利用已知混合物拉曼光谱改善混合物成分识别精度的方法

季明强[1]，朱启兵[1*]，黄敏[1]，张丽文[2]，雷泽民[2]，张恒[2]
[1]江南大学轻工过程先进控制教育部重点实验室，江苏 无锡 214122；
[2]北京卓立汉光仪器有限公司，北京 101102

摘要 构建已知纯净物光谱数据库，计算待识别混合物光谱与数据库中各纯净物光谱的相似度，是利用拉曼光谱技术进行混合物成分分析的常用策略。受测量仪器性能及待测混合物所含成分相互干扰的影响，待识别混合物中所含物质成分的光谱与数据库中对应的纯净物光谱相比会有不同程度的失真，从而给基于纯净物光谱数据库的组分鉴别带来极大困难。针对这一问题，提出了一种使用已知混合物光谱数据来改善混合物成分识别精度的方法。首先利用纯净物拉曼光谱谱峰的位移和半峰全宽信息，将已知混合物的光谱谱峰与其所含有的具体物质对应；然后基于谱峰拉曼位移、半峰全宽和谱峰强度分别构建纯净物、已知混合物和待识别混合物的特征参数，并利用模糊隶属度函数计算待识别混合物光谱与纯净物光谱、已知混合物所含物质光谱的相似度；最终根据光谱相似度确定待识别混合物中含有的疑似组分。基于 204 种纯净物和 8 种已知混合物光谱数据库，对 81 种混合物进行了识别，结果表明：所提方法可降低由光谱失真导致的相似度计算误差，提高识别准确率；相比于纯净物数据库搜索策略，本文方法的识别精度由 76.34%提高到了 92.83%。

关键词 光谱学；拉曼光谱；混合物成分识别；相似度计算；已知混合物；模糊隶属度函数
中图分类号 O433.4 **文献标志码** A **doi**：10.3788/CJL202047.1111001

Method for Improving Identification Accuracy of Components in Mixtures Using Raman Spectra of Known Mixtures

Ji Mingqiang[1], Zhu Qibing[1*], Huang Min[1], Zhang Liwen[2], Lei Zemin[2], Zhang Heng[2]
[1]*Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, Jiangnan University, Wuxi, Jiangsu* 214122, *China*;
[2]*Beijing Zolix Instrument Co., Ltd., Beijing* 101102, *China*

Abstract A common strategy for mixture composition analysis based on Raman spectroscopy is to construct the spectral database of pure substances and calculate the spectral similarity of the mixture to be identified and the pure substances in the database. However, influenced of the performance of the measuring instrument and the mutual interference of the components of the mixture, the spectrum of the substance contained in the mixture to be identified will have different degrees of distortion compared with the corresponding pure substance spectrum in the database, bringing great difficulties in component identification. To address this problem, a method to improve the identification accuracy of components in mixtures using the spectral data of known mixtures is proposed herein. The spectral peak information, including Raman shift and full width at half maximum, of the pure substance in the database is used to study the correspondence of the spectral peaks of the known mixture to the specific substances of the mixture. The spectral feature parameters of the pure substance, the known mixture, and the mixture to be identified are constructed using the Raman shift of the spectral peak, the full width at half maximum, and the peak intensity, respectively and the fuzzy membership function is used to calculate the spectral similarity between the mixture to be identified, the pure substance, and the substances contained in known mixture based on calculated feature parameters. Furthermore, suspected components contained in mixture to be identified are determined based

收稿日期：2020-05-11；**修回日期**：2020-06-05；**录用日期**：2020-06-15
基金项目：国家自然科学基金(61775086)
* **E-mail**：zhuqib@163.com

on the spectral similarity. Based on the spectral database of 204 pure substances and 8 known mixtures, the experimental results for 81 unknown mixtures reveal that the proposed method can reduce the calculation error of similarity caused by spectral distortion and can improve the identification accuracy. Compared with search strategy based on the database of pure substance, the identification accuracy obtained using the proposed method increases from 76.34% to 92.83%.

Key words spectroscopy; Raman spectrum; mixture component identification; similarity calculation; known mixture; fuzzy membership function

OCIS codes 300.6450; 300.6170; 300.6360

1 引　　言

混合物成分鉴别在混合物分析中具有非常重要的意义。待测物质拉曼光谱的谱峰对应于某些特定的分子，因此拉曼光谱也被称为“指纹”光谱。目前，拉曼光谱已被广泛应用于考古学、生物学、物质鉴定等领域[1-6]。

在基于拉曼光谱进行混合物成分识别的方法中，基于光谱数据库的搜索算法获得了广泛应用。其基本原理是构建已知纯净物的拉曼光谱数据库，将待识别物质的光谱与数据库中的光谱逐一进行比较，计算两者的相似度，然后根据相似度的大小最终确定待识别混合物的成分。

刘财政等[7]提出了一种混合物组分识别的新方法。他们首先构建了由 18 种纯净物光谱数据构成的标准物数据库，然后遍历数据库，将混合物参数向量与纯净物参数向量作相关性计算，实现了对混合物组分的识别。

Zhang 等[8]提出了一种基于手持式拉曼光谱仪的反向搜索和非负最小二乘方法。该方法根据拉曼光谱的特征对经典的反向搜索过程进行修改，通过计算待识别混合物与纯净物光谱之间减谱的负比率获得匹配质量。

刘铭晖等[9]针对单一的匹配特征无法全面反映被测样本光谱与谱库光谱相似性这一问题，采用逻辑回归数学模型融合谱峰匹配系数、非负最小二乘匹配系数以及夹角余弦匹配系数，提出了一种新的光谱集成匹配方法，并采用该方法对氨基酸混合物的组分进行了判别。

上述的数据库搜索方法都是基于纯净物拉曼光谱库的搜索策略。在具体应用时，测量仪器自身存在的重复性误差，以及待测混合物中各成分的相互干扰，导致采集的混合物各成分的光谱谱峰相较于数据库中相应物质的谱峰出现了一定程度的失真现象（如谱峰偏移），从而影响了混合物成分鉴别的精度。为了减小拉曼光谱谱峰的偏移，在使用测量仪器时都需要控制测量环境，并对光谱进行校准。但是，对于手持式拉曼光谱仪这一类面向快速检测应用的设备来说，其测量环境往往难以控制，且光谱的校正比较困难，导致在数据库较大的情况下，存在较为严重的误识别。针对这一问题，本文提出了利用实际检测获得的已知混合物的拉曼光谱（这些已知混合物的光谱可以是仪器检测过程中产生的历史记录信息）来协助搜索，从而改善未知混合物识别精度的方法。已知混合物的拉曼光谱中含有相关纯净物的谱峰偏移信息，利用这些信息可以有效抑制谱峰偏移的干扰。相比于单纯的纯净物光谱数据库，本文方法可显著提高未知混合物的识别精度。

2 实验部分

2.1 实验样品与实验仪器

本文首先采用 204 种纯净物构建纯净物光谱数据库，这些纯净物主要包括常见的化学药品及管制品。化学药品购于国药集团化学试剂北京有限公司，其纯度等级均为二级；管制品来源于公安机关，其纯度均在 98%以上。

为了研究已知混合物光谱数据库对识别精度的改善作用，利用乙醇（ethanol）、乙腈（acetonitrile）、丙酮（acetone）、环己烷（cyclohexane）、二丙酮醇（diacetone alcohol）、丙二酸二乙酯（diethyl malonate）配制了 8 种混合物，其中三元混合物 5 种，四元混合物 3 种。考虑到在实际应用中，已知混合物中各组分浓度比的随机性，本文将每种混合物配制成多个浓度比例，其中每种三元混合物有 9 个浓度比（体积比），每种四元混合物有 12 个浓度比，详见表 1。

在构建已知混合物数据库时，在每种混合物中随机抽取一个浓度比的混合物构成已知混合物数据库，即已知混合物数据库的大小为 8。待识别的混合物同样由上述 6 种纯净物按照不同的比例混合而成，其中三元混合物 5 种（每种 9 个浓度比），四元混合物 3 种（每种 12 个浓度比）。表 2 给出了待识别的 81 个混合物的组成及浓度信息。

表 1　不同组分的已知混合物

Table 1　Known mixtures with different components

Mixture type	Component	Volume ratio
Ternary(45)	① ethanol, acetonitrile, acetone ② acetone, cyclohexane, diacetone alcohol ③ cyclohexane, diacetone alcohol, diethyl malonate ④ ethanol, diacetone alcohol, diethyl malonate ⑤ ethanol, acetonitrile, diethyl malonate	7∶2∶1, 5∶3∶2, 4∶3∶3, 2∶1∶7, 3∶2∶5, 3∶3∶4, 1∶7∶2, 2∶5∶3, 3∶4∶3
Quaternary(36)	① acetonitrile, acetone, cyclohexane, ethanol ② acetone, cyclohexane, diacetone alcohol, diethyl malonate ③ ethanol, acetonitrile, diacetone alcohol, diethyl malonate	6∶2∶1∶1, 5∶2∶2∶1, 4∶3∶2∶1, 2∶1∶1∶6, 2∶2∶1∶5, 3∶2∶1∶4, 1∶1∶6∶2, 2∶1∶5∶2, 2∶1∶4∶3, 1∶6∶2∶1, 1∶5∶2∶2, 1∶4∶3∶2

表 2　不同组分的待识别混合物

Table 2　Mixtures to be identified with different components

Mixture type	Component	Volume ratio
Ternary(45)	① acetone, diacetone alcohol, diethyl malonate ② ethanol, acetone, diacetone alcohol ③ ethanol, acetone, cyclohexane ④ ethanol, cyclohexane, diethyl malonate ⑤ acetonitrile, acetone, diacetone alcohol	7∶2∶1, 5∶3∶2, 4∶3∶3, 2∶1∶7, 3∶2∶5, 3∶3∶4, 1∶7∶2, 2∶5∶3, 3∶4∶3
Quaternary(36)	① ethanol, acetone, diacetone alcohol, diethyl malonate ② ethanol, acetonitrile, acetone, diethyl malonate ③ ethanol, acetonitrile, acetone, diacetone alcohol	6∶2∶1∶1, 5∶2∶2∶1, 4∶3∶2∶1, 2∶1∶1∶6, 2∶2∶1∶5, 3∶2∶1∶4, 1∶1∶6∶2, 2∶1∶5∶2, 2∶1∶4∶3, 1∶6∶2∶1, 1∶5∶2∶2, 1∶4∶3∶2

利用北京卓立汉光仪器有限公司生产的型号为 Finder Edge 的手持式拉曼光谱仪(激光发射器的波长为 785 nm,光谱分辨率为 8～10 cm^{-1})采集纯净物和各种混合物的光谱。在光谱采集时,光谱仪的激光功率以及积分时间根据实际情况进行调节(激光功率为 0～300 mW,积分时间为 1.5～2.5 s)。

采集的拉曼光谱在 240～2400 cm^{-1} 拉曼位移范围内具有较多的特征峰,因此本文选取此区域的光谱数据。后续光谱处理与分析均通过 Windows 平台下的 MATLAB R2016a 实现。需要说明的是,纯净物、已知混合物和未知混合物的光谱是由多个拉曼光谱仪于不同时间采集得到的,已知混合物光谱采集于 2019 年 6 月 10 日,未知混合物光谱采集于 2019 年 7 月 15 日。仪器在生产制造过程存在重复误差,因此采集的光谱含有仪器自身的重复误差,这为混合物的识别增加了难度。上述 6 种纯净物的原始拉曼光谱如图 1 所示。

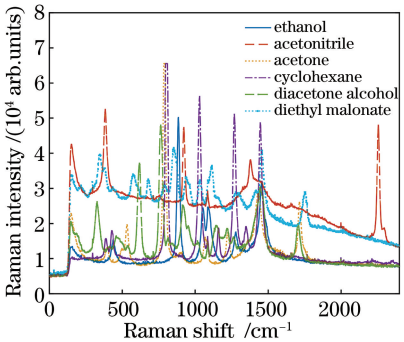

图 1　手持式拉曼光谱仪测得的 6 种纯净物的原始光谱

Fig. 1　Original spectra of six pure substances measured by hand-held Raman spectrometer

2.2　识别算法

混合物识别算法主要包括光谱的预处理、拉曼光谱特征提取和数据库搜索匹配三个环节。

2.2.1　拉曼光谱预处理

受混合物自身特性、仪器性能、环境信息等的影响,实际测得的拉曼光谱数据中会含有噪声[10]以及连续基线[11],这会给后续操作带来较大影响,故而本文采用连续小波变换[12]进行基线去除(小波基函数使用墨西哥帽函数),采用惩罚最小二乘方法[13]进行去噪处理。最后,对光谱数据进行最大值归一化,以便于后续的数据处理。

图 2 为预处理后 6 种纯净物的拉曼光谱图。从图 2 中可以看出,预处理后的光谱数据在保留原始光谱谱峰信息的同时,去除了噪声和基线的干扰。

2.2.2　拉曼光谱特征提取

在理论上,拉曼光谱的谱峰可以用洛伦兹线型

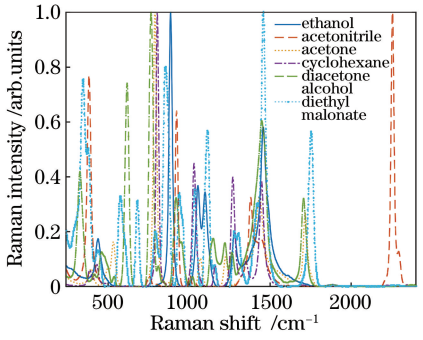

图 2　预处理后 6 种纯净物的拉曼光谱

Fig. 2　Raman spectra of six pure substances after preprocessing

来描述，但是由于仪器精度和混合物自身特性等多种因素的影响，实际测量得到的拉曼光谱的谱峰一般为 Voigt[7] 线型。该函数为洛伦兹线型和高斯线型的卷积，Voigt 函数的数学表达式为

$$I(\lambda)=I_c\left\{\theta\frac{w^2}{(\lambda-\lambda_c)^2+w^2}+(1-\theta)\exp\left[\frac{-(\lambda-\lambda_c)^2}{2w^2}\right]\right\},\tag{1}$$

式中：λ 为拉曼位移；$I(\lambda)$ 为 λ 处的拉曼强度；λ_c、I_c 为谱峰处的拉曼位移及其强度值；w 为半峰全宽；θ 为高斯-洛伦兹系数，该系数的取值范围为(0,1)。

混合物的拉曼光谱中存在较多的重叠峰。由 n 个谱峰构成的重叠峰可以看作是 n 个 Voigt 峰的线性叠加，其数学表达式为

$$I(\lambda)=\sum_{t=1}^{n}I_t\left\{\theta_t\frac{w_t^2}{(\lambda-\lambda_t)^2+w_t^2}+(1-\theta_t)\exp\left[\frac{-(\lambda-\lambda_c)^2}{2w^2}\right]\right\}。\tag{2}$$

对于重叠峰而言，必须对其进行分解，以获取重叠峰的数目及拟合区域。如图 3(a)所示，设$(\lambda^{v}_{i-1},I^{v}_{i-1})$和$(\lambda^{v}_{i+1},I^{v}_{i+1})$分别为谱峰$(\lambda^{p}_{i},I^{p}_{i})$左右两侧的极小值点，设置拉曼强度阈值为 I^{p}，本文中该值取最大峰强的 0.05 倍。若 $I^{v}_{i-1}\leqslant I^{p}$ 且 $I^{v}_{i+1}\leqslant I^{p}$，则认为$(\lambda^{v}_{i-1},I^{v}_{i-1})$至$(\lambda^{v}_{i+1},I^{v}_{i+1})$为单个谱峰待拟合区域；若 $I^{v}_{i+1}>I^{p}$，则继续寻找下一个谱峰的右侧极小值点，直至找到$(\lambda^{v}_{i+n},I^{v}_{i+n})$满足 $I^{v}_{i+n}\leqslant I^{p}$，即得到待拟合的重叠峰区域$(\lambda^{v}_{i-1},\lambda^{v}_{i+n})$和重叠峰个数 n。在拟合区域$(\lambda^{v}_{i-1},\lambda^{v}_{i+n})$内，基于(1)或(2)式，使用 Levenberg-Marquardt(LM)算法[14] 进行谱峰拟合，获取各谱峰的参数值：谱峰处的拉曼位移 λ_c、强度 I_c、半峰全宽 w 和高斯-洛伦兹系数 θ。由于高斯-洛伦兹系数 θ 具有随机性，因此本文将 λ_c、I_c、w 作为谱峰的特征参数。对获得的 N 个谱峰的特征参数，按照其拉曼位移进行排列(从小到大)，获得一组特征向量$[\lambda_1,I_1,\omega_1;\cdots;\lambda_i,I_i,\omega_i;\cdots;\lambda_N,I_N,\omega_N]$，将该特征向量作为该物质的特征参数向量。对纯净物和混合物分别进行上述操作，即可获得每个纯净物或混合物的拉曼光谱特征向量。

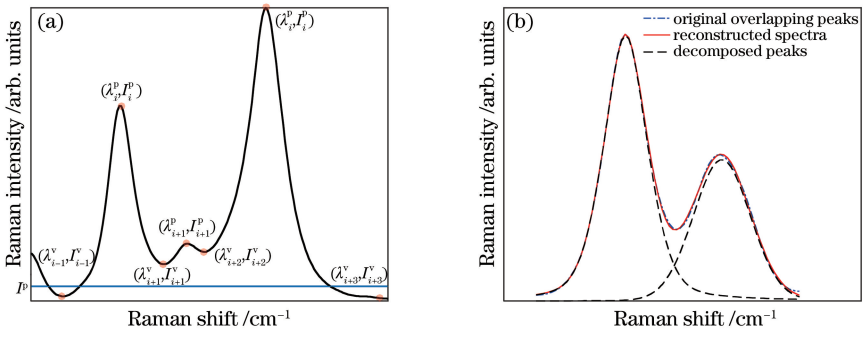

图 3　重叠峰的区域划分与分解。(a)区域划分；(b)分解

Fig. 3　Diagrams of overlapping peak region division and decomposition. (a) Region division; (b) decomposition

图 3(a)为划分重叠峰区域的示意图，图 3(b)为对重叠峰进行谱峰分解的结果。可以看出，LM 算法可以准确地分解出各峰，且分解结果较好。

2.2.3　匹配搜索策略

设待识别混合物 m 的谱峰特征参数向量为$[\lambda^{m}_1,I^{m}_1,\omega^{m}_1;\cdots;\lambda^{m}_i,I^{m}_i,\omega^{m}_i;\cdots;\lambda^{m}_h,I^{m}_h,\omega^{m}_h]$，其中 i 和 h 分别表示待识别物的第 i 个谱峰和谱峰总数。纯净物库中纯净物 p 的特征参数向量为$[\lambda^{p}_1,I^{p}_1,\omega^{p}_1;\cdots;\lambda^{p}_j,I^{p}_j,\omega^{p}_j;\cdots;\lambda^{p}_n,I^{p}_n,\omega^{p}_n]$，$j$ 和 n 分别表示该纯净物的第 j 个谱峰和谱峰总数。对于该纯净物的第 j 个谱峰$[\lambda^{p}_j,I^{p}_j,\omega^{p}_j]$，在待识别物中找到距离该谱峰最近的峰，记其位置索引为 t，$t\in\{1,\cdots,i,\cdots,h\}$，则该谱峰的特征参数为$[\lambda^{m}_t,I^{m}_t,\omega^{m}_t]$。计算$[\lambda^{p}_j,I^{p}_j,\omega^{p}_j]$与$[\lambda^{m}_t,I^{m}_t,\omega^{m}_t]$的相似度值，公式为 $S_j=[S_j(\lambda)+S_j(\omega)]/2$，其中 $S_j(\lambda)$、$S_j(\omega)$ 分别表示纯净物 p 的第 j 个谱峰与待识别物 m 的第 t

个谱峰的拉曼位移和半峰全宽的相似度。考虑到纯净物和待识别混合物可能存在的谱峰偏移和波形变化，本文用模糊隶属度函数计算两者之间的相似度。计算公式为

$$S_j(x)=\begin{cases}1, & x\leqslant l_1\\ \exp\left[\dfrac{-(x-l_1)^2}{2c^2}\right], & l_1<x<l_2\\ 0, & x\geqslant l_2\end{cases},\tag{3}$$

式中：$x=|\lambda_j^{p}-\lambda_t^{m}|$ 或 $x=|\omega_j^{p}-\omega_t^{m}|$；$c$ 为常数；l_1 为计算相似度时 x 的左阈值；l_2 为计算相似度时 x 的右阈值。当 $x\leqslant l_1$ 时，相似度设为 1；当 $l_1<x<l_2$ 时，相似度函数为一个随 x 增大而逐渐下降的高斯函数；当 $x\geqslant l_2$ 时，相似度设为 0。

本文结合所用手持式拉曼光谱仪的性能，将(3)式中的参数设置如下：对于拉曼位移，$l_1=5$，$l_2=15$，$c=5$；对于半峰全宽，$l_1=3$，$l_2=20$，$c=3$。

对纯净物 p 的每个谱峰均进行以上操作，则 p 与 m 的整体相似度值为

$$S^{p,m}=\sum_{j=1}^{n}\frac{I_j^{p}}{\sum_{i=1}^{n}I_i^{p}}\times S_j。\tag{4}$$

式(4)中引入了权重因子。对于强度较大的谱峰，本文赋予较大的权值，反之，赋予较小的权值，从而减小潜在的虚假谱峰(通常强度较小)对相似度计算的影响。

考虑到混合物中多个成分相互干扰导致的谱峰偏移现象，本文引入已知混合物数据库的协助搜索策略，以进一步减小谱峰偏移对相似度计算的影响。若已知混合物库中含有纯净物 p 的混合物个数为 K 个，将混合物记为 $M^1,\cdots,M^k,\cdots,M^K$，则进行如下搜索匹配操作：

1)在已知混合物中，寻找其含有的纯净物 p 的谱峰特征。设第 $k(k=1,\cdots,K)$ 个含有纯净物 p 的已知混合物 M^k 的特征参数向量为$[\lambda_1^{M^k},I_1^{M^k},\omega_1^{M^k};\cdots;\lambda_i^{M^k},I_i^{M^k},\omega_i^{M^k};\cdots;\lambda_h^{M^k},I_h^{M^k},\omega_h^{M^k}]$，其中 i 和 h 分别表示其第 i 个谱峰和谱峰总数。对于纯净物 p 的第 j 个谱峰$[\lambda_j^{p},I_j^{p},\omega_j^{p}]$$(j=1,\cdots,n)$，在已知混合物 M^k 中找到距离纯净物第 j 个谱峰最近的峰，记其位置索引为 q，$q\in\{1,\cdots,i,\cdots,h\}$，谱峰特征参数为$[\lambda_q^{M^k},I_q^{M^k},\omega_q^{M^k}]$。若$[\lambda_q^{M^k},I_q^{M^k},\omega_q^{M^k}]$满足

$$[\lambda_j^{mdb},I_j^{mdb},w_j^{mdb}]=\begin{cases}[\lambda_q^{M^k},I_q^{M^k},w_q^{M^k}], & d\leqslant h_1\\ [\lambda_q^{M^k},I_q^{M^k},w_q^{M^k}], & h_1<d<h_2,S_j(\omega)\geqslant s_\omega\\ \text{null}, & \text{others}\end{cases},\tag{5}$$

则可认为已知混合物的第 q 个谱峰与纯净物 p 的第 j 个谱峰相对应(即其为纯净物第 j 个谱峰的可能偏移)，记录为$[\lambda_j^{mdb},I_j^{mdb},w_j^{mdb}]$。在(5)式中，$d=|\lambda_j^{p}-\lambda_q^{M^k}|$，$\omega=|\omega_j^{p}-\omega_q^{M^k}|$，$S_j(\omega)$为半峰全宽的相似度(根据(3)式计算)，$h_1$、$h_2$ 为许可的谱峰拉曼位移偏移阈值，s_ω 为半峰全宽的相似度阈值(本文取 0.6)。对纯净物 p 所有的谱峰($j=1,\cdots,n$)，按照上述步骤进行操作，可以获得已知混合物 M^k 中所含有的对应于纯净物 p 的谱峰特征向量为$[\lambda_1^{mdb},I_1^{mdb},\omega_1^{mdb};\cdots;\lambda_j^{mdb},I_j^{mdb},\omega_j^{mdb};\cdots;\lambda_n^{mdb},I_n^{mdb},\omega_n^{mdb}]$。

2)利用已知混合物辅助搜索。将构建的已知混合物 $M^k(k=1,\cdots,K)$ 中对应纯净物 p 的谱峰特征向量$[\lambda_1^{mdb},I_1^{mdb},\omega_1^{mdb};\cdots;\lambda_j^{mdb},I_j^{mdb},\omega_j^{mdb};\cdots;\lambda_n^{mdb},I_n^{mdb},\omega_n^{mdb}]$与待识别混合物 m 的谱峰特征向量$[\lambda_1^{m},I_1^{m},\omega_1^{m};\cdots;\lambda_i^{m},I_i^{m},\omega_i^{m};\cdots;\lambda_h^{m},I_h^{m},\omega_h^{m}]$，按照(3)式和(4)式计算相似度，获得其相似度值为 $S^{M^k,m}$。若 $S^{M^k,m}>S^{p,m}$，将 $S^{M^k,m}$ 的值赋给 $S^{p,m}$；否则，$S^{p,m}$ 维持原值。如此，遍历已知混合物库，获得的 $S^{p,m}$ 值即为待识别混合物 m 与纯净物 p 最终的相似度值。

依此方法遍历纯净物库，获得待识别混合物与各纯净物的相似度值，将该相似度值与纯净物对应的序号存储于数组中，对数组依照相似度值从大到小排序。图 4 给出了算法的完整流程图。

3　结果与讨论

3.1　实验结果

本文对 81 组待识别混合物进行了识别，比较了子空间匹配(SM)方法[15]、纯净物数据库匹配(PDM)方法以及本文提出的纯净物和已知混合物数据库匹配(PMDM)方法的性能。SM 算法的相关参数设置同文献[15]，PDM 方法仅在 204 种纯净物库中搜索，而 PMDM 方法在此基础上增加了一个大小为 8 的已知混合物数据库。综合考虑手持式拉曼光谱仪的精度、应用要求，以及待识别混合物包含的组分数目，本文取前 7 个候选物作为最终的识别结果(即若混合物的真实组分在选择的候选物之内，就认为该组分识别成功)。识别准确率定义为正确

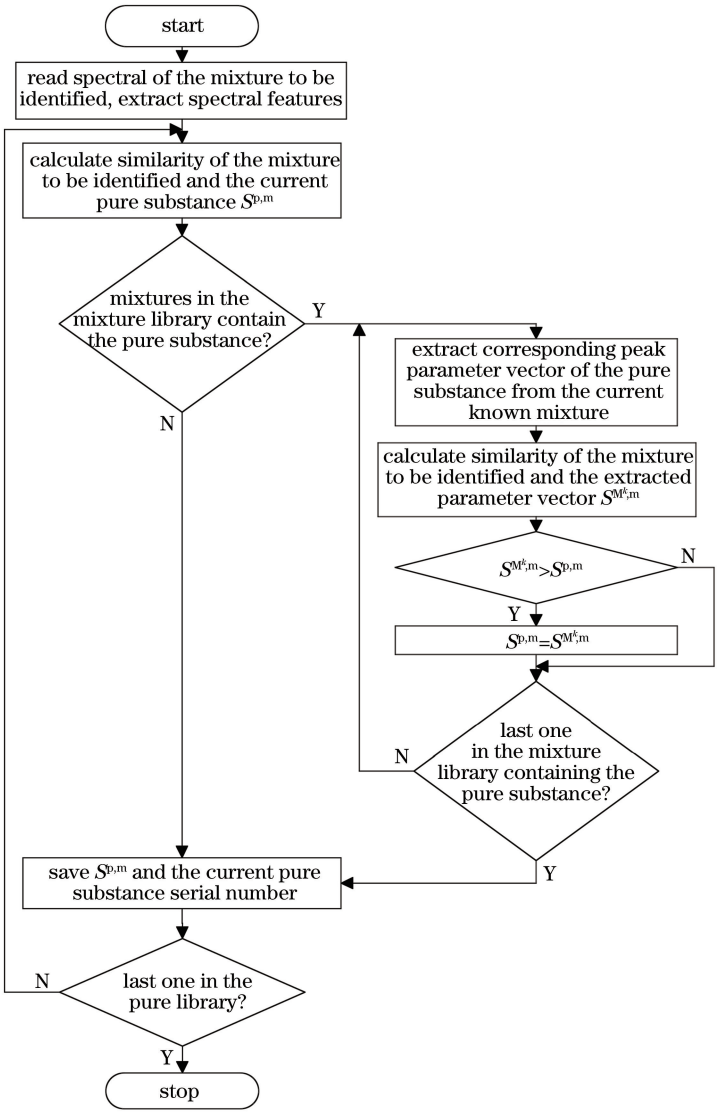

图 4　算法流程图

Fig. 4　Flowchart of algorithm

识别的组分数量占所有待识别混合物总组分数量的百分比。由于已知混合物数据库的构建方式为从 8 组不同浓度比的已知混合物中每组随机抽取一个混合物，考虑到该方式的随机性，取连续 10 次结果的平均值作为最终的识别准确率。

表 3 所示为三种方法对 45 个三元混合物和 36 个四元混合物的识别结果，表 4 为三种方法对混合物单个组分的识别结果。可以看出：在使用 PDM 方法时(无已知混合物数据库)，即使采用了模糊隶属度函数，算法的识别精度依然低于 SM 方法；而 PMDM 方法的整体识别精度相比 PDM 方法提高了约 16 个百分点，并大幅超过了 SM 方法。

表 3　各方法对混合物的识别精度

Table 3　Identification accuracy of each method

Mixture	Identification accuracy			Total number of components
	SM	PDM	PMDM	
Ternary	91.85%(124)	86.67%(117)	96.30%(130)	135
Quaternary	68.75%(99)	66.67%(96)	89.58%(129)	144
Total	79.93%(223)	76.34%(213)	92.83%(259)	279

Notes: The number of components correctly identified in parenthesis.

表 4　单个组分的识别精度

Table 4　Identification accuracy of individual component

Component	Identification accuracy			Total number of components
	SM	PDM	PMDM	
Ethanol	55.56%(35)	74.60%(47)	93.65%(59)	63
Acetonitrile	100%(33)	100%(33)	100%(33)	33
Acetone	80.56%(58)	85.92%(61)	90.28%(65)	72
Cyclohexane	100%(18)	100%(18)	100%(18)	18
Diacetone alcohol	90.20%(46)	58.82%(30)	93.73%(48)	51
Diethyl malonate	78.57%(33)	57.14%(24)	85.95%(36)	42

Notes: The number of components correctly identified in parenthesis.

图 5 为谱峰偏移现象示意图，其中纯净物为丙二酸二乙酯，已知混合物的组分为环己烷、二丙酮醇与丙二酸二乙酯，待识别物的组分为丙酮、二丙酮醇与丙二酸二乙酯。图 5 所示为 900～990 cm^{-1} 波段的光谱图，可以看到，纯净物谱峰的拉曼位移为 947 cm^{-1}，而已知混合物与待识别混合物的拉曼位移分别偏移至 935 cm^{-1} 和 936 cm^{-1}。由于偏移现象的存在，若此时仅采用纯净物库搜索策略，将会给该谱峰相似度的计算带来很大误差，而已知混合物的应用则有效补偿了偏移现象带来的干扰。

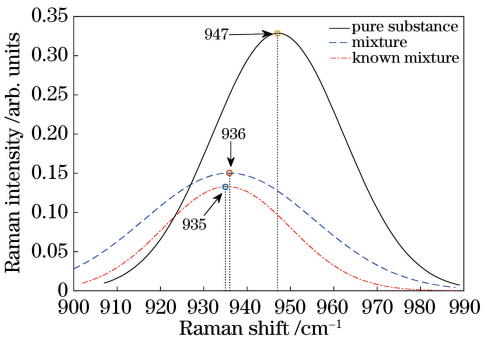

图 5　谱峰偏移示意图

Fig. 5　Schematic of spectral peak shift

由图 5 和表 3～4 可知：相比于仅仅建立纯净物库，对已知混合物的利用可以有效提高待识别物的识别精度。

3.2　已知混合物数据库的大小对识别精度的影响

已知混合物数据库的大小对识别结果具有一定影响，且影响主要体现在混合物数据库的组分构成以及混合物中各组分的浓度上。本文进行了如下实验：从 8 组已知混合物（见表 1）中随机选择 2 组、4 组和 6 组物质（代表已知混合物库中的不同组分构成），再从每组物质中随机抽取不同数量的不同浓度比的混合物（代表已知混合物库中组分浓度的影响），构成已知混合物数据库（其大小等于抽取的组分数量乘以抽取的不同浓度比混合物的数量）。将这些不同大小的子数据库分别应用于 81 个未知混合物的识别，并重复 10 次实验，取识别精度的平均值，结果如表 5。

表 5　不同大小的已知数据库下的识别精度

Table 5　Identification accuracy under different sizes of known databases　%

Number of groups	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
2	84.95	86.74	88.89	88.89	89.25	89.61	90.32	90.68
4	86.74	90.68	93.62	93.62	93.62	94.34	94.70	94.70
6	89.61	93.27	95.06	95.42	96.49	96.58	97.21	97.92

可以看出，随着已知混合物数据库的增大，识别准确率整体上升，并逐渐趋于稳定，说明更多的已知混合物会为未知混合物的识别提供更多的辅助识别信息，提高识别准确率。

在利用拉曼光谱进行混合物成分识别过程中，会产生大量的历史检测数据。在传统的识别方法中，这些历史检测数据会被丢弃。本文所提方法在实际应用中具有重要价值。在实际应用中，只需要定期利用历史检测数据更新已知混合物数据库，就可以达到改善检测精度的目的。

本文方法基于光谱的特征峰参数来计算不同物质之间的相似度值，但如何保证低浓度下弱谱峰的准确检测，并改善相似度的计算准确性是一个需要进一步研究的课题；同时，如何结合测量环境，对已知混合物数据库进行针对性的筛选，以实现已知混合物特征信息的有效挖掘，也是值得研究的问题。

4　结　　论

本文提出了一种使用已知混合物光谱数据库进行辅助搜索的方法。该方法首先将拉曼光谱谱峰的

拉曼位移与半峰全宽作为特征，采用模糊隶属度函数计算光谱特征的相似度；然后通过提取已知混合物光谱中包含的物质的光谱信息，降低拉曼光谱谱峰偏移造成的相似度计算误差。本文对81种混合物进行了组分识别，识别结果表明：相比于单纯的依赖于纯净物数据库的方法，本文方法的识别精度提高到92.83%，验证了本文方法的有效性。

参　考　文　献

[1] Huang S G, Hu J P, Liu M H, et al. Density functional theory calculation and Raman spectroscopy studies of carbamate pesticides[J]. Spectroscopy and Spectral Analysis, 2017, 37(3): 766-771.
黄双根, 胡建平, 刘木华, 等. 氨基甲酸酯类农药的密度泛函理论计算及拉曼光谱研究[J]. 光谱学与光谱分析, 2017, 37(3): 766-771.

[2] Chen S, Guo P, Wan J C, et al. Rapid detecting study of sodium saccharin additive in spirit with SERS[J]. Spectroscopy and Spectral Analysis, 2017, 37(5): 1412-1417.
陈思, 郭平, 万建春, 等. 白酒中糖精钠添加剂表面增强拉曼光谱快速检测研究[J]. 光谱学与光谱分析, 2017, 37(5): 1412-1417.

[3] Xu H D, Lin L L, Li Z, et al. Nephrite origin identification based on Raman spectroscopy and pattern recognition algorithms [J]. Acta Optica Sinica, 2019, 39(3): 0330001.
徐荟迪, 林露璐, 李征, 等. 基于拉曼光谱和模式识别算法的软玉产地鉴别[J]. 光学学报, 2019, 39(3): 0330001.

[4] Liu C, Zang Y C, Zeng H T, et al. Rapid detection of methotrexate and voriconazole in mixtures using surface-enhanced Raman spectroscopy with features matching in wavelet space [J]. Journal of Instrumental Analysis, 2019, 38(6): 668-674.
刘察, 臧颖超, 曾惠桃, 等. 基于小波空间特征匹配及表面增强拉曼光谱技术快速检测混合物中的甲氨蝶呤和伏立康唑[J]. 分析测试学报, 2019, 38(6): 668-674.

[5] Li X L, Zhou R Q, Xu Y F, et al. Spectral unmixing combined with Raman imaging, a preferable analytic technique for molecule visualization [J]. Applied Spectroscopy Reviews, 2017, 52(5): 417-438.

[6] Zhuang X M, Li S Y, Li F, et al. Excess Raman spectroscopy of ammonium sulfate aqueous solution [J]. Acta Optica Sinica, 2018, 38(6): 0630002.
庄欣明, 李申予, 李非, 等. 硫酸铵水溶液的超额拉曼光谱研究[J]. 光学学报, 2018, 38(6): 0630002.

[7] Liu C Z, Zhu Q B, Huang M, et al. Identification of components in mixtures based on Raman spectroscopy[J]. Laser & Optoelectronics Progress, 2019, 56(8): 083004.
刘财政, 朱启兵, 黄敏, 等. 基于拉曼光谱的混合物组分识别方法[J]. 激光与光电子学进展, 2019, 56(8): 083004.

[8] Zhang Z M, Chen X Q, Lu H M, et al. Mixture analysis using reverse searching and non-negative least squares [J]. Chemometrics and Intelligent Laboratory Systems, 2014, 137: 10-20.

[9] Liu M H, Dong Z R, Xin G F, et al. Raman spectrum library matching method based on integrated features [J]. Chinese Journal of Lasers, 2019, 46(1): 0111002.
刘铭晖, 董作人, 辛国锋, 等. 基于集成特征的拉曼光谱谱库匹配方法[J]. 中国激光, 2019, 46(1): 0111002.

[10] He Y, Wang J F. Rapid nondestructive identification of wood lacquer using Raman spectroscopy based on characteristic-band-Fisher-K nearest neighbor [J]. Laser & Optoelectronics Progress, 2020, 57(1): 013001.
何亚, 王继芬. 基于特征波段-Fisher-K近邻的木器漆拉曼光谱的快速无损鉴别[J]. 激光与光电子学进展, 2020, 57(1): 013001.

[11] Liu Y D, Cheng M J, Hao Y, et al. Quantitative analysis of chlorophyll content in citrus leaves by Raman spectroscopy [J]. Spectroscopy and Spectral Analysis, 2019, 39(6): 1768-1772.
刘燕德, 程梦杰, 郝勇, 等. 柑橘叶片叶绿素含量拉曼光谱定量分析方法研究[J]. 光谱学与光谱分析, 2019, 39(6): 1768-1772.

[12] Zhang Z M, Chen S, Liang Y Z, et al. An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy [J]. Journal of Raman Spectroscopy, 2010, 41(6): 659-669.

[13] Eilers P H C. A perfect smoother [J]. Analytical Chemistry, 2003, 75(14): 3631-3636.

[14] Levenberg K. A method for the solution of certain non-linear problems in least squares[J]. Quarterly of Applied Mathematics, 1944, 2(2): 164-168.

[15] Huang P X, Yao Z X, Su H, et al. Spectral pattern recognition of mixed alcohols by means of the method based on judging the subspace coincidence [J]. Journal of Instrumental Analysis, 2013, 32(3): 281-286.
黄培贤, 姚志湘, 粟晖, 等. 基于子空间重合判断的混合醇组分光谱识别方法[J]. 分析测试学报, 2013, 32(3): 281-286.