺度(30 ~ 80 像素之间)的行人检测结果;图 2(c)显示的是各算法对近距离目标即高于 80 像素的行人检测结果;图 2(d)显示了各算法对于部分遮挡目标的检测结果;图 2(e)显示了各算法对于合适尺度目标(高于 50 像素)的检测结果. 由图 2 的所有子图可知,本文提出的联合文本-图像信息的方法在各种尺度以及部分遮挡的行人检测上都优于其基准算法 HOG-SVM 约 5% ~9%. 造成检测效果有显著的提高的原因主要有以下几点:1)所提出的模型中引入了车辆信息,根据行人和车辆在场景图像中的位置关系,可以排除一些处于场景中不可能出现行人区域的候选框,一定程度上降低了行人检测的误检率;2)本文所提出的模型引入了文本信息. 场景的文本描述中包含的关于行人的数量信息,限制了每帧图像的行人检测数量,可以有效降低算法的漏检率;此外文本中关于行人的位置、颜色等信息,实现了对检测结果的验证,进一步降低了算法的误检率. 算法对近距离目标和合适尺度目标,检测效率分别提高了 9% 和 7%,而对小尺度目标和部分遮挡目标,检测准确率只提高 5%,这是由于所有纯视觉检测算法在小尺度目标和遮挡目标检测上都存在一定的局限性,造成在利用 HOG-SVM 获取图像目标候选框时存在一定的漏检和误检. 由图 2(c)可知,HOG-LBP 算法在大尺度目标检测上能取得较好的效果,这是由于 HOG-LBP 特征结合了 HOG 特征和 LBP 特征的优势,但随着目标尺度的减小,HOG-LBP 算法的性能急剧下降,此外由图 2(d)可知,HOG-LBP 特征在部分遮挡行人检测上也展现了较好的性能. 由图 2(d)和(e)可知,HikSvm 算法采用直方图交核训练非线性分类器,其性能在部分遮挡和中距离尺度行人检测上略优于 HOG 算法所采用的线性分类器. 除此之外,LatSVM 算法由于训练了行人的各部件模板,在不同情况的行人检测上也能取得较好的性能. 由图 2(c)可知,本文提出的方法使用基本的 HOG 特征和线性 SVM 分类器,通过融入文本信息和车辆位置信息所取得的检测性能除了在大尺度目标检测中略低于 HOG-LBP 算法外,其余情况下,均等同或者优于其它算法.

为了直观的比较所有行人检测算法的检测性能,本文对所有的检测算法按照检测结果的对数平均漏检率进行排序,即将取得最好的检测效果的算法排在第一位,检测效果最差的放在最后一位. 定义平均排序为各类行人检测算法在不同情况下检测结果的排序平均数,表 1 中



(a) 所有标注行人整体测试结果 (b) 30~80像素行人测试结果 (c) 高于80像素的未遮挡行人测试结果

(d) 部分遮挡行人测试结果 (e) 高于50像素行人测试结果

图2 五种不同情况下Caltech行人数据集测试结果

的结果表明,本文方法在所有检测方法中排序第一,实验结果证明本文方法比其余检测方法更加有效.

文中所提出的模型不仅能在图像信息的基础上联合文本信息提高行人检测的精度,也能通过文中 a_i 变量结合图像信息帮助文中所提的指代消解模型提高指代消解精度. 模型中每个变量 a_i 反映的是图像目标框和文本实体表达之间的共指信息,如果不同的变量选择相同图像目标框,则证明候选目标框所对应的文本实体表达指向图像中同一实体,这些变量在文本中也应存在指代关系. 基于上述推理,本文所提出的方法可以在文本的基础上联合图像信息,增强指代消解模型的消解精度. 表 2 是本文所提出的方法和实体表达对模型之间的消解能力对比. 从表 2 看以得出,联合图像信息后,指代消解模型能提高 Avg_F 值约 4%,消解精度得到显著的提高.

表 1　不同情况下对数平均漏检率(%)及各方法平均排序

检测算法	LatSVM	HOG-LBP	HikSVM	HOG	Ours
高于 80 像素	32	30	42	40	31
30 ~ 80 像素	84	91	87	88	83
高于 50 像素	63	65	66	69	62
部分遮挡	81	80	83	85	80
所有目标	86	88	90	91	85
平均排序	2.2	2.6	4	4.6	1.2

表 2　模型的指代消解能力

模型	Caltech 行人检测数据集描述文本									
	MUC			BCUBED			CEAFE			Avg_F
	P	R	F	P	R	F	P	R	F	F
实体表达对模型	72.3	58.2	64.5	79.4	66.8	72.6	59.5	66.9	63.0	66.7
联合文本和图像信息的模型	80.6	58.9	68.1	89.1	66.3	76	67.6	66.8	67.2	70.5

4　结论

提出一种联合文本和图像信息的马尔科夫随机场模型,以提高交通场景中的行人检测精度. 模型在增加了场景描述文本的加州理工大学行人检测数据集上进行了测评结果表明,相比于纯视觉的行人检测基准方法 HOG-SVM,所提出的方法在不同尺度目标、受遮挡目标检测上效果均有明显的提升. 且该方法在只使用了基本的 HOG 特征和线性 SVM 分类器,性能均等同或者优于其余三种经典的纯视觉行人检测方法. 此外,模型也能联合图像信息提高文本的指代消解模型的精度,在所标注的描述文本上的测评结果表明,模型中的图像对文本的反馈信息能提高指代消解精度约 4%.

参考文献

[1] Enzweiler M, Gavrila D M. Monocular pedestrian detection: Survey and experiments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(12): 2179 – 2195.

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

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. San Diego, USA: IEEE, 2005. 886 – 893.

[4] Wu J, Liu N, Geyer C, et al. C4: A real-time object detection framework[J]. IEEE Transactions on Image Processing, 2013, 22 (10): 4096 – 4107.

[5] 刘威,段成伟,遇冰,等. 基于后验 HOG 特征的多姿态行人检测[J]. 电子学报, 2015, 43(2): 217 – 224.
Liu Wei, Duan Chengwei, Yu Bing, et al. Multi-pose pedestrain detection based on posterior HOG feature[J]. Acta Electronica Sinica, 2015, 43(2): 217 – 224. (in Chinese)

[6] Yan J, Zhang X, Lei Z, et al. Robust multi-resolution pedestrian detection in traffic scenes[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Portland, USA: IEEE, 2013. 3033 – 3040.

[7] Wang X, Han T X, Yan S. An HOG-LBP human detector with partial occlusion handling[A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Kyoto, Japan: IEEE, 2009. 32 – 39.

[8] 苏松志,李绍滋,陈淑媛,等. 行人检测技术综述[J]. 电子学报, 2012, 40(4): 814 – 820.
Su Songzhi, Li Shao zi, Chen Shuyuan, et al. A survey on pedestian detection[J]. Acta Electronica Sinica, 2012, 40(4): 814 – 820. (in Chinese)

[9] Fidler S, Sharma A, Urtasun R. A sentence is worth a thousand pixels[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Portland, USA: IEEE, 2013. 1995 – 2002.

[10] Ramanathan V, Joulin A, Liang P, et al. Linking people in videos with "their" names using coreference resolution

[A]. Computer Vision – ECCV 2014[C]. New York, USA: Springer International Publishing, 2014. 95 – 110.
[11] Kuznetsova P, Ordonez V, Berg A C, et al. Generalizing image captions for image-text parallel corpus[A]. Proceedings of the 51st Annual Meeting of the Association for Computational Linguis-tics [C]. Stroudsburg, USA: ACL, 2013. 790 – 796.
[12] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891 – 2903.
[13] Kong C, Lin D, Bansal M, et al. What are you talking about? text-to-image coreference [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Columbus, USA: IEEE, 2014. 3558 – 3565.
[14] Dollár P, Wojek C, Schiele B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 743 – 761.
[15] 周炫余,刘娟,邵鹏,等. 基于层次过滤的中文指代消解研究[J]. 吉林大学学报(工学版), 2016, 46(4): 1209 – 1215.
Zhou Xuanyu, Liu Juan, Shao Peng, et al. Chinese anaphora resolution based on multi-pass sieve[J]. Journal of Jilin University(Engineering and Technology Edition), 2016, 46(4): 1209 – 1215.
[16] Soon W M, Ng H T, Lim DCY. A machine learning approach to coreference resolution of noun phrases[J]. Computational Linguistics, 2001, 27(04): 521 – 544.
[17] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[J]. Advances in Large Margin Classifiers, 1999, 10(3): 61 – 74.
[18] Tsochantaridis I, Joachims T, Hofmann T, et al. Large margin methods for structured and interdependent output variables[J]. Journal of Machine Learning Research, 2005: 1453 – 1484.
[19] Joachims T, Finley T, Yu C N J. Cutting-plane training of structural SVMs[J]. Machine Learning, 2009, 77(1): 27 – 59.
[20] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627 – 1645.
[21] Maji S, Berg A C, Malik J. Classification using intersection kernel support vector machines is efficient[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Anchorage, USA : IEEE, 2008. 1 – 8.
[22] Marc V, John B, John A, et al. A model theoretic coreference scoring scheme[A]. Proceedings of the 6th Message Unders-tanding Conference[C]. Stroudsburg, USA: ACL, 1995. 45 – 52.
[23] Amit B, Breck B. Algorithms for scoring coreference chains[A]. Proceedings of LREC[C]. Stroudsburg, USA: ACL, 1998. 563 – 566.
[24] Luo X Q. On coreference resolution performance metrics [A]. Proceedings of HLT-EMNLP [C]. Stroudsburg, USA: ACL, 2005. 25 – 32.

作者简介

周炫余 男,1987 年 10 月出生,湖南邵阳人. 现为武汉大学计算机学院博士研究生,从事指代消解、中文自然语言处理、机器学习等有关研究.
E-mail: zhouxuanyu@whu.edu.cn

刘 娟(通信作者) 女,1970 年 2 月出生,湖北武汉人. 教授、博士生导师. 现为武汉大学计算机学院软件所所长,主要从事自然语言处理、图像处理、数据挖掘、机器学习等有关研究.
E-mail: liujuan@whu.edu.cn