基于太赫兹时域光谱技术与 PCA-SVM 的转基因大豆油鉴别研究

张文涛[1,2],李跃文[1,2],占平平[1,2], 熊显名[1,2]

(1. 桂林电子科技大学 电子工程与自动化学院,广西 桂林 541004;
2. 广西高校光电信息处理重点实验室,广西 桂林 541004)

摘 要:太赫兹时域光谱(THz-TDS)技术是基于飞秒超快激光技术的 THz 波段光谱测量新技术,具有较强的光谱分辨本领以及良好的透视性和安全性,在物质检测方面具有广泛的应用价值。文中在采用太赫兹时域光谱技术对转基因大豆油光谱检测的基础上结合主成分分析方法(PCA)及支持向量机(SVM),构建 PCA-SVM 模型对转基因大豆油进行鉴别。首先,从样品在太赫兹波段测得的时域光谱中得到其吸光度光谱;然后,将其作为输入源导入 PCA-SVM 模型中,剔除冗余数据、降低数据维数并鉴别。实验结果表明,所建立的 PCA-SVM 模型能准确识别校验集,可以准确地对转基因大豆油进行鉴别。研究结果表明:太赫兹时域光谱技术可以实现转基因大豆油的快速、无损检测,在食品安全领域有广泛的应用前景。

关键词:转基因大豆油; 太赫兹; 主成分分析; 支持向量机; 无损检测
中图分类号:O433.4 **文献标志码**:A **DOI**:10.3788/IRLA201746.1125004

Recognition of transgenic soybean oil based on terahertz time-domain spectroscopy and PCA-SVM

Zhang Wentao[1,2], Li Yuewen[1,2], Zhan Pingping[1,2], Xiong Xianming[1,2]

(1. Institute of Electrical Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China;
2. Guangxi Colleges and Universities Key Laboratory of Optoelectronic Information Processing, Guilin 541004, China)

Abstract: Terahertz time-domain spectroscopy (THz-TDS) is a new spectroscopic measurement technique based on ultra-fast femtosecond laser technology. THz-TDS technology has been applicated in material detection widely due to its good penetration and safety. In this paper, a method of indentifying transgenic soybean oil with THz-TDS technology was proposed. The methods of the Principle Component Analysis (PCA) and Support Vector Machine (SVM) were based on spectral analysis in the terahertz (THz) range. First, the absorbance spectrum was extracted from the original THz time-domain spectra data of the sample, and then imported into the model of PCA-SVM as an input source import. The utilization of the dimension-reduced data in PCA-SVM model can recognize the validation set accurately, and the

收稿日期:2017-03-05; **修订日期**:2017-04-03
基金项目:国家自然科学基金(61565004);广西自然科学基金(2014GXNSFGA118003);
桂林市科学研究与技术开发课题(20140127-1,20150133-3)
作者简介:张文涛(1976-),男,教授,博士,主要从事光电检测、纳米计量及激光技术方面的研究。Email:uestczrk@126.com

transgenic soybeans can be accurately identified. The research results show that THz-TDS technology can identify the transgenic soybeans rapidly and nondestructively, and which could be widely applied in field of food safety inspection.

Key words: transgenic soybean oil; terahertz; principle component analysis; support vector machine; nondestructive testing

0 引 言

1983 年,世界上最早的转基因作物(烟草)诞生[1]。1996 年转基因作物在美国开始商业化种植,2015 年是其商业化的 20 周年。20 年来,全球转基因作物累计种植面积达到空前的 2×10^8 km^2, 相当于中国总面积(9.6×10^7 km)或美国总面积(9.4×10^7 km^2)的 2 倍多。而其中转基因大豆约占 50%(1×10^8 km^2),其种植面积已经超过中国或美国的国土面积[2]。转基因大豆通常并不直接被人们食用, 而是当作饲料或深加工成各种油制品。含有转基因大豆成分的食用油早已进入人们的日常生活中, 然而到目前为止还没有任何的研究能够完全否定其潜在的危害性。因此,基于对公众食品安全的考虑, 转基因大豆油的检测鉴别具有重要意义。目前转基因产品的检测方法中,应用最为广泛的是聚合酶链式反应 (Polymerase Chain Reaction,PCR)[3-4]。这是一种基于 DNA 的分析检测技术,但加工后的食用油中 DNA 含量较低,且破坏严重,要提取大豆深加工产品中的 DNA,成本较高[5]。还有少数研究通过其他手段提取 DNA,提取过程繁杂,较为耗时[6-7]。

近年来太赫兹时域光谱技术的出现为生物检测提供了一种新方法。太赫兹(terahertz,THz)辐射是频率介于微波和红外波段之间的电磁辐射, 其波长在 3 000~30 μm 之间,其特殊的频段位置(位于电子学和光子学之间的过渡区) 使其在理论研究和实际应用中都具有极高的研究价值[8-10]。有机分子和生物分子在 THz 频段的光谱特征与其大幅度振动及分子间的互作用相关,不同基因片段转录的氨基酸结构不一样,导致在 THz 波段出现的吸收峰位不一样[11]。而且 THz 波的能量非常低(约 1~10 meV),不会对生物大分子造成伤害。这些特性在对生物大分子的检测时显得尤为重要[12-14]。

1 实验装置和技术原理

1.1 实验装置

实验采用的太赫兹时域光谱系统(terahertz time-domain spectroscopy system,THz -TDS) 是美国 Zomega 公司生产的 Z-3 系统,可选择反射或者透射模式。文中实验采用透射模式[15],光谱范围在0.1~3.5 THz 之间,动态范围峰值大于 70 dB,数据采集时间大约需 60 s,上位机界面为 LabVIEW 编译构成。

THz-TDS 装置如图 1 所示,实验采用的超快飞秒激光器由 Toptica Photonics AG 公司生产,中心波

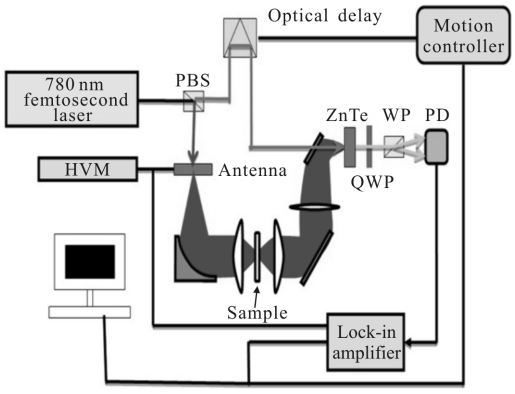


图 1 THz-TDS 原理图

Fig.1 Schematic diagram of the THz-TDS

长 780 nm,脉宽<100 fs,重复频率80 MHz,平均输出功率 140 mW。首先,激光器发出的飞秒激光脉冲经过分束镜(PBS)后分为泵浦光和探测光。泵浦光经过反射后照射到已施加偏压的光电导天线(GaAs)上激发 THz 脉冲,离轴抛物面反射镜将激发的 THz 脉冲聚焦并照射在样品上;探测光在经过延迟、准直和聚焦后与泵浦光同时入射到 THz 探测装置碲化锌(ZnTe)晶体上。最后,对太赫兹电场进行电光取样探测[16],THz 信号经探测器接收并输入到锁相放大器

中。此外,由于水分对 THz 波有非常强的吸收作用,为了减少周围环境的影响,将探测空间的湿度降低至 3%以下,温度稳定在 22~25 ℃。

1.2 实验原理

通过太赫兹时域光谱系统获得时域的参考信号 $E_0(t)$和透射过样品的太赫兹时域信号 $E_{trans}(t)$,对波形进行傅里叶变换,得到其频谱 $E_0(\omega)$和 $E_{trans}(\omega)$。理想状态下样品处于真空中,其前后的介质折射率均为 1,且表现为弱吸收率,即

$$\frac{E_{trans}(\omega)}{E_0(\omega)}=T(\omega)\exp(-i\varphi(\omega)) \tag{1}$$

据公式(1)得到 $T(\omega)$和 $\varphi(\omega)$,继而可以计算出样品在 THz 波段的折射率和吸收系数为:

$$n(\omega)=\phi(\omega)\frac{c}{\omega d}+1 \tag{2}$$

$$\alpha(\omega)=\frac{2}{d}\ln\left(\frac{4n(\omega)}{T(\omega)[n(\omega)+1]^2}\right) \tag{3}$$

式中:c 为光速;d 为样品厚度;$\phi(\omega)$为样品与参考信号的相差;ω 为太赫兹波振动角频率。如果忽略边界损失,公式(3)可近似为:

$$\alpha=\frac{2}{d}\ln\left[\frac{4n}{A(n+1)^2}\right] \tag{4}$$

在实验中,选取的实验样品尽管放置在特制的聚乙烯器皿中,剂量也尽量保持一致,但是样品的厚度仍存在误差,导致在样品的折射率和吸收系数计算时出现误差。为了避免样品厚度误差导致的实验结果误差,采用吸光度(absorbance)分析实验数据。吸光度表示表征材料对光波的吸收程度,是一个无量纲的相对量,计算过程中无需测量样品厚度,计算公式如下:

$$\text{Absorbance}=-\lg\left[\frac{E_{trans}(\omega)^2}{E_0(\omega)^2}\right] \tag{5}$$

得到吸光度光谱后,运用主成分分析(Principal Component Analysis,PCA) 对光谱数据进行处理。PCA 是一种应用非常广泛的数据处理方法,通过对数据的原始特征进行线性变换,可得到能表达原始数据信息的新的综合变量,降低数据维度。通过 PCA 将光谱数据转换成维度尽量少的综合变量,降低数据空间的大小,减少模型训练和测试时间,同时减少噪声干扰,提高映射性能,最终达到改善诊断能力的目的。

利用 PCA 对数据处理之后,运用支持向量机 (Support Vector Machine,SVM) 建模进行分类预测。SVM 是建立在统计学理论中的 VC 维(Vapnik-chervonenkis dimension)理论和结构风险最小化原理基础之上的一种统计学方法。SVM 通过非线性核函数如径向基核函数 (Radial Basis Function,RBF),将输入样本空间映射到高维线性特征空间,具有处理非线性回归问题的能力。SVM 与人工神经网络相比,拥有训练效率高、模型泛化能力强等优点[17]。

此次实验选用市面上 8 种不同的大豆食用油作为探测对象,运用太赫兹时域光谱技术结合主成分分析-支持向量机 (PCA-SVM) 的方法进行分类鉴别。

2 实验结果及分析

实验选取的实验样品为在市面上易获取的 8 种不同的大豆食用油。分别是 1 号:恒大兴安绿色大豆油(非转基因),2 号:九三大豆油(非转基因),3 号:金龙鱼食用调和油(非转基因),4 号:鲁花食用调和油(非转基因),5 号:福临门食用调和油(转基因),6 号:金龙鱼精炼一级大豆油(转基因),7 号:金龙鱼调和油(转基因),8 号:惠宜食用调和油(转基因)。以上调和油样品的原料中均包含大豆成分。每种豆油各制备 12 组样品,共 96 组,其中每种豆油都取其前 8 组样品的光谱作为训练集,后 4 组为测试集。

2.1 光谱分析

图 2 为转基因大豆油样品的太赫兹时域光谱,其中图 2(a)为整体效果,图 2(b)为局部放大。为了降低回波带来的干扰,手动截取太赫兹波的主脉冲,滤除了尾端的回波。从图中可以看出,相对于参考信号,8 种大豆油样品的时域光谱在振幅和相位上都具有一定程度的改变,2 号样品九三非转基因大豆油的相位延迟最久,1 号样品恒大非转基因大豆油的振幅衰减最少。总体上看,包含转基因大豆成分的样品相对于包含非转基因大豆成分的样品在相位上延迟更少,在幅值上衰减更大。聂君扬[18]等人在关于转基因大豆鉴别实验中,观测到:相对于参考信号而言,大豆粉末中含有的转基因大豆成分浓度越高,其太赫兹时域光谱幅值衰减越大,相位延迟越少,与此次实验的观测结果相一致。

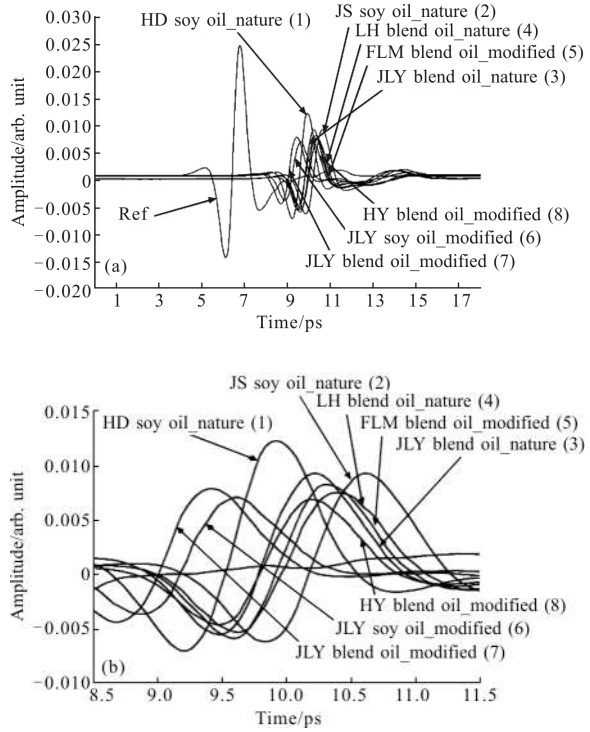


图 2 样品和参考信号的太赫兹时域光谱。(a)滤波后的太赫兹时域光谱;(b)太赫兹时域光谱的局部放大

Fig.2 THz time-domain spectra of the samples and the reference signal. (a) The filtered spectra of THz time-domain; (b) the partial enlargement of THz time-domain spectra

为了进一步研究样品在太赫兹波段内各频率的变化特性,对样品进行鉴别研究,将时域光谱补零后进行傅里叶变换后得到其频谱。图 3(a)为样品在 0.2~1.5 THz 的频谱图,从图中可以观测到所有样品相对于参考信号在各频率上均有较大幅度的衰减,在 0.35 THz 附近有着一个较明显的吸收峰,且各样品的截至频率均在 1.4 THz 左右。将其局部放大后得图 3(b),图中表明在 0.4 THz 之后,相对于参考信号而言,1 号样品恒大非转基因大豆油在各频率上的衰减最小,6 号样品金龙鱼转基因大豆油的波形衰减则最大。整体上看,含有非转基因成分的样品,其各波段的频率相对于含有转基因成分的样品,衰减较小。从图 3(b)中可以看出,除 4 号样品鲁花非转基因调和油的频域光谱波峰在 2.3×10^{-5} 左右,其他含有非转基因大豆成分样品的频域光谱波峰样品的幅值均超过 2.5×10^{-5},在 3×10^{-5} 左右,而含有转基因成分的样品其幅值则不超过 2.5×10^{-5},在 2×10^{-5} 左右。通过以上分析可得,含有转基因大豆成分的样品相对于含有非转基因大豆成分的样品,在太赫兹波段的吸收会更强。

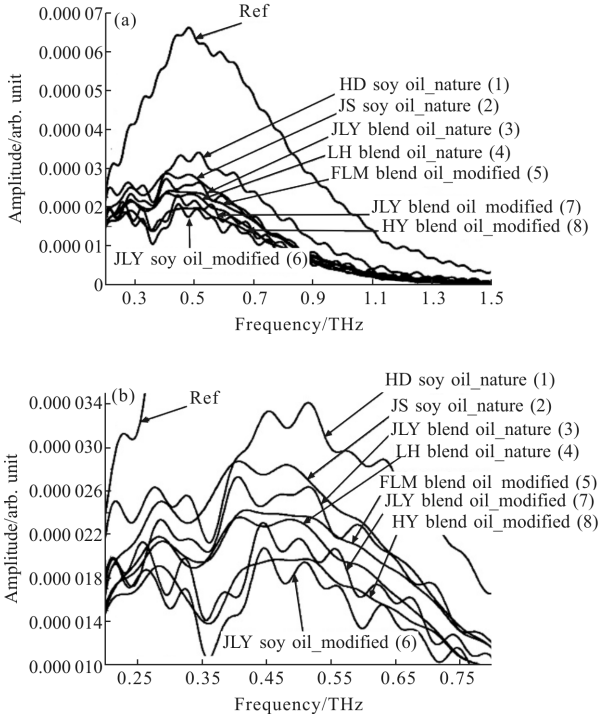
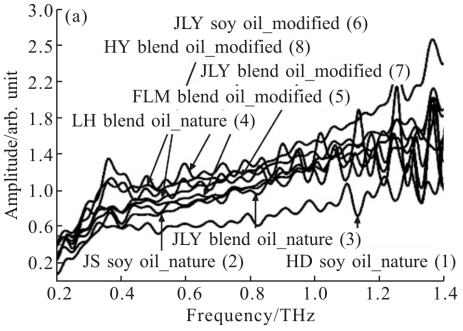


图 3 样品和参考信号的频域谱。(a)样品和参考信号在0.2~1.5 THz上的频域谱,(b)频域谱的局部放大

Fig.3 Frequency spectra of the samples and the reference signal. (a) Frequency spectra of the reference and the samples signal from 0.2–1.5 THz; (b) the partial enlargement of frequency spectra

为了降低实验器皿中每组样品油脂厚度差异引起的干扰,选用吸光度作为参量来处理光谱数据,得到的光谱数据如图 4 所示。其中图 4(a)为各样品在 0.2~1.4 THz 的吸光度光谱,为了更详细地对吸收度光谱进行分析,对图 4(a)进行局部放大,得到图 4(b)。在图 4 中,容易观测到所有样品在 0.35 THz 附近出现一个吸收峰,与频谱中的观测结果相一致。除此之



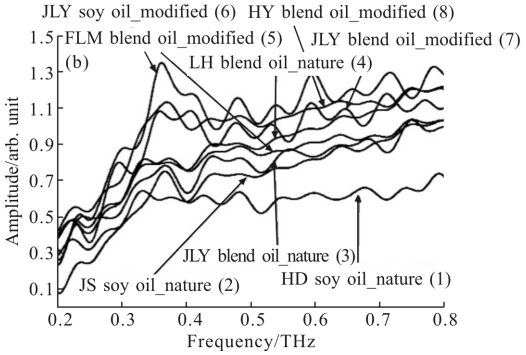


图 4 样品的吸光度谱。(a)样品在 0.2~1.4 THz 上的吸光度谱，(b)吸光度谱的局部放大

Fig.4 Absorbance spectra of the samples. (a) Absorbance spectra of the samples from 0.2–1.4 THz; (b) the partial enlargement of absorbance spectra

外，所有样品的吸光度随着频率的增加在提高。从图4(b)中可以发现，在 0.4 THz 之后，1 号样品恒大非转基因大豆油的吸光度最低，6 号样品金龙鱼大豆油的吸光度最高，且含有转基因大豆成分的样品相对于含有非转基因成分的样品，在太赫兹波段的吸光度会更高，与频谱中观测到的结果相一致。

2.2 建模与鉴别

为了减少光谱的数据冗余，同时降低模型的学习时间，提高建模效率，对样品吸光度光谱中 0.2~1.4 THz 波段的原始数据归一化处理后，进行主成分分析，结果如表 1 所示。从表 1 可以看出，前 8 项主成分的累计方差贡献率达到了 98.912%，可以近似解释所有原变量。可以用这 8 个新变量代替原有光谱数据导入SVM 中进行建模。

表 1 各成分解释总方差

Tab.1 Total variance explained by each component

Component	Initial eigenvalues		
	Total	Variance	Cumulative
PC1	34.709	64.101%	64.101%
PC2	10.456	19.309%	83.410%
PC3	3.107	5.738%	89.148%
PC4	1.957	3.615%	92.763%
PC5	1.506	2.781%	95.544%
PC6	1.044	1.927%	97.471%
PC7	0.607	1.121%	98.592%
PC8	0.173	0.320%	98.912%

实验中，选取径向基核函数为模型的核函数，同时采用交叉验证法自动选择最佳的惩罚参数 c 和核函数参数 g，保证训练与预测效果达到最优。如图 5 所示，实验结果表明当惩罚参数 c 为 2.926 1，核函数参数 g 为 0.052 261 时，实验的交叉验证准确率、训练准确率、测试准确率均为 100%，说明该模型取得较好的鉴别效果，可以对转基因大豆油进行很好的鉴别。

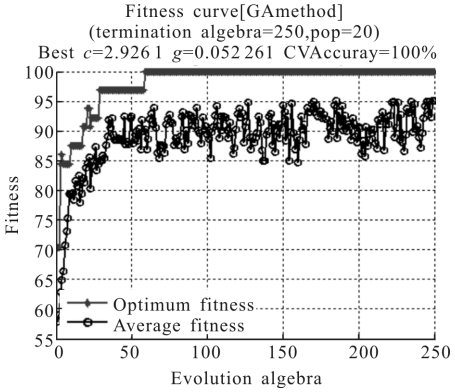


图 5 模型测试结果

Fig.5 Test results of the model

3 结 论

文中以市售 8 种含有转基因或非转基因大豆成分的食用油为研究对象，运用太赫兹时域光谱技术得到样品时域光谱，经滤波，补零，傅里叶变换等数据处理得到其频域光谱和吸光度光谱，并对光谱做了简单的特性分析，然后运用 PCA 方法提取出 0.2~1.4 THz 波段主要特征数据，减少数据冗余。之后，用训练集对 SVM 进行训练，用测试集对训练出的 SVM 模型测试。得到结论如下：

(1) 在太赫兹时域谱中，包含非转基因大豆成分的样品相对于包含转基因大豆成分的样品在相位上延迟更久，在幅值上衰减更少。

(2) 含有转基因大豆成分的样品相对于含有非转基因大豆成分的样品，在太赫兹波段的吸收会更强。

(3) 基于 RBF 核函数的 PCA-SVM 鉴别模型对包含转基因大豆成分的食用油具有很好的鉴别效果。

参考文献：

[1] Zambryski P, Joos H, Genetello C, et al. Ti plasmid vector for the introduction of DNA into plant cells without

alteration of their normal regeneration capacity [J]. ***The EMBO Journal***, 1983, 2(12): 2143-2150.

[2] James C. 20th anniversary (1996 to 2015) of the global commercialization of biotech crops and biotech crop highlights in 2015, ISAAA Brief, No.51 [R]. New York: ISAAA, 2015.

[3] James D, Schmidt A, Wall E, et al. Reliable detection and identification of genetically modified maize, soybean, and canola by multiplex PCR analysis [J]. ***Journal of Agricultural and Food Chemistry***, 2003, 51(20): 5829-5834.

[4] Liu Yong, Qian Honghu, Zhu Ling, et al. Nonuniformity correction for fluorescence imaging of microfluidic real-time PCR [J]. ***Optics and Precision Engineering***, 2013, 21(8): 2161-2168. (in Chinese)
刘勇, 钱鸿鹄, 朱灵, 等. 微流控实时荧光聚合酶链式反应成像非均匀性的校正 [J]. 光学 精密工程, 2013, 21(8): 2161-2168.

[5] Huang Kunlun, Luo Yunbo. Detecting genetically modified soybean roundup ready ingredient in foodstuffs by nested PCR and semi-nested PCR [J]. ***Journal of Agicultural Biotechnology***, 2003, 11(5): 461-466. (in Chinese)
黄昆仑, 罗云波. 用巢式和半巢式 PCR 检测转基因大豆 Roundup Ready 及其深加工食品 [J]. 农业生物技术学报, 2003, 11(5): 461-466.

[6] Cheng Hongmei, Peng Yufa, Jin Wujun, et al. A simple and rapid method for isolation of DNA from and detection of transgene sequences in soybean oil [J]. ***Scientia Agricultura Sinica***, 2007, 40(5): 1069-1072. (in Chinese)
程红梅, 彭于发, 金芜军, 等. 一种快速, 简便提取大豆油 DNA 的方法及转基因大豆油的检测 [J]. 中国农业科学, 2007, 40(5): 1069-1072.

[7] Fischer B M, Walther M, Jepsen P Uhd. Far-infrared vibrational modes of DNA components studied by terahertz time-domain spectroscopy[J]. ***Phys Med Biol***, 2002, 47(S1): 3807-3814.

[8] Tian Li, Jin Weiqi, Cai Yi, et al. Imaging area and contrast of THz focal plane array CW transmission imaging system [J]. ***Optics and Precision Engineering***, 2015, 23 (8): 2164-2170. (in Chinese)
田莉, 金伟其, 蔡毅, 等. THz 焦平面连续波透射成像系统的成像面积及对比度 [J]. 光学 精密工程, 2015, 23(8): 2164-2170.

[9] Zhang Jianfa, Yuan Xiaodong, Qin Shiqiao. Tunable terahertz and optical metamaterials [J]. ***Chinese Optics***, 2014, 7(3): 349-364. (in Chinese)
张检发, 袁晓东, 秦石乔. 可调太赫兹与光学超材料 [J]. 中国光学, 2014, 7(3): 349-364.

[10] Li Chenyu, Yang Zhou, Zhou Qingli, et al. Influence of structures on optical modulation in terahertz metamaterials [J]. ***Infrared and Laser Engineering***, 2016, 45(7): 0703002. (in Chinese)
李晨毓, 杨舟, 周庆莉, 等. 结构对太赫兹超材料光调控特性的影响[J]. 红外与激光工程, 2016, 45(7): 0703002.

[11] Hosobuchi M, Komatsu M, Xie X, et al. Measurements of THz absorption peaks in photo-degraded polyethylene and their assignment by quantum chemical calculations[C]//2013 IEEE Conference on Electrical Insulation and Dielectric Phenomena, 2013: 1046-1049.

[12] Wang Fang, Qiu Dajian, Xia Hongyan, et al. Preliminary exploration on identification of probiotics in terahertz time-domain spectroscopy [J]. ***Infrared and Laser Engineering***, 2016, 45(7): 0703001. (in Chinese)
王芳, 仇大剑, 夏红岩, 等. 太赫兹时域光谱技术在识别鉴定菌制剂中的初探 [J]. 红外与激光工程, 2016, 45(7): 0703001.

[13] Fischer B M, Helm H, Jepsen P U. Chemical recognition with broadband THz spectroscopy [J]. ***Proceedings of the IEEE***, 2007, 95(8): 1592-1604.

[14] Pickwell E, Cole B E, Fitzgerald A J, et al. Simulation of terahertz pulse propagation in biological systems [J]. ***Applied Physics Letters***, 2004, 84(12): 2190-2192.

[15] Platte F, Heise H M. Substance identification based on transmission THz spectra using library search [J]. ***Journal of Molecular Structure***, 2014, 1073(S1): 3-9.

[16] Ye Quanyi, Yang Chun. Recent progress in THz sources based on photonics methods[J]. ***Chinese Optics***, 2012, 5(1): 1-11. (in Chinese)
叶全意, 杨春. 光子学太赫兹源研究进展 [J]. 中国光学, 2012, 5(1): 1-11.

[17] Wang Shuzhou, San Ye. A survey on training algorithms for support vector machine[J]. ***CAAI Transactions on Intelligent Systems***, 2008, 3(6): 467-475.

[18] Nie Junyang, Zhang Wentao, Xiong Xianming, et al. Recognition of transgenic soybeans based on terahertz spectroscopy and PCA-BPN network [J]. ***Acta Photonica Sinica***, 2016, 45(5): 530001.