第31卷，第10期
2011年10月

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

Vol. 31，No. 10，pp2834-2838
October，2011

# 基于光谱分析技术的黄瓜与茎叶识别研究

王海青，姬长英*，陈坤杰

南京农业大学工学院，江苏省高等学校智能化农业装备重点实验室，江苏 南京 210031

**摘 要** 为了能够快速实时地识别温室中的黄瓜，研究了黄瓜和其茎叶的近红外反射光谱特性。利用近红外光谱仪在室内共采集138个样本（黄瓜46个，茎46个，叶46个）的反射光谱，进行Savitzky-Golay平滑后，抽取光谱中的108个样本作为校正集，采用偏差权重法选择信息量较大的光谱波段690～950 nm进行研究。在主成分分析(PCA)的基础上，结合马氏距离建立识别模型，剔除了7个异常样本。用剩余的101个样本进行偏最小二乘法建模，对校正集之外的30个样本进行预测。结果显示预测值和实际值的相关性达0.994 1，正确识别率达100%。说明黄瓜、茎和叶的近红外反射光谱特性之间有一定差异，可以用近红外光谱技术进行鉴别，为黄瓜识别提供了一种新的方法和思路。

**关键词** 光谱分析；黄瓜识别；主成分分析；偏最小二乘法；马氏距离法

**中图分类号**：S123；TH744.1 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2011)10-2834-05

## 引 言

在传统农业生产过程中，收获环节耗功耗时，所以研究如何提高收获效率是重中之重。机械化生产可以取代人们繁重的劳动，是提高劳动效率的首选。对于智能化黄瓜收获机器人而言，关键在于如何快速识别黄瓜。与利用图像处理的黄瓜识别[1,2]相比，基于光谱分析技术的黄瓜识别方法具有对研究对象本身色彩不敏感，处理方法简单，实时性好等优势。

在自然光条件下，由于黄瓜与其果柄、茎和叶颜色相近，用机器视觉处理采集到的黄瓜图像，分割和识别目标存在一定难度，而且效果并不理想。光谱技术可反映扫描对象内部结构和化学成分及含量的差异，可用于区别颜色相近的生物。

由于光谱技术自身的优势，将光谱技术应用于黄瓜识别是一个可行而重要的研究方向。由于果实表面的反射光谱特征是果实物质的一种固有特征，在不同的光谱波段内表现出不同的分布状态。在可见光范围内黄瓜与其茎叶的反射系数相近，反射曲线可能重叠或交叉在一起，很难区分。但是在近红外波段黄瓜的反射系数大于其茎叶的反射系数，且在该区域光谱反射率比较稳定，容易区分。本文拟研究黄瓜及其茎叶在600～1 099 nm范围内的光谱特性，寻求切实可行的方法识别树上黄瓜。

## 1 样本光谱数据获取与预处理

### 1.1 样本光谱数据获取

#### 1.1.1 样本来源

在南京林大农业发展有限公司，于2010年5月16号上午9:00—10:00采集成熟黄瓜果实，及对应其植株的叶片和茎三种各46个样本，共计138个样本。

#### 1.1.2 样本光谱的获取

实验所用近红外检测系统为SupNIR-1000近红外光谱分析仪(杭州聚光科技有限公司)，光谱分辨率1 nm，波长范围600～1 099 nm，扫描时间为80 ms。系统采用与光谱仪配套的外部光源，为了尽量模拟真实的环境，实验在自然光下进行(2010年5月16日13:00—14:00)。光纤探头置于被测物表面上方10～15 mm，聚光世达近红外分析仪测量分析软件采集到光谱数据为ripx格式，光谱仪设置为一次测量得一组样本的反射率，采集每种样本46组数据作为该样本的原始光谱，共获取138组光谱数据。为了使光谱和样本更好地相关，应用数学关系式：$A=\log(1/R)$(其中$A$为吸光度，$R$为反射率)，将反射率数据转换为吸光度数据。

#### 1.1.3 数据集

获取的138个样本，其中的108个样本作为校正集用来

**收稿日期**：2010-12-06，**修订日期**：2011-03-08

**基金项目**：国家(863计划)项目(2006AA10Z259)资助

**作者简介**：王海青，1981年生，南京农业大学工学院博士研究生 e-mail：whq_nj@126.com

*通讯联系人 e-mail：chyji@njau.edu.cn

建立预测模型。剩余的 30 个样本作为验证集，评估预测模型的性能。为了能够建立良好的预测模型，校正集中包括 36 个黄瓜样本，36 个茎样本，36 个叶样本。验证集包含黄瓜、茎和叶三类，每类 10 个样本。

### 1.2　光谱数据预处理

为了提取光谱的特征信息，去除高频随机噪声、样本不均匀，加强被分析信号权重，提高有用信号比例，需要对原始光谱数据预处理。经初步分析，所有光谱都采用 Savitzky-Golay(平滑点数为 2)进行平滑后，校正集样本进行主成分判别可达到完全正确，为稳定、可靠的数学模型的建立奠定基础。因此本文采用 Savitzky-Golay 进行平滑的预处理方法。所用数据在 Matlab 7.0 和 Unscrambler X 10.0.1 统计软件中进行。

## 2　光谱的处理方法

### 2.1　光谱波段的选择

理论上由近红外光谱仪所得光谱数据均可以用于光谱分析和建模，但是，不同波段贡献的特征信息有较大差异，剔除过多的冗余信息，选择贡献率较大的光谱波段，进行分析和建模，可降低运算开销，避免无关变量对判别分析结果的影响。所以在实际分析和建模过程中，采用偏差权重法选取谱段，即根据偏差大的变量比偏差小的变量特征更明显的原则，选用偏差较大的波长区域作为光谱特征输入量。

### 2.2　主成分分析

主成分分析应用线性变换，在不丢失主要光谱信息的前提下选择维数较少的新变量来代替原来较多变量，是实现压缩光谱数据维数的一种有效方法[3, 4]，可以得出各个变量对各主成分的贡献，同时获取样本光谱的主成分得分。

### 2.3　马氏距离判别法

在定性分析中，一般把光谱异常对应的样本定义为异常样本，认为对建模有较大影响的样本点。结合以上主成分分析所得样本的主成分得分，引入马氏距离鉴定校正集中的异常样本。通过主成分分析求得样本的主成分得分矩阵，计算各个样本的马氏距离($M_i$)，并设置离群阈值($M_{\text{th}}$)实现异常样本的鉴定和剔除[5]。

计算各样本的马氏距离方法如下

$$D_i^2 = (t_i - \overline{T})\boldsymbol{M}^{-1}(t_i - \overline{T})'$$

$$\overline{T} = \left(\sum_{i=1}^{m} t_i\right)/m$$

其中 $\boldsymbol{M}$ 为校正集光谱主成分得分矩阵的协方差阵；$t_i$ 为样本 $i$ 的主成分得分向量；$\overline{T}$ 为 $m$ 个校正集样本的平均得分矩阵；$D_i$ 为校正集样本 i 的马氏距离。

检验校正集中的异常样本存在的阈值计算公式如下

$$D_{\text{th}} = D_m + e \cdot \sigma_d$$

给定阈值调整权重系数 $e$，$D_m$ 和 $\sigma_d$ 分别为 $m$ 个样本马氏距离的平均值和标准差。

凡满足 $M_i \geqslant M_{\text{th}}$，认为校正集中第 $i$ 个样本是异常样本，予以剔除；反之 $M_i < M_{\text{th}}$，认为校正集中样本 $i$ 的光谱在主成分空间中相似。

### 2.4　PLS 交互验证回归模型

偏最小二乘法(partial least squares，PLS)是多元线性回归、典型相关分析和主成分分析的有机结合，较传统的回归分析、主成分分析有着巨大的优势，使模型精度、稳定性和实用性都得到提高[6]。PLS 算法分析高度共线性的数据集表现出了良好的效果[7-9]，适用于独立变量数大于样本数的数据集分析。但是在建模过程中，主成分数会影响到建模的精度，过少的主成分数可能导致模型“欠拟合”，而过多的主成分数不能建立精确的模型，容易发生“过拟合”，所以利用交互有效性原则[10]来确定最佳主成分数。其主要思想是利用预测光谱残差平方和(PRESS) 确定主成分数，进而评价模型的预测能力。一般来说，当 PRESS 值最小时确定的主成分数，建立的模型是最优的，但是由于 PRESS 值不能总是达到完全最小[11]。所以在研究过程中需要在 PRESS 值尽量小的情况下，选择合适的主成分数。

## 3　结果与讨论

### 3.1　样本光谱

实验所得光谱曲线如图 1 所示。在红光区域(600～680 nm)茎和黄瓜果实的光谱曲线发生了交叉或重叠。在可见光波段与近红外波段的过渡区，吸光度陡然增加，720 nm 处出现吸光度峰值。在 720～1 099 nm 的近红外波段内，茎和叶的吸光度明显高于黄瓜果实的吸光度且趋于平缓。这是由于黄瓜果实和茎叶各自内部结构的差异造成的：一方面黄瓜内部含有籽粒为中空结构，且含水率高，与茎叶的内部结构及含水率有明显的差异；另一方面黄瓜植株属双子叶植物，其叶片是双面叶，具有腹面(近轴面)和背面(远轴面)，微观上腹面的栅栏组织具有相当的细胞空隙，背面的海绵组织则形成了较大的细胞间隙，部分光发生了透射或散射。当样本曲线数量加大时，黄瓜光谱曲线会和茎叶光谱曲线交叠在一起，则区分相对复杂，所以选择合理的近红外波段进行识别是必要的。

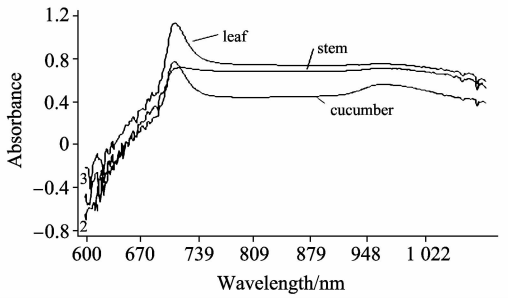


**Fig. 1　Absorbance spectral of cucumber, stem and leaf**

### 3.2　光谱波段相关性选择

本文选用光谱的标准差作为偏差的输入量，权重系数为 1 的偏差权重法，选择光谱特征与实验样本相关性较大的波段。图 2 所示为校证集光谱数据经预处理之后的标准差分布图，从图中信息可知光谱在首段由于噪声干扰，光谱特征存在不稳定性。偏差权重较大且平稳的区域主要在 690～950

nm 范围内，说明光谱特征信息在 690～950 nm 范围内比较显著。

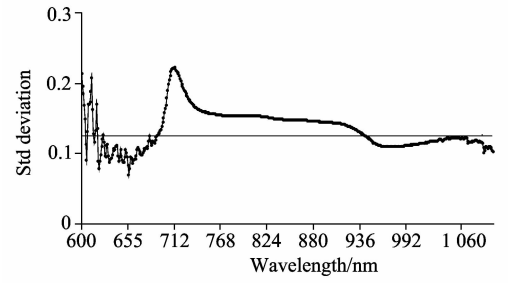


**Fig. 2 Spectral standard deviation of 108 samples in 600～1 099 nm band**

### 3.3 异常样本的剔除

对校正集在 690～950 nm 波段内做主成分分析(PCA)。三类样本的 PCA 结果如图 3 所示，主成分(principal component)中的 PC-1 和 PC-2 的贡献率分别为 86%和 13%；前两个主成分的贡献率达到 99%，可以满足定性分析的要求。

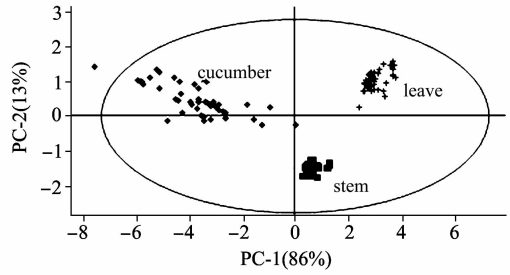


**Fig. 3 Classification result using first and second principal component (PC) of cucumber, stem and leaf**

◆: Cucumber; ■: Stem; +: leaf

从图 3 中可以看到，叶和茎的样本聚集区较小，比较集中。而黄瓜样本在图中比较分散，造成这种差异的原因为：一方面黄瓜表面各部分水分量的不同使黄瓜样本的吸光度有所差异；另一方面黄瓜表面粗糙，有较密集的小刺，影响了入射角，造成了一定的吸光度误差。为了在后续偏最小二乘法建立的判别模型精度更高，应用马氏距离法鉴别异常样本。

本文设置阈值调整权重系数 $e$ 为 2，校正集中的 108 个样本的主成分马氏距离如图 4 所示。从图中可明显识别出样本 1，5，12，23，31，32，81 的马氏距离大于设定阈值，这 7 个光谱异常作为对建模有较大影响的异常样本予以剔除。

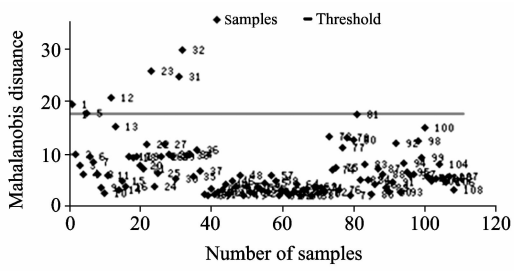


**Fig. 4 Mahalanobis distance distribution of 108 samples**

根据马氏距离结合 PCA 的分析结果，校正集中的 7 个异常样本被剔除，用剩下的 101 个样本作为定标样本建立定性分析预测模型。

### 3.4 PLS 预测模型

本文暂时选择建模主成分总数 20，计算 PRESS 值与各主成分数的对应关系，结果如图 5 所示，PRESS 值随着主成分数从 1 到 7 的变化过程中，持续减小，当主成分数大于 7 时，PRESS 值没有了明显的变化。考虑到 2.4 节中的叙述的情况，交互验证法得到建模最佳主成分数为 7。

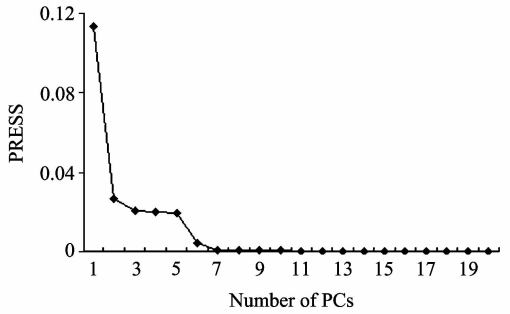


**Fig. 5 Plot of PRESS value vs number of principal components**

选择 7 个主成分建立 PLS 预测模型。用相关系数(correlation coefficient)，建模标准差(standard error of calibration, SEC)，交互验证标准差(standard error of cross validation, SECV)，作为模型精度和稳定性的参考标准。当相关系数较大，其余参数较小时，模型是最优的。

表 1 显示 101 个样本在 690～950 nm 光谱波段的 PLS 模型的建模和交叉验证结果。其中建模和交叉验证的相关系数分别为 0.993 71 和 0.993 32，表明这个模型对数据的拟合良好。SEC 为 0.022 04，SECV 为 0.022 71，两者差值较小，充分说明该模型线性特征良好，没有出现过度拟合与欠拟合，应该具有良好的预测能力。

**Table 1 Results of modeling and cross validation of PLS**

| Parameter | Calibration | Cross validation |
|---|---|---|
| Element | 101 | 101 |
| Correlation | 0.993 71 | 0.993 32 |
| Bias | $4.3\times10^{-5}$ | −0.000 03 |
| SEC | 0.022 04 | |
| SECV | | 0.022 71 |

### 3.5 预测结果

利用 PLS 模型来预测验证集中的 30 个样本，结果如图 6 所示，横坐标为参考值数据，由于定性分析中没有相应的化学指标作为应变量 $Y$，人工设置 $Y$ 值的参考量，一个数值区域代表研究对象的一类：0.1～0.3 为黄瓜类，0.3～0.4 为茎类，0.4～0.6 为叶子类；纵坐标为预测值数据。预测结果的相关系数为 0.994 1，SEP 为 0.025 37，Bias 为−0.009 59 表明预测值与真实值之间有极大的相关性和较低的预测误差。从图 6 中可以看到，预测值分别落在参考分类的区域内，可以 100%地对果实和茎叶进行区分，说明了模型有较强的

识别能力和较好的识别效果。

## 4　结　论

在实验室条件下利用光谱分析技术能对颜色相近的黄瓜和其茎叶进行识别。利用 PCA 对校正集分析，前两个主成分累计贡献率达 99%，清晰可靠地区分三类研究对象，利用样本的主成分得分结合马氏距离法剔除 7 个异常样本。剩余的 101 个样本作为定标建模样本，交互验证法得到最佳主成分数为 7 并建立 PLS 模型，对验证集预测。PLS 模型的预测结果显示预测值和实际值有很高的相关度，通过人工设定参考值区域，识别率达到 100%。能够快速区分黄瓜果实、茎和叶。与传统的计算机视觉相比，本文提出的光谱识别黄瓜果实具有一定特色，为黄瓜采摘，喷药等智能作业的识别工作提供了依据，是一种有效的方法。

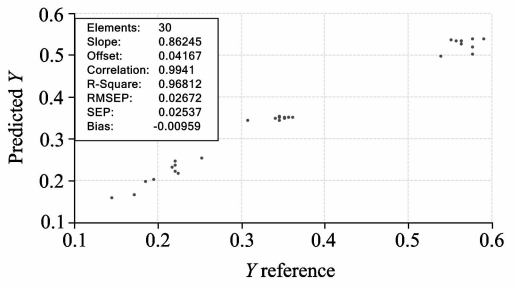


**Fig. 6　Prediction result using PLS method**

## References


[1] Yang qing-hua, Qi li-yong, Bao Guanjun, et al. New Zealand Journal of Agricultural Research, 2007, 50(5): 989.

[2] YUAN Ting, ZHANG Jun-xiong, LI Wei, et al(袁　挺，张俊雄，李　伟，等). Transactions of the Chinese Society of Agricaltural Machinery(农业机械学报), 2009, 40(8): 170.

[3] Eui-Cheol Shin, Brian D. Craft, Ronald B. Pegg, et al. Food Chemistry, 2010, 119: 1262.

[4] Abdulhamit Subasi, Ismail Gursoy M. Expert Systems with Applications, 2010, 37: 8661.

[5] CHEN Bin, ZOU Xian-yong, ZHU Wen-jing(陈　斌，邹贤勇，朱文静). Journal of Jiangsu University • Natural Science Ediction(江苏大学学报 • 自然科学版), 2008, 29(4): 278.

[6] ZENG Jiu-sun, LIU Xiang-guan, LUO Shi-hua, et al(曾九孙，刘祥官，罗世华，等). Journal of Zhejiang University • Science Edition(浙江大学学报 • 理学版), 2009, 36(1): 34.

[7] Delalieux S, van Aardt J, Keulemans W, et al. European Journal of Agronomy, 2007, 27: 130.

[8] Jonesa C D, Jonesb J B, Leea W S. Computers and Electronics in Agriculture, 2010, 74: 329.

[9] Ana M Aguilera, Manuel Escabias, Cristian Preda, et al. Chemometrics and Intelligent Laboratory Systems, 2010, 104: 289.

[10] Efron B, Gong G. The American Statistician, 1983, 37(1): 44.

[11] Eike Luedeling, Adam Hale, Minghua Zhang, et al. International Journal of Applied Earth Observation and Geoinformation, 2009, 11: 247.


# Research on Identification of Cucumber, Stem and Leaf Based on Spectrum Analysis Technology


WANG Hai-qing, JI Chang-ying*, CHEN Kun-jie

Key Laboratory of Intelligent Agricultural Equipment of Higher Education Institute in Jiangsu Province, College of Engineering, Nanjing Agricultural University, Nanjing　210031, China



**Abstract**　To be able to quickly identify the cucumber real time, the present paper studied the near infrared reflectance characteristics of cucumber, stem and leaf. Spectral reflectance of 138 samples (46 cucumbers, 46 stems and 46 leaves) was collected using near infrared spectroscopy in the band range of 600～1 099 nm indoor. After Savitzky-Golay smoothing preprocessing, random 108 spectral samples were put forward as calibration set. The weighted deviation method was used for choosing the spectral bands 690～950 nm that include much more information. The samples were analyzed by PCA method to extract the principal component scores, combining the Mahalanobis distance method the recognition model was established, and seven abnormal samples were excluded. The partial least squares (PLS) model was established by remaining 101 samples spectra of calibration set, which was used for predicting the validation set (30 samples except of the calibration set). The result shows that the correlation of the predicted value and the actual value reaches up to 0.994 1, and the correct recognition rate is 100%. This significantly illustrates that the near infrared spectral reflectance characteristics are different among the cucumbers, stems and leaves, which

can be successfully applied to recognition of cucumber by the method. The developed technique can provide a new method for cucumber identification.

**Keywords** Spectral analysis; Cucumber recognition; Principal component analysis; Partial least squares; Mahalanobis distance

* Corresponding author

(Received Dec. 6, 2010; accepted Mar. 8, 2011)

## 深切哀悼《光谱学与光谱分析》首席顾问王大珩先生逝世

两弹一星功勋奖章”获得者，中国科学院院士、中国工程院院士，国际宇航科学院院士，著名光学家，我国近代光学工程的重要学术奠基人、开拓者和组织领导者，杰出的战略科学家、教育家王大珩先生，因病于 2011 年 7 月 21 日 13 时 02 分在北京逝世，享年 96 岁。

王大珩先生是中国光学学会创始人，历任中国光学学会第一、二、三届理事长，第四、五、六届名誉理事长。光学界的同仁怀着无比崇敬和悲痛的心情，缅怀王先生的高尚人格和丰功伟绩。王大珩先生是热爱祖国、潜心科研、提携后辈的楷模，他的高风亮节，永远照亮后辈的征程。我们怀着深切哀悼的心情表示，要继承王先生的遗志，赶超世界先进水平，完成中国光学界的历史责任。

王大珩先生 1915 年 2 月 26 日出生在日本，祖籍江苏吴县。1936 年毕业于清华大学物理系。1938 年考取“庚款”留学生，在英国伦敦大学帝国理工学院应用光学专业学习，获理学硕士学位。1941—1942 年在英国雪菲尔德大学玻璃制造技术系，攻读博士学位。1942—1948 年在英国昌司玻璃公司从事光学玻璃研究工作。1948 年回国。1949—1951 年任大连大学教授，应用物理系主任。1951 年到中国科学院工作，1952—1983 年在长春从事光学仪器与工程研究，1983 年调至中国科学院技术科学部工作。

王大珩先生历任中国科学院长春光学精密机械研究所所长，中国科学院长春分院院长，中国科学院技术科学部主任，中国科学院空间科学技术中心主任；解放军总装备部科学技术委员会顾问；长春光学精密机械学院院长，哈尔滨科学技术大学校长；中国科学技术协会副主席，北京市科学技术协会主席，中国光学学会理事长，中国仪器仪表学会理事长，中国计量测试学会理事长，中国高技术产业化研究会理事长等职。

王大珩先生 1978 年 10 月加入中国共产党。曾当选为中国共产党第十二次全国代表大会代表，全国人民代表大会第三、四、五、六届代表，全国人民政协第三、七届委员。

王大珩先生在光学与光学工程研究和组织领导工作中做出了杰出贡献。他主持 150 工程，领导研制我国第一台靶场装备大型精密光学跟踪电影经纬仪；主持 718 工程，领导研制我国第一台激光红外电视电影经纬仪和船体变形测量系统，为发展我国的尖端武器做出了杰出贡献。曾荣获国家科技进步特等奖，名列首位；何梁何利基金科学与技术成就奖、国家“两弹一星功勋奖章”；国家“863 计划”特殊贡献先进个人称号。

王大珩先生远见卓识，从战略高度上思考并联合其他科学家对国家科学技术发展提出多项重大建议。关于跟踪研究战略性高技术发展的建议，最后成为国家《高技术研究发展计划纲要》(简称“863”计划)，使发展高科技成为实现我国科技现代化的一项重要战略部署。这些建议为国家科技决策发挥了积极作用，产生了深远影响。

王大珩先生也是杰出的教育家，他是中国光学、仪器仪表和计量科教事业的奠基人之一。在他领导的研究所以及他创办的院校，为国家培养了一大批科技英才。

王大珩先生的学术思想和他对国家光学、仪器、计量科学事业的贡献以及对国家科技发展战略的重大建议等等，都将载入史册。

王大珩先生是《光谱学与光谱分析》首席顾问，《光谱学与光谱分析》的刊名是王老所起，《光谱学与光谱分析》在 1982 年向中国科协报批时，困难很大，于是王老就亲自找时任中国科协主席的周培源院士汇报情况并得到批复；以后凡是有光谱的学术会议，王老总是力争出席，他总是提出一些发展方向性的问题，指导《光谱学与光谱分析》不断前进。王老时刻关注科技期刊，他联合几位老院士为我国科技期刊的困境向上级呼吁，最后得到国家财政部、国家科委的认可和支持，由中国科协执行的资助科技期刊政策。

我国科技界将永远怀念王老，他为祖国、为人民辛劳了一辈子，王老您安息吧！

(摘自《中国光学期刊网》)