基于自适应中文分词和近似SVM的文本分类算法

冯 永 李 华 钟 将 叶春晓
(重庆大学计算机学院 重庆 400030)

摘 要 中文分词的难点在于处理歧义和识别未登录词,传统字典的匹配算法很大程度上是依靠字典的代表性而无法有效地识别新词,特别是对于各种行业领域的知识管理。基于二元统计模型的分词算法能很好地适应不同的语料信息,且时间和精度都能满足文本知识管理的应用需要。近似支持向量机是将问题归结成仅含线性等式约束的二次规划问题,该算法的时间复杂度和空间复杂度比传统SVM算法的均有降低。在利用自适应分词算法进行分词的基础上,再利用近似支持向量机进行文本分类。实验表明,该方法能够自动适应行业领域的知识管理,且满足文本知识管理对训练时间敏感和需要处理大量文本的苛刻环境要求,从而具备较大的实用价值。

关键词 自适应中文分词,近似支持向量机,文本分类,知识管理
中图法分类号 TP182 **文献标识码** A

Text Classification Algorithm Based on Adaptive Chinese Word Segmentation and Proximal SVM

FENG Yong LI Hua ZHONG Jiang YE Chun-xiao
(College of Computer Science, Chongqing University, Chongqing 400030, China)

Abstract New words recognition and ambiguity resolving are key problems in Chinese word segmentation. The result of traditional dictionary-based matching algorithm largely depends on the representative of the dictionary so that it can not recognize new words effectively, especially in some professional domains. Chinese word segmentation method in this dissertation is based on 2-gram statistical model and can meet the requirements of application in accuracy and efficiency respectively. PSVM takes classification as a linear equality quadratic programming problem. This dissertation describes a text classification algorithm based on adaptive chinese word segmentation and PSVM, which has faster training speed and smaller memory requirements advantages. Several data sets of experiments showed that the classification algorithm can automatically adapt to knowledge management of some professional domains and has better classfication performance under the condition of time-sensitive.

Keywords Adaptive chinese word segmentation, Proximal support vector machines, Text classification, Knowledge management

在文本知识管理系统中的知识获取、存储和检索及共享等关键处理过程中都需要使用到分词和文本分类技术。中文分词是机器翻译、分类、主题词提取以及信息检索的重要基础。面向文本知识管理的中文分词主要考察其是否有助于提高知识文本信息检索的准确度。难点主要表现在对新词的识别和歧义的解决,这对行业知识新词的识别尤为重要。传统的字典匹配分词其分词性能受限于词典的完备性,从而无法适应现实日益发展的领域知识管理需求。本文从统计理论出发,采用一种能自适应中文语料和领域的分词方法进行分词,然后利用近似支持向量机将文本分类问题归结为仅含线性等式约束的二次规划问题,从而降低训练文本的复杂度和算法时空复杂度。

1 中文分词和文本分类技术

1.1 中文分词技术

中文分词的难点在于处理歧义[1]和识别未登录词。目前国内比较权威的汉语分词系统所采用的分词方法,主要有3种类型[2,3]:基于字典匹配的分词法、基于语料统计的分词法、语义分词法。

从现有文献分析,取得较好效果的基于词典匹配的分词法主要有最大匹配法(MM)[4]、逆向最大匹配法(RMM, OMM, IMM)[5]、双向匹配法、最佳匹配法(OM)。基于词典匹配的分词法,实现简单,实用性强,但该分词法的最大缺点就是词典的完备性不能得到保证。

基于统计的方法是基于汉字同时出现来组成有意义的词的概率,可以用一阶马尔科夫假设和独立性假设来进行分词处理[6,7]。其中具有代表性的方法有互信息、N-gram、最大熵等。基于语料统计的分词方法有许多优点:降低了未登录词的影响,只要有足够的训练文本就易于创建和使用。

部分分词算法采用规则和统计相结合的办法,可以降低

到稿日期:2009-05-09 返修日期:2009-06-30 本文受重庆市自然科学基金(2008BB2183),中国博士后科学基金(20080440699),国家社会科学基金(ACA07004-08)资助。

冯 永(1977-),男,副教授,主要研究方向为知识发现等,E-mail:fengyong@cqu.edu.cn;李 华(1962-),女,副教授,主要研究方向为网络教育等;钟 将(1974-),男,副教授,主要研究方向为知识管理等;叶春晓(1973-),男,副教授,主要研究方向为网络安全等。

统计对语料库的依赖性，充分利用已有的词法信息，同时弥补规则方法的不足[8-10]。

1.2 文本分类技术

文本分类是把一个或者多个预先指定的类别标号自动分配给未分类文本的过程，广泛应用于信息处理、数据挖掘、机器学习、知识管理等领域[11,12]。

一般文本分类需要以下几个步骤：

Step1 获取进行分类的文本集。

Step2 选择文本分类模型。常见的分类模型有 k 最近邻(k-Nearest Neighbor，kNN)[13]、支持向量机(SVM)[14]、朴素贝叶斯分类器(NB)[15]、决策树分类器(Decision Tree)、BP 神经网络(BP Neural Networks)。

Step3 将文本集按照所选分类模型建立每个文本的特征向量。

Step4 用训练数据集构建文本分类器。

Step5 用测试数据集评估文本分类，并根据评估结果调整文本分类器的参数以进行优化。

普遍认为，文本分类的效果和数据集本身的特点(如有的数据集包含噪声，有的分布稀疏，有的字段和属性相关性强)有关系。目前，认为不存在某种方法能够完全适合于各种特点的数据集。

2 算法的理论基础

2.1 n 元(N-gram)统计模型原理

N 元语法的基本思想是一个单词的出现与其上下文中出现的其他单词密切相关。一个句子可以看成一个有联系的字符串序列，可以是字序列，也可以是已知的词构成的词序列。对于一个句子 $w_1w_2\cdots w_k$ 的出现概率用 $P(W)$ 来表示，有：

$$\begin{aligned}P(W)&=P(w_1w_2\cdots w_k)\\&=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_n|w_1w_2\cdots w_{n-1})\\&=\prod_{i=1}^{n}P(w_i|w_1w_2\cdots w_{i-1})\end{aligned}\tag{1}$$

从字的角度来看，该模型认为第 k 个词的出现与前面 $k-1$ 个词相关。为了预测 w_k 的出现概率，就必须知道前面所有词的出现概率，其计算过于复杂甚至是不可能的。

由此可见，N-gram 方法实际上把分词问题转化为求最佳的分词组合 $w_1w_2\cdots w_k$，使得 $P(W)$ 的值最大。

如果假设 w_k 只与其前面出现的 $n-1$ 个词有关，就是 N 元模型。比如只与前面的两个词有关，则称该语言模型是三元模型。公式简化为：

$$P(W)\approx P(w_1)P(w_2|w_1)\prod_{i=3}^{k}P(w_i|w_{i-2}w_{i-1})\tag{2}$$

公式中的概率参数均可以通过大规模语料库来进行计算。

$$P(w_i|w_{i-2}w_{i-1})\approx\frac{count(w_{i-2}w_{i-1}w_i)}{count(w_{i-2}w_{i-1})}\tag{3}$$

其中，$count(L)$ 表示字串 L 在整个语料库中出现的累计次数。

二元语法模型，也叫一阶马尔科夫链，即：

$$P(W)\approx\prod_{i=1}^{k}P(w_i|w_{i-1})\tag{4}$$

本文分词阶段使用 2-gram 二元模型。

2.2 文本分类理论基础

若文本集中的每个文本必须属于且只能属于一个类别，即只能为文本指定一个类标号，那么这种分类称为单标号文本分类(Single-Label)。若文本集中的每个文本可以属于一个或多个类，那么这种分类称为多标号文本分类(Multilabel Text Categorization)。本文的方法，既能支持单标号文本分类，也能支持多标号文本分类。

本文采用保持(holdout)评估法评估分类模型。给定的数据集随机划分为两个独立部分：一个作为训练集；另一个作为测试集。通常训练集占 2/3，测试集占 1/3。利用训练集导出分类模型，再以分类模型对测试集的分类准确率来评估分类模型，如图 1 所示。

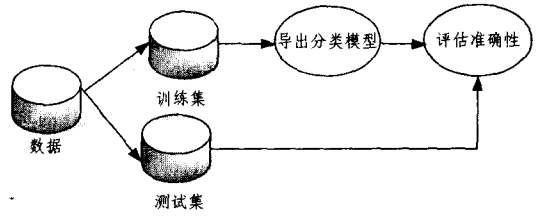

图 1 保持评估法

3 基于自适应中文分词和 PSVM 的文本分类

基于上述理论基础，本文提出一种能够很好地适应各种语料信息并可降低训练文本复杂度及算法时空复杂度的文本分类方法。

3.1 分词的预处理

预处理的主要目的是将长句划分为多个子句，从而降低分词的复杂度。本文的方法有 3 种：

1)利用有限状态机识别待分词文本中最常见的数字、日期以及域名等并以它们为标志将句子划分为子句。例如，利用有限状态机识别年份，如图 2 所示。

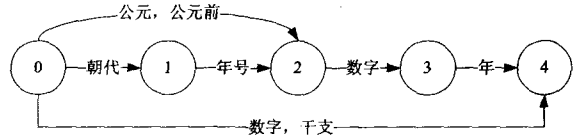

图 2 识别年份的有限状态机

2)对于用上述方法无法处理的长句，本文定义一个连词集合，以基于连词分隔的方式进行处理，这样可以进一步降低长句的概率。本文只使用常出现在语句中部的连接词作为有意义的语句划分词。

3)对于以上两种方法都无法处理的长句，则采用分治的方法，直接将句子划分为 t 个子句(假设除最后一个子句外的前面子句长度为 k)，先对各个子句进行分词，最后归并其结果。其思路如图 3 所示。

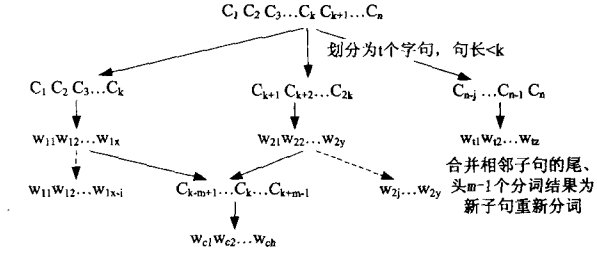

图 3 “分而治之”长句划分思想

3.2 自适应中文分词

使用预处理方法将待处理文本中的长句划分为子句后，对每个子句采用 2-gram 算法，步骤如下。

输入：经过预处理后的文本文件 $s=s_1 s_2 \cdots s_n$；$s_i=c_{i1} c_{i2} \cdots c_{ij}$ 其中，c_{ij} 均为单字；从语料库加工的词频字典。

处理流程：

Step1　用二级 Hash 表加载词频字典并做数据平滑。

Step2　使用词长优先获得二元切分路径。

Step3　使用深度优先算法选择最优路径。

输出：由 s_i 最优路径分词结果 $w_{i1} w_{i2} \cdots w_{ik}$ 组成的 s 分词结果。

分词的后处理主要是用简单的词性搭配规则对 2-gram 切分结果进行歧义发现和处理。将 2-gram 分词结果的词性重新组合为一个二元模型，设定词性搭配阈值进行筛选(本文采用 10)，发现可能产生歧义的近邻二元词，并重新进行切分。

3.3　基于降维的近似支持向量机学习算法

近似支持向量机 PSVM 使用一个超平面 $w \cdot x+b=0$ 来分割正类和负类，但其参数 w 和 b 是通过求解另一个优化问题(如下所示)决定的：

$$\begin{cases}\min \dfrac{1}{2}(||w||^2+b^2)+C\sum\limits_{i=1}^{m}\xi_i^2 \\ s.t.\ y_i(w \cdot x_i+b)+\xi_i=1\end{cases} \tag{5}$$

PSVM 分类器的目标可总结为：使正类尽量靠近 $w \cdot x+b=1$，使负类尽量靠近 $w \cdot x+b=-1$，使两个超平面 $w \cdot x+b=1$ 和 $w \cdot x+b=-1$ 之间的间隔尽量大。

根据式(5)，近似支持向量机的训练和学习过程可以看成一个线性等式约束二次规划问题，本文采用一种基于降维和分块矩阵的文本分类算法。首先介绍一般的等式约束问题的降维 K-T 条件：

$$(ECP)\begin{cases}\min\ f(x) \\ s.t.\ h(x)=0\end{cases} \tag{6}$$

$$f:R^n \to R, h:R^n \to R^m, m \leqslant n, P=n-m$$

$$P(x)=\left(\frac{\partial f(x)}{\partial x_1},\frac{\partial f(x)}{\partial x_2},\cdots,\frac{\partial f(x)}{\partial x_P}\right)^T$$

$$Q(x)=\left(\frac{\partial f(x)}{\partial x_{P+1}},\frac{\partial f(x)}{\partial x_{P+2}},\cdots,\frac{\partial f(x)}{\partial x_n}\right)^T$$

$$N(x)=\begin{pmatrix}\frac{\partial h_1(x)}{\partial x_1} & \frac{\partial h_1(x)}{\partial x_2} & \cdots & \frac{\partial h_1(x)}{\partial x_P} \\ \frac{\partial h_2(x)}{\partial x_1} & \frac{\partial h_2(x)}{\partial x_2} & \cdots & \frac{\partial h_2(x)}{\partial x_P} \\ \vdots & \vdots & \vdots & \\ \frac{\partial h_m(x)}{\partial x_1} & \frac{\partial h_m(x)}{\partial x_2} & \cdots & \frac{\partial h_m(x)}{\partial x_P}\end{pmatrix}$$

$$M(x)=\begin{pmatrix}\frac{\partial h_1(x)}{\partial x_{P+1}} & \frac{\partial h_1(x)}{\partial x_{P+2}} & \cdots & \frac{\partial h_1(x)}{\partial x_n} \\ \frac{\partial h_2(x)}{\partial x_{P+1}} & \frac{\partial h_2(x)}{\partial x_{P+2}} & \cdots & \frac{\partial h_2(x)}{\partial x_n} \\ \vdots & \vdots & \vdots & \\ \frac{\partial h_m(x)}{\partial x_{P+1}} & \frac{\partial h_m(x)}{\partial x_{P+2}} & \cdots & \frac{\partial h_m(x)}{\partial x_n}\end{pmatrix}$$

定理 1　设 $x^0 \in R^n$ 是 ECP 的最优解，f 和 h 连续可微；若矩阵 $M(x^0)$ 是非奇异，则有：

$$P(x^0)=[M(x^0)^{-1}N(x^0)]^T Q(x^0) \tag{7}$$

定理 2　假定有 $h(x^0)=0$；矩阵 $M(x^0)$ 是非奇异；$P(x^0)=[M(x^0)^{-1}N(x^0)]^T Q(x^0)$，则 x^0 是 ECP 的一个 K-T 点。

推论 1　如果 $x^0 \in R^n$ 是方程组

$$P(x)=[M(x)^{-1}N(x)]^T Q(x) \tag{8}$$

的解，使得 $M(x^0)$ 非奇异，则 x^0 是 ECP 的一个 K-T 点。

相对于经典的 K-T 条件，由于式(8)不含对应的 Lagrange 乘子，方程的维数得以降低 m 维(等式约束个数)，因此称式(8)为等式约束问题 ECP 的降维 K-T 条件。

关于线性等式约束的二次规划问题 EQP 及基于降维的二次规划算法，在前期研究[16]中已经报道，这里不再赘述。

对于近似支持向量机的学习过程，可以看成是式(5)所对应的线性等式约束的二次规划问题，该式可以转换为矩阵形式：

$$\begin{cases}\min \dfrac{1}{2}(w^T,b^T,\zeta^T)G(w^T,b^T,\zeta^T)^T \\ s.t.\ (A_1,A_2,A_3)(w^T,b^T,\zeta^T)^T=e\end{cases} \tag{9}$$

其中，$G=\begin{pmatrix}E_n & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & CE_m\end{pmatrix}$，$A_1=\begin{pmatrix}y_1w_1 & \cdots & y_1w_n \\ \vdots & \ddots & \vdots \\ y_mw_1 & \cdots & y_mw_n\end{pmatrix}$，$w=\begin{pmatrix}w_1 \\ \vdots \\ w_n\end{pmatrix}$，$\zeta=\begin{pmatrix}\zeta_1 \\ \vdots \\ \zeta_m\end{pmatrix}$，$A_2=\begin{pmatrix}y_1 \\ \vdots \\ y_m\end{pmatrix}$，$A_3=E_m$，$E_m$ 表示 m 阶单位矩阵，e 表示 $m+n+1$ 维列向量，C 为式(5)中权系数。

记 $A=(A_1,A_2,A_3)$，$x=(w^T,b^T,\zeta^T)^T$，那么式(9)可以转换为线性等式约束的二次规划问题，故可以代入到基于降维的二次规划算法中求解其对应的最优解 x。

由于传统的近似支持向量机采用基于 K-T 条件进行求解，计算复杂度为 $O(n+m)^3$，其中 m 表示训练样本的个数，n 表示训练数据集属性的维数。本文的训练方法计算时间包括降维处理时间和 n 个 n 变量的方程组求解时间，因此计算复杂度为 $O(n^3+m^3)$。空间复杂度相应地由传统 PSVM 算法所需的 $O(n+m)^2$ 降为 $O(n^2+m^2)$。当 m 和 n 接近时，计算复杂度可以降低为原来的 1/4 左右，而空间复杂度则减少一半左右。

3.4　文本分类过程

文本分类过程如图 4 所示。

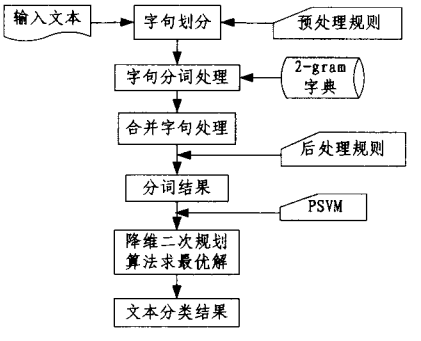

图 4　文本分类过程

4　实验与分析

构建文本分类实验主要有文本预处理、文本特征构建、文本特征选取、对训练集进行学习训练以构造分类器、分类测试和评估等几个步骤。文本预处理主要是分词，在分词的过程中去除停用词。本文的自适应分词方法将与 ICTCLAS 进行比较，ICTCLAS 是能够免费获取使用的最好分词系统。

实验是在 CPU 为 Intel Core2 Duo 1.80 的机器上完成的，内存为 1G，操作系统为 Windows Server 2003。

实验采用的 SVM 为 SVMTorch。它是由瑞士 IDIAP 机构开发的。IDIAP 机构是一个半私有且非营利性的研究所，隶属于瑞士联邦科技研究院。SVMTorch 专门为大规模分类问题量身定制，它可以直接处理多种分类问题，可以从 http://www.idiap.ch/~bengio/projects/SVMTorch.html 上免费获得。

为了验证本文方法的可行性和有效性，本文使用搜狐实验室提供的已分类的语料库。选取其中教育、财经、健康、体育、军事等 5 类文本，每类文本 1990 个，共计 9950 个文本。实验过程中每一类文本训练一个分类器，训练过程中任选 1000 个样本文件作为训练过程中的正例，其它类别中挑出的 4000 个样本为反例。在另外的 990 个文档中进行测试。实验过程中的中文分词技术采用了第 2 节提出的自适应分词方法，为了进一步提高分词精度，实验过程中又添加了 50 篇语料，即每一类文本中采用人工方式选择了 10 篇典型的语料。

本文在实验过程中通过统计方法，在各类文档数据中采用文档特征词的方法，筛选 1500 左右个名词和动词来构成文档的特征向量。文本特征权重分别采用 TFIDF、信息增益、期望交叉熵、互信息 4 种方法来精心测试算法的精度。

4.1 搜狗分类数据集上的分类性能比较

Sogou 语料是较为常见的与工程应用类似的中文语料，为了综合考虑分类精度和召回率以便对比分析，我们对表 1 的结果进行加工，再用宏 F1 进行对比分析。计算方法如下：

$$Macro\_F1=\sum_{i=1}^{m}\frac{N_i}{N}\times F1_i=\sum_{i=1}^{m}\frac{N_i}{N}\times\frac{2\times P_i\times R_i}{P_i+R_i} \qquad (10)$$

其中，N_i 为第 i 类的测试文档数量，在本实验中取值均为 1000，N 为测试文档总数，在本实验中取值为 5000，P 和 R 分别是第 i 类的准确率和召回率，m 为类别数，本实验中取值为 5。

表 1 PSVM 与 KNN，SVM 在自适应/ICTCLAS 分词基础上的性能比较

分词/分类方法	性能	教育	财经	健康	体育	军事
ICTCLAS/KNN	准确率	0.637	0.681	0.730	0.749	0.792
	查全率	0.624	0.716	0.723	0.714	0.812
自适应/KNN	准确率	0.694	0.720	0.767	0.791	0.823
	查全率	0.691	0.744	0.765	0.787	0.837
ICTCLAS/SVM	准确率	0.850	0.768	0.770	0.821	0.843
	查全率	0.823	0.816	0.763	0.774	0.872
自适应/ SVM	准确率	0.881	0.812	0.799	0.862	0.879
	查全率	0.873	0.843	0.798	0.821	0.900
ICTCLAS/PSVM	准确率	0.799	0.642	0.708	0.789	0.884
	查全率	0.734	0.751	0.691	0.765	0.854
自适应/ PSVM	准确率	0.845	0.712	0.757	0.838	0.910
	查全率	0.793	0.796	0.733	0.813	0.877

整理后的 3 种算法的宏 F1 如表 2 所列。

表 2 PSVM 与 KNN，SVM 在自适应/ICTCLAS 分词基础上的 F1 比较

分词/分类方法	Macro_F1
ICTCLAS/KNN	0.718
自适应/KNN	0.737
ICTCLAS/SVM	0.809
自适应/ SVM	0.855
ICTCLAS/PSVM	0.760
自适应/ PSVM	0.813

实验结果表明：基于自适应分词的 PSVM，SVM 和 KNN 获得的准确率、查全率和 F1 明显优于基于 ICTCLAS 分词的 3 种分类器；对于 SVM 而言，在几个数据集上存在准确率和查全率略低的现象，但是 PSVM 的效率高于 SVM 的，处理大量文本将会获得明显优势。

4.2 不同权值模型下的自适应/PSVM 的分类性能

为了便于指导工程应用，本文在不同的文档特征权重选择方法下做了对比实验。实验结果如表 3 所列。

表 3 自适应/PSVM 在不同权值模型下的性能比较

权值模型	性能	教育	财经	健康	体育	军事
TFIDF	准确率	0.845	0.712	0.757	0.838	0.910
	查全率	0.793	0.796	0.733	0.813	0.877
信息增益	准确率	0.841	0.720	0.734	0.823	0.837
	查全率	0.787	0.776	0.741	0.825	0.865
期望交叉熵	准确率	0.839	0.719	0.725	0.819	0.859
	查全率	0.792	0.770	0.747	0.794	0.886
互信息	准确率	0.861	0.736	0.764	0.840	0.925
	查全率	0.813	0.809	0.762	0.833	0.917

计算宏 F1 结果如表 4 所列。

表 4 自适应/PSVM 在不同权值模型下的 F1 比较

权值模型	Macro_F1
TFIDF	0.785
信息增益	0.771
期望交叉熵	0.782
互信息	0.821

实验结果表明：对于文档特征权值模型，采用互信息的方法比传统的 TFIDF 方法要好，而期望交差熵方法和信息增益方法不存在明显改进，甚至可能出现分类性能下降的情况。因此，在实际的工程应用中使用本文的文本分类算法，采用互信息权值的向量空间模型，能够达到快速且有效的分类效果。

结束语 本文提出的文本分类方法在分词阶段克服了传统的字典匹配分词法和一般统计分词法的弱点，使分词处理歧义和识别新词方面有较大的改善；在文本分类阶段采用基于降维近似支持向量机分类算法 PSVM，该算法的时间复杂度由传统算法的 $O(n+m)^3$ 降低为 $O(n^3+m^3)$，空间复杂度由 $O(n+m)^2$ 降低为 $O(n^2+m^2)$。该方法与已有的主要分类算法相比，有较好的性能，尽管较标准 SVM 算法的精度有所下降，但训练时间比标准 SVM 算法的短，可以满足文本知识管理系统应用中对训练时间敏感、需要处理大量文本的苛刻条件，从而具备较大的实用价值。

参 考 文 献

[1] 梁南元. 书面汉语自动分词系统 CDWS[J]. 中文信息学报，1987，2(2)：101-106

[2] 孙茂松，邹嘉彦. 汉语自动分词研究评述[J]. 当代语言学，2001，3(1)：22-32

[3] 冯书晓，徐新，杨春梅. 国内中文分词技术研究新进展[J]. 情报杂志，2002，11：29-30

[4] 黄德根，朱和合，王昆仑，等. 基于最长次长匹配的汉语自动分词[J]. 大连理工大学学报，1999，39(6)：831-835

[5] 吴胜远. 一种汉语分词方法[J]. 计算机研究与发展，1996，33(4)：306-311

[6] 冯冲，陈肇雄，黄河燕，等. 基于 Multigram 语言模型的主动学习中文分词[J]. 中文信息学报，2006，20(1)：50-58

[7] 曹勇刚，曹羽中，金茂忠，等. 面向信息检索的自适应中文分词系统[J]. 软件学报，2006，17(3)：356-363

（下转第 293 页）

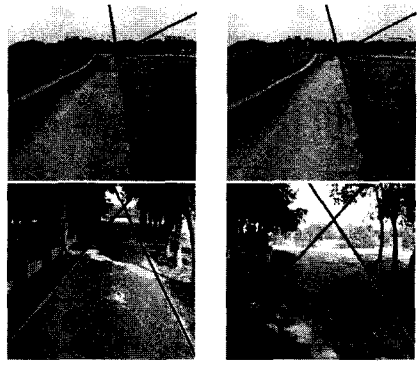

图 6　普通道路实验环境

结束语　主动导航车辆中利用视觉系统算法理解车辆的环境非常重要。为了在图像中更精确地检测因为较多光照阴影、边缘信息模糊等因素造成影响的道路边缘，本文提出了一种新的方法。该方法以直线变形模型为基础，在先验知识的辅助下定义后验概率分布描述道路特征结构，利用粒子群优化算法搜索与图像匹配的最优模板。同时利用车辆运动的动态模型和粒子滤波算法，综合预测信息，达到了对道路边缘检测较为准确的估计。实验证明，本文算法实现了较为准确的道路边缘或者行道线的检测，能够有效地减少噪声因素的影响，具有较好的实用性。

参 考 文 献

[1] Chapuis R, Aufrere R, Chausse F. Accurate road following and reconstruction by computer vision[J]. IEEE Transactions on Intelligent Transportation Systems, 2002, 3(4): 261-270

[2] Li Qing, Zheng Nan-ning, Cheng Hong. Springrobot: a prototype autonomous vehicle and its algorithms for lane detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2004, 5(4): 300-308

[3] Wang Yue, Teoh E K, Shen Dinggang. Lane detection using B-Snake[J]. Image and Vision Computing, 2004, 22(4): 269-280

[4] Park J W, Lee J W, Jhang K Y. A lane-curve detection based on an LCF[J]. Pattern Recognition Letters, 2003, 24(14): 2301-2313

[5] Bcher T, Curio C, Edelbrunner J, et al. Image processing and behavior planning for intelligent vehicles[J]. IEEE Transactions on Industrial Electronics, 2003, 50(1): 62-75

[6] Crisman J D, Thorpe C E. UNSCARF: A Color Vision System for the Detection of Unstructured Roads[C]// Proc. of IEEE Int'l Conf on Robotics and Automation. 1991: 2496-2501

[7] Crisman J D, Thorpe C E. SCARF: A Color Vision System that Tracks Roads and Intersections[J]. IEEE Trans. on Robotics and Automation, 1993, 9(1): 49-58

[8] 孙涵. 基于红外图像的道路识别与运动目标跟踪[D]. 南京, 2005

[9] Thorpe C, et al. Vision and navigation for the Carnegie-Mellon NAVALAB[J]. IEEE Trans. Pattern Analysis and Machine Intelligence, 1988, 10(3): 362-73

[10] Zhang Jin-you, Naqel H. Texture - based segmentation of road images[C]// IEEE Intelligent Vehicles apos 94 Symposium. Oct 1994: 260- 265

[11] Eberhart R C, Kennedy J. A new optimizer using particle swarm theory[C]// Proc. 6th Int. Symp. Micro Machine and Human Science. 1995: 39-43

[12] Engelbrecht A P, Ismail A. Training product unit neural networks[J]. Stability Control: Theory Appl., 1999, 2(1): 59-74

[13] Isard M, Blake A. Condensation-condition density propagation for visual tracking[J]. International Journal of Computer Vision, 1998, 29: 5-28

[14] Arulampalam M S, et al. A tutorial on particle filters for online nonlinear/non-gaussian Bayesian tracking[J]. IEEE Trans. Signal Processing, 2002, 50(2): 174-188

[15] Doucet A, et al. On sequential Monte Carlo sampling methods for Bayesian filtering[J]. State Compute, 2000, 10(3): 197-208

(上接第 254 页)

[8] 张华平，刘群. 基于 N-最短路径方法的中文词语粗分模型[J]. 中文信息学报，2002，16(5)：1-7

[9] Goh C-L, Asahara M, Matsumoto Y. Chinese Word Segmentation by Classification of Characters[J]. Computational Linguistics and Chinese Language Processing, 2005, 10(3): 381-396

[10] Wang Zhuoran, Liu Ting. Chinese Unknown Word Identification Based on Local Bigram Model[J]. International Journal of Computer Processing of Oriental Languages, 2005, 18(3): 185-196

[11] Sebastiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34(1): 1-47

[12] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859

[13] 李荣陆，胡运发. 基于密度的 kNN 文本分类器训练样本裁剪方法[J]. 计算机研究与发展，2004，41(4)：539-545

[14] 庄东，陈英. 基于加权近似支持向量机的文本分类[J]. 清华大学学报：自然科学版，2005，45(S1)：1787-1790

[15] Kim S-B. Some Effective Techniques for Naive Bayes Text Classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1457-1466

[16] 钟将，温罗生，冯永，等. 基于近似支持向量机的 Web 文本分类研究[J]. 计算机科学，2008，35(3)：167-169

(上接第 281 页)

[5] Liao S X, Pawlak M. On Image Analysis by the Methods of Momnets[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(3): 254-266

[6] Flusser J. Moment Invariants in Image Analysis[J]. Proceedings of World Academy of Science, Engineering and Technology, 2006, 11(2): 196-201

[7] Chong C W, Raveendran P, Mukundan R. Translation invariants of Zernike moments[J]. Pattern Recognition, 2003, 36(9): 1765-1773

[8] Xu Dong, Li Hua. Geometric moment invariants [J]. Pattern Recognition, 2008, 41(1): 240-249

[9] 丁险峰，吴洪，张宏江. 形状匹配综述[J]. 自动化学报，2001，27(5)：678-694

[10] 姚玉荣，章毓晋. 利用小波和矩进行基于形状的图象检索[J]. 中国图像图形学报，2000，5(3)：206-210

[11] 夏永泉，刘正东，杨静宇. 不变矩方法在区域匹配中的应用[J]. 计算机辅助设计与图形学学报，2005，17(10)：2152-2156

[12] 云挺，顾磊，吴慧中. 基于 Zernike 矩的区域匹配方法[J]. 中国图象图形学报，2008，13(8)：1517-1524

[13] 叶斌，彭嘉雄. 伪 Zernike 矩不变性分析及其改进研究[J]. 中国图象图形，2003，8(3)：246-252

[14] http://www.ist.temple.edu/~hbling/

[15] Zhang D S, Lu G. Review of Shape Representation and Description Techniques[J]. Pattern Recognition, 2004, 37(1): 1-19

[16] Iqbal Q, Aggarwal J K. Applying Pecepture Grouping to Content-based Image Retrieval: Building images[C]// IEEE Workshop on Content-based Aeeess of Image and Video Libraries. Fort Collins, CO, 1999: 42-48