太赫兹光谱数据库的建立和使用

王凌辉[1] 王迎新[1] 刘圆圆[2] 赵自然[1]

([1]清华大学工程物理系粒子技术与辐射成像教育部重点实验室，北京 100084
[2]环境保护部核与辐射安全中心，北京 100082)

摘要 为了将太赫兹光谱分析技术应用于物质识别领域，需要建立太赫兹波段的光谱数据库，并研究合适的数据库使用方法，以鉴别未知物质。光谱获取采用自行搭建的太赫兹时域光谱测量系统，通过小波变换去除基线和噪声等干扰信息，建立起含有 20 种典型有机物的光谱数据库。使用该数据库识别未知物质时，分成两步：1) 用径向基函数神经网络算法判断未知物质是否在数据库中；2) 若在数据库中，采用基于纠错输出编码的支持向量机多类算法鉴别物质种类。测试结果表明，对库内物质识别率为 96.7%，对库外物质也有较好的预测和推断能力，识别率为 93.2%。提出的太赫兹光谱数据库建立和使用方法，对系统噪声等干扰因素有很好的抑制作用，可以应用到实际场合。

关键词 光谱学；太赫兹光谱数据库；小波变换；径向基函数神经网络；支持向量机；纠错输出编码

中图分类号 O443.4 **文献标识码** A **doi**：10.3788/CJL201239.0815002

Establishment and the Usage of Terahertz Spectral Database

Wang Linghui[1] Wang Yingxin[1] Liu Yuanyuan[2] Zhao Ziran[1]

([1] *Key Laboratory of Particle & Radiation Imaging, Ministry of Education, Department of Engineering Physics, Tsinghua University, Beijing* 100084, *China*
[2] *Nuclear and Radiation Safety Center, Ministry of Environmental Protection of P. R. China, Beijing* 100082, *China*)

Abstract Terahertz spectroscopy provides a new way for material identification. Investigation on establishment and usage methods of terahertz spectral database is needed so as to distinguish unknown substances. In order to establish the database, terahertz spectra of twenty organic materials are measured by own built terahertz time-domain spectroscopy system and the interference information of baseline and noise are removed by wavelet transform. The usage of database is divided into two steps: 1) determine whether the unknown substance is in the database through the radial basis function neural network; 2) identify the material if it is in the database by multi-class support vector machine. The fault tolerance of the algorithm is improved combined with error-correcting output coding to handle the multi-class problem with recognition rate of 96.7%. The network also has a good prediction of materials outside database with recognition rate of 93.2%. The establishment and usage methods of the database suppress the system noise and can be applied to practical situations.

Key words spectroscopy; terahertz spectral database; wavelet transform; radial basis function neural network; support vector machine; error-correcting output coding

OCIS codes 070.4790; 070.5010; 300.6495

1 引 言

太赫兹(THz)波通常指频率在 0.1～10.0 THz 之间的电磁辐射，从频域上看，属于远红外波段，作为电磁波谱上的最后一段空隙，是近 20 年来国内外科学家研究的热点。THz 波有很多优良的特性，包括对非极性物质有很强的穿透能力、光子能量低等。大多数极性分子在 THz 波段内共振吸收能量，形成各自的指纹谱，这为无损检测和物质识别提供了一种新的有效途径。

在 THz 光谱识别领域，国内机构对识别方法进

收稿日期：2012-02-27；**收到修改稿日期**：2012-03-21

基金项目：科技部项目(2010DFR10250)和清华大学自主科研专项(2010THZ05)资助课题。

作者简介：王凌辉(1988—)，男，硕士研究生，主要从事太赫兹光谱处理方面的研究。E-mail：wlh6543210@126.com

导师简介：赵自然(1977—)，男，博士，副研究员，主要从事粒子信息获取与处理、辐射成像等方面的研究。E-mail：zhaozr@mail.tsinghua.edu.cn(通信联系人)

行了一定程度的研究，但建立 THz 光谱数据库方面的工作，仍处于起步阶段。如赵晶晶[1]对四种爆炸物建立了光谱数据库。国外研究机构在这方面的研究进行得较早，美国伦斯勒理工学院 THz 研究中心，研究了含有 17 种爆炸物的 THz 光谱数据库[2]；美国标准技术组织(NIST)在其网站上公布了部分食品和药物的 THz 光谱[3,4]；欧洲“THz-Bridge”工程给出了蛋白质等生物分子的 THz 光谱[3]；2007 年，日本的 RIKEN 和 NICT 的 THz 数据库公诸于世，并逐年扩充，如今已有 500 多种可供查询的物质[4]，有利于各个机构对 THz 光谱分析技术的研究与利用。国外组织在各自的研究领域建立了不同类别的 THz 光谱数据库，但没有统一标准，测量系统主要有 THz 时域光谱(THz-TDS)系统和傅里叶变换红外光谱分析仪(FTIR)，两者在带宽、分辨率和信噪比方面各有特点。现有数据库对于光谱吸收峰的信息有很重要的参考价值，但是在进行实际物质鉴别时，往往需要全谱信息。所以针对特定的应用领域和光谱测量系统，若感兴趣的物质不在现有的数据库中，或已有数据不能满足实际系统要求，研究人员需要重新建立光谱数据库。

在光谱识别方法方面，研究人员通常将可见光、近红外[5]等波段范围的识别算法应用到 THz 领域，以寻找适合 THz 光谱特点的识别方法。如首都师范大学的翟爽等[6]对环三次甲基三硝铵(RDX)、2，4-二硝基甲苯(DNT)、2，4，6-三硝基甲苯(TNT)和环四次甲基四硝铵(HMX)四种爆炸物使用曲线极值算法分析，识别率为 98%；贾燕等[7]在氮气环境下实验，用误差反向传播(BP)神经网络算法，对 9 种毒品训练和识别，正确率为 89%。Hoshina 等[8]在邮件识别系统中，采用互相关系数进行识别。Liang 等[9]使用自组织特征映射神经网络对 6 种毒品正确识别。陈艳江[10]对中药进行识别，结果表明支持向量机(SVM)优于三层 BP 神经网络算法。上述研究的局限性在于，数据库规模有限，若未知物质不在数据库中，将会产生误判。

本文针对 20 种有机物，讨论 THz 光谱数据库的建立方法，并将该数据库应用于物质识别。获取光谱时，使用小波变换预处理，可有效去除基线和噪声等干扰因素。识别过程分成两步：1) 使用径向基函数(RBF)神经网络算法，判断未知物质是否在数据库中，排除数据库外物质干扰；2) 然后采用基于纠错输出编码(ECOC)的 SVM 多类算法，进行物质识别，取得了较为理想的结果。

2　THz 光谱数据库的建立

在利用 THz 光谱分析和鉴别物质时需要一个光谱数据库，为了确定未知物质成分，需要将未知的光谱与数据库内的已知光谱进行比较，分析未知物质种类。该数据库与研究领域相关，往往需要自行建立。在测量 THz 光谱的过程中，由于环境背景辐射、测量系统噪声、样品颗粒散射等因素的影响，直接得到的 THz 光谱中存在大量干扰信息，影响数据库的质量和识别的效果。因此，需要对直接测量得到的光谱进行预处理，如图 1 所示，尽量滤除系统干扰，提高光谱数据库质量，有利于系统的识别。

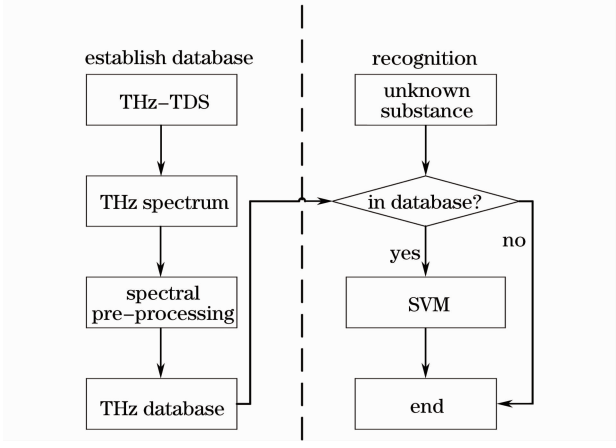

图 1　THz 光谱数据库建立和物质识别流程图

Fig. 1　THz spectral database establishment and material identification process

2.1　样品制备

建立样品光谱数据库，选择 20 种有机物，包括：白氨酸、半胱氨酸、苯甲酸、对甲苯甲酸、对氯苯胺、二苯甲酮、谷氨酸、果糖、海藻糖、抗坏血酸、邻苯二甲酸、酪氨酸、麦芽糖、木糖、苏氨酸、腺苷、烟酰胺、乙二胺四乙酸、蔗糖和组氨酸，涵盖了酸类、糖类、胺类和酮类等，具有一定的代表性。

制作片状样品时，称量 70 mg 固体样本和 140 mg聚乙烯粉末，在碾钵中碾碎并混合均匀，取出 120 mg 混合物放入压片机中，在 10 MPa 压力下压片 10 min，可得到片状样品，直径为 13 mm，厚度约为 1 mm。

2.2　THz-TDS

THz-TDS 系统结构如图 2 所示。飞秒激光器中心波长为 800 nm，脉宽100 fs，重复频率80 MHz，经分光镜分成两束，一束作为抽运光，透过椭球面镜的孔射向低温生长的砷化镓 LT-GaAs 光电导天线激发 THz 脉冲，THz 脉冲经过椭球面镜反射后穿过样品，聚焦于探测晶体ZnTe；另一束作为探测光被反射镜

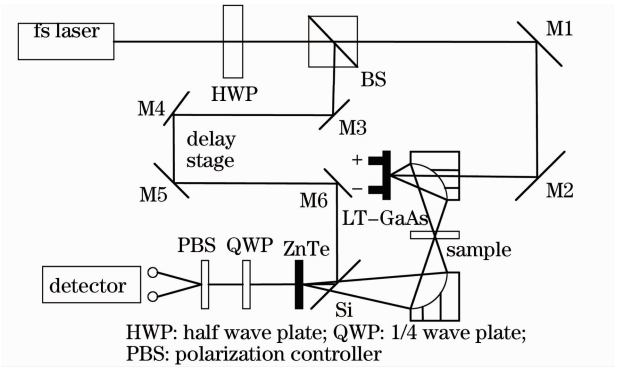

图 2 THz-TDS 系统示意图

Fig. 2 Schematic diagram of THz-TDS system

反射，同样照射到探测晶体上。THz 脉冲的电场通过线性电光效应调制 ZnTe 折射率，使得探测激光偏振状态发生改变。移动平移台改变探测光光程，实现对 THz 脉冲的时域扫描，获得完整的 THz 波形。

THz-TDS 系统测量在室温、空气环境下进行，相对湿度约为 25%，系统带宽 0.3～2.2 THz，频谱分辨率为 40 GHz，满足固体样本的识别要求。

2.3 光谱预处理

受样本颗粒散射等因素影响，THz 光谱会出现基线漂移现象。测量环境背景干扰会使光谱出现噪声，尤其是空气中水蒸气的吸收，在某些特定频率处

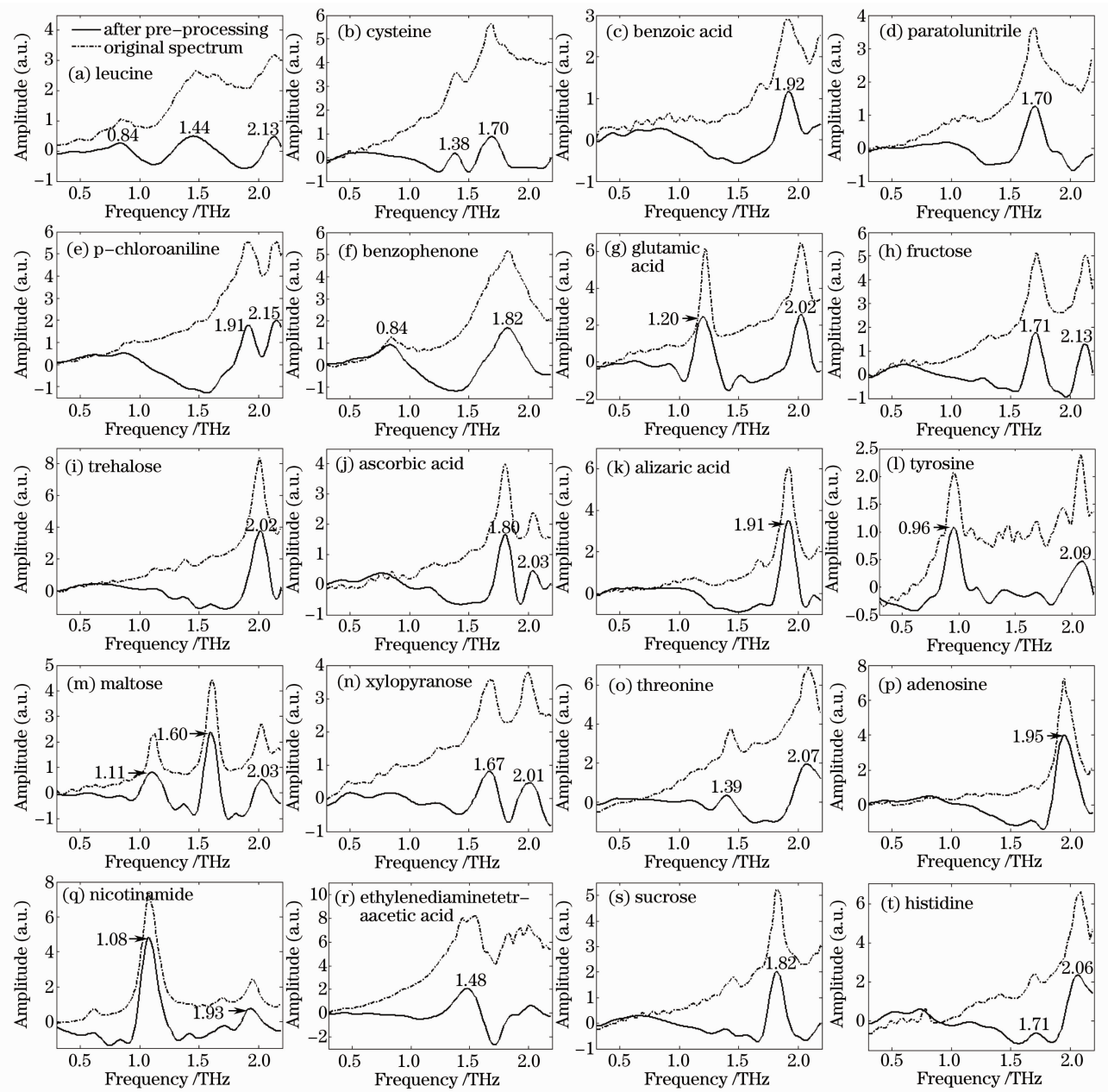

图 3 20 种物质 THz 光谱

Fig. 3 THz spectra of twenty materials

(如 1.1 THz、1.7 THz 和 2.2 THz 等),光谱的信噪比相对较低,噪声明显。若直接对测量的 THz 光谱建立数据库,大量无用的干扰信息将影响建库质量,进一步影响数据库的使用效果。

采用小波变换去除低频的基线和高频的噪声信息,在使用小波变换过程中,需要解决两个问题:1) 小波基函数的选择;2) 如何进行分解和重构。小波基函数的选择方法目前没有统一的准则,通常视具体的应用场合而定。在预处理过程中,小波基函数需要满足两个约束条件:1) 小波基函数平滑,也就是不同尺度下重建后的小波连续,没有阶跃点;2) 小波分解后的低频分量,与光谱能量走势相同,且低频起始端尽量趋近于 0。在对低频分量选择的同时,决定了小波分解和重构的方法。

通过对不同光谱,不同小波基函数的尝试发现,正交小波 Db6、Db9 和 Db10 都可满足要求。选择 Db9 小波为小波基函数,对 THz 光谱进行 6 尺度 Mallat 分解,第 6 尺度下的低频分量表示基线信息,第 1～3 尺度下的高频分量代表噪声信息,将其舍去,保留第 4～6 尺度下的高频分量。对数据库中 20 种物质处理后的结果如图 3 所示,可以看出,基线被有效地去除,高频波动被抑制,保留了有效信息。由于基线分量中包含直流成分,舍去后出现小于零的现象,对光谱的识别没有影响。将 THz 光谱与 THz 数据库[4]以及文献[11,12]比较,吸收峰所处位置基本一致。

3　THz 光谱数据库的使用

将 THz 光谱数据库应用到物质识别领域,判别未知物质种类,如图 1 所示。识别过程分为两步:1) 判断未知物质是否在数据库中。由于数据库规模有限,未知物质不在数据库中的情况广泛存在,使用基于 RBF 神经网络算法排除库外物质,尽量避免误判的可能性。2) 若未知物质在数据库中,采用基于 ECOC 的多类 SVM 算法,判别物质种类。SVM 具有最小化结构风险特征,判别多类问题时与 ECOC 相结合,进一步提高算法容错能力,取得了较好的结果。

3.1　鉴别物质是否在数据库中

使用 RBF 神经网络鉴别是否在数据库中,该网络由输入层、隐藏层和输出层三部分组成,如图 4 所示。输入层为 20 个样本数据 $\boldsymbol{P}=\{p_1,p_2,\cdots,p_{20}\}$;隐藏层由 20 个节点组成,计算样本数据 p_i 与节点 C_j 距离的高斯映射;输出层一个节点,采用线性传输函数,目标输出为 $\boldsymbol{T}=\{t_1,t_2,\cdots,t_{20}\}$。

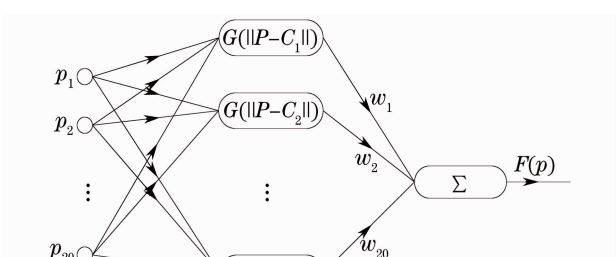

图 4　RBF 神经网络结构

Fig. 4　Structure of RBF neural network

对该网络训练,寻找函数满足 $F(p_i)=t_i=1$,$1\leqslant i\leqslant 20$。将所有 20 个样本输入分别作为隐节点的中心 $C_i=p_i$,$1\leqslant i\leqslant 20$,目标求解权值 w_i,$1\leqslant i\leqslant 20$ 满足线性方程组:

$$\sum_{j=1}^{N} w_j G(\|\ p_i-p_j\ \|)=t_i,1\leqslant i\leqslant 20.$$

设第 j 个隐节点在第 i 个样品的输出为 $\varphi_{ij}=G(\|\ p_i-p_j\ \|)$,上述方程组可表示成 $\boldsymbol{\Phi W}=\boldsymbol{T}$,若 $\boldsymbol{\Phi}$ 可逆,可求出方程组的解为 $\boldsymbol{W}=\boldsymbol{\Phi}^{-1}\boldsymbol{T}$。在 THz 光谱识别中,每种物质都有各自的指纹谱,$\boldsymbol{\Phi}$ 必定可逆,线性方程组有解。

分析未知物质是否在数据库中,难点在于缺少数据库外物质作为训练样本,因为库外物质是未知和无限的,无法用有限的库外物质代替模拟训练。RBF 神经网络将数据库内的物质映射到 1,隐藏层的高斯函数将数据库外物质映射到 0,对未知物质有很好的预测效果。

3.2　鉴别物质种类

使用基于 ECOC 的 SVM 算法鉴别物质种类。SVM 是一种统计学习方法,在最小化样本点误差的同时,提高了模型的泛化能力,而且没有数据维度的限制[13]。对于一个线性可分的两类问题,在数学上,最终可以转化为求解不等式约束下的最小值[14]。对于非线性可分的情况,处理的思想是将低维空间线性不可分问题,映射到高维空间,转化成线性可分问题,在高维空间中构造最优分类超平面进行分类。

SVM 算法的学习规则,决定了它适合解决两类问题[14],延伸到处理多类问题时,通常构造多个二值分类器。ECOC 提供了一个可行的方案,它在编码过程中引入冗余数据,当数据由于干扰等因素出现错误时,可以对错误位纠正,恢复出正确的数据,纠错能力和编码的设计有关[15]。

对于一个 N 类分类问题,设计编码矩阵 $\boldsymbol{C}_{N\times M}$,

M表示M位编码，即M个SVM二值分类器，编码矩阵第i行C_i表示第i类物质的编码，C_{ij}表示第i类物质的j位编码为0或者1，用以SVM分类。对样本数据的训练，就是对M个SVM分类器训练的过程，将原始的数据空间，映射到编码空间。对未知样本测试时，将样本光谱数据投影到编码空间，计算与各个类别编码间的距离，距离最小的类别即为未知样本所属类别。即使样本在数据空间投影过程中出现了某些位的错误，ECOC依然能够恢复出正确的信息，划分到正确的类别，从而使得算法的容错能力大大提高。

4 数据库使用结果

4.1 鉴别物质是否在数据库处理结果

选择20种有机物作为数据库内样品，7种有机物作为数据库外的测试样品，在不同时间、相对湿度(20%～30%)、背景辐射等环境下测量，得到数据库内物质100个样本，数据库外物质28个样本。选取数据库内40个样本，取平均后得到20个输入样本，对隐藏层含20个节点的RBF网络训练，余下的88个样本用来测试，输出结果为了方便观察，通过符号函数调整，结果如图5所示。

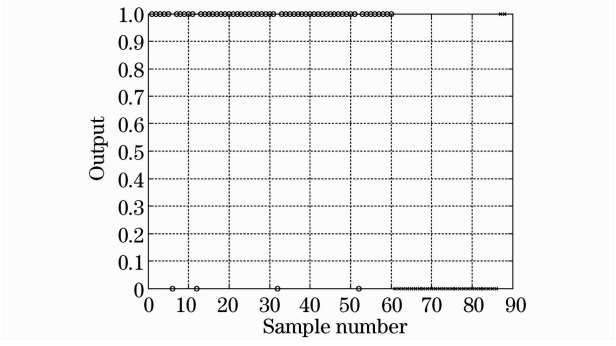

图5 RBF神经网络测试结果

Fig. 5 Testing results of RBF neural network

图5中'O'表示数据库内物质，'X'代表数据库外物质，横坐标为样本编号，纵坐标为网络输出值，1代表判定结果为数据库内物质，0代表数据库外物质。从结果中可以看出，数据库内60个样本有4个被识别错误，原因是系统测量时，水蒸气吸收、噪声干扰等环境因素影响，造成测量谱与节点间距离相对较大而形成误判。在数据库外物质训练样本缺失的情况下，库外28个样本仅有2个误识别，说明网络对库外物质有很好的推断能力。RBF神经网络算法，对该88种样本正确识别82种，识别率为93.2%，可用于排除数据库外物质。

4.2 鉴别物质种类结果

RBF神经网络算法已经排除了数据库外的物质干扰，本部分将研究对数据库内物质分类的效果。实验数据和RBF神经网络保持一致，对数据库内100个样本，选择同样的40个样本训练SVM网络，60个样本用以测试。ECOC矩阵大小为20×63，使用63个二值分类器将训练集分成20个类，具有较好的容错性。

样本训练是对63个SVM设计的过程，使用径向基核函数处理63个非线性可分的两类问题，训练成功后，确定了原始光谱数据空间向编码空间映射的方案，配合ECOC，解决了20个类的SVM分类问题。

60个测试样品，正确识别58个，识别率为96.7%。目前获得的THz光谱，由于硬件和成本因素的制约，带宽不够宽、分辨率不够高、高频处信噪比不够好是普遍存在的现象。在实验平台上测量的THz光谱，2.2 THz范围内，很多物质只有一个谱峰，峰宽较宽，大部分物质的吸收峰集中在1.5～2.2 THz之间，部分物质吸收峰十分接近，较难分辨。基于ECOC的SVM多类算法在设计过程中，引入了风险最小化和纠错机制，在处理THz光谱识别问题上取得了很好的结果，同时具有较好的泛化能力，适合应用在实际场合。

4.3 隐藏样品识别结果

THz辐射对绝大部分非极性物质穿透能力很强，包括塑料、纸张、纤维和布等常用遮挡材料。THz光子能量低，约为4 meV，目前普遍认为对人体没有伤害，这些特性决定THz-TDS技术在隐藏物非接触测量领域有很好的应用前景[16]。

实验讨论在纸张(信封)遮挡情况下，使用THz光谱数据库识别的结果。在数据库中随机选择3种

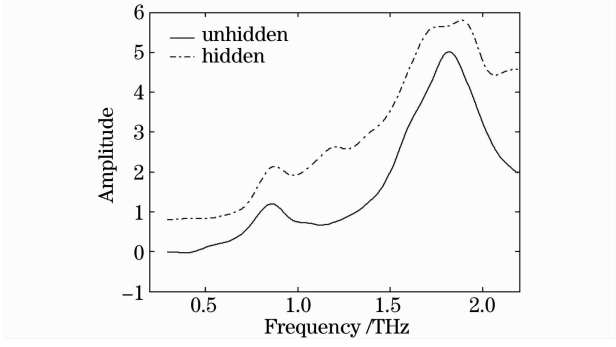

图6 纸张遮挡前后二苯甲酮光谱对比

Fig. 6 Spectral contrast of benzophenone before and after paper block

物质，对甲苯甲酸、二苯甲酮和酪氨酸，分别测量它们在没有遮挡和装在信封情况下的 THz 光谱，图 6 和图 7 分别给出了二苯甲酮和酪氨酸的 THz 光谱数据库识别结果，为更好地对比，滤除了光谱的高频噪声，保留了低频基线，并进行了平移处理。酪氨酸遮挡后的 THz 光谱，有很明显的基线漂移现象，二苯甲酮在 1.2 THz 处多了一个弱吸收峰，1.8 THz 处的主峰宽度变宽，可能是信封纸中微量水的吸收与隐藏物质光谱叠加的结果。

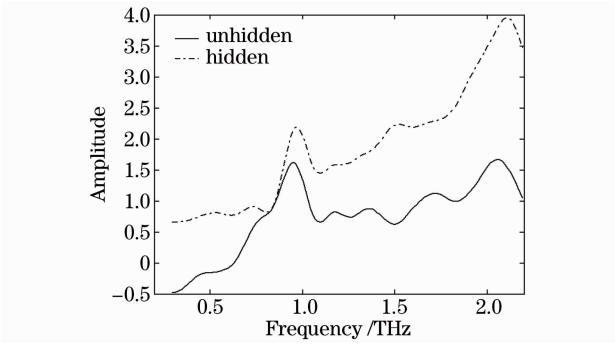

图 7　纸张遮挡前后酪氨酸光谱对比

Fig. 7　Spectral contrast of tyrosine before and after paper block

对隐藏样品检测，会受到遮挡物干扰的影响，通常表现为基线漂移、弱吸收峰数量和位置的改变以及主峰宽度和幅度的变化。基线漂移可以通过去基线算法加以抑制，对吸收峰的干扰由于缺少未知遮挡物的信息，难以滤除。故隐藏样品的识别率很大程度取决于识别算法的设计，基于 ECOC 的 SVM 算法具有较好的容错和纠错能力，对三个遮挡样品识别得到了正确的结果，说明讨论的数据库搭建和使用方法具有很好的适应性，在隐藏样品识别领域具有一定的应用潜力。

5　结　　论

对 20 种有机物，讨论了 THz 光谱数据库的建立方法，并将其应用于光谱识别领域。实验结果验证了建库方法以及光谱数据库使用方法的可行性。最后，对隐藏样品进行讨论，由于基于 ECOC 的 SVM 算法具有很强的容错能力，在存在遮挡物干扰的情况下，依旧能够正确识别。所讨论的建库和用库方法，在 THz 光谱识别领域有很好的应用前景。

参　考　文　献

1 Zhao Jingjing. Research for Explosive Identification Based on THz Database[D]. Beijing: Capital Normal University, 2009
赵晶晶. 基于 THz 光谱数据库的爆炸物分类识别研究[D]. 北京：首都师范大学，2009

2 J. Chen, Y. Chen, H. Zhao *et al*.. Absorption coefficients of selected explosives and related compounds in range of 0.1～2.8 THz[J]. *Opt. Express*, 2007, **15**(19): 12060～12067

3 I. Hosako, N. Sekine, K. Fukunaga *et al*.. At the dawn of a new era in terahertz technology[J]. *Proceedings of the IEEE*, 2007, **95**(8): 1611～1623

4 Terahertz Database[DB]. OL. http://www.thzdb.org

5 Zhang Haihong, Zhang Shujuan, Wang Fenghua *et al*.. Study on fast discrimination of seabuckthorn juice varieties using visible-nir spectroscopy[J]. *Acta Optica Sinica*, 2010, **30**(2): 574～578
张海红，张淑娟，王凤花 等. 应用可见-近红外光谱快速识别沙棘汁品牌[J]. 光学学报，2010，**30**(2)：574～578

6 Qu Shuang, Ge Qingping. Explosives identification methods based on terahertz spectroscopy[J]. *Journal of Capital Normal University*, 2009, **30**(2): 6～8
翟　爽，葛庆平. 基于太赫兹光谱曲线识别的爆炸物识别方法研究[J]. 首都师范大学学报，2009，**30**(2)：6～8

7 Jia Yan, Chen Sijia, Li Ning *et al*.. Identification of terahertz absorption spectra of illicit drugs using back propagation neural networks[J]. *Chinese J. Lasers*, 2007, **34**(5): 719～722
贾　燕，陈思嘉，李　宁 等. 利用误差逆传播神经网络法识别几种毒品的太赫兹光谱[J]. 中国激光，2007，**34**(5)：719～722

8 H. Hoshina, Y. Sasaki, A. Hayashi *et al*.. Noninvasive mail inspection system with terahertz radiation[J]. *Society for Applied Spectroscopy*, 2009, **63**(1): 81～86

9 M. Liang, J. Shen, Q. Wang *et al*.. Identification of illicit drugs by using SOM neural networks[J]. *J. Phys.* (*D*), 2008, **41**(13): 135306

10 Chen Yanjiang. Chinese Traditional Medicine Recognition by Support Vector Machine Terahertz Spectrum[D]. Beijing: Capital Normal University, 2009
陈艳江. 基于支持向量机的中药太赫兹光谱鉴别[D]. 北京：首都师范大学，2009

11 Zhu Dechong, Zhang Liangliang, Zhao Yaqin *et al*.. Terahertz boardband spectroscopic investigation of amino acid[J]. *Chinese J. Lasers*, 2011, **38**(s1): s111008
祝德充，张亮亮，赵亚琴 等. 氨基酸的超宽带太赫兹光谱[J]. 中国激光，2011，**38**(s1)：s111008

12 Wang Weining, Li Hongqi, Zhang Yan *et al*.. Correlations between Terahertz spectra and molecular structures of 20 standard α-amino acids[J]. *Acta Physico-Chimica Sinica*, 2009, **25**(10): 2074～2079
王卫宁，李洪起，张　岩 等. 20 种 α-氨基酸的太赫兹光谱及其分子结构的相关性[J]. 物理化学学报，2009，**25**(10)：2074～2079

13 G. Zhen, Z. Qian, Q. Yang *et al*.. The combination approach of SVM and ECOC for powerful identification and classification of transcription factor[J]. *BMC Bioinformatics*, 2008, **9**: 282

14 Bian Zhaoqi, Zhang Xuegong. Pattern Recognition[M]. Beijing: Tsinghua University Press, 2000. 284～303
边肇祺，张学工. 模式识别[M]. 北京：清华大学出版社，2000. 284～303

15 Liu Liangfu, Wang Xiaoping. Text classifier based on support vector machine and output coding[J]. *Computer Applications*, 2004, **24**(8): 32～34
刘良赋，王小平. 基于支持向量机和输出编码的文本分类器研究[J]. 计算机应用，2004，**24**(8)：32～34

16 Li Qi, Yao Rui, Ding Shenghui *et al*.. Experiment on 2.52 THz transmission-mode imaging for concealed objects[J]. *Chinese J. Lasers*, 2011, **38**(7): 0711001
李　琦，姚　睿，丁胜晖 等. 遮挡物的 2.52 THz 透射成像实验研究[J]. 中国激光，2011，**38**(7)：0711001

栏目编辑：李文喆