第42卷，第10期
2022年10月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42，No. 10，pp3174-3179
October，2022

一种基于中值滤波和统计学的拉曼光谱 spike 剔除算法

叶瑞乾[1]，何 浩[1]，郑 鹏[1]，徐梦溪[2]，王 磊[1*]

1. 厦门大学航空航天学院，福建 厦门 361101
2. 厦门大学化学化工学院，固体表面物理化学国家重点实验室，能源材料化学协同创新中心(iChEM)，福建 厦门 361101

摘 要 拉曼光谱是一种已广泛应用于化学、生物学和物理学的技术。然而拉曼光谱仪的电荷耦合器件很容易受到宇宙射线的影响，从而产生随机的窄带宽、高强度的 spike。在真实样品中出现概率较低，约为千分之一，但一旦出现将严重降低信号对比度。该研究提出一种实用的 spike 剔除算法。该算法对中值滤波后的数据与原始数据作差，得到偏差数据。用分位数的方法将偏差数据从小到大排序，取中间 99%数据作为真实数据作高斯分布拟合。根据 spike 强度高，稀疏的特性，以光谱中高强度数据的出现概率作为阈值标准剔除 spike。最后以中值滤波结果带入原始数据代替 spike，从而最大程度还原样本原始信息且不需任何调试参数。以加入不同强度 spike 的拉曼光谱作为验证对象，实验结果表明本算法对 spike 检测与去除的灵敏度可以高达 99.5%。本算法同时适用于一维拉曼光谱、二维拉曼图像和三维拉曼数据立方体，且算法表现随着维度的增加而提高，一维 spike 剔除算法能检测超过最大峰强度 40%的 spike，而在三维拉曼数据立方体中，超过峰值 20%的 spike 即能被检测出。用该算法对40 000条真实拉曼光谱进行处理，可以在不扭曲真实信号的情况下有效地剔除 spike，进一步证明了算法的实用性。

关键词 拉曼光谱；宇宙射线；中值滤波；spike 剔除

中图分类号：O657.37 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)10-3174-06

引 言

拉曼光谱能原位地实时探测多物质的化学组分[1]，被广泛应用于细胞生物和电化学[2]、食品安全[3]、临床医学[4-5]等。然而，绝大多数分子的拉曼散射效应极其微弱，因而拉曼光谱中经常包含大量噪声。其中 spike 是一种在随机位置上出现的宽度极窄、峰高很强的单向尖峰信号[6]。一般认为 spike 的产生是因为宇宙射线被拉曼光谱仪中广泛使用的电荷耦合器件(CCD)[7]所捕获，宇宙射线可以直接激发 CCD 像素，产生的 spike 的强度有时比正常的拉曼峰高出几个数量级，会显著降低光谱的对比度，必须在进一步分析之前被剔除。

目前已经有一些文献介绍不同的 spike 剔除方法，可以归结于两类，首先是基于硬件的实现。例如，Zhao[8]用直缝扩展狭缝高度以提升信号强度，并分析由此带来的图像弯曲问题，加以校正，将 CCD 图像分成多个条带，通过条带之间的比较可以识别和剔除 spike；Huang[9]等将 CCD 高度所可以容纳的光纤打包到光纤束中，并以特殊的方法排列，以修复被 spike 损坏的图像。然而，复杂的仪器操作和额外成本的增加使基于硬件的 spike 剔除方法的应用较少。第二类剔除方法是基于软件的 spike 剔除方法。这些方法大致可以分为两种类型。第一种类型是使用滤波器对信号进行处理，包括有鲁棒性的平滑滤波器[10]、加权移动窗口滤波器[11]、小波变换[6,12-13]等。这些算法是基于以下假设，即与真实的拉曼峰相比，spike 的带宽更小且强度更高。然而，在实际应用中，一些反映真实分子信息的窄峰会被误认为是 spike。此外，使用滤波器可能会让真实的拉曼光谱失真。第二种是基于概率统计原理的剔除方法，包括上限频谱数据矩阵[14]、鲁棒求和方法[15]、基于分布设置阈值[16-17]和差值谱图检测方法[18]。这些方法成立的理论基础是在两个连续光谱中的同一位置发生 spike 的概率非常小。由于需要采集多次，因此，这些方法需要较长的处理时间来搜索 spike，而且参数的选择也很复杂。同时，这两类 spike 剔除方法的处理对象是基于某条特定的光谱，对于有成千上万条光谱的拉曼成像数据立方体，其应用效率较低且保真性较差。随着拉曼成像技术

收稿日期：2021-08-19，**修订日期**：2022-02-25

基金项目：国家自然科学基金项目(21373173)资助

作者简介：叶瑞乾，1996年生，厦门大学航空航天学院硕士研究生 e-mail：yeruiqian@stu.xmu.edu.cn

*通讯作者 e-mail：wanglei33@xmu.edu.cn

的发展，针对大量光谱数据的批量 spike 剔除方法是未来的发展要求。

因此，本文提出一种实用的 spike 剔除方法，其原理是使用中值滤波器处理原始数据，并对中值滤波后的数据和原始数据的偏差作统计分析来寻找 spike，并只替换这些 spike 值，以保证数据的高保真度，中值滤波的参数设置较于上述方法更加简单且有效。同时该算法有大量数据的处理能力，不仅适用于一维拉曼谱和二维拉曼图像的 spike 剔除，还适用于拉曼光谱成像得到的三维数据立方体。实验证明，当 spike 强度超过最高峰值强度 20%时，该算法可以准确地识别 spike，灵敏度为 99.5%。与目前 spike 识别方法[14,19]相比，该算法具有较高的灵敏度和较小的检测极限，为拉曼光谱分析人员进行后续数据降维及机器学习算法等应用提供了一个重要的光谱预处理工具。

1　实验部分

1.1　通过中值滤波器和统计分析剔除 spike

spike 可以定义为一维、二维或者三维数据内相邻小区域的异常值。中值滤波器是以当前数据点为中心的小窗口的中值来代替每个数据点，其原理可以用来剔除这些异常值，但是非 spike 的信号也将随着滤波操作而失真。因此，我们开发了一种分两步进行的 spike 剔除方法，第一步是定位 spike，第二步是只对被定位的 spike 执行中值滤波。这种策略可以较好地去除 spike，并保留原始信号的高保真度。定位 spike 是其中的关键，可以根据中值滤波后的数据和原始数据作偏差，通过设置阈值来识别 spike。将原始数据表示为 M，滤波后的数据表示为 N，偏差表示为 $E(E=M-N)$。E 的分布可以近似表示为高斯函数[式(1)]

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1}$$

式(1)中，x 是随机变量，$f(x)$是概率密度，μ 是期望，σ 是标准差。

根据 3-σ 规则，99.7%的数据在以期望 μ 为中心，标准方差 σ 的 3 倍区域内。这个特征可以用来识别 spike，因为 spike 通常稀疏地分布在测量数据中，并且具有比正常信号高得多的强度。这些 spike 可以通过根据 E 的分布性质设置一个阈值来识别。该阈值将把 E 中大部分较小的数据点识别为正常值，而当 E 中的一个数据点的值超过该阈值时，它将被识别为 spike。如果原始数据 M 中没有 spike，则 E 中的大部分值将非常小，拟合分布的期望将接近 0。但如果 spike 存在，使用式(1)直接拟合，拟合曲线将因为异常值的原因偏离它的真实分布。为解决这个问题，本文按升序排序 E，并使用中间数据而不是使用 E 中的所有值来拟合分布。我们使用分位数方法将 E 切成一个系列连续区域。分位数的函数可以被描述为式(2)

$$\tau=P(X\leqslant x_\tau)=F(x_\tau) \tag{2}$$

式(2)中，X 是随机变量，$F(x)$是连续分布函数，P 是概率，τ 是分位数。分位数指满足条件 $P(X<x_\tau)$的点。

根据分位数的概念将 E 中的数据点以概率的方式划分为 100 个份，通过选择合适的 τ，取中间的 99 个部分作为拟合分布所用的数据，中间 99%的数据已剔除强度很大的 spike 和测量结果为负的 CCD 坏点，可以表现真实拉曼信号特征，经过实验测试，取 99%以上的中间数据会引入个别异常值，取 99%以下的数据，对拟合的分布提升不明显，且可能会丢失一些真实信号。原始数据和经过排序后的中间 99%数据的拟合分布结果如图 1 所示。

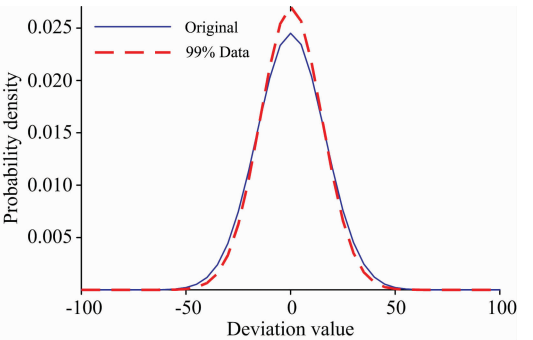

图 1　原始数据和中间 99%的数据的分布

Fig. 1　The distribution with the original data and middle 99% data

图 1 是使用全部数据及中间 99%的数据做拟合的概率密度函数图，其中使用 99%的数据的分布更瘦更高，意义为数据更加集中，异常值和正常值的距离更远，且大部分的偏差 E 在 0 左右，99%的数据拟合后左右两端比使用全部数据的拟合更窄，意味着正常值的数据更加集中，阈值的设置更加简单。

通过基于概率密度函数(PDF)的原理，为 E 的真实分布设置一个阈值，可以很容易地识别 spike。该原理可以表示为式(3)

$$P(X\leqslant x)=F_X(x)=\int_{-\infty}^{x}f(t)\mathrm{d}t \tag{3}$$

式(3)中，X 是随机变量，x 是实数，P 为概率，$f(t)$为概率密度函数。

PDF 函数可以通过设定一个数来获取数据超过该数的概率，但由这种直接方法设置的阈值将有两个缺陷。一是阈值选择的难度：关于 spike 和正常值之间的距离尚不清楚，因此没有参考依据。另一个缺陷是缺乏自动化和通用性：针对不同的拉曼数据需要设置不同的阈值，这对分析者来说很不方便且设置的值与使用者的经验有关。为了解决这些问题，我们采用基于逆累积分布函数(ICDF)的反向思想来直接设置概率阈值，通过设定某个概率 $P(X\geqslant x)$，可以根据该分布函数自动确定一个阈值 x，数据超过该值为满足这个概率$(X\geqslant x)$。因此，在不同的拉曼数据中将根据概率自动调整不同的绝对阈值 x 以剔除 spike。为了方便起见，本文以下部分中设置的阈值指的均是概率阈值 P。

在通过阈值设置检测到 spike 之后，下一步是用表示真实拉曼特征的值来替换原始数据 M 中的 spike。最简单、直接的方法是用线性插值来填充空位置。但这种方法引入了伪影，并会因拉曼数据的复杂性而导致错误的结果。应用非线

性函数进行插值可以使替换完成的拉曼数据更逼近真实数据。然而，一些非线性函数会引入更多的变量和未知的参数，这将使这个过程变得极其复杂，错误的参数甚至会降低拉曼特征的准确性。一种简单有效的方法是使用中值滤波后的数据 N 中的值代替原始数据 M 中的 spike。中值将当前像素中的信号与前后像素连接起来，确保该值的可靠性。此外，它仅需要少量的参数来确定，易于操作。

整个算法的基本思想是在偏差分布 E 中找到超过阈值的数据点，并将其标记为 spike。然后去除原始数据 M 中的这些 spike，并填充对应位置的滤波后的 N 值，具体步骤如表 1 所示。拉曼数据 M 可以是一维拉曼光谱、二维拉曼图像和三维拉曼数据立方体，取决于维度 D。Median(M, W)表示对数据 M 应用窗口大小为 W 的中值滤波器，Fit(E)代表对分布 E 作曲线拟合，$M(I)$是数据 M 在位置 I 的索引值，I 可以是(i)，表示一维数据，(i, j)表示二维数据，(i, j, k)表示三维数据。

表 1 算法流程图

Table 1 Flowchart of algorithm

	Input raw Raman data M, Window size W, Threshold θ, Dimension D;
1	Data N=Median(M, W);
2	Deviation $E=M-N$;
3	Distribution F=Fit(middle 99% E);
4	Value $V_1=F(\theta)$, $V_2=F(1-\theta)$;
5	If $E(I)<V_1$ or $E(I)>V_2$
6	Replace $M(I)$ by $N(I)$;
7	Endif
	Output Data M

1.2 参数优化

参数优化包括中值滤波器的窗口大小的选择和阈值的设定。这些参数应根据拉曼数据的噪声水平、强度和锐度的差异来选择，以在不影响真实拉曼特征的情况下准确地去除 spike。本文针对三种维度下不同窗口尺寸和不同阈值进行实验。在此，以一维拉曼谱为例，演示如何选择合适的参数。拉曼光谱中不同参数下的结果如图 2 所示。

图 2a 为未经 spike 剔除算法的一维拉曼光谱数据，使用较大的窗口，有可能无法完全剔除图中的 spike(如图 2b 所示)，或对图中的真实拉曼峰造成影响(如图 2d)，使用较小的窗口尺寸的结果如图 2e 所示，由于 spike 窄带宽，高强度特征，小窗口识别更成功。因此，我们选择 3 作为一维拉曼光谱的窗口尺寸。

在实验中，我们设置阈值 1×10^{-3} 和 1×10^{-5} 并进行比较。尽管阈值相差两个数量级但只有一个位于 2 000 波数附近的 spike 没有被检测到。引入分位数的方法确保了拟合分布是使用正常值，扩大了正常值和 spike 的差异，让阈值选择空间足够大。

二维拉曼图像和三维拉曼数据立方体的参数选择与上述操作相似。对于二维图像，将使用二维中值滤波器作用于原始数据。从一维拉曼光谱到二维拉曼图像的扩展不仅仅是从

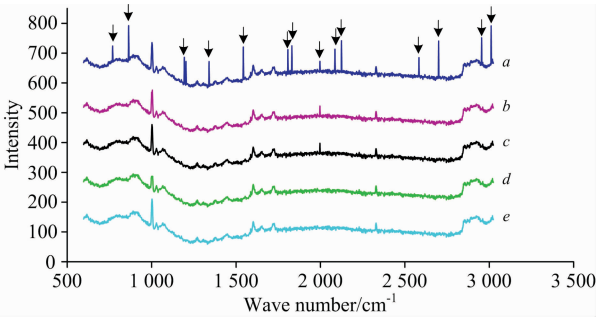

图 2 不同窗口尺寸和阈值的一维 spike 剔除

a：原始拉曼光谱；b：窗口尺寸为 7，阈值为 1×10^{-5} 的 spike 剔除过程；c：窗口尺寸为 3，阈值为 1×10^{-5} 的 spike 剔除过程；d：窗口尺寸为 7，阈值为 1×10^{-3} 的 spike 剔除过程；e：窗口尺寸为 3，阈值为 1×10^{-3} 的 spike 剔除过程

Fig. 2 Spike removal of 1-dimensional spectra with different window size and threshold

a: Raw spectrum; b: Spike removal with window size 7, threshold 1×10^{-5}; c: Spike removal with window size 3, threshold 1×10^{-5}; d: Spike removal with window size 7, threshold 1×10^{-3}; e: Spike removal with window size 3, threshold 1×10^{-3}. Arbitrary offsets were applied on the spectra for better visualization

行扩展到列的更改，这涉及到滤波器在不同对象上的操作。在一维数据中，对不同波数下的信号强度进行滤波，是光谱维度上的滤波。在二维中，同一波数上的不同像素信号被滤波，对于采集过程，属于空间上的滤波。原始数据与中值滤波后的数据之间的偏差是二维分布，因此设置的阈值也是二维数据，对应于图像的行和列。为了简化拟合分布，我们首先将二维偏差数据重组为一维偏差数据。对扩展的一维数据的分析与前面描述的一维拉曼光谱数据相同。本文选择 3×3 作为窗口大小，1×10^{-6} 作为设定的阈值用来剔除二维拉曼图像中的 spike。

对于三维拉曼数据立方体，数据的中值滤波器和偏差也将是三维的。三维数据的扩展既需要考虑拉曼数据采集的时间，也需要考虑不同的波数，首先是在深度方向上，然后是在宽度方向上，最后是在高度方向上。这将极大增加数据的数量，并使偏差分布更接近真实分布，也更准确。

根据参数的选择，对于三维数据立方体，信号的变化受到多维的限制。因此，一个小的滤波窗口足以去除 spike，也可以减少程序的处理时间。由于数据量的扩展，阈值设置为 1×10^{-6}。

2 结果与讨论

为了测试该算法在实际应用中的性能，使用真实的实验数据集进行了测试。我们使用 Raman 11 线扫描快速成像共聚焦拉曼光谱仪器(Nanophoton, Japan)用于扫描生物细胞。该实验总共收集 100 线的光谱，每线包含 400 条光谱，光谱深度为 1 340 个波数。收集到的数据可以以完整光谱的形式进行区分，共获得 40 000 条光谱，每条光谱包含 1 340 个波数。如果使用每次采集的光谱数和采集的数量作为区分条

件，可以获得 1 340 张大小为 100×400 的拉曼图像。此外，我们还可以根据采集深度、采集宽度和采集线数来构建一个大小为 1 340×400×100 的三维拉曼数据立方体。

为了更好地分析该算法的有效性，我们还在真实的实验数据中引入了不同数量的、不同强度的人工 spike 来进行验证实验。由于一维拉曼光谱、二维拉曼图像和三维拉曼数据立方体原理的不同，我们在不同的条件下进行了实验。以下分析中列出了具体的实验参数。

2.1　一维 spike 剔除

一维 spike 剔除的处理结果如图 3 所示。Spike 出现概率较小，但是强度很大，远超过其他波峰，该算法很好地剔除了 spike，同时完整保留了 1 000 cm^{-1}处的强拉曼峰(细胞成分：苯丙氨酸)。

作为验证，测试了 20 个光谱，其中有 10 个手动添加的人工 spike。增加的 spike 位置是随机的，强度为峰值的 0.1～0.4 倍。实验结果见图 4。

用该算法进行了 10 次独立的验证实验，并将检测结果进行平均，最终结果如表 2 所示。其中 Ratios 表示峰值的比例(0.1～0.4 倍)，Predicted 表示算法所预测出来的 spike 数目，True 为真实 spike 数量，Percentage 是算法的准确度。

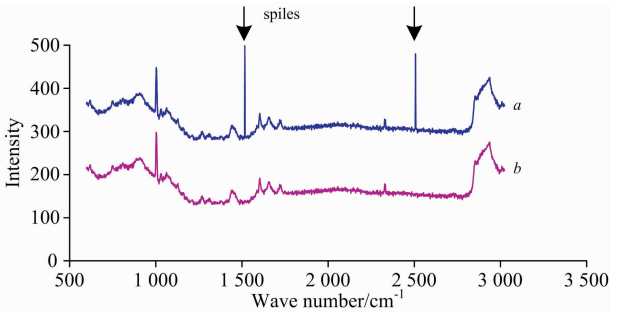

图 3　一维拉曼光谱 spike 剔除

a：原始拉曼光谱；

b：经过 spike 剔除算法后的拉曼光谱

Fig. 3　1-dimensional Raman spectra spike removal process

a: Raw Raman spectrum; *b*: Raman Spectrum after 1-dimensional spike removal. Arbitrary offsets were applied on the spectra for better visualization

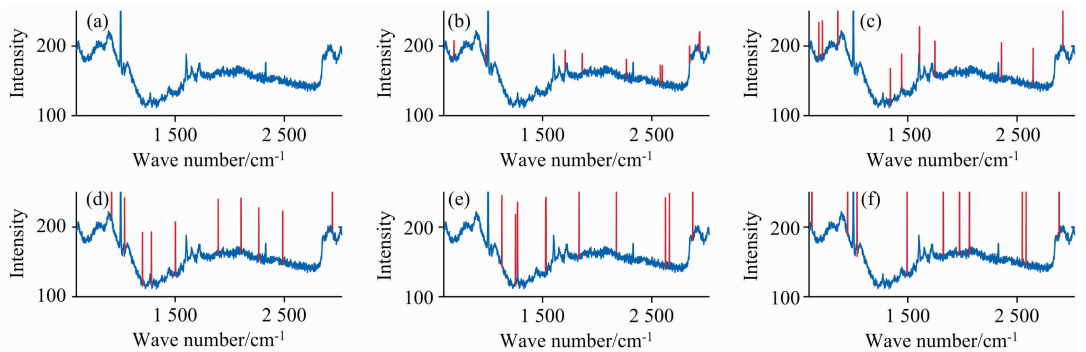

图 4　一维拉曼光谱(加人工 spike)

(a)：经过 spike 剔除后的拉曼光谱；(b)：增加 0.1 倍人工 spike 的拉曼光谱；(c)：增加 0.2 倍人工 spike 的拉曼光谱；(d)：增加 0.3 倍人工 spike 的拉曼光谱；(e)：增加 0.4 倍人工 spike 的拉曼光谱；(f)：增加 0.5 倍人工 spike 的拉曼光谱

Fig. 4　1-dimensional spectrum with artificial spikes

(a): Raman spectrum after spike removal; (b): Raman spectrum with 0.1 times artificial spikes; (c): Raman spectrum with 0.2 times artificial spikes; (d): Raman spectrum with 0.3 times artificial spikes; (e): Raman spectrum with 0.4 times artificial spikes; (f): Raman spectrum with 0.5 times artificial spikes

表 2　一维拉曼光谱 spike 检测

Table 2　1-dimensional Raman spectra spike detection

Ratios	Predicted	True	Percentage/%
0.1	175	200	87.5
0.2	195	200	97.5
0.3	197	200	98.5
0.4	199	200	99.5

实验结果所示，在 spike 强度较低情况下(峰值的比值较低)，spike 剔除算法的性能较差，因为人工 spike 可能由于其随机位置而落入信号的低处，使得叠加的信号太小，无法被检测到。当该比率增加到 40%时，检测灵敏度可达到 99.5%。参考文献[15]的检测灵敏度为 97.5%，参考文献[19]中检测超过峰值 50%的 spike 识别算法，灵敏度可以达到 99%，与它们相比，本文介绍的方法可以检测强度更弱的 spike，且灵敏度有较大提升。

2.2　二维 spike 剔除

图 5(a)和(b)为波数在 2 940 cm^{-1}的真实拉曼图像的二维 spike 剔除处理结果。其中 *Y* 轴是线数，*X* 轴是光谱条数，构成 100×400 的拉曼图像。颜色代表相对值的大小，蓝色代表小数值，黄色代表高数值。未剔除 spike 前，高强度的 spike 使图像其他位置的特征对比度降低。在剔除 spike 后，严重污染的原始拉曼图像可以显示图像的特征。

对于二维验证实验，我们从 1 340 张图像中随机选择 20 张，并在每张图像中选择 10 个随机位置设置 spike。然后将每个选定位置的值设定为原本值加上当前图像最大像素值的 0.1 到 0.3 倍作为 spike。

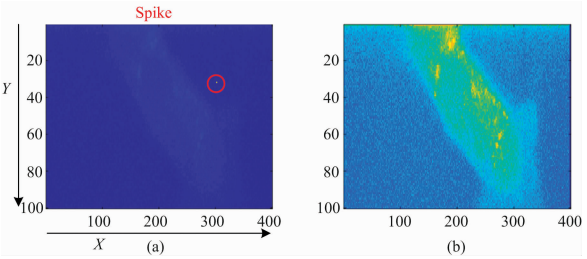

图 5 二维拉曼图像 spike 剔除

(a)：原始拉曼图像；

(b)：经过二维 spike 剔除算法处理后的拉曼图像

Fig. 5 2-dimensional Raman image spike removal process

(a)：Raw Raman image；

(b)：Raman image after 2-dimensional spike removal

通过增加不同的比率，我们得到了不同的检测结果。为减少验证实验的随机误差，进行了 10 个独立验证实验，将检测结果进行平均，最终结果如表 3 所示。

表 3 二维拉曼图像 spike 检测

Table 3 2-dimensional Raman image spike detection

Ratios	Predicted	True	Percentage/%
0.1	174	200	87
0.2	194	200	97
0.3	199	200	99.5

对于二维 spike 剔除方法，当 spike 强度超过峰值的 30%时，可以有效地检测 spike。与参考文献[19]中的二维 spike 检测算法相比，本文介绍的方法可以在 spike 强度较小的情况下检测和剔除 spike。

二维 spike 检测灵敏度高于一维，通过二维 spike 剔除方法可以去除具有较低强度的 spike。

2.3 三维 spike 剔除

三维拉曼数据立方体包含三维坐标信息和当前位置的值，它表示四维信息。在图中难以表示四维信息。因此，在使用了三维 spike 剔除方法后，我们可以简单地使用图 6 来显示三维拉曼数据立方体 spike 剔除的效果。

拉曼数据立方体可以表示为 1 340 张二维拉曼图像，或 40 000 条一维拉曼光谱。因此，用三维方法剔除 spike 后，可以直接剔除一维和二维的 spike。图 6 中的立方体是指一个包含 1 340 个拉曼图像的真实样本数据立方体。图像上的点表示某条拉曼光谱，图的下部显示该光谱的 spike 剔除结果。图中选取三个数据立方体切片作为效果说明，分别对应波数在 1 000，2 130 和 2 880 cm^{-1}时的拉曼图像。

对于三维验证实验，我们随机选择 20 个波数上的图像，并在随机位置添加 10 个 spike。spike 的强度是拉曼数据立方体最高值的 0.1～0.2 倍。计算结果如表 4 所示。

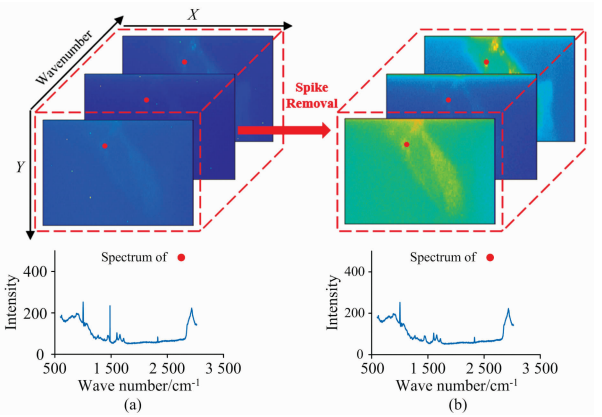

图 6 三维拉曼数据立方体 spike 剔除

(a)：原始拉曼数据立方体；

(b)：经过三维 spike 剔除算法处理后的拉曼数据立方体

Fig. 6 3-dimensional Raman data cube spike removal process

(a)：Raw Raman data cube；

(b)：Raman data cube after 3-dimensional spike removal

表 4 三维拉曼数据立方体 spike 检测

Table 4 3-dimensional Raman data cube spike detection

Ratios	Predicted	True	Percentage/%
0.1	186	200	93
0.2	199	200	99.5

当强度大于峰值的 20%时，可剔除所有 spike。与一维和二维 spike 剔除相比，三维方法可以检测更低强度的 spike，提高数据处理能力。这是一种新的针对大量拉曼数据的 spike 剔除方法，在其他文献中很少被提及。

灵敏度随着维度的增加而增加有两个原因。一方面，数据的特征更加明显。另一方面，维数的增加使数据量迅速增加，数据拟合出的分布更准确，spike 剔除算法更灵敏。

3 结 论

随机的窄带宽、高强度的 spike 容易附着在拉曼光谱上导致信号的失真，对后续光谱分析造成困难。本文提出一种 spike 剔除算法，可以根据原始拉曼数据和中值滤波后数据之间的偏差拟合高斯分布并设置概率阈值来剔除 spike。该方法不仅适用于一维拉曼光谱，而且也适用于二维拉曼图像和三维拉曼数据立方体，检测灵敏度将随着维度的增加而提高。当 spike 比峰值高 20%时，该算法可以在保留原始信号的真实性的基础上达到 99.5%的灵敏度。该算法阈值设置简单，适用不同维度的拉曼数据，能有效剔除低强度的 spike，有望成为拉曼光谱分析人员重要的预处理工具。

References

[1] YAN Jing-wen, LI Ying, CHEN Liang, et al(闫静文，李 颖，陈 靓，等). Journal of Atmospheric and Environmental Optics(大气与环境光学学报). 2017, 12(2): 128.

[2] REN Bin(任 斌). Optics & Optoelectronic Technology(光学与光电技术), 2021, 19(4): 1.

[3] LIANG Ying-fang, ZHOU Hua-lan, WANG Yan, et al(梁营芳，周化岚，王　燕，等). Physical Testing and Chemical Analysis Part B: Chemical Analysis(理化检验：化学分册), 2020, 56(4): 487.
[4] TIAN Hui-yan, LIU Yu, HUANG Jiao-qi, et al(田晖艳，刘　羽，黄姣祺，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(10): 3021.
[5] LIU Xia, HUO Ya-peng, KANG Wei-jun, et al(刘　厦，霍亚鹏，康维钧，等). Chinese Science Bulletin(科学通报), 2020, 65(15): 1448.
[6] Schulze H G, Turner R F. Applied Spectroscopy, 2013, 67(4): 457.
[7] Tian Y, Burch K S. Applied Spectroscopy, 2016, 70(11): 1861.
[8] Zhao J. Applied Spectroscopy, 2003, 57(11): 1368.
[9] Huang Z, Zeng H, Hamzavi I, et al. Optics Letters, 2001, 26(22): 1782.
[10] Lee J S, Cox D D. Computational Statistics & Data Analysis, 2010, 54(12): 3131.
[11] Katsumoto Y, Ozaki Y. Applied Spectroscopy, 2003, 57(3): 317.
[12] Silveira Jr L, Bodanese B, Zangaro R A, et al. Instrumentation Science Technology, 2010, 38(4): 268.
[13] Veneri G, Federighi P, Rosini F, et al. Journal of Neuroscience Methods, 2011, 196(2): 318.
[14] Zhang D, Ben-Amotz D. Applied Spectroscopy, 2002, 56(1): 91.
[15] Takeuchi H, Hashimoto S, Harada I. Applied Spectroscopy, 1993, 47(1): 129.
[16] Uckert K, Bhartia R, Michel J. Applied Spectroscopy, 2019, 73(9): 1019.
[17] Sharan T S, Sharma S, Sharma N. Applied Spectroscopy, 2021, 88(1): 117.
[18] LI Sheng, DAI Lian-kui(李　晟，戴连奎). The Journal of Light Scattering(光散射学报), 2011, 23(3): 188.
[19] Ryabchykov O, Bocklitz T, Ramoji A, et al. Chemomtrics and Intelligent Laboratory Systems, 2016, 155: 1.

A Spike Removal Algorithm Based on Median Filter and Statistic for Raman Spectra

YE Rui-qian[1], HE Hao[1], ZHENG Peng[1], XU Meng-xi[2], WANG Lei[1]*
1. School of Aerospace Engineering, Xiamen University, Xiamen　361101, China
2. State Key Laboratory of Physical Chemistry of Solid Surfaces, Collaborative Innovation Center of Chemistry for Energy Materials (iChEM), College of Chemistry and Chemical Engineering, Xiamen University, Xiamen　361101, China

Abstract　Raman spectroscopy is a promising technique widely used in chemistry, biology, and physics. However, as the key part of the Raman spectrometer, the charge-coupled device is vulnerable to cosmic rays, resulting in a random narrow bandwidth and a high-intensity spikes. It will cause a significant reduction in signal contrast. In this paper, we propose a practical spike removal algorithm. Firstly, the algorithm obtains deviation data by separating the median filtered data from the original data. Then, deviation data is sorted from small to large by quantile method, and the intermediate 99% data are selected for Gaussian distribution fitting. Considering the characteristics of high-intensity and sparsity of the spike, the occurrence probability of high intensity data in the spectra is used as the threshold standard to remove spike. Finally, the spikes are replaced by new data using median filtered at corresponding positions. This algorithm restores the original sample information without any debugging parameters. Different intensities of spikes are added in Raman spectra to verify the algorithm, and the experimental results show that this algorithm's sensitivity can reach 99.5%. Besides, this algorithm is applicable for one-dimensional Raman spectra, two-dimensional Raman images and three-dimensional Raman data cubes, and the performance improves with the increase of dimensionality. Specifically, the one-dimensional spike removal algorithm can detect spikes exceeding 40% of the maximum peak intensity. The Raman data cubes can be detected exceeding 20% of the peak value. The algorithm is used to process 40 000 real Raman spectra and can effectively remove spikes without distorting the real signal, proving the algorithm's practicability.

Keywords　Raman spectroscopy; Cosmic rays; Median filter; Spike removal

(Received Aug. 19, 2021; accepted Feb. 25, 2022)

* Corresponding author