基于太赫兹时域光谱系统的橡胶分类识别*

殷贤华　王　宁　陈晶溪
（桂林电子科技大学电子工程与自动化学院 桂林 541004）

摘　要：基于太赫兹时域光谱（THz-TDS）系统对4种橡胶样品进行检测，分别采用核主成分分析（KPCA）和核典型相关分析（KCCA）方法对橡胶太赫兹光谱进行特征提取，引入PCA和CCA作为对比，再结合支持向量机（SVM）建立分类模型，对橡胶进行分类识别，最后以偏最小二乘判别法（PLS-DA）的识别结果作为参考。结果表明，SVM结合特征提取方法可以对橡胶的光谱进行分类识别，KPCA-SVM对吸收谱的分类效果最佳，而PLS-DA对折射谱的分类效果要优于SVM，同时，KPCA对光谱的特征提取效果要优于标准的KCCA方法。实验为橡胶的识别分析提供了新的方法。

关键词：橡胶；KPCA；KCCA；支持向量机；PLS-DA
中图分类号：O433.4　TP391　TN209　　**文献标识码**：A　　**国家标准学科分类代码**：510.20

Rubber classification and recognition based on THz time-domain spectroscopy system

Yin Xianhua　Wang Ning　Chen Jingxi
（School of Electrical Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China）

Abstract: Based on the terahertz time-domain spectroscopy system, 4 kinds of rubber samples were detected. Comparing with PCA and CCA, Kernel principal component analysis (KPCA) and kernel canonical correlation analysis (KCCA) were carried out on the feature extraction of rubber terahertz spectrum. The classification model was established by support vector machine (SVM) to classify the rubber samples. Finally, the recognition results of partial least squares (PLS-DA) are used as the reference. The experimental results show that SVM can be used to classify the spectrum of rubber combined with the feature extraction methods. The classification effect of KPCA-SVM on the absorption spectrum is the best, and PLS-DA is better than SVM on refraction spectrum classification. Meanwhile, KPCA is better than the standard KCCA method for the feature extraction of the spectrum. The experiment provides a new method for the identification and analysis of rubber.

Keywords: rubber; KPCA; KCCA; support vector machine; PLS-DA

1　引　言

橡胶已经成为工业生产和人类生活中重要的材料。随着工业的发展和化学技术的提高，橡胶的种类日益丰富，在工业领域得到广泛的应用。在橡胶的生产和产品分析中，对不同橡胶材料的检测和鉴别是一项重要和复杂的过程，针对橡胶的检测技术和标准仍存在较大差异。太赫兹波（THz）在电磁波谱中位于微波和红外之间[1-2]，由于许多物质在太赫兹波段包含丰富的物理和化学信息，所以THz检测技术在物质检测和鉴别等方面受到广泛的关注和研究。

近几年，国内外的学者对太赫兹检测橡胶的课题进行了研究。德国的PETERS O等人[3]将THz-TDS系统引入到橡胶生产挤出机的监控中，研究了添加剂在橡胶混合生产过程中对太赫兹信号的影响；日本的HIRAKAWA Y等人[4]通过THz时域光谱系统检测了天然橡胶、丁腈橡胶、丁苯橡胶和其他包括炭黑、氧化锌、硫等添加剂的吸收光谱，证明了太赫兹可以作为无损检测橡胶的一种新方法；国内的苗青等人[5]研究了氯丁橡胶、丁腈橡胶和三元乙丙橡胶在0.2～1.8 THz频段的光学性能和光谱特性。

本研究针对4种硫化橡胶产品进行了太赫兹光谱检测实验，分别利用几种特征提取算法（PCA、CCA、KPCA、KCCA）结合SVM模型对吸收光谱和折射率光谱的数据

收稿日期：2016-03
* 基金项目：广西自然科学基金项目（2015GXNSFBA139252）、广西自动检测技术与仪器重点实验室基金项目（YQ15104）资助

进行了分类识别研究，取得了较好的分类效果，并与 PLS-DA 的分类结果作对比，为橡胶的太赫兹光谱分析提供了一定参考价值。

2 实验部分

2.1 实验装置与样品制备

实验装置采用美国 Zomega 公司研制的 Z-3 太赫兹时域光谱系统（THz-TDS），激光器使用德国 TOPTICA Photonics AG 公司的超快飞秒光纤激光器 FemtoFiber pro NIR[6]。系统装置原理如图 1 所示。激光器中心波长为 780 nm，平均功率大于 140 mW，脉冲宽度小于 100 fs，重复频率为 80 MHz。激光束作为光源被 λ/2 波片分为泵浦光和探测光，泵浦光激发大孔径 LT-GaAs 光导天线产生 THz 脉冲，探测光利用电光采样原理探测 THz 波的电场强度，探测元件为 ZnTe 电光晶体。通过扫描探测激光脉冲和 THz 脉冲的相对时间延迟，获得 THz 脉冲随时间变化的电场波形，并从中提取吸收系数等物理参量。系统的光谱范围是 0.1～3.5 THz，频谱分辨率好于 5 GHz，信噪比大于 70 dB，数据采集时间约 1 min。为减少空气中水分对 THz 波的吸收，THz-TDS 系统的光路部分装在一个密闭的箱体内，实验时充以干燥空气使湿度小于 2%，测量在室温环境下进行。

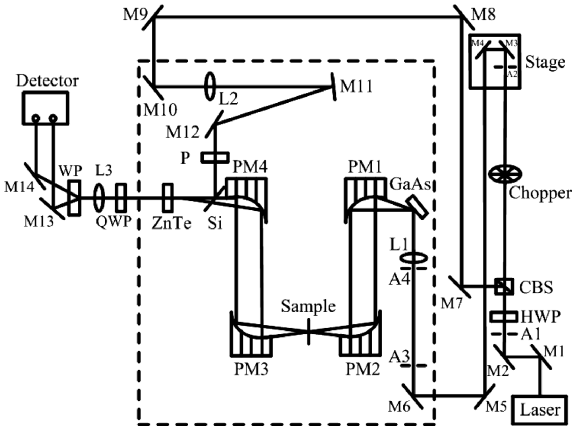


图 1 透射式 THz-TDS 系统原理

实验所用的橡胶样品包括天然橡胶（NR）、三元乙丙橡胶（EPDM）、氯丁橡胶（CR）和丁腈橡胶（NBR），它们均是硫化加工后的产品，颜色为黑色。经测量，NR、EPDM、CR 和 NBR 4 种橡胶的厚度分别是 1.08 mm、1.03 mm、1.20 mm、1.20 mm。为便于实验，每个样品的长和宽都裁剪到 10 mm 左右。

2.2 数据处理

通过 THz-TDS 系统可以获得透过样品的太赫兹波的时域信号，包括了幅度和相位信息。然后采用 Dorney 和 Duvillatet 等人提出的利用 THz-TDS 技术提取材料光学参数的模型，对实验获得的参考信号和样品信号进行快速傅里叶变换得到它们的频率谱 $E_{ref}(\omega)$ 和 $E_{sam}(\omega)$，根据 $E_{ref}(\omega)$ 和 $E_{sam}(\omega)$ 就可以得到样品的折射率 $n(\omega)$ 和吸收系数 $\alpha(\omega)$。$n(\omega)$ 和 $\alpha(\omega)$ 可以通过以下式获得：

$$n(\omega)=\frac{c\varphi(\omega)}{\omega d}+1 \tag{1}$$

$$\alpha(\omega)=\frac{2}{d}\ln\frac{4n(\omega)}{A(\omega)(n(\omega)+1)^2} \tag{2}$$

式中：c 是光速，$\varphi(\omega)$ 和 $A(\omega)$ 分别是样品信号和参考信号的相位差和振幅比，ω 是太赫兹波振动的角频率，d 是样品的厚度。

由于仪器自身的影响，所测频域光谱信号的两端信噪比比较低，所以剔除掉两端信息，保留中间信噪比高的区域进行分析。Z-3 系统的有效光谱探测范围是 0.1～3.5 THz，取其中的 0.3～1.6 THz 频段的光谱数据进行分析。

实验获得的几种橡胶样品的吸收系数光谱曲线和折射率光谱曲线如图 2 和图 3 所示。每个样品的光谱数据是通过对样品进行 3 次扫描，然后取平均值获得。由图 2 可知，4 种橡胶样品在 0.3～1.0 THz 范围内随频率近似呈线性增加的趋势，在 1.0～1.6 THz 范围内出现较大的振荡，吸收峰更多，说明在这个频段，材料对太赫兹波的吸收最强。4 种橡胶样品在 0.3～1.6 THz 之间的最大特征吸收峰的位置分别为 1.27 THz、1.26 THz、1.22 THz 和 1.21 THz，这 4 个最大特征吸收峰的位置非常接近，不易观察区分。另外，因为硫化橡胶在工业加工过程中的添加成分很多是相同的，如硫和炭黑等，所以成品的太赫兹光谱某些区域的峰值比较接近，难以区分，需要采用光谱特征提取方法，建立相关分类模型进行鉴别。

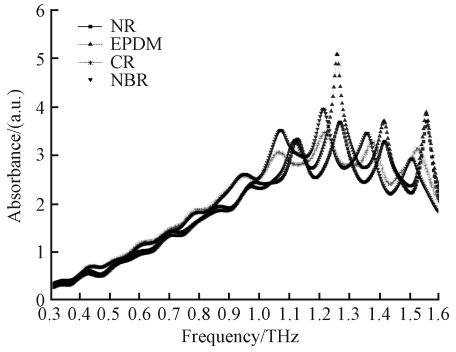


图 2 橡胶样品的吸收系数谱

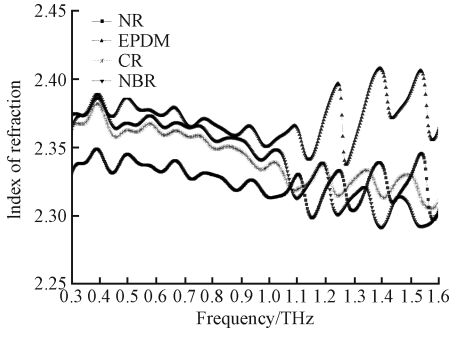


图 3 橡胶样品的折射率谱

3 特征提取和分类方法

3.1 特征提取方法

太赫兹光谱曲线数据往往存在数据量较大、维数较高、吸收特征不明显等问题，需要对光谱数据进行特征提取，降低数据维数，这样可以简化数据、减少冗余信息，同时抑制部分噪声的干扰，对光谱的分类识别有重要影响。本文主要采用核主成分分析法(KPCA)和核典型相关分析法(KCCA)，同时引入 PCA 和 CCA 作为对比。

KPCA 是 PCA 的改进算法，采用非线性方法提取主成分[7]，即 KPCA 通过非线性变换将输入数据空间映射到高维特征空间，使非线性问题转化为线性问题，然后在高维空间中利用 PCA 方法提取主成分，在保持原数据信息量的基础上达到降维目的。与 PCA 相比，KPCA 能抽取更多的主分量，有效捕捉数据的非线性特征，且对原始数据的分布没有要求。

KCCA 是在典型相关分析(CCA)的基础上引入核函数改进而来[8]。CCA 主要处理两个随机矢量之间的相互依赖关系，但是只能提取数据的线性特征。当原始数据的特征存在非线性关系时，使用 KCCA 进行特征提取比 CCA 更加有效。

KCCA 的基本思想是通过非线性映射将 R^d 空间的原始数据样本 x 由低维空间映射到高维特征空间 F，在高维空间中运用 CCA 方法进行特征提取，在保持原数据信息量的基础上实现降维目的[8]。设非线性映射 $\varphi: R^d \rightarrow F$，$x \rightarrow \varphi(x)$。定义原始数据矩阵 $\boldsymbol{X}=(x_{ij})^{\mathrm{T}}$，$x_{ij} \in R^d$，表示第 i 个类别中的第 j 个样本。矩阵 $\boldsymbol{Y}$ 表示每个样本所属类别。经过非线性映射 φ 后，$\boldsymbol{X}$ 变为 $\boldsymbol{X}_{\varphi}=[\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_n)]^{\mathrm{T}}$，$\boldsymbol{Y}$ 保持不变，样本经过非线性变换后在特征空间中的内积运算，可使用满足 Mercer 条件的正定核函数 $\boldsymbol{k}(x,y)=\boldsymbol{\varphi}(x)^{\mathrm{T}}\boldsymbol{\varphi}(y)$。本文采用高斯径向基(RBF)核函数，$\boldsymbol{k}(x,y)=\exp(\frac{-\|x-y\|^2}{2\sigma^2})$。

样本矩阵 $\boldsymbol{X}_{\Phi}$ 定义矩阵 $\boldsymbol{K}=\boldsymbol{X}_{\varphi}X_{\varphi}^{\mathrm{T}}$，$n\times n$ 对称阵 K 的第 i 行第 j 列元素是 $K_{ij}=k(x_i,x_j)$。KCCA 的目的是求解两个投影向量 a_{Φ} 和 b，使如下的相关系数最大：

$$\max_{a_{\varphi},\beta}(r(a_{\varphi},b))=\boldsymbol{a}_{\varphi}^{\mathrm{T}}\boldsymbol{X}_{\varphi}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{b} \tag{3}$$

$$\text{s.t.}\quad \boldsymbol{a}_{\varphi}^{\mathrm{T}}\boldsymbol{X}_{\varphi}^{\mathrm{T}}\boldsymbol{X}_{\varphi}\boldsymbol{a}_{\varphi}=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{b}=1 \tag{4}$$

参考文献[8]提供的推导过程，最后可以得到一个非线性的特征提取投影方程为：

$$\text{feature}(z)=(\alpha_1,\alpha_2,\cdots,\alpha_{C-1})^{\mathrm{T}}\boldsymbol{X}_{\varphi}\boldsymbol{\Phi}(z)=(\alpha_1,\alpha_2,\cdots,\alpha_{C-1})^{\mathrm{T}}\boldsymbol{K}_z$$

式中：$\boldsymbol{K}_z$ 表示 n 维的核函数。$C-1$ 是矩阵 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 的秩，$\boldsymbol{\alpha}_k$ 是对应的特征矢量，$k=1,2,\cdots,C-1$。

3.2 支持向量机

支持向量机(SVM)是VAPNIKV首先提出的一种基于统计学习理论的机器学习算法，针对有限样本情况，通过结构风险最小化(SRM)原则提高学习机泛化能力，实现经验风险和置信范围的最小化，将线性不可分问题通过核函数映射到高维空间，通过建立最优分类平面使其线性可分。

SVM 的分类原理[9]是在特征空间中构造分类超平面，分类函数用 $f(x)=\text{sgn}(\boldsymbol{\omega}^{\mathrm{T}}x+b)$ 表示，$\boldsymbol{\omega}^{\mathrm{T}}x+b=\pm 1$ 表示两个平行超平面，分类间隔是 $2/\|\boldsymbol{\omega}\|^2$，分类的最合适标准是使分类间隔达到最大，即求 $\|\boldsymbol{\omega}\|$ 最小。对于非线性问题引入核函数，求最优分类面的问题可转化为对偶化问题，即为：

$$\min Q(\alpha)=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum_{i=1}^{n}\alpha_i \tag{6}$$

$$\text{s.t.}\quad \alpha_i\geqslant 0,\quad \sum_{i=1}^{n}y_i\alpha_i=0,(i=1,2,\cdots,n) \tag{7}$$

最优分类函数变为：

$$f(x)=\text{sgn}[\sum_{i=1}^{n}\alpha_i^* y_i k(x_i,x)+b^*] \tag{8}$$

式中：$k(x_i,x)$表示满足 Mercer 条件的核函数。任选一支持向量，b 可由下式求出：

$$y_i[\sum_{i=1}^{n}\alpha_i^* y_i k(x_i,x)+b^*]=1 \tag{9}$$

3.3 偏最小二乘判别分析

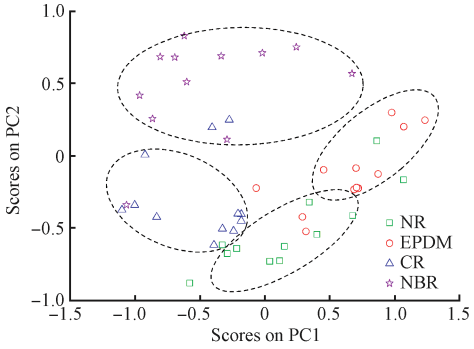
偏最小二乘判别分析(PLS-DA)是在 PLS 回归基础上的统计判别分析方法，具有 PLS 特征提取和降噪等优点，可以对输入变量进行有效降维，并获得良好的分类效果，适合样本观测数少，解释变量数多的情况。

利用 PLS-DA 方法对太赫兹光谱进行分类的原理是[10]：1)根据样本实际不同类别建立校正集样本的分类变量；2)利用 PLS 回归方法建立 THz 光谱数据与分类变量之间的关系模型；3)根据分类变量和光谱特征数据的 PLS 模型，对验证集样本的分类变量值进行预测，判定样本的类别，最后与验证集样本实际类别进行对比得出分类的正确率。

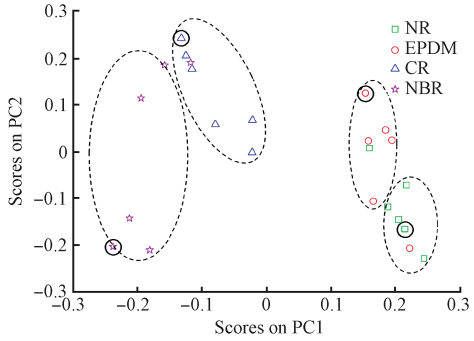
4 实验结果与分析

实验的软件平台采用 MATLAB 2012b，使用 Libsvm 工具箱建立 SVM 的分类模型。利用 THz-TDS 系统和数据处理算法获取 4 种橡胶样品(NR、EPDM、CR、NBR)的太赫兹吸收光谱数据和折射光谱数据，取频率在 0.6～1.6 THz之间光谱数据的 300 个样本点作为实验原始数据。样品一共 48 个，每种橡胶 12 个，其中 28 个样品作为训练集，其他 20 个作为测试集，通过特征提取方法提取解释性更强的特征数据，再利用 SVM 对特征数据进行分类识别。经过反复测试，应用于本实验的 C-SVM 的惩罚参数 C 和核函数参数 σ 分别设置为 2 和 1，核函数采用高斯径向基(RBF)，此时分类模型的分类效果基本处于最佳状态。

针对4种橡胶的吸收光谱数据，分别使用KPCA和KCCA方法，求出特征数据的2D主成分得分图，如图4(a)所示为KPCA的前两个投影数据的主成分得分图，如图4(b)所示为是KCCA的主成分得分图，图中圆圈标出的是KCCA的训练集数据在特征空间的投影点。从图中可以看出，不同种样品吸收光谱的特征数据呈现出不同的区域分布情况。从KPCA的得分图中，特征数据的分布区域更为广阔，某些重叠区域的特征数据可能对分类的准确率产生影响；从KCCA的得分图中，训练集的特征投影被集中于4个象限的不同点上，通过分类实验看出，离训练点距离较远的预测投影点会对分类的结果产生影响。



(a) KPCA特征数据主成分得分图



(b) KCCA特征数据主成分得分图

图4 KPCA和KCCA特征数据主成分得分图

KPCA方法是通过提取样本中解释信息最多的主成分进行特征筛选，保留了原始数据中较多的有用信息，适用的范围更广；KCCA理论上等价于KPCA与CCA的组合，是基于协方差的特征降维方法，通过对输入空间数据集的熵估计有贡献的KPCA轴方向上的投影达到转换和降维，这些轴不一定对应核矩阵中最大的特征值和特征向量，适用于两个随机矢量所在的样本空间协方差矩阵非奇异的情况。另一方面，KPCA和KCCA都没有用到已知样本的类别信息，在特征提取中可能会丢失掉对分类有用的鉴别信息。

分别采用4种特征提取方法(PCA、CCA、KPCA、KCCA)对原始光谱进行了降维处理，将特征数据引入SVM识别模型，求出分类识别的正确率。作为对比，利用PLS-DA对原始光谱数据进行识别分析。如表1所示为不同降维数下，采用几种特征提取方法后SVM的识别精度，其中PLS-DA是原始光谱未经特征提取而进行识别的精度。经测试，当特征提取的维数为3，采用KPCA-SVM的识别精度达到了最佳(80%)，而PLS-DA在特征维数为7时可以达到最大75%。经过PCA、KPCA和KCCA方法后的特征数据的SVM识别精度均高于未进行特征提取而直接进行SVM分类的精度(45%，表中未给出)。从表中也可以看出，特征维数越高并不能代表提取出了原始光谱数据更多的有效信息。

表1 吸收光谱分类结果 (%)

维数	PCA-SVM	CCA-SVM	KPCA-SVM	KCCA-SVM	PLS-DA
3	70	30	80	60	65
5	60	30	55	55	55
7	55	35	55	60	75
10	65	30	50	55	65
15	75	35	55	50	50

针对橡胶样品的THz折射率光谱进行分类实验的结果如表2所示。当特征维数较低到3维时，利用PLS-DA进行光谱识别的正确率达到了100%，PCA-SVM的识别率达到95%，而KPCA-SVM和KCCA-SVM的识别率都较低，说明了KPCA和KCCA没有利用类别信息的弊端；随着特征维数增大，PLS-DA的识别率降低，KPCA-SVM识别率升高到85%，优于KCCA-SVM的识别效果。

表2 折射率光谱分类结果 (%)

维数	PCA-SVM	CCA-SVM	KPCA-SVM	KCCA-SVM	PLS-DA
3	95	60	75	70	100
5	70	60	70	75	85
7	90	75	65	75	90
10	85	75	85	70	85
15	90	60	85	75	70

通过实验结果，橡胶样品的折射率光谱的分类效果要优于吸收光谱；另外，特征维数的选择对SVM和PLS-DA的分类识别效果有较大影响，采用合适的降维维数会使分类效果达到最佳，另外，PLS-DA算法对数量较少且特征维数低的样本的分类效果较好。

5 结 论

本文分别采用PCA、CCA、KPCA和KCCA方法对橡胶的吸收光谱和折射率光谱进行了特征值提取，对KPCA和KCCA降维的特征数据主成分得分图进行了分析，引入SVM模型，对4种橡胶样品进行了分类识别研究，并对比PLS-DA的识别效果。实验表明，不同的特征降维数和特征提取算法对SVM识别结果有重要影响，采用KPCA-SVM在对吸收光谱分类识别中表现最好，而PLS-DA对折射光谱的分类可以达到100%的准确率，KPCA对光谱的特征提取效果要优于KCCA。考虑到实验样本数较少及存在的实验误差，针对特征维数与特征提取算法的关系有待进一步研究。实验为橡胶材料的太赫兹光谱分类识别应用提供了一定的参考。

参 考 文 献

[1] 姜万顺，邓建钦. 太赫兹测试测量技术与仪器研究进展[J]. 国外电子测量技术，2014，33(5)：20-23.

[2] 赵国忠，申彦春，刘影. 太赫兹技术在军事和安全领域的应用[J]. 电子测量与仪器学报，2015，29(8)：1097-1101.

[3] PETERS O，SCHWERDTFEGER M，WIETZKE S，et al. Terahertz spectroscopy for rubber production testing [J]. Polymer Testing，2013(32)：932-936.

[4] HIRAKAWA Y，OHNO Y，GONDOH T，et al. Nondestructive evaluation of rubber compounds by terahertz time-domain spectroscopy[J]. Journal of Infrared，Millimeter，and Terahertz Waves，2011，32(12)：1457-1463.

[5] 苗青，田璐，赵昆，等. 三种橡胶材料的太赫兹光谱研究[J]. 现代科学仪器，2011(5)：110-113.

[6] 陈涛，李智，莫玮，等. 太赫兹时域光谱的药物多组分同时定量测定[J]. 光谱学与光谱分析，2013，35(5)：1220-1225.

[7] 陈如清. 基于KPCA-MVU的噪声非线性过程故障检测方法[J]. 仪器仪表学报，2014，35(12)：2673-2680.

[8] 李志农，张芬，何旭平. 基于小波-KCCA的非线性欠定盲分离方法研究[J]. 仪器仪表学报，2014，35(3)：601-606.

[9] 李太福，易军，苏盈盈，等. 基于KCCA的虚假邻点判别的非线性变量选择[J]. 仪器仪表学报，2012，33(1)：213-220.

[10] 吴哲君，赵忠华，唐雷. 基于SVM的行人步态实时分类方法[J]. 电子测量技术，2015，38(7)：41-44.

作 者 简 介

殷贤华，1974年出生，副教授，测控技术与仪器系主任。主要研究方向为自动测试总线与系统、集成电路测试理论和技术等。

E-mail：8455890@qq.com

王宁(通讯作者)，1990年出生，研究生。主要研究方向为自动测试系统、太赫兹检测技术。

E-mail：wangningngu@126.com