

氧气 A 吸收带偏最小二乘基线拟合方法

李晋华¹, 张敏娟¹, 张德龙²

¹中北大学山西省光电信息与仪器工程技术研究中心, 山西 太原 030051;

²中国科学院长春光学精密机械与物理研究所应用光学国家重点实验室, 吉林 长春 130039

摘要 为了研究被测目标在氧气 A 带的自发射光谱与透射率的关系, 根据透射率的计算方法, 提出了采用偏最小二乘回归法对氧气 A 带的光谱进行基线拟合, 搭建了实验系统, 验证了该方法拟合基线的准确度。首先, 以黑体辐射理论为依据, 给出氧气 A 带平均透射率计算方法, 以实际被测目标光谱为研究对象, 以带外数据为依据, 利用偏最小二乘法拟合被测目标在氧气 A 带的基线。为提高拟合精度, 剔除了测量奇异点, 并利用基线拟合不确定度来评价偏最小二乘回归法拟合基线的准确度。为了验证该方法的准确性, 以卤素灯为光源, 在 0~130 m 范围内, 获得不同距离处的光谱曲线, 以及相同距离不同分辨率下的光谱曲线, 将各种曲线分别进行基线拟合, 分析各自的标准偏差。结果表明, 同一距离处不同分辨率下的平均标准偏差为 0.23%, 随着分辨率的降低, 基线拟合不确定度变小, 信噪比增大; 基线拟合不确定度还与测试设备的分辨率有关, 分辨率越高, 带外信息基线拟合不确定度越大, 反之, 带外基线拟合不确定度越小。

关键词 光谱学; 氧气 A 带; 基线拟合; 偏最小二乘回归; 拟合不确定度

中图分类号 O433.4

文献标识码 A

doi: 10.3788/CJL201744.0811001

Baseline Fitting of Partial Least Squares for Oxygen A Absorption Band

Li Jinhua¹, Zhang Minjuan¹, Zhang Delong²

¹Engineering and Technology Research Center of Shanxi Provincial for Optical-Electric Information and Instrument, North University of China, Taiyuan, Shanxi 030051, China;

²State Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of sciences, Changchun, Jilin 130039, China

Abstract In order to research the relation between self-emission spectrum and transmittance of target at oxygen A-band, the method for fitting the baseline of oxygen A-band spectrum based on the partial least square regression (PLSR) is proposed according to the transmittance calculation method. The experimental system is set up to verify the accuracy of baseline fitting. Firstly, the calculation method of oxygen A-band average transmittance is given according to the blackbody radiation theory. The spectrum of the target to be measured is used as study object. The PLSR is used to fit the baseline of the target to be measured at oxygen A-band according to the out-of-band data. For the sake of improving the fitting accuracy, the singular point of measurement is eliminated. The baseline fitting uncertainty is used to evaluate the fitting accuracy of PLSR. In order to verify the accuracy of this method, the spectra with different distances and different resolutions at the same distance are measured in the range of 0~130 m using the halogen light. These spectra are fitted and each standard deviation is analyzed. The results show that the average standard deviation is 0.23% with different resolutions at the same distance. The baseline fitting uncertainty becomes smaller and signal-to-noise ratio increases with the resolution reducing. In addition, the baseline fitting uncertainty is related to the resolution of the test equipment. The higher the resolution is, the larger the uncertainty of out-of-band data baseline fitting becomes, whereas, the smaller the out-of-band baseline fitting becomes.

Key words spectroscopy; oxygen A-band; baseline fitting; partial least squares regression; fitting uncertainty

OCIS codes 330.6170; 010.1285; 010.1300; 010.1320

收稿日期: 2017-02-17; **收到修改稿日期:** 2017-04-10

基金项目: 国际科技合作项目(2013DFR10150)、国家自然科学基金仪器专项基金(61127015)、国家自然科学基金青年科学基金项目(61505180)

作者简介: 李晋华(1977—), 女, 博士, 讲师, 主要从事信号获取与处理方面的研究。E-mail: ljh4961@163.com

1 引 言

氧气 A 带的基线拟合是依据黑体辐射理论,利用红外目标辐射光谱经大气衰减后测得的吸收光谱拟合出无衰减时的光谱曲线,即利用吸收带外的光谱数据拟合基线^[1-2]。测量基线的拟合精度直接影响到测距的精度,利用带外光谱拟合基线主要考虑其是否为最佳拟合^[3]。目前,关于光谱的基线拟合方法有插值法^[4]、多项式拟合^[5]、分段线性拟合方法^[6]等,针对氧分子 A 吸收带内光谱波动大的情况,带外数据的基线拟合都存在误差平方过大的问题,这将影响光谱测量的精度,进而影响到光谱透射率的精度。而回归分析法根据测量数据分析变量之间的相互关系,更适用于氧分子 A 带基线的拟合。目前常用的回归拟合方法有多元线性回归(MLR)、主成分回归(PCR)、最小二乘回归和偏最小二乘回归(PLSR)等方法。最常用的是最小二乘回归方法^[7],但最小二乘回归法的不稳健性会导致基线拟合偏差较大。使用 MLR 的前提是自变量必须相互独立,否则会出现多重共线性的问题,使得回归系数不稳定,导致计算误差变大^[8]。PCR 方法运算速度比 MLR 慢,且 PCR 仅考虑了自变量的信息,没考虑因变量的解释作用,会造成信息的丢失^[9]。而 PLSR 既考虑了自变量的信息,又考虑了因变量对自变量的解释作用,稳定性强^[10]。在相同环境,光谱仪分辨率为 8 cm^{-1} 时,多次测量分别采用上述方法进行基线拟合,利用上述不同方法获取的拟合基线标准偏差分别为:最小二乘回归为 0.32% ,MLR 法为 0.51% ,PCR 法为 0.42% ,而 PLSR 法为 0.23% 。综上所述,采用偏最小二乘曲线拟合方法拟合氧分子 A 吸收带的基线,使数据点均匀分布在基线两侧,不仅考虑所计算的带外数据与拟合基线方差为最大,还考虑了需拟合的基线与带外数据相关性最大,有效区分系统信息和噪声,提高系统建模的准确性。

2 目标光谱的带平均透射率计算模型

氧分子光谱区间 $\Delta\omega$ 间隔内的带平均透射率^[11] $\overline{T_\omega}$ 定义为:

$$\overline{T_\omega} = \frac{I_m(\omega)}{I_0(\omega)} = \frac{1}{\Delta\omega} \int T(\omega) d\omega = \frac{1}{\Delta\omega} \int_{\omega_i}^{\omega_f} \exp[-\tau(\omega)] d\omega, \quad (1)$$

式中 $I_m(\omega)$ 是经衰减后测得的光谱强度, $I_0(\omega)$ 为光源辐射强度, $T(\omega)$ 为单谱线透射率, $\Delta\omega$ 是谱带宽度, $\tau(\omega)$ 为光学厚度。

如图 1 所示为 1 cm^{-1} 下测得氧分子近红外波段的吸收光谱示意图。带平均透射率的基本思想是,通过探测器得到目标辐射光谱 $I_m(\omega)$, 利用带外的光谱数据拟合基线 $I_b(\omega)$, 即可得到被测目标光谱的透射率, 可以表示为:

$$\overline{T_\omega} = \frac{I_m(\omega)}{I_b(\omega)}. \quad (2)$$

根据(2)式,目标光谱 $I_m(\omega)$ 可通过光谱仪测得,算法的关键就在于基线 $I_b(\omega)$ 的计算,基线拟合的准确度直接影响到透射率的精度。而测量光谱的带外数据的质量会给基线拟合带来较大影响,所以首先要对带外数据进行筛选,以达到最佳拟合。

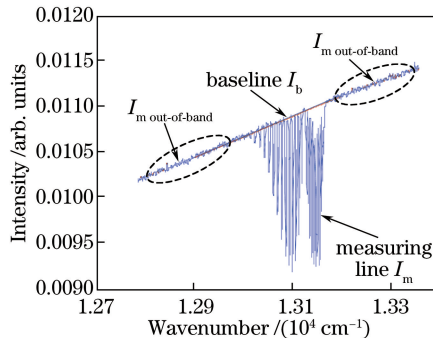


图 1 被测光谱、基线拟合示意图

Fig. 1 Fitting diagram of measured spectrum and baseline

3 带外数据的基线拟合方法

3.1 PLSR 数学模型

根据比尔定律,光强 Y 与测量光谱 X 符合下述关系:

$$Y = XB + G, \quad (3)$$

式中 X 为带外波数, Y 为带外光强, B 为回归系数矩阵, G 为误差矩阵。

将光强 Y 和波数 X 同时进行主成分分解得到 2 个数学模型, 分别可以表示为:

$$X = TP^T + E, \quad (4)$$

$$Y = UQ^T + F, \quad (5)$$

式中 P, Q 分别为 X, Y 的载荷矩阵; T, U 分别为 X, Y 的得分矩阵; E, F 分别是 X, Y 的误差矩阵。

同时考虑 T 和 U 间的线性关系, 使其最大程度地相关。 T 和 U 中的列分别是 X 和 Y 的主成分。 设 u 为 U 的某一列, 对应 T 的一列为 t , u 与 t 之间的线性关系为:

$$u = \beta t + e, \quad (6)$$

式中 β 为系数, 是标量; e 为误差向量。

(3)~(6)式构成 PLSR 的基本数学模型。(4)式和(5)式为外部模型,(3)式为内部模型。外部模型各自进行拟合误差分解,内部模型则把两个外部模型通过最大程度的线性相关紧密联系。 X 通过拟合误差投影得到新变量 t , Y 投影得到新变量 u 。二者通过内部模型建立线性关系。

经过不断迭代计算,若回归方程达到精度要求,则算法中止。否则继续提取第二对光谱信息,直到满足精度要求为止。

氧气吸收带外数据利用 PLSR 的方法可以得到被测目标光谱的基线,但是目标光谱测量过程中不可避免会出现奇异点,它会给 PLSR 相对较高的权重。因此过多地强调拟合基线和这些奇异点的关系,会导致拟合出现偏差,据此做出的基线可能会和实际情况有相当大的差距。因此必须剔除带外数据奇异点,获得新的回归参数,以期达到最佳拟合。

3.2 奇异点的剔除

基于上述基线拟合方法,采用 Kennard-Stone (K-S)选择法^[12]进行奇异点剔除。

基于带外光谱特征计算每个数据点与相邻数据点之间的欧氏距离^[13],在带外均匀选择样本,其中,欧氏距离为:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (7)$$

式中 d_{ij} 为相邻两个数据点的距离, x_i, x_j 为相邻的两个带外数据点。

设有 z 个带外光谱数据,从中选择 n 个校正数作为基线拟合的依据,K-S 方法选择样本的过程为:

- 1) 首先计算所有数据两两之间的欧式距离 d_{ij} , 选择距离最远的两个数据 z_1 和 z_2 进入校正集;
- 2) 计算剩余 $z-2$ 个数据点与所选择的这两个样本 z_1 和 z_2 之间的距离并各取其最小值 $\min(d_i, z_1, d_i, z_2)$, 然后选取其中最大值 $\max[\min(d_i, z_1, d_i, z_2)]$ 对应的一个样本 z_3 进入校正集;
- 3) 计算剩余 $z-3$ 个数据与所选择的这三个样本 z_1, z_2 和 z_3 之间的距离并各取其最小值 $\min(d_i, z_1, d_i, z_2, d_i, z_3)$, 然后选取其中最大值 $\max[\min(d_i, z_1, d_i, z_2, d_i, z_3)]$ 对应的一个样本 z_4 进入校正集;
- 4) 重复上述过程,直至选中 n 个校正样本。

利用上述方法剔除了奇异点前后的效果如图 2 所示,奇异点剔除前后的残差如表 1 所示。图 2 中,一系列蓝色吸收谱线上方接近水平的蓝线为剔除奇异点之前拟合的基线,上方黑色虚线为剔除奇异点之后拟合的基线。从表 1 中可知,奇异点剔除前带外数据拟合平均残差为 1.68×10^{-4} ,而剔除奇异点后的残差仅剩 1.16×10^{-4} ,带外数据几乎均匀分布于基线两侧。

综上所述,奇异点剔除前后基线拟合的曲线发生了变化,模型中存在的奇异点明显使残差增大,所以剔除奇异点后能有效提高基线拟合的准确性。利用相应的评价标准来判断是否达到最佳基线拟合。

3.3 基线拟合评价标准

采用基线拟合的方式估算原始光谱信息,不可避免地会引入测量结果的不确定性及误差。引入带平均

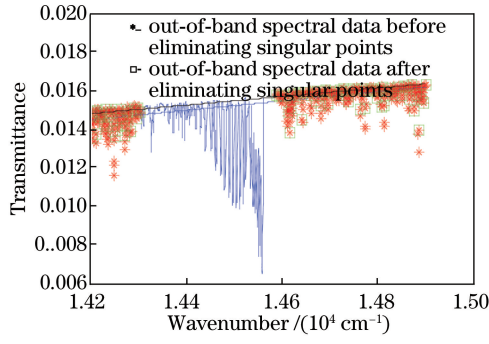


图 2 奇异点剔除前后拟合的基线

Fig. 2 Fitting baselines before and after eliminating singular points

吸收率的标准偏差来说明基线拟合的不确定度。

表 1 奇异点剔除前后的平均残差

Table 1 Mean residual before and after eliminating singular points

	Before	After
Mean residual	1.68×10^{-4}	-1.16×10^{-4}

根据(2)式可得带平均吸收率为:

$$\bar{A} = 1 - \bar{T} = 1 - \frac{\bar{I}}{\bar{I}_b}, \quad (8)$$

式中,分子为被测目标辐射光强算术平均值,而分母是带外数据拟合基线的算术平均值计算带平均吸收率时,带外数据拟合基线至关重要,而基线拟合方法有很多,其评价标准可用统计分析方法计算不确定度表示^[14],多次重复测量得到被测目标强度的平均值为:

$$\bar{I}(\omega) = \frac{1}{n} \sum_{i=1}^n I_i. \quad (9)$$

总的带平均吸收率的不确定度用 $\sigma_{\bar{A}}$ 表示,是(8)式各变量的函数,可以表示为:

$$\sigma_{\bar{A}} = \sqrt{\left(\frac{\partial \bar{A}}{\partial \bar{I}} \sigma_I\right)^2 + \left(\frac{\partial \bar{A}}{\partial \bar{I}_b} \sigma_{I_b}\right)^2}, \quad (10)$$

式中, σ 表示带平均吸收率的标准偏差,对(8)式求偏差,带平均吸收率的标准偏差变为:

$$\sigma_{\bar{A}} = \frac{1}{\bar{I}_b} \sqrt{\sigma_I^2 + \left(\frac{\bar{I}}{\bar{I}_b}\right)^2 \sigma_{I_b}^2}, \quad (11)$$

将(8)式代入上式可得:

$$\sigma_{\bar{A}} = \frac{1}{\bar{I}_b} \sqrt{\sigma_I^2 + (1 - \bar{A})^2 \sigma_{I_b}^2}, \quad (12)$$

式中, σ_I 和 σ_{I_b} 分别代表吸收带内和带外的标准偏差,而 σ_I 的计算必须已知光谱吸收带的先验知识,为解决此问题, σ_I 假设近似等于带外辐射的标准偏差。即利用带外数据通过低阶多项式拟合获得基线, σ_I 就等于相应带外基线与原始带外数据之差的标准偏差,可以表示为:

$$\sigma_I = \sqrt{\frac{1}{N} \sum_{i=1}^N (\Delta x_{i,\text{out}} - \overline{\Delta \mu_{\text{out}}})^2}, \quad (13)$$

式中, N 为测量光谱点数, $\Delta x_{i,\text{out}}$ 为测得带外测量光谱与拟合带外光谱值的差, $\overline{\Delta \mu_{\text{out}}}$ 为带外光谱差的平均值。

σ_{I_b} 是预测误差的标准偏差,预测误差是预测值与实际值之间的差异,是评价预测准确度的一个重要因素,可以表示为:

$$\sigma_{I_b} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\Delta x_{i,\text{in}} - \overline{\Delta \mu_{\text{in}}})^2}. \quad (14)$$

基线拟合是带平均透射率计算的关键,以(13)式为依据来评价带平均透射率计算的准确度,为预测模型误差的计算提供依据。

4 实验结果

为了验证氧气 A 带基线拟合方法精度对带平均透射率的影响,在空气温度为 19.8 °C,标准大气压下,在 0~130 m 内任取 9 个距离点进行测量,以 500 W 石英卤素灯作为光源,输入端采用直径为 165 mm 望远镜,采用不同分辨率分别多次扫描取平均后获得目标的衰减光谱。

为了分析基线拟合不确定度与不同分辨率的关系,在 46.4 m 处取不同分辨率下光谱数据进行氧分子 A 带的基线拟合,根据(12)式计算不同分辨率下的拟合标准偏差,结果如表 2 所示。

表 2 氧气 A 带 46.4 m 处基线拟合参数

Table 2 Baseline fitting parameters of oxygen A-band at 46.4 m

Resolution / cm^{-1}	Standard deviation σ_A /%	Signal-to-noise ratio /dB
1	0.3	476.3
2	0.23	627.5
4	0.23	633.8
8	0.23	635.9
16	0.19	796.1
32	0.2	817.3

可见,同一距离点处不同分辨率下的平均标准偏差为 0.23%,随着分辨率的降低,基线拟合不确定度变小,信噪比增大;但并非分辨率越小越好,当分辨率降低到 32 cm^{-1} 时,虽然信噪比最高,但是不确定度却增加了。

在 9 个不同距离点处,不同分辨率下获得拟合不确定度与距离的关系如图 3 所示,随着距离的增大,拟合不确定度总体呈增大的趋势。但是在 8 cm^{-1} 和 16 cm^{-1} 时,随着距离的增大,拟合不确定度比较平稳,上升缓慢,尤其是 8 cm^{-1} 时,在不同距离处波动相对更小,变换更加平稳。不同分辨率下,9 个距离点处的基线拟合不确定度如表 3 所示。

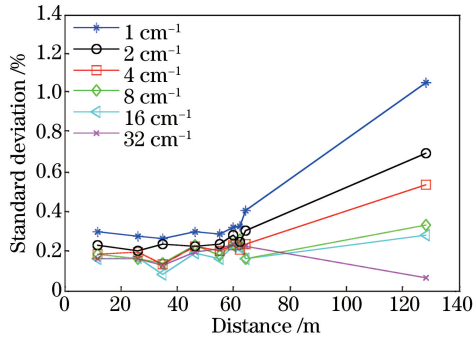


图 3 带平均吸收率基线拟合标准偏差

Fig. 3 Baseline fitting standard deviation with band average absorptivity

随着被测距离的增大,拟合不确定度随分辨率的降低而减小,总的拟合不确定度不超过 0.4%。理论和实验证明,带平均透射率的计算方法还可消除吸收带内外仪器和散射带来的误差。

依据吸收带外数据拟合带内基线,计算带平均透射率的方法不仅是可行的,而且可以不依赖目标先验知识而获取目标光谱透射率。由表 3 可知,基线拟合不确定度还与测试设备的分辨率有关,分辨率越高,带外信息拟合基线不确定度越大,反之,带外基线拟合不确定度减小。

5 结 论

本文以红外目标在氧分子 A 带的吸收峰为研究对象,利用带外数据对带内光谱进行基线拟合。为了减小基线拟合精度对带平均透射率的影响,首先剔除光谱采集中产生的奇异点,采用偏最小二乘拟合方法进行基线拟合。搭建了实验系统,通过对不同距离点处的吸收光谱拟合基线,计算了不同点、不同分辨率下的拟

合不确定度。结果证明,同一距离点处不同分辨率下的平均标准偏差为 0.23%,随着分辨率的降低,基线拟合不确定度变小,信噪比增大;随着被测距离的增大,拟合不确定度随分辨率的降低而减小,总的拟合不确定度不超过 0.4%。该结论为后续测距系统的误差分析提供了依据。

表 3 不同距离、不同分辨率的基线拟合不确定度

Table 3 Baseline fitting uncertainty with different distances and resolutions

%

Distance /m	1 cm ⁻¹	2 cm ⁻¹	4 cm ⁻¹	8 cm ⁻¹	16 cm ⁻¹	32 cm ⁻¹
12	0.30	0.23	0.19	0.19	0.17	0.16
26	0.28	0.20	0.20	0.17	0.17	0.16
35	0.27	0.23	0.13	0.14	0.08	0.13
46	0.30	0.23	0.23	0.23	0.19	0.20
55	0.29	0.23	0.20	0.18	0.16	0.21
60	0.32	0.29	0.23	0.24	0.22	0.24
62	0.33	0.25	0.21	0.26	0.22	0.25
64	0.41	0.31	0.24	0.16	0.16	0.22
128	1.06	0.70	0.54	0.33	0.28	0.07
Average uncertainty	0.40	0.30	0.24	0.21	0.18	0.18

参 考 文 献

[1] Hawks M R, Perram G P. Passive ranging of emissive targets using atmospheric oxygen absorption lines[C]. SPIE, 2005, 5811: 112-122.

[2] Lü Hang, Liao Ningfang, Wu Wenmin, *et al.* Interferogram baseline correction method based on self-adaptive differential filtering[J]. Acta Optica Sinica, 2015, 35(10): 1030002.
吕 航, 廖宁放, 吴文敏, 等. 基于自适应差分滤波的干涉图基线校正方法[J]. 光学学报, 2015, 35(10): 1030002.

[3] Liu Xia, Wang Yunfeng. Research of automatically piecewise polynomial curve-fitting method based on least-square principle[J]. Science Technology and Engineering, 2014, 14(3): 55-58.
刘 霞, 王运锋. 基于最小二乘法的自动分段多项式曲线拟合方法研究[J]. 科学技术与工程, 2014, 14(3): 55-58.

[4] Zong Pengfei, Wang Zhibin, Zhang Jilong, *et al.* Study on baseline fitting method based on passive infrared range measurement[J]. Laser Technology, 2013, 37(2): 174-177.
宗鹏飞, 王志斌, 张记龙, 等. 基于红外被动测距的基线拟合算法研究[J]. 激光技术, 2013, 37(2): 174-177.

[5] Zhang Yufeng, Dai Jingmin, Lu Xiaodong, *et al.* Correction methods of temperature drift for infrared spectral emissivity measurement system[J]. Spectroscopy and Spectral Analysis, 2016, 36(9): 3003-3007.
张宇峰, 戴景民, 陆晓东, 等. 红外光谱发射率测量系统的温漂修正方法[J]. 光谱学与光谱分析, 2016, 36(9): 3003-3007.

[6] Qin Zhaojun, Tao Zhanhua, Liu Junxian, *et al.* Baseline correction of Raman spectrum based on piecewise linear fitting [J]. Spectroscopy and Spectral Analysis, 2013, 33(2): 383-386.
覃赵军, 陶站华, 刘军贤, 等. 分段式线性拟合校正拉曼光谱基线漂移[J]. 光谱学与光谱分析, 2013, 33(2): 383-386.

[7] Diao Weifeng, Liu Jiqiao, Zhu Xiaopeng, *et al.* Study of all-fiber coherent Doppler lidar wind profile nonlinear least square retrieval method and validation experiment[J]. Chinese J Lasers, 2015, 42(9): 0914003.
刁伟峰, 刘继桥, 竹孝鹏, 等. 全光纤相干多普勒激光雷达非线性最小二乘风速反演方法及实验研究[J]. 中国激光, 2015, 42(9): 0914003.

[8] Dowell F E, Pearson T C, Maghirang E B, *et al.* Reflectance and transmittance spectroscopy applied to detecting fumonisin in single corn kernels infected with *Fusarium verticillioides*[J]. Cereal Chemistry, 2002, 79(2): 222-226.

[9] 杜一平, 潘铁英, 张玉兰. 化学计量学应用[M]. 北京: 化学工业出版社, 2008.

[10] Li Jinhua, Yang Zhiliang, Wang Zhaoba, *et al.* The corn content measurement with near infrared diffuse transmission [J]. Infrared Technology, 2013, 35(11): 732-736.
李晋华, 杨志良, 王召巴, 等. 近红外漫透射技术检测玉米成分[J]. 红外技术, 2013, 35(11): 732-736.

[11] Li Jinhua, Wang Zhibin, Chen Yuanyuan, *et al.* Near infrared oxygen A-band atmospheric transmittance calculation

[J]. *Laser & Infrared*, 2013, 43(10): 1142-1145.

李晋华, 王志斌, 陈媛媛, 等. 近红外氧气 A 带大气透射率的计算[J]. *激光与红外*, 2013, 43(10): 1142-1145.

[12] Danielsson P E. Euclidean distance mapping[J]. *Computer Graphics and Image Processing*, 1980, 14(3): 227-248.

[13] 李慎安. 《测量不确定度评定与表示》计量技术规范问与答(上)[J]. *计量技术*, 2002, (11): 46-48.

[14] Cao Hongyan. Uncertainty in measurement and its evaluation in analysis and testing Part 5 Some question which need to claim attention in the evaluation of uncertainty in measurement[J]. *Metallurgical Analysis*, 2006, 26(4): 89-94.

曹宏燕. 分析测试中测量不确定度及评定 第五部分 测量不确定度评定中要注意的一些问题[J]. *冶金分析*, 2006, 26(4): 89-94.