文章编号：1000-7032(2015)11-1335-07

利用油品紫外荧光特性的多光谱成像检测

韩仲志[1,2]，万剑华[1*]，刘 杰[2]，刘康炜[1,3]

（1. 中国石油大学(华东) 地球科学学院，山东 青岛 266580；
2. 青岛农业大学 理学与信息科学学院，山东 青岛 266109；
3. 中国石化青岛安全工程研究院，山东 青岛 266071）

摘要：利用石油及其产品具有的紫外荧光特性，搭建了一套紫外诱导多光谱成像系统。该系统主要由 3 个紫外诱导光源、8 个滤波片和 1 个彩色 CCD 相机组成。采集了 6 种油品的多光谱图像，以有效光斑的 24 个颜色分量均值作为特征，提出了一种联合熵最大化的独立分量分析特征优化方法。K 均值聚类和支持向量机识别结果表明，较改进前的 ICA 方法，该方法的特征优化性能得到了有效提高，油种识别率达到了 92.3%。

关 键 词：紫外诱导；多光谱成像；联合熵独立分量分析；油品检测
中图分类号：O439 **文献标识码**：A **DOI**：10.3788/fgxb20153611.1335

Multispectral Imaging Detection Using The Ultraviolet Fluorescence Characteristics of Oil

HAN Zhong-zhi[1,2], WAN Jian-hua[1*], LIU Jie[2], LIU Kang-wei[1,3]

(1. *School of Geosciences, China University of Petroleum, Qingdao* 266580 *China*;
2. *Information College, Qingdao Agricultural University, Qingdao* 266109 *China*;
3. *Sinopec Safety Engineering Institute, Qingdao* 266071 *China*)

* *Corresponding Author, E-mail: wjh*66310@163.*com*

Abstract: Based on the UV fluorescence phenomena of oil and its products, a multispectral imaging system was constructed. This system was composed of 3 UV excitation light sources, 8 optics filters and a CCD camera. Using this system, multi-spectral images of 6 kinds of oil were collected. The mean of 24 color features of effective light spots was used as the feature set. Then, a novel method called maximize the joint entropy of independent component analysis (ICA) was proposed for K-mean cluster and SVM recognition. It is proved that this method is better than traditional ICA for feature optimized, and the identification rate is 92.3%. This result has positive significance for oil detection.

Key words: UV excitation light; multi-spectral imaging; joint entropy of independent component analysis; oil identification

收稿日期：2015-08-11；**修订日期**：2015-09-16
基金项目：国家自然科学基金(31201133)；青岛市科技发展计划(14-2-3-52-nsh)资助项目

1 引 言

对油品的识别探测是石油工业的重要内容。传统对油品的探测，主要是基于以高效液相色谱法（High performance liquid chromatography，HPLC）为代表的生化方法[1]，该类方法的缺点是检测速度慢、代价高。近年来光谱分析法[2-3]成为油品鉴别的新兴手段，受到普遍关注。

Kim 等[4]最早应用近红外光谱进行石油产品的分类，他们利用主分量分析（Principle component analysis，PCA）和贝叶斯分类器实现了柴油、煤油、粗汽油等 6 个油种的识别；王丽等[5]利用近红外光谱技术自行配制了 56 个汽油、柴油、润滑油油样，进行油品种类鉴别。油品在紫外激发下具有荧光现象，可根据这一特性探测海洋溢油[6]。在实验室情况下，王春艳等[7]提出使用基于浓度参量同步荧光光谱技术，可以实现实验室不同油的类型及不同油源原油的准确分类。尹晓楠[8]利用小波分析方法分析了 4 大类 6 种油品的三维荧光光谱，并对油品种类进行了识别研究。Xie 等[9]将高光谱成像技术应用到对油的种类识别中，可以正确探测油是否掺假。然而上述文献技术中使用的仪器复杂，不能做到现场、快速的原位探测。

多光谱成像技术由于其结构简单、仪器携带方便而得到了广泛的应用。蔡健荣等[10]在可见光和近红外区域选择 5 个特征波长滤波片，采集得到 5 幅滤波后的图像，并利用光谱角分类算法完成了柑橘识别，准确度达到 96%。雷鹏等[11]利用有机磷农药的荧光特性，结合多光谱成像技术和荧光激发技术，研发了叶片农药残留快速检测的多光谱荧光成像系统，有机磷农药氧乐果会在 440 nm 附近产生显著的荧光发射，并且荧光发射光谱峰值与农药浓度具有良好的线性关系。上述研究为多光谱成像探测提供了一种便捷、可行的方法。

本文利用石油及其产品的紫外荧光特性，搭建了一套紫外诱导的多光谱成像系统，并基于该系统采集了油样的图像，提出了一种基于联合熵的核独立分量分析方法进行图像特征提取，最后通过聚类和识别技术探讨了搭建的系统与提出方法的有效性。

2 系统搭建与数据采集

2.1 多光谱成像系统搭建

实验使用的多光谱图像采集系统主要由紫外诱导光源、滤波片组、彩色 CCD 相机及计算机组成，图 1 为该系统的示意图。

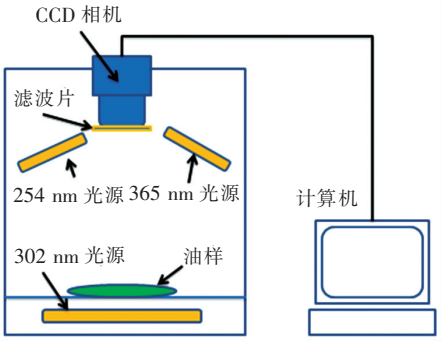

图 1 紫外诱导多光谱成像系统示意图

Fig. 1 Multispectral fluorescence imaging system

紫外诱导光源为紫外分析仪上所带的光源，该仪器为上海光豪分析仪器有限公司生产的 ZF-90B 型暗箱式多功能紫外线反射透射仪，其中 254 nm、365 nm 光源分别为 2 只 15 W 的紫外灯管，302 nm 为 8 只 15 W 的紫外灯管，可见光为 2 W 日光灯。302 nm 为透射测量，254 nm 和 365 nm 为反射测量。

滤光片共有 8 片，分别为：365，420，450 nm（带宽：10 nm；峰值波长透光率 T_{max}：60% ~65%）和 404，410，435，546，577 nm（带宽：20 nm，峰值波长透光率 T_{max}：70%）。将滤波片通过支架安装在相机镜头前方。所使用的相机为 Sony HDR-XR520。

2.2 样品采集

实验样品共 6 种油样（样品编号：1：汽油；2：柴油；3：煤油；4：机油；5：原油；6：花生油）。

实验时用吸管取样品 18 mL，滴定到圆形玻璃培养皿中。将样品放置在相机下，打开紫外光源，进行图片采集，采样时相机距离样品 26.0 cm。分别采集 254 反射、302 透射、365 反射紫外光激发情况下，在镜头前分别加装 8 种滤光片的图像，不同光源、不同滤光片下对每个样品采集 2 张图像，拍摄时选择微距模式。共采集样品图像 288 幅（6 个油种 ×3 个波长光源 ×8 种滤波片 ×2 副）。

对采集的图像进行编号，如 4 机油，光源 254 nm，滤波片 577 nm，第 1 张，编号为 4-254-577-1.jpg。图 2 为机油在 3 种紫外光 8 种波长下的图像。

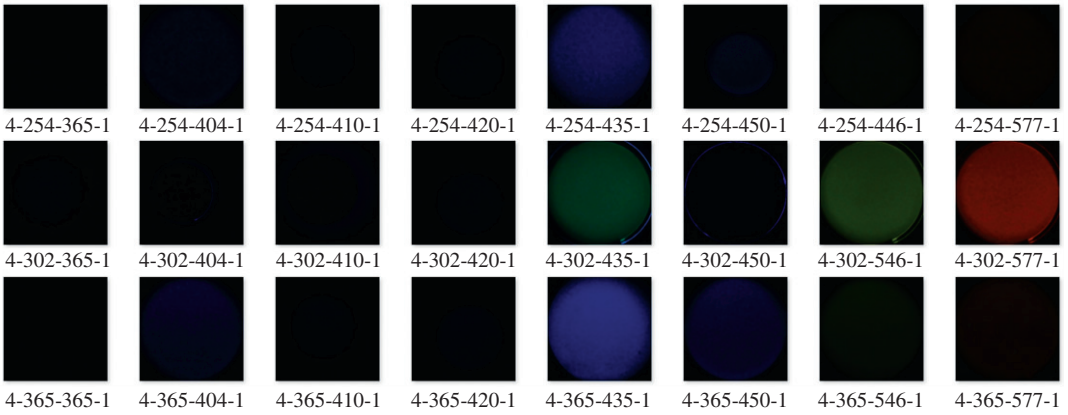

图 2　机油在 3 种紫外光 8 种波长滤波片下的照片，不同紫外激发下，可以看出部分波段下荧光现象较为明显，用彩色相机拍摄时表现出不同的颜色。

Fig. 2　Images of engine oil of 8 wavelengths Edmund Optics under 3 UV lights. Under different excitation UV light, it is found that the fluorescence is obvious relatively under some bands. It presents different color shot by camera.

2.3　图像预处理与特征提取

多光谱成像系统采集的图像大小为 3 000 × 4 000 = 1.2 × 10^7 像素，图像中包括了背景、玻璃皿等信息，应该去除。图像预处理主要是对油样的有效光斑信息进行提取。使用略小于玻璃皿的内接矩形将有效光板分割出来，大小为 500 × 500，然后将图像分割成 100 × 100 的子图像 25 副。由于每个油样重复了两次拍摄，所以可以得到 50 副子图像，也就是 50 个样本。求得 50 个样本的 24 个特征，这 24 个特征为：RGB 和 HSV 共 6 个分量的均值、方差、偏度、峰度。由此得到 6 个油种、个油种 50 个样本、每个样本 24 个特征的每特征矩阵，保存到 Excel 表格中，以备后面识别使用。Matlab2010a 编程自动实现上述过程。

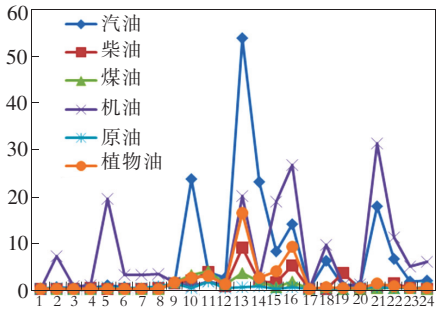

图 3　图像的均值表，RGB 和 HSV 共 6 个分量的均值、方差、偏度、峰度，共得到 24 个特征，依次编号为 1 ~ 24。这里的植物油以花生油为代表。

Fig. 3　Mean of features of different oil images, the mean, variance, skewness, kurtosis of RGB and HSV, which are numbered from 1to 24. Here, the peanut oil represents the plant oil.

3 种激发波长，每个波长 8 种滤波片情况下拍摄到的光斑图像的均值分布图如图 3 所示，图 4 为 6 种油样特征数据分布的箱形图。均值表和箱型图反应了光斑数据的大致分布。

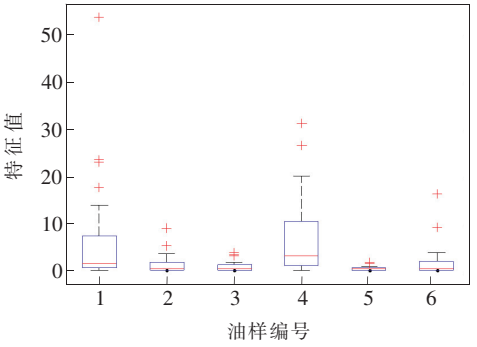

图 4　6 种油样数据的箱型图，反应了数据位置和分散情况的关键信息，编号为 1 和 4 的油品差异比较明显。

Fig. 4　Boxplot of 6 kinds of oil samples, indicating the distribution and dispersion of the data. The difference of class 1 and class 4 is obviously

3　联合熵独立分量分析

3.1　ICA 与 K-ICA

独立分量分析[9]（ICA，Independent component analysis）最早用来解决盲源分离问题[10]，是一种基于数据高阶统计量（四阶统计量）的非高

斯信号处理方法。由于图像满足亚高斯分布，所以可用 ICA 对图像进行特征优化，从中找到最为独立的特征，这样就从一定程度上减少了数据之间的冗余，提高了数据的可分性。

核独立分量分析[11]算法是在特征优化过程中利用了核函数的思想，核函数的作用是避免计算高维变换，直接用低维度的参数带入核函数来计算高维度的向量的内积。可选择的核函数有高斯核函数(Gaussian)、多项式核函数(Poly)、埃尔米特核函数(Hermite)，本文使用的高斯核函数，当然用户也可以根据自己的需要创建核函数。

另外在做数据的独立分量分析之前需要先对数据进行白化和中心化，可选择使用主分量分析(PCA)对特征数据进行去相关。独立分量分析虽然能在最大程度上找到独立分量，但是并没有好的方法对独立分量的顺序和幅值方向进行标记，这往往会导致幅值是负的情况出现，需要增加一些后处理手段进行校正。

3.2 独立分量分析的改进

上述独立分量分析算法所分离的独立分量(ICs)是无序的，第一个分离的分量可能并不是最重要的，所以需要通过一个准则来对独立分量的重要性进行排序，这里使用负熵来衡量其重要性，则最大化负熵的 IC 将会被首先分离出来。

$$N_g(Y) = H(Y_{Gauss}) - H(Y), \quad (1)$$

这里的 Y_{Gauss} 是与 Y 具有相同方差的随机高斯信号，$H(\cdot)$ 是随机变量的微分熵。第 i 个 ICs 的熵值为：

$$H(i) = -\sum_{i\in X} P(x)\log_2[P(x)], \quad (2)$$

然而即使将最大化负熵的独立分量首先提取出来，也并不能代表前几个独立分量的信息贡献率最大。这里我们进一步将联合熵最大化处理。一般来说联合熵越大，则信息越丰富。两个独立分量的联合熵为：

$$H(x,y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2[P(x,y)], \quad (3)$$

这里使得联合熵最大的即为最优特征组合。下文中将基于联合熵的独立分量分析方法表示为 H-ICA，其中 H 表示熵的数学符号。

4 结果与讨论

4.1 K 均值聚类分析

聚类是根据样本特征间的相似程度，相近的样本首先聚为一类。图 5(a)、(b)、(c)是分别对特征矩阵未优化、ICA 优化、H-ICA 优化后，使用 K 均值(K-means)聚类的结果，从图 5 可以看出，原始特征的聚类效果不好，经过 ICA 特征优化后，出现了类间聚集的趋势，经过 H-ICA 变换后聚类效果最好，数据的可分性明显增强，特征优化后三维图上显示数据在三维空间可分性明显增强。这里的初始聚类中心是系统随机生成的。

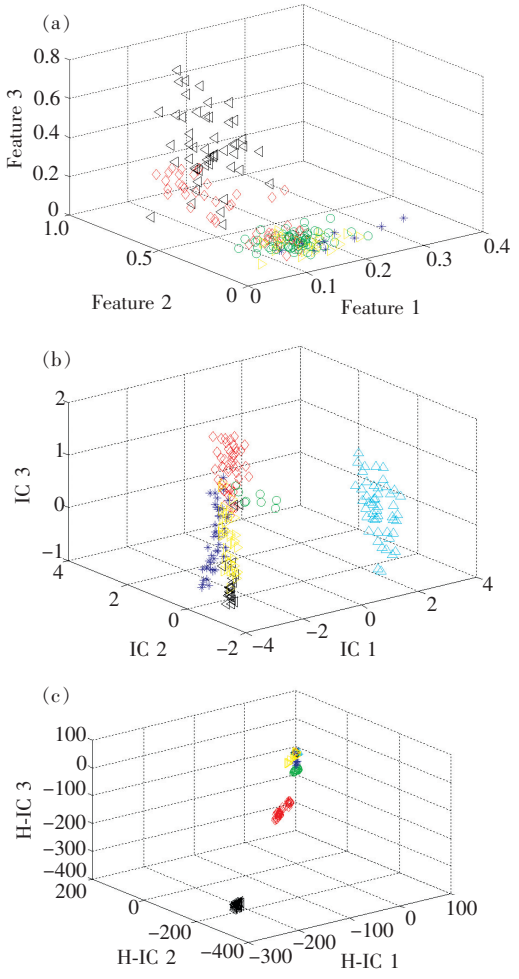

图 5 3 种方法的聚类效果图。其中红色 o 为第一类，蓝色 o 为第二类，绿色 * 为第三类，黑色 o 为第四类，黄色 * 为第五类，粉色 * 为第六类。(a)原始特征 K 均值聚类；(b)ICA 特征优化后聚类；(c)H-ICA 特征优化后聚类。

Fig. 5 Cluster result of 3 methods. Here, red symbol "o" indicates class 1, blue symbol "o" indicates class 2, green symbol " * " indicates class 3, black symbol "o" indicates class 4, yellow symbol " * " indicates class 5, and pink symbol " * " indicates class 6. (a) Cluster of original features. (b) Cluster of ICA. (c) Cluster of H-ICA.

由于 ICA 算法本身的顺序和幅值的不确定性[12-13]，使得 ICA 特征优化后样本的顺序发生了改变，不能对样品的序号识别时做到一一对应，所以系统的输出并不能按照识别序号从小到大排列，需要对识别结果进行后处理，按照从小到大排列，识别结果按照投票方式进行，取本类样品识别最多的结果作为此类的识别结果进行排序。表 1 为 3 种优化方法的 K 均值聚类识别结果，但总体来看，只有少量的样本聚类错误，绝大部分样本聚类结果正确。

表 1　聚类结果

Table 1　Cluster result

Methods	错误	正确	正确率/%
未优化	127	173	57.7
ICA 优化	45	262	87.3
H-ICA 优化	48	265	88.3

4.2　SVM 识别

选用的特征为对未优化、经 ICA 优化和经 H-ICA 优化后，选用支持向量机进行识别。选用的核函数是径向基 RBF 核函数，其中的两个参数 C 和 gamma 可由系统网格法寻优给出[15]。训练和测试过程采用 5 折交叉验证法，即随机选取 50 组数据中的 40 组作为训练集，10 组作为测试集。图 6 是该方法的识别相对误差，可以看出 SVM 识别模型训练集误差比测试误差要小得多。H-ICA 算法的相对误差较小。

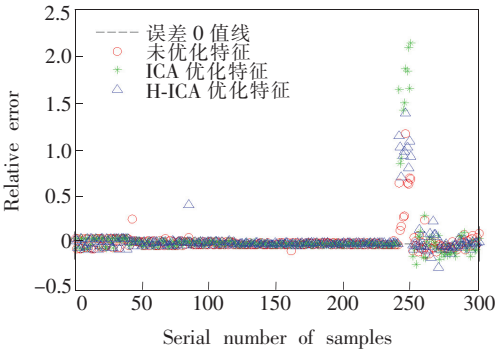

图 6　SVM 识别相对误差。经 H-ICA 算法优化后的特征的相对误差变小。

Fig. 6　Relative error of SVM recognition. For the feature optimized using H-ICA algorithm, the relative error declined.

我们主要采用 3 个参数指标：平均平方误差（Mean squared error, MSE）[16]、平均相关系数（Squared correlation coefficient, R^2）[17]、识别率（Correct recognition rate, CRR, %），来比较 SVM 识别模型的预测性能。MSE 越小、R^2 越接近于 1，模型的性能越好。CRR 用来评价衡量模型的识别性能。模型性能结果见表 2。

表 2　SVM 模型预测性能指标

Table 2　Ultimate performance index of SVM model

样本集合（测试项目）	测试集 MSE	训练集 MSE	测试集 R^2	训练集 R^2	测试（CRR）	训练（CRR）	平均（CRR）
未优化	0.022 1	0.000 3	0.832 0	0.997 2	0.995 8	0.716 7	0.856 3
ICA	0.008 7	0.000 2	0.941 1	0.998 5	0.995 8	0.800 0	0.897 9
H-ICA	0.004 3	0.000 1	0.965 3	0.999 3	1.000 0	0.850 0	0.925 0

表 2 中的模型预测性能指标对模型的性能进行了量化，从表中不难看出 SVM 的总体性能较好，训练集的性能普遍比测试集好，这是可以理解的。测试集预测性能表明了模型的泛化能力，从识别率上看，对预测集经优化后，识别率有所提高，但测试集的识别率有所下降，H-ICA 的整体性能优于 ICA。

4.3　讨　论

刘晓华等[18]利用诱导荧光技术对油种进行了识别研究。他们以 355 nm 紫外激光诱导 9 种常见机油样品发射荧光，共采集 450 组荧光光谱数据，其中 360 组数据用于分类训练，90 组数据用于识别。分析发现，不同机油的荧光光谱特征有明显差异。他们利用主成分分析结合聚类分析法实现了对 90 组待识别光谱数据的快速识别，识别率可达 90% 以上。他们的结论与本文一致，但本文使用的是不同波长的紫外光源诱导荧光，并使用了一种简单的多光谱相机，仪器更为便宜、简单，而且本文涉及的油种更为广泛。

高光谱技术具有更精细的光谱分辨率，目前高光谱成像技术已广泛应用于物质分析中，同样也可以应用于油品分析中[9]。然而高光谱成像

需要配合高精度的液晶可调式滤波器或者光栅可调滤波模块,成像的速度慢,代价高。准确来讲,多光谱成像是高光谱成像的一种合理的简化,它通过面向对象特征的几个波段成像从而实现大致高光谱的物质分辨能力,提高了成像速度,更适合于现场应用。但这是以牺牲识别精度为代价的,也就是说成像波段的减少会增大识别误差。现实中使用的技术应该在效率与精度之间寻找一合理的平衡,从识别效果上来看,多光谱成像可以满足对油品的识别需要。

数据分布箱形图反映了范围的不同。在进行识别时,由于特征量纲的不同,数据之间没有可比性,因此需要对数据进行归一化,将所有数据都归一化到相同的范围。然而这样操作隐含着一个前提假设是,各个特征对品种识别的贡献是相同的。事实上,各个特征对品种识别的贡献率是不同的,因此可考虑特征加权,例如单特征 ROC 曲线下面积可作为权值的一种参考,方法有待进一步研究。

本文是使用了每个样品的 24 个颜色均值作为识别的特征,ICA 和 H-ICA 也是将特征优化为 24 个而进行的识别,这从本质上没有发挥 ICA 和 H-ICA 的优势。这两种方法的特征优化可将特征优化为比较少的几个特征,这样在做识别时可充分利用这些特征进行识别,从而提高了效率。交叉验证法(Cross validation,CV)是广泛采用的模型验证方法,本文采用的是 5 折交叉验证,没有特别说明,本文是其中 1 次的结果。目前独立分量分析已经应用于近红外物质的组分解析中[19],结合油品的多光谱成像数据,对物质组分进行精确解析是下一步要做的工作。

诚然本文在进行油种识别时使用了 6 个油品,所采集的样本数量较少且是在实验室情况下的识别。如果增加更多的实测现场数据,研究将更有意义。系统聚类与识别不同的是,聚类并不能直接给出识别的结果,需要对聚类结果与品种之间进行对应,本文是采用投票的方法对聚类结果与品种间建立对应关系,将本类中最多的族作为类别品种,当然也可以通过模式识别后处理的方式对聚类结果进行自动化判别。

5 结 论

基于一台自制的紫外可见光区的多光谱相机,利用实验室环境下 3 个紫外光源(254,302,365 nm)采集了 6 个油品(汽油、柴油、煤油、机油、原油、花生油)在 8 个滤波片(365,404,410,420,435,450,546,577 nm)下的图像。以采集到的图像光斑 24 个颜色分量均值作为特征进行油种的识别。提出了一种通过联合熵最大化的 ICA 方法进行特征优化,特征优化后的 K 均值聚类和 SVM 识别的效果,较未优化、ICA 优化效果都有一定程度的提高。本文提供了一种快速对油种进行识别的方法,该方法对油种的快速识别和分析具有积极意义。

参 考 文 献:

[1] Bonaccorsi I L, McNair H M, Brunner L A, *et al*. Fast HPLC for the analysis of oxygen heterocyclic compounds of citrus essential oils [J]. *J. Agric. Food Chem.*, 1999, 47(10):4237-4239.

[2] Hahn S, Yoon G. Identification of pure component spectra by independent component analysis in glucose prediction based on mid-infrared spectroscopy [J]. *Appl. Opt.*, 2006, 45(32):8374-8380.

[3] Wang G, Sun Y, Ding Q, *et al*. Estimation of source spectra profiles and simultaneous determination of poly component in mixtures from ultraviolet spectra data using kernel independent component analysis and support vector regression [J]. *Anal. Chim. Acta*, 2007, 594(1):101-106.

[4] Kim M, Lee Y H, Han C. Real-time classification of petroleum products using near-infrared spectra [J]. *Comput. Chem. Eng.*, 2000, 24(2):513-517.

[5] Wang L, Zuo L, He Y, *et al*. Recognition of simulated spilled oil at sea by near-infrared [J]. *Spectrosc. Spect. Anal.* (光谱学与光谱分析), 2004, 24(12):1537-1539 (in Chinese).

[6] Kessler J D, Valentine D L, Redmond M C, *et al*. A persistent oxygen anomaly reveals the fate of spilled methane in the deep Gulf of Mexico [J]. *Science*, 2011, 331(6015):312-315.

[7] Wang C Y, Shi X F, Li W D, *et al*. Concentration dependent synchronous fluorescence oil spill fingerprinting identifica-

tion based on principal component analysis and support vector machine [J]. *J. Instrum. Anal.* (分析测试学报), 2014, 33(3):289-295 (in Chinese).

[8] Yin X N. Studies on The Identification of Oil Types Base on 3D Fluorescence Spectroscopy and Wavelet Analysis [D]. Qingdao: Ocean University of China, 2012 (in Chinese).

[9] Xie C, Wang Q, He Y. Identification of different varieties of sesame oil using near-infrared hyperspectral imaging and chemometrics algorithms [J]. *PloS one*, 2014, 9(5):1-8.

[10] Cai J R, Wang J H, Chen Q S, *et al*. Recognition of mature citrus in natural scene using spectral imaging and spectral angle mapper [J]. *Acta Photon. Sinca* (光子学报), 2009, 38(12):3171-3175 (in Chinese).

[11] Lei P, Lyu S B, Li Y, *et al*. Multispectral fluorescence imaging technology for pesticide residues detection [J]. *Chin. J. Lumin.* (发光学报), 2014, 35(6):748-753 (in Chinese).

[12] Hyviirinen A, Karhunen J, Oja E. *Independent Component Analysis* [M]. New york: Wiley and Sons., 2001.

[13] Hyviirinen A, Oja E. A fast fixed-point algorithm for independent component analysis [J]. *Neur. Comput.*, 1997, 9(7): 1483-1492.

[14] Bach F R, Jordan M I. Kernel independent component analysis [J]. *J. Mach. Learn. Res.*, 2003, 3:1-48.

[15] Chang C C, Lin C J. LIBSVM: A library for support vector machines [EB/OL]. 2011-10-05. http://www.csie.ntu.edu.tw/-cjlin/libsvm.

[16] Hansen B E. The integrated mean squared error of series regression and a Rosenthal Hilbert-space inequality [J]. *Econom. Theory*, 2015, 31(2):337-361.

[17] Soja M J, Persson H, Ulander L M H. Estimation of forest height and canopy density from a single InSAR correlation coefficient [J]. *IEEE Geosci. Remote Sens. Lett.*, 2015, 12(3):646-650.

[18] Liu X H, Chen S Y, Zhang Y C, *et al*. Rapid recognition of common machine oils based in laser induced fluorescence [J]. *Spectrosc. Spect. Anal.* (光谱学与光谱分析), 2014, 34(8):2148-2151 (in Chinese).

[19] Fang L M, Lin M. A method of near infrared multi-component analysis based on independent component analysis and neural networks model [J]. *Chin. J. Anal. Chem.* (分析化学), 2008, 36(6):815-818 (in Chinese).

韩仲志(1981 -),男,山东莒南人,副教授,2006 年于广西师范大学获得硕士学位,主要从事光学信息处理方面的研究。

E-mail: hanzhongzhi@qau.edu.cn

万剑华(1963 -),男,山东单县人,教授,博士生导师,2001 年于武汉大学获得博士学位,主要从事石油分析与探测方面的研究。

E-mail: ywjh66310@163.com