第 37 卷 第 24 期
Vol.37 No.24

计 算 机 工 程
Computer Engineering

2011 年 12 月
December 2011

·人工智能及识别技术· 文章编号：1000—3428(2011)24—0169—04 文献标识码：A 中图分类号：TP306

基于短语和依存句法结构的中文语义角色标注

徐 靖，李军辉，朱巧明，李培峰

(苏州大学计算机科学与技术学院，江苏 苏州 215006)

摘 要：提出一种基于短语和依存句法结构的中文语义角色标注(SRL)方法。联合短语句法特征和依存句法特征，对句法树进行剪枝，过滤句法树上不可能担当语义角色的组块短语单元和关系结点，对担当语义角色的组块或节点进行角色类别标注。基于正确句法树和正确谓词的识别结果表明，该方法的 SRL 性能 F1 值为 73.53%，优于目前国内外的同类系统。

关键词：语义角色标注；短语句法特征；依存句法特征；名词性谓词；句法结构

Chinese Semantic Role Labeling Based on Phrase and Dependency Syntactic Structure

XU Jing, LI Jun-hui, ZHU Qiao-ming, LI Pei-feng

(School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

【Abstract】This paper proposes a Chinese Semantic Role Labeling(SRL) based on phrase and dependency syntactic structure. Combining the features of phrase and dependency syntactic structure, syntax tree are pruned, filtrates group piece phrases unit and relationship nodes of syntax tree which can't assume semantic role, labels role category for group piece and nodes which assume semantic role. Recognition results show that the nominal SRL approach achieves the performance of 73.53% in F1-measure on golden parse trees and golden predicates.

【Key words】Semantic Role Labeling(SRL); phrase syntactic feature; dependency syntactic feature; nominal predicate; syntactic structure

DOI: 10.3969/j.issn.1000-3428.2011.24.057

1 概述

语义角色标注(Semantic Role Labeling, SRL)是目前自然语言处理的一个热点研究课题之一，是目前语义分析的一种实现方式。所谓语义角色标注，就是对于给定句子，采用“谓词-论元角色”的结构形式，对句中的每个谓词(动词性谓词或名词性谓词等)标注出句子中谓词的相应语义角色成分，包括核心的语义角色(如施事者、受事者等)和附属语义角色(如地点、时间、方式、原因等)。SRL 标注的语义角色对回答 5W (Who, What, When, Where 和 Why)问题提供了强有力的支持。这使其应用非常广泛，包括问答系统、指代消解、信息检索、机器翻译等领域，具有广泛的应用前景。

根据利用的句法结构信息不同，可以将现有的 SRL 分为 3 个类别：基于组块的 SRL，基于短语的 SRL 和基于依存关系的 SRL。基于组块的 SRL 建立于浅层句法分析的基础上，由于不能获得全部句法分析信息，其性能有限；基于短语的 SRL 建立在句法分析的基础上，以句法成分为标注单元，性能较好。随着依存分析研究的不断深入，基于依存关系的 SRL 研究也越来越受到关注。然而，已往的 SRL 研究大多局限于某种句法结构信息，而忽略了各种句法结构信息之间的互补、探索联合不同句法结构的 SRL 研究。

本文研究了联合短语句法分析和依存句法分析的中文名词性谓词语义角色标注。在基于短语 SRL 的基础上，进一步融合依存关系的特征，分析基于依存关系的 SRL 的常用特征对基于短语的 SRL 的作用。

2 相关工作

由于中文 NomBank 发布得比较晚和其语料的标注实例少，因此中文的名词性谓词的 SRL 研究也相对较少。在英文方面，文献[1]以 NomBank 为实验语料，将基于动词性谓词的英文 SRL 方法移植于名词性谓词的英文 SRL，并探索了大量与英文名词性谓词相关的特征。在中文方面，文献[2]利用大规模语料库中文 NomBank，展开了中文名词性谓词的 SRL，在使用正确和自动句法树情况下，性能 F1 值分别取得 71.6%和 48.3%。文献[3]进一步探索了中文名词性谓词 SRL，该文还尝试了借助动词性谓词的标注实例，用来扩展名词性谓词 SRL 的训练集规模，以期望提高名词性谓词 SRL 性能。不过实验结果并未如愿，其原因在于动词性和名词性谓词标注实例中的特征值差异非常明显。文献[4]在传统动词性谓词 SRL 相关特征的基础上，进一步提出了名词性谓词 SRL 相关的特征集，并且探索了中文动词性谓词 SRL 对中文名词性谓词 SRL 的影响。

除了以上描述的基于短语句法树的 SRL 外，基于依存句法树的 SRL 已成为目前研究的热点，然而其相关研究大多集中在动词性谓词 SRL 上进行。文献[5]采用基于依存分析的方法实现语义角色标注，所使用的依存树是由短语句法树转化而来。文献[6]在 CTB 转换语料和 CoNLL 2009 中文语料上，研究了基于依存句法分析的中文语义角色标注研究，基于正确谓词和自动谓词的情况下，在 CTB 语料上分别得到系统性

基金项目：国家自然科学基金资助项目(90920004, 60970056, 60873150)；江苏省自然科学基金资助项目(BK2008160)；江苏省高校自然科学重大基础研究基金资助项目(08KJA520002)

作者简介：徐 靖(1986－)，男，硕士研究生，主研方向：自然语言处理；李军辉，博士研究生；朱巧明，教授、博士生导师；李培峰，副教授

收稿日期：2011-08-18 **E-mail：**xujing0103@163.com

能 F1 值 84.30%和 81.02%，在 CoNLL 语料上分别得到系统性能 F1 值 81.68%和 81.33%。然而，联合多种句法结构的研究相对很少，文献[7]分别训练了基于短语结构和依存关系的 SRL 系统，使用这 2 个系统的输出作为附加特征，用基于块结构的 SRL 系统标注。比起单一的标注，结果得到了很大的提高，而该文献也提到了联合使用多种句法分析也是极其复杂的。

3 联合短语和依存句法结构的中文名词性谓词

3.1 实验语料

英文NomBank语料库采用与PropBank一致的标注框架，由于中文 NomBank 继承了英文 NomBank 的标注框架，因此与 PropBank 的标注框架基本一致。图 1 给出中文 NomBank 的标注实例。其中，谓词“扩大”包含 2 个核心语义角色，分别为“青岛(Arg0)”和“企业规模(Arg1)”。

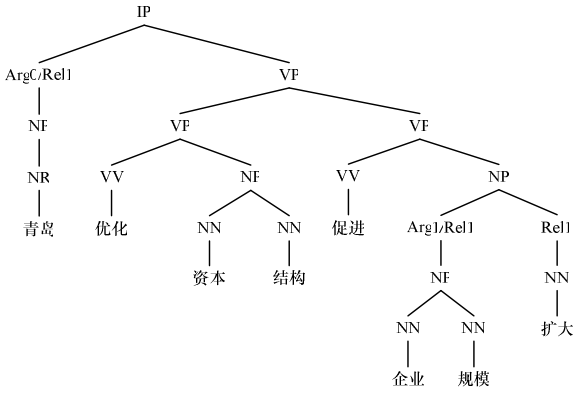

图 1 名词性谓词“扩大”的语义角色标注实例

在本文中，由于要获得依存关系的特征，因此将中文 NomBank 的语料通过 Penn2Malt 工具转换为 CoNLL 2008 评测中所使用的语料的形式。图 2 是图 1 中实例转换后的依存关系树。在图 2 中，W 表示单词，R 表示依存关系，G 表示词性，粗体字表示谓词及各个谓词所对应的角色。

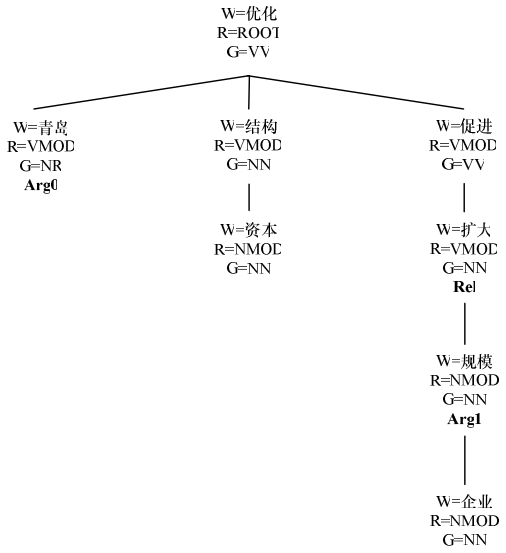

图 2 语义角色标注实例对应的依存树

在本文中，SRL 的特征集包含两部分：基准特征和扩展特征。把基于短语的特征看作是基准特征，把基于依存关系的特征看作是扩展特征。

3.2 系统流程

在进行语义角色标注时，给定某棵句法树和其中的名词性谓词，本文系统共分为 3 个步骤，但由于短语句法树和依存句法树关注的句法结构不一样，因此在相应的步骤处理上也不大一样。如图 3 所示：首先是角色剪枝(预处理)，其次是角色识别，最后是角色分类，并标注其类型。

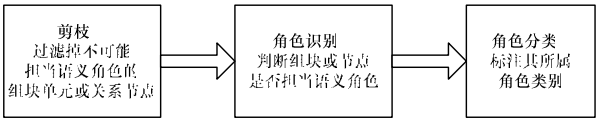

图 3 语义角色标注步骤

预处理主要是对句法树进行剪枝。在短语句法分析时，过滤掉句法树上最不大可能承担语义角色的组块短语单元，本文参照文献[3-4]的剪枝规则；在依存分析时，过滤掉句法树上最不可能承担语义角色的关系结点，本文参照文献[5-6]的剪枝方法。

3.3 基于短语的特征

在特征的选取上，由于角色识别和角色分类阶段的任务性质不一样，因此选用的部分特征可能在角色识别时有利，而在角色分类时起反作用，反之亦然[8]。本文选取的短语特征参照文献[2,4]中的特征。

3.4 基于依存关系的特征

由于依存树表示的是单词之间的依赖关系，缺少组块短语信息，因此如果使用基于依存关系的特征标注组块短语，首先必须把每一个组块短语映射到依存树上，映射为一个或几个合适的单词。在依存分析时，本文使用每一个组块短语的中心词代表相应的组块短语。

在基于依存关系特征的选取上，本文参照文献[6]基于依存树的动词性谓词 SRL 特征，制定了适合本文名词性谓词 SRL 的候选依存关系特征集合。

由于在 SRL 中路径特征一直是最重要的一个特征，因此在依存特征选择上，“依存路径”是一个必不可少的特征。对于谓词与论元结点位置关系也是一个重要的信息，选用了特征“家族成员”，因为在一些通用的句子语境中，谓词与论元结点之间的位置关系很固定，比如例句“上海浦东开发与法制建设同步”，对于名词性谓词“建设”，作为它的论元角色的，通常其论元角色是“建设”的孩子结点。

在SRL中，谓词周围的信息被提取出来作为重要的特征，因为这些特征能体现上下文的相关信息，对其判断论元趋向于哪一类角色有着重要的作用，所以本文选用了特征“谓词孩子的词性链”和“谓词兄弟的词性链”。由于依存树关注于单词之间的依存关系，而本文的思想是借助这种依存关系，融合短语句法和依存句法结构之间的信息互补，因此本文选用“依存关系+依存关系前一个词”、“依存关系+依存关系后一个词”；而“子类框架”特征是体现了谓词上下文信息和依存关系信息。

因为在基于短语句法树结构句法分析的 SRL 系统中，大量的研究都表明中心词特征对 SRL 系统性能贡献很大，而本文的依存词相当于短语结构句法分析中的中心词，通过短语句法结构转换成依存句法结构，获取中心词周围的依存信息，实现短语句法和依存句法结构之间的信息互补，本文选用了特征“依存词+依存词前一个词”和“依存词+依存词后一个词”。

以图 2 为例，当前结点为“W=青岛”，当前谓词为“扩大”，对应的特征如下(特征不存在使用 NULL 代替)，为了下文的表示方便，用{*D*1~*D*9}表示这 9 个特征，如表 1 所示。

表 1 候选依存特征集合

特征	特征含义	特征举例
依存路径(*D*1)	依存句法树上当前论元结点到谓词的路径，即途经结点的依存关系	VMOD->ROOT->VMOD->VMOD
家族成员(*D*2)	当前论元结点与当前谓词结点的家族关系，如 father、siblings、C 等	C
谓词的孩子的词性链(*D*3)	当前谓词的所有孩子结点的词性组成的链	NN
谓词的兄弟的词性链(*D*4)	当前谓词的所有兄弟结点的词性组成的链	NULL
子类框架(*D*5)	当前谓词结点出发的所有孩子结点的依存关系链	VMOD->NMOD
依存关系+依存关系前一个词(*D*6)	当前依存关系的类型+依存关系前一个词	VMOD+NULL
依存关系+依存关系后一个词(*D*7)	当前依存关系的类型+依存关系后一个词	VMOD+优化
依存词+依存词前一个词(*D*8)	当前依存词+依存词的前一个词	青岛+NULL
依存词+依存词后一个词(*D*9)	当前依存词+依存词的后一个词	青岛+优化

由于短语句法结构特征关注于多单词组块短语的依赖关系，而依存特征关注于各个单词之间的依存关系，因此为了在自动句法分析时也能保持对短语句法树和依存句法树结构特征的一致性，采用中心词把基于短语句法特征和基于依存句法特征融合起来。考虑到支持向量机(Support Vector Machine, SVM)分类器对特征值表示方法的要求，所有特征值均转化为数字。

3.5 特征选择

本文参照文献[1,4]的特征选择，具体的特征选择过程描述如下：(1)按照文献[4]划分的角色识别特征集为基本特征集合，执行特征选择算法，从依存特征集{*D*1~*D*9}中找到给角色识别带来更佳效果的特征集合；(2)固定由上述得到的最佳特征集为角色识别特征集合，以文献[4]划分的角色分类特征集为基本特征集合，执行上述算法，从依存特征集{*D*1~*D*9}中找到给角色分类带来更佳效果的特征集合。

4 实验结果与分析

参照文献[2]的实验数据划分，取中文 NomBank 中的 648 个文件(chtb_081.fib~chtb_899.fid)用作训练集，40 个文件(chtb_041.fib~chtb_080.fid)用作开发集，72 个文件(chtb_001.fib~chtb_040.fid 和 chtb_900.fid~chtb_931.fid)用作测试集。其中，训练集、开发集、测试集所包含的名词性谓词数分别为 8 642 个、731 个、1 124 个。

在语义角色标注任务中，使用 SVM Light 工具包作为分类模型，由于 SVM Light 分类器本质上是一个二元分类器，采用一对多时，将其重新封装为多元分类器。在训练时，采用线性核，训练参数 C 值设置为 0.22，由于在测试阶段，SVM Light 输出的是测试样例超平面的距离，本实验采用 Sigmoid 函数将其转化为概率值。

4.1 加入依存的特征结果

在实验时，本文首先建立一个基于短语特征的系统，称为基准系统(本文的基准系统与文献[4]系统相类似)；使用 3.5 节描述的特征选择方法从依存特征集{*D*1~*D*9}中分别为角色识别和角色分类选取有效的特征，基于开发集，特征{*D*2, *D*5, *D*1}先后被选入角色识别特征集体，而特征{*D*6, *D*3, *D*7}先后被选入角色分类特征集。表 2 给出了基于开发集的新选入的每个特征在基准系统上语义角色标注的贡献(没有选入的特征，其性能没有列出来)。结果采用准确率(P)、召回率(R)、调和平均(F1)值表示其相应的性能。

表 2 依存特征加入基准系统上的结果(开发集)

性能指标	基准系统	加入识别特征			加入分类特征(固定识别特征)		
		*D*1	*D*2	*D*5	*D*3	*D*6	*D*7
P/(%)	77.45	78.53	78.99	78.84	79.84	79.45	78.89
R/(%)	68.29	67.91	67.61	67.37	67.52	67.81	67.76
F1 值/(%)	72.58	72.83	72.86	72.66	73.16	73.17	72.90

从表 2 可以看出，添加有效的依存特征后，中文名词性谓词的语义角色标注性能 F1 值得到了提高。从选入的特征可以看出，加入了与依存关系有关的特征对系统的性能贡献很大(依存关系相关的特征(*D*5、*D*6、*D*7)等)，特别是特征 *D*6、*D*7 表达了当前依存分析的上下文特征，对提供上下文信息有着重要作用，对捕获短语句法树中的结构化信息有着重要的作用。从 *D*1 特征被选入可以看出，不管是在短语句法分析或依存句法分析或是两者联合的句法分析中，路径都是一个重要特征。从 *D*2 特征选入可以看出，在中文句法中，大多数单词之间的位置关系是很固定的。从短语句法树转换为依存树后，*D*3 特征很好地表达了谓词的上下文信息。

由于依存句法中的依存词相当于短语结构句法分析中的中心词，因此它们的作用也是类似的，选入依存词相关的特征的本意是想表达中心词周围更丰富的信息，但由于在短语特征中心词已经包含了一些相关的信息，因此这些特征(*D*8, *D*9)作用不大，对实验结果没有提高。而 *D*4 也没有被选入特征，因为经统计，在依存树中，大部分谓词周围的兄弟结点都是空结点，所以不能表现出谓词周围相关的信息，如图 2 中的谓词“扩大”。

4.2 系统结果及与原有系统结果的比较分析

在基准系统上添加了经特征选择选取的相关依存特征后，得到了在测试集上的系统性能，为了便于比较，表 3 列出与本文系统采用相同的数据集的已有系统的结果。

表 3 实验结果对比(测试集) (%)

系统	P	R	F1 值
文献[3]系统	69.70	73.70	71.60
文献[4]系统	77.51	68.40	72.67
本文系统	79.97	68.05	73.53

从表 3 可以看出，本文系统的性能相比文献[3-4]，F1 值分别提高了约 2.0 和 0.9 个百分点，说明比起单一使用从短语句法树上抽取出来的特征，联合基于短语和依存关系特征能使中文名词性谓词 SRL 的性能得到提高，从而验证了联合这 2 种句法分析树的方法是有效的。

从文献[5]使用依存分析的方法实现语义角色标注可以看出，谓词相关信息的重组对性能影响很大，在这里也得到了很好的体现。以图 1 和图 2 上的谓词“扩大”为例：NP 组块“企业规模”被标注为 Arg1，论元和谓词之间的路径为：“NP->NP->NN”；但在图 2 中，单个的单词“规模”被标注为 Arg1，而论元和谓词之间的路径是靠依存关系“VMOD->NMOD”联系，相比于短语结构的路径特征，这样标注更详细。因此，在 SRL 时这 2 种句法结构可以实现信息互补。

然而，从 F1 值提高的程度来看，幅度并不是很大，一方面的原因是使用 MaltParser 工具进行转换时，转换的结果存在一定误差，也就影响了抽取出来的特征结果；但其最主要

原因是名词性谓词的角色识别很困难，根据文献[9]中描述的名词性谓词的角色标注原则，即使某个名词为动词的派生词，并不是该名词的所有修饰成分都将被标注为该名词的语义角色。这使得判断某个修饰成分是否为该谓词的角色时，需要考虑它们之间的语义信息；而根据文献[4]在开发集中的统计数据看出，在动词性谓词的兄弟结点中，96%的结点被标注为角色，而在名词性谓词的兄弟结点中，仅有 40%的结点被标注为角色，这增加了名词性谓词的角色识别的难度。

在基于正确句法树和正确谓词识别的 F1 值性能上比较，相比于中文动词性谓词 SRL 性能 F1 值 92%[2]，中文名词性谓词 SRL 仍然是一项艰巨的任务。其主要原因为：(1)中文 NomBank 的标注实例少；(2)名词性谓词与其角色之间的结构更灵活和复杂等。

4.3 基于自动句法分析的中文名词性谓词

为便于性能比较，表 4 列出了基于正确句法树和正确谓词、自动句法树和自动谓词识别情况下(自动谓词识别采用文献[10]描述的系统，基于正确句法树，自动谓词识别性能 F1 值为 91.64%)，测试集上语义角色标注的性能。本文选用句法分析器为 Berkerley Parser。在实验中，以正确的分词结果作为句法分析的输入。Berkerley Parser 使用中文 CTB5.1 的数据，采用与语义角色标注实验划分一致的训练集和开发集，在测试集上，基于正确分词的句法性能 F1 值为 80.96%。

表 4 自动句法分析的 SRL 性能(测试集) (%)

句法分析器	P	R	F1 值
金标准	79.97	68.05	73.53
Berkerley Parser	59.51	51.98	55.49

从表 4 可以看出，在基于自动句法树和自动谓词的识别情况下，中文名词性谓词 SRL 性能 F1 值为 55.49%，相比文献[10]的 F1 值 56.19%有所下降，其主要原因：(1)由于自动短语句法树本身的性能不理想；(2)自动短语句法树转换成依存树后，从文献[6]可知，中文依存句法分析性能是影响中文依存语义分析的关键因素，反过来也说明，中文依存语义分析依赖于句法分析。从表 4 看出，相比于基于正确句法树和正确谓词，中文名词性谓词 SRL 的研究具有非常大的挑战性，也说明了中文短语句法分析和依存句法分析还远不够完善。

5 结束语

本文在短语句法结构特征的中文名词性谓词 SRL 的基础上，联合了依存句法结构特征，提出一种新的中文语义较色标注方法。相比于传统的单一的基于短语结构句法分析的中文名词性谓词 SRL，该系统使用依存句法分析的结果，添加上相应的依存特征，是对中文名词性谓词 SRL 的一种新的尝试。今后将选取更丰富、有效的依存特征，并且与短语的特征进行组合和筛选。

参考文献

[1] Jiang Zhengping, Ng H T. Semantic Role Labeling of NomBank: A Maximum Entropy Approach[C]//Proc. of Conference on Empirical Methods in Natural Language Processing. Sydney, Australia: ACM Press, 2006: 138-145.

[2] Xue Nianwen. Labeling Chinese Predicates with Semantic Roles[J]. Computational Linguistics, 2008, 34(2): 225-255.

[3] Xue Nianwen. Semantic Role Labeling of Nominalized Predicates in Chinese[C]//Proc. of HLT-NAACL'06. New York, USA: ACM Press, 2006: 431-438.

[4] Li Junhui, Zhou Guodong, Zhao Hai, et al. Improving Nominal SRL in Chinese Language with Verbal SRL Information and Automatic Predicate Recognition[C]//Proc. of Conference on Empirical Methods in Natural Language Processing. Singapore: ACM Press, 2009: 1280-1288.

[5] Hacioglu K. Semantic Role Labeling Using Dependency Trees[C]//Proc. of the 20th International Conference on Computational Linguistics. Geneva, Switzerland: [s. n.], 2004.

[6] 王步康, 王红玲, 袁晓红, 等. 基于依存句法分析的中文语义角色标注[J]. 中文信息学报, 2010, 24(1): 25-29.

[7] Sameer P, Wayne W, Martin J H. Towards Robust Semantic Role Labeling[J]. Computational Linguistics, 2008, 34(2): 289-310.

[8] Xue Nianwen, Palmer M. Calibrating Features for Semantic Role Labeling[C]//Proc. of EMNLP'04. Barcelona, Spain: [s. n.], 2004.

[9] Xue Nianwen. Annotating the Predicate-argument Structure of Chinese Nominalizations[C]//Proc. of LREC'06. Genoa, Italy: [s. n.], 2006: 1382-1387.

[10] 李军辉, 周国栋, 朱巧明, 等. 中文名词性谓词语义角色标注研究[J]. 软件学报, 2011, 22(8): 1725-1737.

编辑 陆燕菲

(上接第 168 页)

参考文献

[1] Miezianko R, Pokrajac D. Detecting and Recognizing Abandoned Objects in Crowded Environments[C]//Proc. of AVSS'07. London, UK: [s. n.], 2007.

[2] Dalley G, Wang Xiaogang, Grimson W. Event Detection Using an Attention Based Tracker[C]//Proc. of AVSS'07. London, UK: [s. n.], 2007.

[3] Bayona A, San Miguel J C, Martínez J M. Comparative Evaluation of Stationary Foreground Object Detection Algorithms Based on Background Subtraction Techniques[C]//Proc. of AVSS'08. Genova, Italy: [s. n.], 2009.

[4] 杨 涛, 李 静, 潘 泉, 等. 一种基于多层背景模型的前景检测算法[J]. 中国图象图形学报, 2008, 13(7): 1303-1308.

[5] 杨 珺, 史忠科. 基于改进单高斯模型法的交通背景提取[J]. 光子学报, 2009, 38(5): 1293-1296.

[6] 曹昌霞, 沈小艳, 毕 胜, 等. 基于自适应更新率的滑动平均背景更新算法[C]//全国第 18 届计算机技术与应用学术会议论文集. 合肥: 中国科学技术大学出版社, 2007.

[7] Koller D, Weber J, Huang T. Toward Robust Automatic Traffic Scene Analysis in Real-time[C]//Proc. of International Conf. on Pattern Recognition. Jerusalem, Israel: [s. n.], 1994.

[8] 李全民, 张运楚. 自适应混合高斯背景模型的改进[J]. 计算机应用, 2007, 27(8): 2014-2017.

[9] Arsic D, Hofmann M, Schuller B, et al. Multi-camera Person Tracking and Left Luggage Detection Applying Homographic Transformation[C]//Proc. of AVSS'07. London, UK: [s. n.], 2007.

[10] 梁 华. 多摄像机视频监控中运动目标检测与跟踪[D]. 长沙: 国防科学技术大学, 2009.

[11] Pets Metrics[DB/OL]. (2007-08-01). http://www.petsmetrics.net.

编辑 顾姣健