

基于样本优化和主成分分析的多通道拉曼光谱重建及其快速成像

范贤光^{1,2}, 刘 龙¹, 支瑜亮¹, 康哲铭¹, 夏 宏¹, 张佳杰¹, 王 昕^{1,2*}

1. 厦门大学航空航天学院仪器与电气系, 福建 厦门 361005
2. 传感技术福建省高等学校重点实验室, 福建 厦门 361005

摘 要 拉曼成像是一种无损伤、无需标记的光谱成像技术,在生物医学领域得到了广泛的应用。然而,由于大多数生物样本中的自发拉曼信号都很弱,为了获得较好的成像结果,需要较长的时间来获取高信噪比的拉曼光谱,严重影响了拉曼成像的时空分辨率,阻碍了其在快速动态体系中的应用。多通道拉曼成像是解决这一问题的有效途径之一,在多通道拉曼成像技术中,完整拉曼光谱的标定-重建算法是关键。目前,适用于光谱重建的算法有伪逆法、Wiener估计算法等,这些方法虽然简单且易于实现,但是在应用于多通道拉曼成像时,一方面易受噪声、振动等非线性因素的直接干扰,另一方面在多通道拉曼成像中,数量相对较少的训练样本和坏样本的存在均很容易影响重建效果。为解决这两类因素的影响,本文提出了一种基于训练样本优化和主成分分析(PCA)的拉曼光谱重建算法。首先,利用滤光片理论响应矩阵函数计算训练样本的模拟窄带测量值,借助 Wiener估计重建完整拉曼光谱,得到重建光谱的模拟窄带测量值,比较样本与重建光谱的窄带测量值,完成训练样本的优化;然后,基于多项式回归,拓展优化处理后的窄带测量值,降低非线性因素的干扰;最后,利用主成分分析,提取训练样本主要信息,完成转移矩阵的计算,并引入归一化处理,实现拉曼光谱的快速重建。在试验中,选取有机玻璃(PMMA)作为实验样本,利用伪逆法、Wiener估计算法和本算法,分别完成拉曼光谱重建。采用均方根误差,评价拉曼光谱的重建精度。结果证明,该算法优于传统算法,为拉曼成像技术进一步在快速动态体系中的应用提供了理论支持。

关键词 多通道拉曼成像;训练样本优化;PCA;光谱重建

中图分类号: O657.37 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2020)08-2495-05

引 言

拉曼光谱作为一种无损快速检测技术,在材料、化工、生物医学、环保等领域已有着广泛的应用^[1-3]。然而,大多数生物、药材等样本中的自发拉曼信号都较弱且易受荧光背景和噪声的干扰,因此采谱时间往往较长,导致常规拉曼成像的动态性能较差,严重阻碍了拉曼光谱成像技术在快速动态体系中的应用。

目前,常用的拉曼成像技术有两大类^[4-5]:自发拉曼成像和非线性拉曼成像。自发拉曼成像主要包括扫描模式和宽场模式。扫描模式可获得完整的三个维度信息,但是扫描时间过长,难以观察快速变化的动态过程;宽场模式可获取更好的动态性能,但是只能收集两个空间维度的单波长光谱。非线性拉曼成像主要包括表面增强拉曼散射成像和针尖增强

拉曼光谱成像。这两种成像技术具备高空间分辨率和快速成像的优势,但是其对样品有一定的损伤且仪器结构复杂、造价昂贵,一般由实验室自行搭建,用于前沿科学研究,商品化仪器相对较少,推广难度大。

因此,开发低成本且快速成像的技术至关重要。多通道拉曼成像技术^[6]是解决这一问题的有效途径之一,其基本原理如图1所示。硬件上,通过宽场模式激发样品,将散射光同时经过多个不同参数的滤光通道,耦合于面阵CCD的不同感光区域,从而得到多通道窄带数据;软件算法上,基于事先标定好的转移矩阵(不同样本的转移矩阵不同)将多通道数据重建为完整的拉曼光谱,从而实现拉曼成像。

由图1可知,多通道拉曼光谱成像的核心是拉曼光谱标定-重建算法。目前,常用光谱重建算法有伪逆法^[7-8]、Wiener估计算法^[9]等。现有的这些算法来源于光谱反射率重建领域,并非专门的多通道拉曼重建算法,要么没有考虑非线性

收稿日期: 2019-07-06, 修订日期: 2019-11-20

基金项目: 国家自然科学基金项目(21874113), 国家重大科研仪器研制项目(21627811)资助

作者简介: 范贤光, 1980年生, 厦门大学航空航天学院仪器与电气系教授 e-mail: fangxg@xmu.edu.cn

* 通讯联系人 e-mail: xinwang@xmu.edu.cn

性因素的影响,要么没有对训练样本进行优化处理且计算量过大。所以,本文提出一种专用的多通道拉曼光谱重建算法,首先对训练样本进行优化处理,排除坏样本的干扰;然后对得到的多通道窄带测量值进行多项式回归拓展,降低非线性因素的影响;最后引入归一化处理,并依据主成分分析^[10-11](principal component analysis, PCA)简化计算,完成拉曼光谱快速重建。同时,在仿真与实验中对伪逆法、Wiener 估计算法等进行了验证,并与本文算法进行了比较。

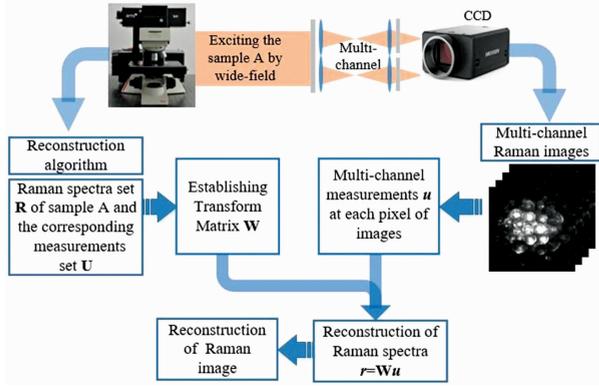


图 1 多通道拉曼成像原理图

Fig. 1 The theory of multi-channel Raman spectroscopy imaging

1 算法

1.1 光谱重建问题阐述

假设待测样本的拉曼光谱为向量 r_1 ($N \times 1$ 维), N 取决于拉曼光路的分光光栅参数和线阵 CCD 的像素点个数。通过滤光片组产生相应窄带测量值由 u_1 表示,二者之间的关系如式(1)

$$u_1 = Tr_1 + n \quad (1)$$

其中, n 表示 ($M \times 1$ 维) 窄带测量中的随机噪声, T ($M \times N$ 维, 其中 M 为滤光片数量) 表示滤光片组的光谱响应函数矩阵, u_1 是 $M \times 1$ 向量。

快速重建拉曼光谱过程就是建立 r_1 和 u_1 之间的逆向数学模型,即转移矩阵 W ($N \times M$ 维)。通过多通道拉曼成像系统,获得实测窄带测量值 u , 进而重建完整拉曼光谱 r , 如式(2)

$$r = Wu \quad (2)$$

1.2 基于 PCA 的拉曼光谱重建

针对图 1 中测试样品 A, 在开始建立重建算法前, 借助常规拉曼光谱仪完成对样品 A 的多次测量, 获取训练样本集 $R = \{r_1, r_2, \dots, r_i, \dots, r_k\}$ ($N \times k$ 维, 其中 k 为样本数量)。根据式(1), 不计随机噪声, 完成窄带通滤光片从整个拉曼光谱中的特征峰波段内选择性地检测拉曼信号, 得样本模拟窄带测量值集 $U = \{u_1, u_2, \dots, u_i, \dots, u_k\}$ ($M \times k$ 维)。至此, 求解转移矩阵 W 的准备工作已完成, 具体求解如下。

1.2.1 预处理阶段

依据 Wiener 估计思想, 待测样品 A 的训练样本 R 数目

越多, 重建完整拉曼光谱的精度会相应提高。但是, 训练样本增多, 不可避免的会有差异过大的样本存在, 故将样本 R 进行优化处理, 具体操作如下:

(1) 实测训练样本 R 带有荧光背景, 做去基线处理, 得训练样本集 $\tilde{R} = \{\tilde{r}_1, \tilde{r}_2, \dots, \tilde{r}_i, \dots, \tilde{r}_k\}$ 。借助式(1), 得到去荧光背景的窄带测量值 $\tilde{U} = \{\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_i, \dots, \tilde{u}_k\}$ 。

(2) 根据 Wiener 估计, 由 \tilde{U} 中任一元素 \tilde{u}_i , 快速重建出拉曼光谱, 并归一化处理, 得重建光谱 \tilde{r}'_i , 代入式(1), 求得重建光谱的窄带测量值 \tilde{u}'_i ;

(3) 计算 \tilde{u}_i 和 \tilde{u}'_i 中对应元素的比值 λ (近似为 1), 并设定阈值 $\lambda_0 \leq \lambda \leq \lambda_1$ 。将不在阈值内的样本直接剔除, 实现优化训练样本 R 的目的, 得 \bar{R} ; 同时优化与之相对应的模拟窄带测量值集 \bar{U} 。两者之间的关系为

$$\bar{R} = W\bar{U} \quad (3)$$

实测多通道窄带拉曼图像, 窄带测量值会受到各种非线性因素(诸如, 噪声/振动)影响。为降低干扰, 引入多项式回归^[12]。取 \bar{U} 中任一元素, 记作 $\bar{u}_i = \{u_{i1}, u_{i2}, u_{i3}, u_{i4}\}$ (4 为滤光片数目), 对 \bar{u}_i 作二阶多项式回归, 得

$$\bar{u}_i = \{1, u_{i1}, u_{i2}, \dots, u_{i4}, u_{i1}u_{i1}, u_{i1}u_{i2}, \dots, u_{i1}u_{i4}, u_{i2}u_{i2}, \dots, u_{i2}u_{i4}, \dots, u_{i3}u_{i4}, u_{i4}u_{i4}\} \quad (4)$$

从而, 将模拟窄带测量值集合 \bar{U} 拓展 \hat{U} 。

1.2.2 校准阶段

训练样本集 R 经上述优化后, 样本容量相对较大, 在简化计算的同时尽可能保留训练样本的有用信息, 采用 PCA 对优化后的样本 \bar{R} 进行数据压缩。首先, 求解其主要成分 V

$$\bar{R} = VA \quad (5)$$

其中

$$A = V^+ \bar{R}, V^+ = (V^T V)^{-1} V^T \quad (6)$$

式(6)中, 上标“ T ”表示矩阵的转置, 上标“ -1 ”表示方阵的逆矩阵。

联立式(3), 式(5)和式(6), 得

$$W = VA(\hat{U})^+ \quad (7)$$

1.2.3 测试阶段

根据校准阶段求得转移矩阵 W , 联立式(2), 完整由窄带测量值快速重建完整拉曼光谱, 整体流程如图 2 所示。

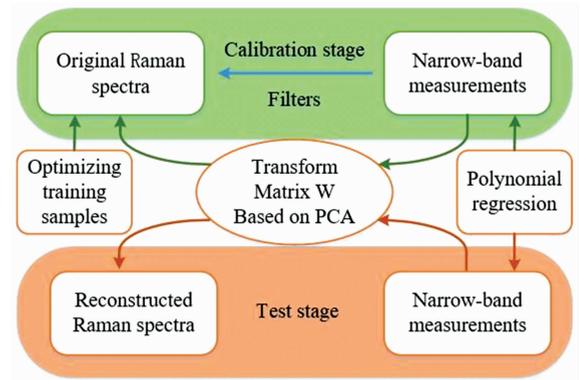


图 2 基于 PCA 的拉曼光谱重建流程图
Fig. 2 The process of Raman spectrum reconstruction based on PCA

2 实验部分

本文以有机玻璃 (polymethyl methacrylate, PMMA) 作为实验样品, 借助常规拉曼光谱仪 (532 nm 激光光源), 完成对样品的多次测量, 获取训练样本集 $\mathbf{R} = \{r_1, r_2, \dots, r_i, \dots, r_k\}$, 其中 k 代表测量次数。原始拉曼光谱带有荧光背景, 为训练样本优化做准备, 对其作荧光去除。本文采用 B 样条去除荧光背景, 处理结果如图 3 所示。

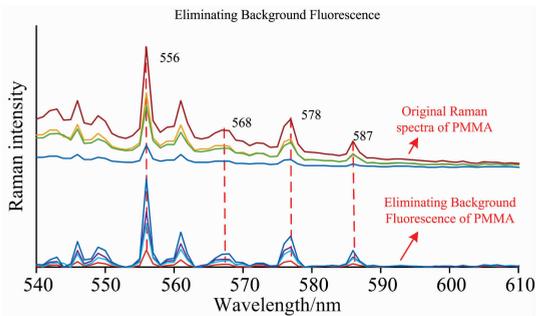


图 3 PMMA 原始和去荧光背景拉曼光谱
Fig. 3 Original spectra and eliminating background fluorescence of PMMA

由图 3 可知, PMMA 特征谱峰波长集中在 540~600 nm 范围, 依据确定滤光片的两个经验法则^[13], 选取四款滤光片 (索雷博公司, 中心波长分别为 560, 570, 580 和 590 nm, 如图 4 所示) 的光谱响应函数用以计算窄带测量值。根据式 (1), 求得样本窄带测量值集 $\mathbf{U} = \{u_1, u_2, \dots, u_i, \dots, u_k\}$, 经优化处理得 $\hat{\mathbf{U}}$ 。

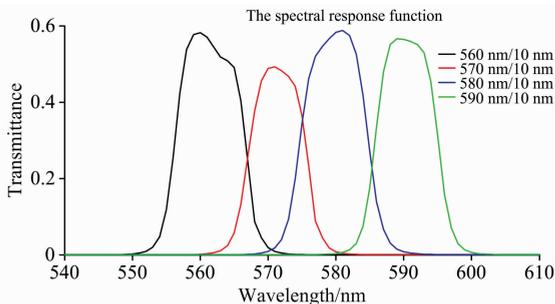


图 4 四款滤光片光谱响应函数
Fig. 4 The spectral response functions of four filters

取优化后的窄带测量值集 $\hat{\mathbf{U}}$ 中任一元素 \hat{U}_i , 联立式 (2) 和式 (7) 快速重建出 PMMA 的完整拉曼光谱 \hat{r}_i , 经归一化处理后与实际光谱 r_i 比较。同时, 为比较本文重建算法的优劣, 分别采用伪逆法, Wiener 估计算法完成光谱重建, 重建结果如图 5 所示。

由图 5 可知, 由于仿真中没有噪声、荧光等干扰因素, 本文算法和伪逆法, Wiener 算法的重建光谱 \hat{r}_i 与实际光谱 r_i 之间的吻合度均较好。这里, 采用 RMSE 评价拉曼光谱的重建精度

$$RMSE = ((\hat{r}_i - r_i)^T (\hat{r}_i - r_i) / N)^{1/2} \quad (8)$$

基于训练样本优化的 PCA 求得重建光谱 \hat{r}_i 与样本 \mathbf{R} 的平均 RMSE 为 0.04, 本文算法的重建效果比其他两种方法略有提高。具体重建 RMSE 如表 1 所示, 表中 r_1, r_2, r_3 代表随机选择的部分训练样本。

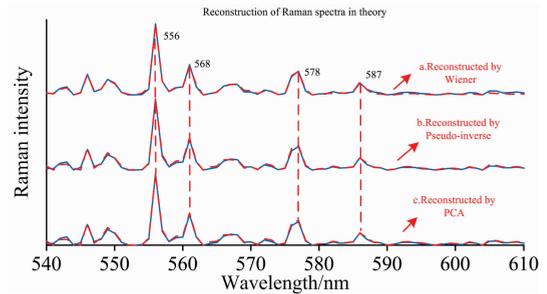


图 5 三种方法对 PMMA 拉曼光谱的仿真重建效果对比
Fig. 5 Comparison among spectra reconstruction of PMMA based on three algorithms in theory

表 1 三种算法重建光谱与训练样本的 RMSE

| Methods | r_1 | r_2 | r_3 | Average |
|----------------|---------|---------|---------|---------|
| Wiener | 0.107 5 | 0.049 8 | 0.067 7 | 0.075 0 |
| Pseudo-inverse | 0.112 4 | 0.055 8 | 0.072 6 | 0.080 3 |
| PCA | 0.066 5 | 0.048 9 | 0.004 5 | 0.040 0 |

现为验证本文提出的拉曼光谱重建算法, 基于实验室搭建的多通道拉曼窄带图像采集系统, 获得 PMMA 的四组滤光片对应的拉曼窄带图像, 如图 6 所示。

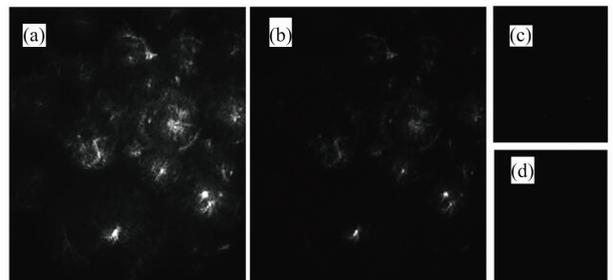


图 6 PMMA 多通道窄带图像

(a): 560/10 nm 滤光片测得窄带图像; (b): 570/10 nm 滤光片测得窄带图像; (c): 580/10 nm 滤光片测得窄带图像; (d): 590/10 nm 滤光片测得窄带图像

Fig. 6 The multi-channel narrow-band image of PMMA

(a): Narrow-band image taken by the filter 560/10 nm; (b): Narrow-band image taken by the filter 570/10 nm; (c): Narrow-band image taken by the filter 580/10 nm; (d): Narrow-band image taken by the filter 590/10 nm

由图 6 分析可知, 鉴于 PMMA 的最高特征峰主要集中在波长 555~565 nm 之间 (如图 3 可知), 所以 560/10 nm 滤光片测得窄带图像最亮, 其他顺次变暗。由于测试样本是 PMMA 固体粉末颗粒, 颗粒密度分布不均, 所以得到的窄带图像呈现出凹凸感。

根据图 6, 求得任一像素点对应的四通道窄带测量值 u 。联立式(2)和式(7), 基于训练样本优化的 PCA 算法, 即可实现拉曼光谱的快速重建。同时, 依据伪逆法, Wiener 估计算法, 也完成拉曼光谱重建, 如图 7 所示。

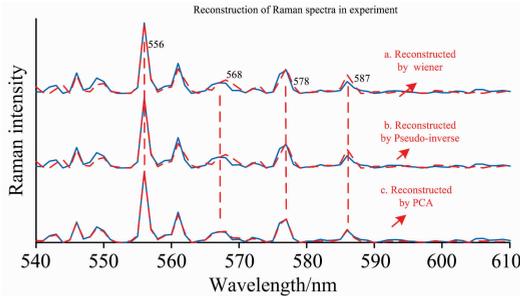


图 7 三种方法对 PMMA 拉曼光谱的实验重建效果对比
Fig. 7 Comparison among spectra reconstruction of PMMA based on three algorithms in experiment

首先, 定性分析三条重建光谱: 三种算法均完成了 PMMA 拉曼光谱的重建, 但是伪逆法和 Wiener 估计算法在完成光谱重建时, 重建光谱在波长 568 nm 的峰值出现偏离, 在波长 587 nm 的峰值拔高。而本文提出的样本优化的 PCA 重建算法, 不会出现上述问题。

然后, 定量分析三条重建光谱: 基于式(8), 分别求解出三条重建光谱的 RMSE, 如表 2 所示。由表 2 可知, 训练样本优化的 PCA 算法的 RMSE 最小。对比表 1 可知, 实验重建拉曼光谱的 RMSE, 比理论 RMSE 偏大, 其原因应为实测多通道窄带数据中含有杂散光和噪声的干扰。

表 2 实验重建光谱与训练样本的 RMSE

Table 2 RMSE between reconstructed spectra and training samples in experiment

| Algorithm | Wiener | Pseudo-inverse | PCA |
|-----------|---------|----------------|---------|
| RMSE | 0.076 7 | 0.100 8 | 0.055 3 |

最后, 根据本文重建算法, 完成对图 6 多通道拉曼窄带图像某一指定区域重建拉曼图像, 具体如图 8(a)所示。该区域包含 $131 \times 131 (1.43 \times 1.43 \text{ mm}^2)$ 个像素点。重建该区域拉曼图像的步骤, 具体如下:

- (1) 提取该区域内 131×131 个像素点的对应的多通道窄带响应值;
- (2) 根据本文提出的重建算法, 联立式(2)和式(7), 得到 131×131 条完整的重建拉曼光谱;
- (3) 对上述重建光谱在波长 560/10 nm 内, 执行积分, 得到一个 131×131 的矩阵, 即为重建后的拉曼图像, 如图 8(d)所示。

对比验证本文算法重建拉曼图像的质量, 依据伪逆法和

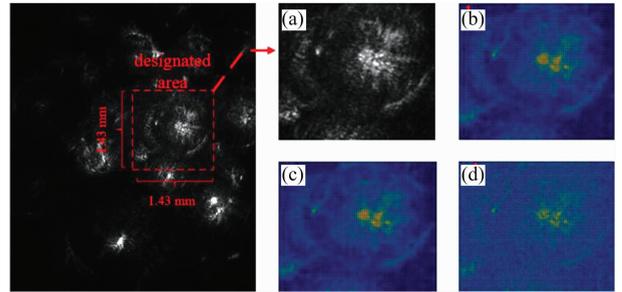


图 8 PMMA 多通道窄带图像重建拉曼图像

- (a): 待重建拉曼图像的区域; (b): 伪逆法重建拉曼图像; (c): Wiener 估计法重建拉曼图像; (d): PCA 重建拉曼图像

Fig. 8 The reconstruction images of PMMA based on multi-channel narrow-band measurements

(a): The area of Raman image to be reconstructed; (b): The reconstruction of Raman image based on pseudo-revers; (c): The reconstruction of Raman image based on Wiener estimation; (d): The reconstruction of Raman image based on PCA

Wiener 估计算法重建拉曼图像, 如图 8(b)和 8(c)所示。实验中, 图 8(a)的中心位置 PMMA 的浓度仅比周围略高, 图 8(b)和(c)在该位置明显过亮, 其原因在于激发激光在该处的干扰。由图 8 分析可知, 基于样本优化的 PCA 重建拉曼图像图 8(d)中间亮度相对较弱, 在一定程度上克服了激光干扰, 更好地展现了 PMMA 的拉曼图像。同时, 本文提出的重建算法, 实现 131×131 个像素点重建拉曼图像所需的时间远低于常规的点扫描和线扫描成像模式(如表 3 所示)。

表 3 不同拉曼成像形式所需成像时间

Table 3 The time of Raman imaging among different imaging formats

| Imaging mode | Point scanning | Line scanning | Multi-channel imaging (exposure time+calculation time) |
|--------------|----------------|---------------|--|
| Time/s | 858.05 | 6.55 | 0.05+0.58 |

3 结 论

基于样本优化和主成分分析, 提出了一种快速拉曼光谱重建方法。该方法首先通过优化训练样本, 剔除差样本干扰; 然后, 引入多项式回归, 降低非线性因素影响; 最后利用 PCA, 经归一化处理, 完成拉曼光谱重建。仿真和实验结果表明, 本文算法比传统算法更好地完成了光谱重建, 不会出现重建光谱波峰的偏移和突变。同时, 借助自行开发的多通道拉曼成像仪器, 利用本文算法完成了 PMMA 的拉曼光谱重建和成像。因此, 本文的研究成果, 有力地完善了多通道拉曼成像技术, 进而推进了多通道拉曼成像技术在快速动态体系中的应用。

References

- [1] Christoph Krafft, Larysa Shapoval, Stephan B. Sobottka, et al. Biochimica et Biophysica Acta-Biomembranes, 2006, 1758(7): 883.

- [2] Christoph Krafft, Gerald Steiner, Claudia Beleites, et al. *Journal of Biophotonics*, 2009, 2: 13.
- [3] Popp J, Krafft C, Mayerhofer T. *Optik & Photonik*, 2011, 6(4): 24.
- [4] Shona Stewart, Ryan J Priore, Matthew P Nelson, et al. *Annual Review of Analytical Chemistry*, 2012, 5: 337.
- [5] Zheng Xiaoshan, Zong Cheng, Xu Mengxi, et al. *Small*, 2015, 11: 3395.
- [6] Wei D, Chen S, Ong Y H, et al. *Optics Letters*, 2016, 41(12): 2783.
- [7] Pan Yujian, Luo Guoqing, Jin Huayan, et al. *Signal Processing*, 2018, 153: 47.
- [8] Zhang X L, Shao L Y, Yu Z W, et al. *Optics Communications*, 2004, 239: 79.
- [9] Katrasnik J, Pernus F, Likar B. *Applied Spectroscopy*, 2010, 64(11): 1265.
- [10] Koch H, Polepil S, Eisen K, et al. *Physical Chemistry Chemical Physics*, 2017, 19(45): 30533.
- [11] Halstead J E, Smith J A, Carter E A, et al. *Environmental Pollution*, 2018, 234: 552.
- [12] WENG Shi-zhuang, YUAN Bao-hong, ZHENG Shou-guo, et al(翁士状, 袁宝红, 郑守国, 等). *Spectroscopy and Spectral Analysis(光谱学与光谱分析)*, 2018, 38(2): 454.
- [13] Chen S, Ong Y H, Liu Q. *J. Raman Spectrosc.*, 2013, 44: 875.

Fast Reconstruction for Multi-Channel Raman Imaging Based on and Sample Optimization and PCA

FAN Xian-guang^{1, 2}, LIU Long¹, ZHI Yu-liang¹, KANG Zhe-ming¹, XIA Hong¹, ZHANG Jia-jie¹, WANG Xin^{1, 2*}

1. Department of Instrumental and Electrical Engineering, School of Aerospace Engineering, Xiamen University, Xiamen 361005, China
2. Fujian Key Laboratory of Universities and Colleges for Transducer Technology, Xiamen 361005, China

Abstract Raman imaging is a noninvasive, label-free spectral imaging technique that has been widely used in the biomedical field. However, the spontaneous Raman signals of most biological samples are weak. It takes a long time to obtain an image with a high signal-to-noise ratio, which seriously affects the spatial and temporal resolution of Raman imaging and hinders its application in fast dynamic systems. The multi-channel Raman imaging is one of the effective ways to solve this problem. The reconstruction of the full Raman spectrum is the key in this system, and the corresponding algorithm of reconstruction is needed to be developed. At present, the algorithms capable for spectral reconstruction are pseudo-inverse and Wiener estimation. Although these methods are simple and easy to be carried out, they are susceptible to nonlinear factors such as noise and vibration when applied to the system of multi-channel. On the other hand, the numbers of the training sample are relatively small, and the bad sample affects the reconstruction in the system of multi-channel. In order to solve those problems, we propose an algorithm based on sample optimization and principal component analysis (PCA). Firstly, the simulated narrow-band measurements of the training samples are calculated by using the spectral response function of the filter, and the full Raman spectra are reconstructed by Wiener estimation, and then the simulated narrow-band measurements of the reconstructed spectra are achieved. The sample gets Optimized by comparing the simulated narrow-band measurements of the sample and the reconstructed spectra. Second, the interference of nonlinear factors is reduced by introducing the polynomial regression and expanding the optimized narrow-band measurements. At last, the main information of training samples is extracted, and the calculation is reduced by using PCA, and the transform matrix is completed. At the same time, the normalization is introduced to realize the reconstruction of Raman spectra. In the experiment, the polymethyl methacrylate is selected as the experimental sample, and the Raman spectrum is reconstructed by pseudo-inverse, Wiener estimation and our algorithm. The root means square error is used to evaluate the accuracy of the reconstructed spectra. The result proves that our algorithm is significant. It provides theoretical support for the further application of Raman imaging technology in fast dynamic systems.

Keywords Multi-channel image; Sample optimization; PCA; Spectral reconstruction

* Corresponding author

(Received Jul. 6, 2019; accepted Nov. 20, 2019)