# 一种改进的 N-FINDR 高光谱端元提取算法

赵春晖 齐 滨* 王玉磊
(哈尔滨工程大学信息与通信工程学院 哈尔滨 150001)

**摘 要**：光谱端元提取是对高光谱数据进一步分析的重要前提。在各种端元提取算法中，N-FINDR 算法因其全自动和选择效果较好等优点受到了广泛的关注。然而样本的排序对该算法的端元提取会造成一定影响，并且传统 N-FINDR 算法需要根据端元的个数进行降维处理，从而限制了该算法的应用。实际高光谱数据中存在的同一地物在高维空间中非紧密团聚现象也对端元提取增加了难度。为此该文提出改进的算法停机准则和数据特征预处理方法，并使用支持向量机对提取到的端元进行二次提取。实验结果表明，改进的停机准则进一步增加了由端元向量组组成的凸体体积。数据特征预处理和基于支持向量机的二次端元提取分别提升了数据的可分性和提取到端元的精度。

**关键词**：图像处理；光谱端元提取；N-FINDR 算法；改进的停机准则；特征预处理；支持向量机
**中图分类号**：TP751.1 **文献标识码**： A **文章编号**：1009-5896(2012)02-0499-05
**DOI**: 10.3724/SP.J.1146.2011.00575

# An Improved N-FINDR Hyperspectral Endmember Extraction Algorithm

Zhao Chun-hui Qi Bin Wang Yu-lei
(*College of Information and Communication Engineering*, *Harbin Engineering University*, *Harbin* 150001, *China*)

**Abstract**: Spectral endmember extraction is an important pretreatment for the further analysis of hyperspectral data. Regarding many kinds of endmember extraction algorithms, N-FINDR algorithm is widely utilized for its full-automation and better endmember extraction performance. However, the order of the samples has a certain effect on the endmember extraction, and traditional N-FINDR algorithm also needs to reduce the dimensionality based on the number of the endmembers, which will limit its application. In the actual hyperspectral data, the incompact clustering of the same species presented in the high dimensional space also increases the difficulty of endmember extraction. So this paper proposed an improved stop rule and the pretreatment of the features, and utilizing Support Vector Machine (SVM) to conduct the second endmember extraction. Experiments show that the improved stop rule further increased the volume of the convex polyhedron composed of the endmembers. The pretreatment of the features and the second SVM endmember extraction increase the separability of the data and the precision of the extracted endmembers respectively.

**Key words**: Image processing; Spectral endmember extraction; N-FINDR algorithm; Improved stop rule; Pretreatment of the features; Support Vector Machine (SVM)

## 1 引言

随着图像处理技术的发展，高光谱图像由于其丰富的波谱信息得到越来越广泛的应用。“端元”被定义为数据中代表类别特征的理想化纯数据[1]。提取高光谱数据中的纯光谱技术称为光谱端元提取，是高光谱数据解混和其他高光谱数据分析实施的前提。近年来，多种端元提取算法相继发展起来，代表性的算法有 N-FINDR 算法[2]，SGA 算法，像素纯度索引，迭代误差分析，等等。N-FINDR 算法是基于搜寻凸体体积最大化的经典端元提取算法，由于其无参数，选择效果好而备受欢迎，但该算法需要降维处理，可能会忽略图像中存在的小目标端元，造成提取的端元不完整。目前已有一些文献对 N-FINDR 算法提出改进，文献[3]引入虚拟维数的概念来确定提取端元的数目，并用迭代误差分析法确定初始端元的选择。文献[4]提出基于线性最小二乘支持向量机的距离计算方法，取代原始 N-FINDR 算法中凸体体积的计算。文献[5]提出利用一个与数

2011-06-14 收到，2011-09-29 改回
国家自然科学基金(61077079)，教育部博士点计划基金(20102304110013)和哈尔滨市杰出学术带头人基金(2009RFXXG034)资助课题
*通信作者：齐滨 qibinwinter@gmail.com

据维数无关的高维单形体体积公式计算凸体体积，避免了 N-FINDR 算法需要降维处理的要求。

由于提取到的端元被作为其他高光谱数据处理方法的先验知识，因此提取到的端元是否能代表这一类地物的特征，对后续处理算法的精度起着至关重要的作用。为此本文在深入分析凸体体积计算公式和支持向量机(SVM)[6−8]模型的基础上，提出新的 N-FINDR 算法停机准则，高光谱数据特征预处理以及基于支持向量机的端元二次提取算法。

## 2 N-FINDR 端元提取算法

高光谱图像的所有像素，在高维空间形成一个凸体，每一个端元对应于凸体的一个顶点，因此端元提取过程转化为提取相应的凸体顶点，使得由这些顶点组成的凸体体积最大，而非端元点像素则存在于凸体的内部、棱上或面上。$p$ 个像素 $\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_p$ 张成的凸体体积为

$$V(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_p)=\frac{\left|\det\begin{bmatrix}1 & 1 & \cdots & 1\\ \boldsymbol{r}_1 & \boldsymbol{r}_2 & \cdots & \boldsymbol{r}_p\end{bmatrix}\right|}{(p-1)!} \tag{1}$$

其中 $\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_p$ 为 $\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_p$ 降维至 $(p-1)$ 维后所对应的向量，$\det[\cdot]$和$||$分别为行列式算子和绝对值算子，降维的目的是保证行列式的计算得以实施。篇幅所限，具体的迭代过程参见文献[3]。文献[5]提出的凸体体积计算公式为

$$V(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_p)=\frac{1}{p!}\sqrt{\left|\boldsymbol{A}_p^{\mathrm{T}}\boldsymbol{A}_p\right|} \tag{2}$$

其中 $\boldsymbol{A}_p=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_p)$，由于 $\boldsymbol{A}_i^{\mathrm{T}}\boldsymbol{A}_i(i=1,\cdots,p)$ 一定是方阵，所以式(2)对任何维数的高光谱数据都适用，不需要降维处理。

## 3 支持向量机原理

支持向量机的基本思想是寻求一个最优分类超平面，使得两类样本能够尽可能地被分开，并且这两类物体间的距离尽可能地远。更多关于支持向量机的理论请参考文献[9] 。支持向量机的判别函数为

$$f(x)=\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x)+b^*=(w^*\cdot x)+b^* \tag{3}$$

其中 $w^*=\sum_{i=1}^{n}\alpha_i x_i y_i$, $0\le\alpha_i\le C$ 。

## 4 改进的端元提取算法

### 4.1 改进的 N-FINDR 算法停机准则

原始 N-FINDR 算法将数据降维后，随机选取一组像素作为初始端元，计算以这组端元作为顶点的凸体体积，将图像中的每个像素依次替换现有端元向量组中的端元，并用式(1)计算每次端元替换后的体积。当替换后的体积大于替换前的体积时，用这个像素替换向量组中对应的端元，组成新的端元向量组，并更新当前体积值，重复上述过程，直到数据中所有的像素均被计算一次，最终的端元向量组被认为是该图像的端元。设高光谱图像的端元数为 $p$，有 $N$ 个像素点，原始算法的停机准则为图像中的每个像素点都遍历地进行 $p$ 次体积运算，总的体积运算次数为 $N\times p$ 。这种停机准则的一大弊端在于先进行体积计算的端元失去与最可能成为端元的像素进行体积计算的机会，即若某个像素为端元像素，但是较早的被用来进行体积运算，并在体积运算中与非端元像素组成的凸体体积没有端元替换前大，因而没有被选为端元像素，那么该像素将失去成为端元的可能。然而若将数据中的像素点遍历的计算所有端元组合可能，那么将计算 $C_N^p$ 次，显然这个计算量过于庞大。为此本文提出改进的 N-FINDR 算法停机准则，对于每一次整体像素遍历体积计算，将是否有像素点更替端元作为停机准则，即若在本次所有像素遍历体积计算中，端元被更替，那么所有的像素都再进行一次遍历的体积运算，直到没有端元被更替时算法停止。实验显示，停机准则改进后，得到的凸体体积比原始 N-FINDR 算法进一步增大，为此付出的代价是增加了计算时间，但通常情况下所有像素点搜索两次均能收敛到最优解。

### 4.2 特征预处理

实际高光谱图像中，由于噪声的影响及邻近地物类别的光谱相似性，同一类别像素在高维空间并没有呈现理想化的紧密团聚。若采用原始 N-FINDR 算法，由于降维的维数受端元个数的限制，有时不能正确地提取每个地物的端元。若采用文献[5]提出的高维单形体体积计算公式，由于过多的冗余特征限制，形成的凸体并不能很好地凸显不同地物间的差别，也限制了端元的提取。为此本文提出对原始高光谱图像进行特征预处理，选取方差变化较大的特征进行体积计算，方差计算公式如下式

$$\sigma_i^2=\frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\overline{x}_i)^2,\quad i=1,\cdots,N \tag{4}$$

其中 $N$ 为原始高光谱数据的特征(波段)个数。选取方差变化较大的特征，组成特征空间。方差变化较大的特征即为不同地物差异性较大的波段，这样选择出来的特征组成的特征空间将使不同地物之间的距离最大，增加了不同地物在特征空间中的可分性，提高了准确提取每个地物类别端元的精度。

### 4.3 基于支持向量机的端元二次提取

原始 N-FINDR 算法和文献[5]提出的体积计算方法均将体积运算作为端元提取的唯一准则，但在

实际高光谱图像中，由于噪声的影响，提取的端元往往存在于真实端元附近，为此本文提出将支持向量机的高维数据空间变换理论引入端元提取中，即某一类的端元应该在空间变换后与其他类别的端元有着最大的距离，借此提高端元提取的精度。支持向量机使用核函数对样本数据进行非线性映射，也增加了样本在高维空间的可分性。$p$ 个端元组成的高维空间凸体 $\sum_p$ 的体积为

$$V\left(\sum_p\right)=\frac{1}{p-1}V\left(\sum_{p-1,i}\right)h_i,\ i=1,2,\cdots,p \qquad (5)$$

其中 $\sum_{p-1,i}$ 为缺少第 $i$ 个顶点，由 $(p-1)$ 个顶点组成的凸体，$h_i$ 为第 $i$ 个顶点到 $\sum_{p-1,i}$ 的距离，即凸体 $\sum_p$ 对应于第 $i$ 个顶点的高。若端元个数为 $p$，对于 N-FINDR 算法选取出来的端元，在整个数据空间中为每个端元选取 $k$ 个最近邻域像素(欧氏距离)，形成 $p$ 个类别，构造 $p$ 个支持向量机，设每个支持向量机将某一类别定义为+1，而将其他类别定义为-1。对于某个+1 类别，选取这一类别中与其他-1 类别样本点构成凸体体积最大的样本点为该类别的端元向量，相应的体积计算公式为

$$V\left(\sum_{k(p-1)+1,i}\right)=\frac{1}{k(p-1)}V\left(\sum_{k(p-1),i}\right)h_i,\ i=1,2,\cdots,k \qquad (6)$$

其中 $\sum_{k(p-1)+1,i}$ 为支持向量机中，标号 $-1$ 的 $k(p-1)$ 个样本点加上标号+1 的第 $i$ 个样本点组成的凸体，$\sum_{k(p-1),i}$ 为标号-1 的 $k(p-1)$ 样本点组成的凸体。

为便于说明，以一个由 3 个端元组成的高维数据进行阐述。规定 3 个端元分别为 $A,B,C$，每个端元选取 3 个最近邻域像素，分别为 $A_1,A_2,A_3;B_1,B_2,B_3;C_1,C_2,C_3$ (图 1)。若要计算类别 1 中的某个样本点 $A_i,i=1,2,3$ 与 $B_1B_2B_3;C_1C_2C_3$ 组成凸体的体积，由于凸体 $\sum_{B_1B_2B_3C_1C_2C_3}$ 对于任意 $A_i$ 点，体积都是固定的，因此仅需要计算由 $A_i,i=1,2,3$ 到凸体 $\sum_{B_1B_2B_3C_1C_2C_3}$ 的高 $h_i$ 即可。又因为点 $A_i$ 到最优超平面的距离 $A_iO_i$ 正比于点 $A_i$ 到 $\sum_{B_1B_2B_3C_1C_2C_3}$ 的高，因此原始的体积计算问题转化为求解点 $A_i,i=1,2,3$ 到最优超平面的最大值问题。

点 $A_i$ 到最优超平面的距离为

$$A_iO_i=\sum_{j=1}^{k\cdot p}\beta_j^*y_jK(x_j,x_i)+b^*,\ i=1,2,\cdots,k \qquad (7)$$

对于第 1 个类别，选取的端元点为

$$A=\left\{A_i \mid A_iO_i=\max(A_jO_j),j=1,2,\cdots,k\right\} \qquad (8)$$

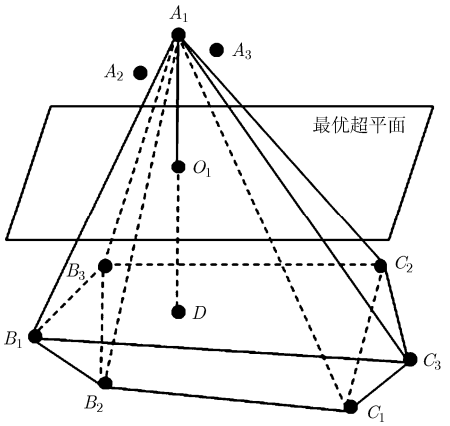


图 1 支持向量机的距离测算

以上算法的具体算法步骤如下：

步骤 1　设高光谱图像为 $\{\boldsymbol{x}_i\},\boldsymbol{x}_i\in R^N$，端元个数为 $p$，特征提取 $M$ 维。计算每一波段方差 $\sigma_i^2$，从原始图像 $\{\boldsymbol{x}_i\}$ 中选取方差最大的 $M$ 个波段组成提取特征后的数据 $\{\boldsymbol{z}_i\}$。从 $\{\boldsymbol{z}_i\}$ 中随机选取 $p$ 个向量组成初始端元向量组 $\{\boldsymbol{e}_i\}$，计算由初始端元组成的体积。设 EM 为端元更新次数，并将 EM 置零。

步骤 2　从 $\{\boldsymbol{z}_i\}$ 中依次选取向量，逐个替换端元向量组 $\{\boldsymbol{e}_i\}$ 位置，计算相应的体积，若体积变大，则用新计算的体积替换原有体积，用向量 $\boldsymbol{z}_i$ 替代相应端元 EM=EM+1，直到遍历 $\{\boldsymbol{z}_i\}$ 中的所有向量。

步骤 3　判断 EM 是否为零，若为零，则遍历所有向量后，没有向量可以使凸体体积进一步增大，转至步骤 4，否则将 EM 置零，重复步骤 2。

步骤 4　将获得的 $p$ 个端元 $\{\boldsymbol{e}_i\}$ 选取其在原始空间 $\{\boldsymbol{x}_i\}$ 中对应的向量组成端元向量组。每个端元在 $\{\boldsymbol{x}_i\}$ 中选取 $k$ 个最近邻域，构成 $p$ 个支持向量机，每个支持向量机选取某一类别为+1 其他类别为-1，将距离-1 类别最远的+1 类别向量作为最终端元向量选取出来，算法结束。

## 5　仿真实验

将修改停机准则后的算法定义为 SN-FINDR，使用支持向量机进行端元二次提取的算法定义为 SVMN-FINDR。文献[5]提出的体积计算方法定义为 TN-FINDR；第 1 组实验重点比较 SN-FINDR 和 SVMN-FINDR 所选择端元形成的凸体体积的大小，以及所选出的端元与理论点间的误差。实验用 2 维人工合成数据(图 2)，点(-15,0), (0,20), (15,0)为理论端元，被随机产生的归一化系数合成 10000 个点，附加 0 均值，方差为 2 的高斯噪声。图中显示改进停机准则后，N-FINDR 算法提取的端元向量组中有一个端元被更替过，得到的凸体体积较原始 N-

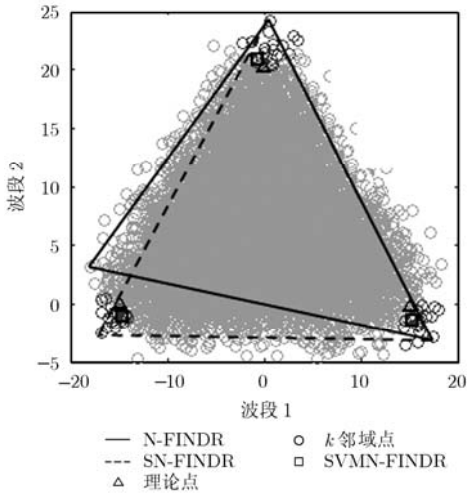


图 2 实验 1 中合成数据

FINDR 算法提高 7.47%。在 SN-FINDR 提取到的端元基础上附加支持向量机进行二次端元提取，设 $k$ 邻域为 20，从表 1 中可以看出，本文算法提取到的端元将与理论点间的均方根误差降低为 N-FINDR 算法的 1/4，搜索到的端元更接近理论值。

**表 1 实验 1 中不同算法搜索到端元组成的体积及误差比较**

| EM 提取方法 | N-FINDR | SN-FINDR | SVMN-FINDR |
|---|---|---|---|
| 体积 | 141.8671 | 152.4598 | 109.8215 |
| 误差 | 4.0360 | 3.6260 | 1.0944 |

第 2 组实验旨在比较 N-FINDR, TN-FINDR 和 SVMN-FINDR 算法提取出来的端元用于线性光谱混合模型时的解混效果。为获得监督评价，采用人工合成数据进行测算。随机产生两个 1×12 的向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$，分别附加均值为 0，方差为 0.5 的随机噪声，生成两个 1200×12 的矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$。将矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 按照不同的比例合成矩阵 $\boldsymbol{C}$，合成的规则如下：$\boldsymbol{C}$ 的 1 至 200 数据由 100%的 $\boldsymbol{A}$ 的 1 至 200 数据组成；$\boldsymbol{C}$ 的 201 至 400 数据由 80%的 $\boldsymbol{A}$ 的 201 至 400 数据合成 20%的 $\boldsymbol{B}$ 的 201 至 400 数据组成，以此类推，$\boldsymbol{A}$ 的合成比例由 100%递减到 0%，每次递减 20%，相应的 $\boldsymbol{B}$ 的合成比例由 0%递增到 100%，每次递增 20%。特征预处理过程中，特征空间维数为 5，支持向量机的 $k$ 邻域样本点为 20，解混算法采用直接全约束最小二乘法。各类别的解混分量如图 3 所示，相应的解混误差如表 2 所示，从表中可以看出，本文提出的算法较 N-FINDR 和 TN-FINDR 提取端元的解混误差要小，解混结果更接近理论值。

**表 2 实验 2 中不同算法搜索到端元的解混误差比较**

| EM 提取算法 | N-FINDR | TN-FINDR | SVMN-FINDR |
|---|---|---|---|
| 解混误差 | 0.0057 | 0.0024 | 0.0021 |

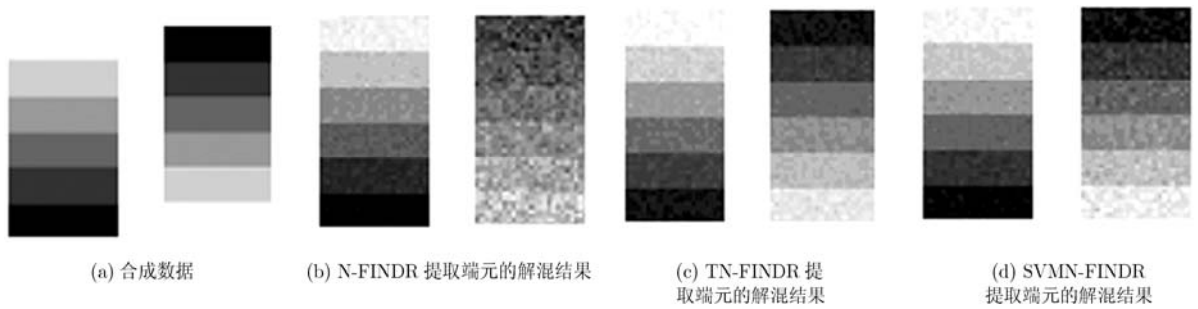


图 3 各类别的解混分量

第 3 组实验对高光谱图像中，由于噪声影响，造成同一类物体在高维空间中呈现不紧密聚类现象的端元提取进行研究。选取 AVIRIS 高光谱图像对算法进行评价，由于 AVIRIS 图像从高空拍摄，像素的空间分辨率较低(20 m×20 m)，像素混合概率很大，并且有些地物仅含有几十个像素，不能完全代表该地物的光谱属性，因此选取图像中像素个数较多的 9 种地物作为样本测试不同算法的端元提取性能。表 3 给出了 9 种不同地物的名称以及每种地物所包含的像素个数。特征预处理过程中，特征空间的维数为 10，支持向量机的 $k$ 邻域为 20。分别应用 N-FINDR, TN-FINDR 和 SVMN-FINDR 进行端元提取，提取结果如表 4 所示。从表 4 中可以看出 N-FINDR 和 TN-FINDR 都仅能提取到来自 6 种不同地物的端元，SVMN-FINDR 算法提取到来自 8 种不同地物的端元，唯一没有提取到的端元为地物 6(大豆 1)。由于 AVIRIS 图像拍摄于 6 月，正处于大豆的生长早期，地物 6，地物 7，地物 8 所反映的光谱地物极其相似，在这些客观原因下，造成该类别地物的端元提取存在一定误差。

**表 3 实验 3 中各地物名称及像素个数**

| | 地物 1 | 地物 2 | 地物 3 | 地物 4 | 地物 5 | 地物 6 | 地物 7 | 地物 8 | 地物 9 |
|---|---|---|---|---|---|---|---|---|---|
| 地物名称 | 玉米 1 | 玉米 2 | 牧场 | 灌木 | 干草 | 大豆 1 | 大豆 2 | 大豆 3 | 乔木 |
| 像素个数 | 1434 | 834 | 497 | 747 | 489 | 968 | 2468 | 614 | 1294 |

**表 4 实验 3 中不同算法提取的端元分布**

| | 端元 1 | 端元 2 | 端元 3 | 端元 4 | 端元 5 | 端元 6 | 端元 7 | 端元 8 | 端元 9 |
|---|---|---|---|---|---|---|---|---|---|
| N-FINDR | 地物 9 | 地物 5 | 地物 1 | 地物 8 | 地物 7 | 地物 7 | 地物 7 | 地物 4 | 地物 8 |
| TN-FINDR | 地物 3 | 地物 7 | 地物 8 | 地物 5 | 地物 4 | 地物 8 | 地物 9 | 地物 9 | 地物 7 |
| SVMN-FINDR | 地物 9 | 地物 8 | 地物 1 | 地物 7 | 地物 3 | 地物 8 | 地物 2 | 地物 4 | 地物 5 |

## 6 结论

本文改进了 N-FINDR 算法的停机准则，修改停机准则后，搜索到的凸体体积较原方法进一步增大。提出了高光谱数据特征预处理，选取方差较大的波段进行体积计算，增强了不同物体在高维空间的的可分性。引入了支持向量机的多类分类模型，在原始搜索到的端元基础上，进行二次端元提取，提高了端元提取的精度。本文提出的 3 种端元提取改进方法可单独使用，也可以任意复选方式组合，在实际情况中可以根据需要进行选择。

## 参 考 文 献


[1] Schowengerdt A R. Remote Sensing: Models and Methods for Image Processing. San Diego: Academic Press, 1997: 120–121.

[2] Winter E M. N-finder: an algorithm for fast autonomous spectral endmember determination in hyperspectral data. SPIE Conference on Imaging Spectrometry V, Denver, Colorado, 1999, 3753: 266–275.

[3] Plaza A and Chang C I. An improved N-FINDR algorithm in implementation. Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI, Bellingham, WA, 2005, 5806: 298–306.

[4] 王立国, 张晶, 刘丹凤, 等. 从端元选择到光谱解混的距离测算方法. 红外与毫米波学报, 2010, 29(6): 471–475.
Wang Li-guo, Zhang Jing, Liu Dan-feng, *et al.*. Distance measurement based methods from endmember selection to spectral unmixing. *Journal of Infrared and Millimeter Waves*, 2010, 29(6): 471–475.

[5] 耿修瑞, 赵永超, 周冠华. 一种利用单形体体积自动提取高光谱图像端元的算法. 自然科学进展, 2006, 16(9): 1196–1200.
Geng Xiu-rui, Zhao Yong-chao, and Zhou Guan-hua. An automatic hyperspectral image endmember extraction algorithm utilized the volume of simplex. *Nature Science Process*, 2006, 16(9): 1196–1200.

[6] 汪延华, 田盛丰, 黄厚宽. 特征加权支持向量机. 电子与信息学报, 2009, 31(3): 514–518.
Wang Yan-hua, Tian Sheng-feng, and Huang Hou-kuan. Feature weighted support vector machine. *Journal of Electronics & Information Technology*, 2009, 31(3): 514–518.

[7] 陶剑文, 王士同. 具有磁场效应的大间隔支持向量机. 电子与信息学报, 2011, 33(5): 1055–1061.
Tao Jian-wen and Wang Shi-tong. Maximal margin support vector machine with magnetic field effect. *Journal of Electronics & Information Technology*, 2011, 33(5): 1055–1061.

[8] 田江, 顾宏. 孤立点一类支持向量机算法研究. 电子与信息学报, 2010, 32(6): 1284–1288.
Tian Jiang and Gu Hong. Outlier one class support vector machines. *Journal of Electronics & Information Technology*, 2010, 32(6): 1284–1288.

[9] Li Kun-lun, Luo Xue-rong, and Jin Ming. Semi-supervised lerning for SVM-KNN. *Journal of Computers*, 2010, 5(5): 671–678.



赵春晖：　男，1965 年生，博士，教授，博士生导师，研究方向为非线性信号处理及图像处理.

齐　滨：　男，1985 年生，博士生，研究方向为高光谱图像分类.

王玉磊：　女，1986 年生，博士生，研究方向为高光谱图像检测.