文章编号:1004-4213(2011)06-0847-5

基于优选样本的 KPCA 高光谱图像降维方法

王瀛, 郭雷, 梁楠
(西北工业大学 自动化学院,西安 710129)

摘 要:降维是高光谱图像常用的预处理手段,而核主成份分析通过非线性映射能够挖掘数据的高阶统计特性,是目前较常使用的特征提取方法.本文提出了一种基于优选样本的核主成份分析高光谱图像降维方法,算法挑选参与核主成份分析运算的样本时兼顾整幅高光谱图像的统计特性,以与全图能量分布相近的最小样本集为最终选择样本.本算法由 IDL7.0 实现,并在实际高光谱图像 Cuprite 上进行实验.结果表明,在大幅缩短运算时间的同时,降维效果优于传统的核主成份分析方法.

关键词:高光谱图像; 核主成份分析; 非线性映射; 迹; 降维
中图分类号:TP751.1 **文献标识码**:A **doi**:10.3788/gzxb20114006.0847

0 引言

高光谱遥感图像可以在大量连续的光谱波段上反映出成像区域所反射、吸收以及放射的电磁波能量[1],包含极其丰富的光谱信息和空间信息,因此在地物识别[2]、分类[3]、小目标提取和探测[4]等方面都有着不可比拟的优势.通常情况下,一幅高光谱遥感图像包含几十或者上百个光谱波段的信息[5],丰富的光谱信息为各种应用提供了保障,同时,由于光谱之间的高度相关性,以及光谱数远远大于地物数的客观现实,也增加了数据处理的难度,甚至会放大结果的不确定性,比如 Hughes 现象[6],因此高光谱遥感图像的降维操作已经成为很重要的预处理手段.核主成份分析(Kernel Principal Components Analysis, KPCA)[7]是目前应用较为广泛的一种高光谱图像降维方法.作为传统的主成份分析(Principal Components Analysis, PCA)[8]算法的非线性形式,KPCA 兼顾考察了高光谱图像的非线性和高阶统计特性,同时具备较小运算量,是处理遥感图像的有力工具.针对一幅具体高光谱遥感图像来说,在使用 KPCA 做降维操作时,核函数的选择[9]、参量的设定[10]以及参与运算样本集的选取,是整个流程中的重要部分,直接关系到最终结果.

本文提出了一种确定参与 KPCA 运算样本集的方法,考虑了整幅图像的内在特性,并提出了一种从 KPCA 处理角度衡量样本集优劣的指标.实验证明,以本文算法抽取的样本集作为输入进行的 KPCA 操作,运算量更低,与具体图像更为适配且能达到较好的降维效果.

1 KPCA 算法原理

为了克服传统 PCA 方法不能反映数据的非线性特征以及只考虑二阶统计特性的缺陷[11],Schökopf 等人提出了 KPCA 算法[12],KPCA 算法可以看作是 PCA 算法的非线性版本.KPCA 通过非线性映射将原数据映射到高维甚至是无限维的非线性空间,然后在非线性空间中对数据进行 PCA 操作,兼顾考虑了高光谱遥感图像的非线性特性和高阶统计特性.

给定一组高光谱数据向量集$\{x_1,x_2,\cdots,x_M\}$,其中$\{x_k\in R^N,k=1,\cdots,M\}$,设存在非线性映射$\Phi$将原始高光谱数据映射到特征空间$F$

$$\{\Phi:R^N\rightarrow F,x\rightarrow\varphi(x)\} \tag{1}$$

其中特征空间F是再生核 Hilbert 空间(Reproducing Kernel Hibert Space, RKHS)[13],可能是高维甚至无限维,KPCA 算法就是通过选择合适的核函数在原数据空间中完成特征空间中的主成分分析.

由于对特征空间F的维数没有限制,因此映射后数据集$\{\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_M)\in F\}$的协方差

基金项目:国家自然科学基金(No. 60802084)资助
第一作者:王瀛(1976-),男,高级实验师,博士研究生,主要研究方向为高光谱遥感图像处理. Email:wangying@henu.edu.cn
导 师:郭雷(1956-),男,教授,博导,主要研究方向为模式识别与图像处理. Email:lguo@nwpu.edu.cn
收稿日期:2010-11-27;**修回日期**:2011-03-01

矩阵

$$\{\bar{\boldsymbol{C}}=\frac{1}{M}\sum_{i=1}^{M}\phi(x_i)\phi(x_i)^{\mathrm{T}}\} \tag{2}$$

可能是超高阶或者无限阶. 通过选取合适的核函数 k，将对角化 $\bar{\boldsymbol{C}}$ 的操作转化为求核矩阵 $\boldsymbol{K}$ 的特征问题. 核矩阵 $\boldsymbol{K}$ 的形式为

$$\boldsymbol{K}=\begin{Bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_M) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_M) \\ \vdots & \vdots & \cdots & \vdots \\ k(x_M,x_1) & k(x_M,x_2) & \cdots & k(x_M,x_M) \end{Bmatrix} \tag{3}$$

解决核矩阵 $\boldsymbol{K}$ 的特征值问题

$$\lambda\alpha=\boldsymbol{K}\alpha \tag{4}$$

将特征向量标准化

$$\|\alpha\|_2=1/\lambda \tag{5}$$

按特征值降序排列，相应的特征向量就是核主成分，计算原数据在核主成分上的投影需再次利用核函数

$$\{\phi_{kpc}^{j}(x)=\sum_{i=1}^{M}\alpha_i^j k(x_i,x)\} \tag{6}$$

求特征空间 F 中的协方差矩阵 $\bar{\boldsymbol{C}}$ 时需要将样本去均值化，同样可以通过处理核矩阵 $\boldsymbol{K}$ 完成

$$\boldsymbol{K}_c=\boldsymbol{K}-1_M\boldsymbol{K}-\boldsymbol{K}1_M+1_M\boldsymbol{K}1_M \tag{7}$$

式中$(1_M)_{ij}=1/M$.

KPCA 算法中采用的核函数是在原数据空间处理非线性映射数据的保障，是算法的核心部分. 目前常用的核函数有多项式核函数

$$k(x,y)=(\langle x,y\rangle+d)^p \tag{8}$$

和高斯核函数

$$k(x,y)=\left(\exp\left(-\frac{\|x-y\|^2}{2\delta^2}\right)\right)^p \tag{9}$$

2 基于优选样本集的 KPCA 算法

在 KPCA 算法中，核函数选择、参量设定以及参与运算的样本集是决定算法最终效果的关键因素. 不同的核函数和参量隐性决定了输入空间到特征空间的映射，而好的样本集能更接近一幅遥感图像的本质. 实践证明，在核函数和参量固定的前提下，参与运算的样本集不同，KPCA 输出的结果不尽相同，相应的，对后续的降维、分类、小目标探测等工作都会产生深远的影响.

2.1 样本集对 KPCA 的影响

通过分析 KPCA 算法可以发现，KPCA 实际上是在特征空间 F 中进行的主成分分析，由于映射关系 Φ 的非线性特征，因此特征空间 F 中的主成分分析，相对于传统的 PCA，可以挖掘数据的非线性本质和高级统计特性. 特征空间 F 可能是超高维或者无限维，直接对角化 F 上协方差矩阵 $\bar{\boldsymbol{C}}$ 是一项很困难的工作，而利用矩阵的奇异值分解和再生核 Hilbert 空间的性质，可以将对 $\bar{\boldsymbol{C}}$ 的操作转化为对核矩阵 $\boldsymbol{K}$ 的操作. 设用于产生核矩阵 $\boldsymbol{K}$ 的样本数为 m，由公式(3)可以看出，核矩阵 K 的阶数为 m×m，而一副高光谱遥感图像可能包含几十万个光谱向量，因此如何选择合适的样本产生核矩阵 $\boldsymbol{K}$ 是一个值得探讨的问题. 一个好的样本集可以用较少量的样本尽可能地代表整幅图像的特征，而由此产生的核矩阵 $\boldsymbol{K}$ 以及后续的 KPCA 降维就能获得较好的结果. 从计算量代价的角度来看，较少的样本数代表较低的计算量，所以，一个好的样本集应该包含较少的样本数量、尽量反应整幅图像的特征、产生的核矩阵 $\boldsymbol{K}$ 以及 KPCA 的结果应该大致接近以整幅图像全部光谱向量为样本集产生的结果.

2.2 样本集的考量标准

主成分分析通过对角化数据集的协方差矩阵寻找新的彼此正交的性质指标，原始数据在这些少量的新性质指标上的投影可以保留大部分能量. KPCA 通过核矩阵 $\boldsymbol{K}$ 隐性地完成特征空间 F 上的主成分分析，组成核矩阵 $\boldsymbol{K}$ 的样本向量集是全部光谱向量的一部分，从能量的角度来看，以样本集和全部光谱向量之间能量比来作为考量样本集优劣的标准是可行的.

设整幅高光谱图像的协方差矩阵为 $\boldsymbol{C}$，样本集的协方差矩阵 $\boldsymbol{C}_m$，二者之间的能量比定义为

$$E_{\mathrm{ratio}}=\frac{\sum\lambda_{\boldsymbol{C}_m}}{\sum\lambda_{\boldsymbol{C}}}\times100=\frac{\mathrm{tr}(\boldsymbol{C}_m)}{\mathrm{tr}(\boldsymbol{C})}\times100 \tag{10}$$

式(10)利用了协方差矩阵的性质和矩阵迹与特征值之间的关系.

2.3 样本集的选择算法

本文依照如下步骤选择样本集：

1)设样本集包含的样本数为 m，整幅图像的光谱向量总数为 M.

2)计算整幅高光谱图像的均值向量 x_{mean}.

3)计算每个光谱向量和 x_{mean} 之间的欧式距离 x_i^{Euc}.

4)确定划分跨度

$$d=\frac{x_{\max}^{\mathrm{Euc}}-x_{\min}^{\mathrm{Euc}}}{n} \tag{11}$$

5)统计 $x_i^{\mathrm{Euc}}\in[x_{\min}^{\mathrm{Euc}},x_{\min}^{\mathrm{Euc}}+j\times d)$ 对应的光谱向量数，记为 NUM_j，其中 $j=1,\cdots,n$.

6)求出针对于距离区间 j 的比率

$$\mathrm{THR}_j=\frac{\mathrm{NUM}_j}{M} \tag{12}$$

7)计算在每一距离区间应取得的样本数

$$\mathrm{num}_j = m \times \mathrm{THR}_j \tag{13}$$

8)利用某种挑选原则，在落入距离区间 j 的光谱向量中挑选相应 num_j 个样本.

本文选择样本集的目标是用尽可能少的样本数近似反映整幅高光谱图像的能量分布，使得公式(10)给出的能量比接近于 100. 算法从均值向量 x_{mean} 开始，以欧式距离 x_i^{Euc} 作为光谱向量和均值向量相似度的测量，然后将整个距离范围分为 n 个距离区间，以落入某距离区间光谱向量数和整幅图像光谱向量总数的比值作为该区间在最终样本集所占的比率，最后根据事先设定的样本集数目结合各个距离区间的比率分区间挑选样本. 本算法的优点在于，首先是以能量比做为样本集的考量标准，因为无论主成分分析或者 KPCA 算法的目标都是尽量保持能量；其次，考虑了光谱间的相似度，使所选的样本能够包含更广的光谱范围；最后，依据能量比 E_{ratio}，可以反馈回来确定样本数 m 和距离区间数目 n，优化样本集.

3　实验与分析

实验采用 ENVI 软件自带的 AVIRIS 于美国内华达州获取的 Cuprite 图像，空间分辨率为 400×350，包含 50 个波段，数据为经过大气校正的反射谱. 本节首先从各个方面比较传统 KPCA 采用随机法选出的样本集与本文提出的优选法选出的样本集之间的优劣. 然后用不同的样本集产生核矩阵 $\boldsymbol{K}$，在原始数据上运行 KPCA 算法. 比较不同样本集对 KPCA 算法结果的影响，并且评估实际降维效果.

本文以能量比 E_{ratio} 作为考量样本集优劣的标准，E_{ratio} 越接近 100，表示样本集从能量角度来讲越接近整幅高光谱图像，所使用的随机数由 IDL7.0 自带过程 RANDOM 产生，根据不同的种子将产生不同的随机数列，优选法在某一距离区间中也采用随机法挑选样本. 为了进行多方面的比较，考虑了三种情况：1)样本数固定为一个较小的值，比较不同随机数列下两种样本选择方法的优劣；2)样本数递增，随机数列固定，比较样本数增加对两种样本选择方法的影响；3)随机法产生的大样本集和优选法产生的小样本集之间的比较. 每种情况采用 75 个样本集进行对比. 结果如图 1，图中优选样本集与整幅图像的能量比曲线用实线表示，随机样本集的能量比曲线用虚线表示.

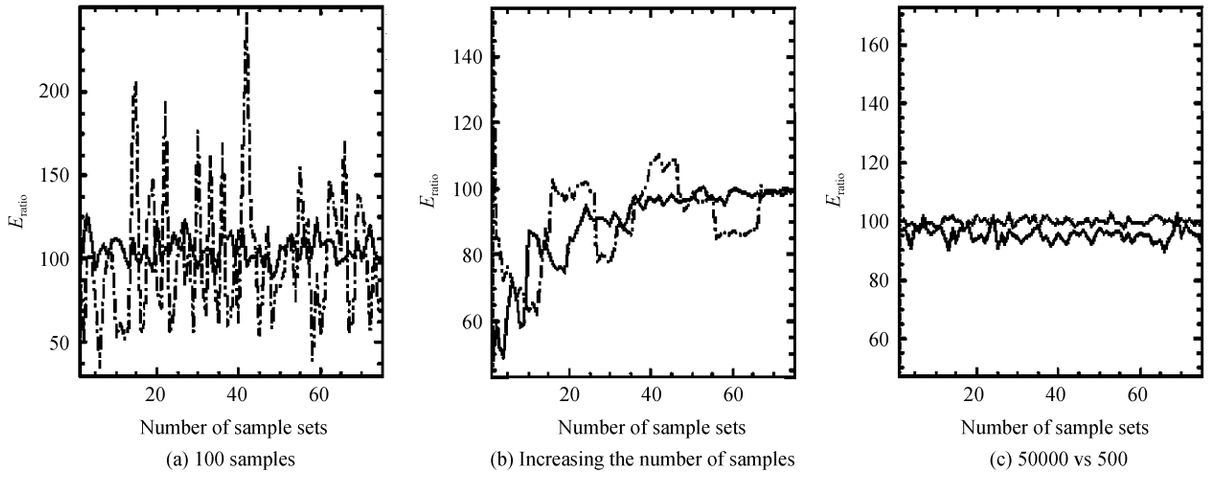


图 1　样本集比较

Fig. 1　Comparison of sample sets

从图 1(a)中可以看出在样本数较小的情况下(本例中样本数为 100)直接采用随机法生成的样本集所反映的能量与原图相差较大，而通过优选法产生的样本集，由于充分考虑了整幅图像的统计特性并强制选择不同的光谱向量作为样本，使得样本集在样本数很少的情况下，可以近似地在能量角度上反映整幅图像；图 1(b)的样本数最少为 10，最大为 70 000(整幅图光谱向量总数的一半)，以样本数递增排列，由结果可以看出，随着样本数的增加，两种方法生成的样本集对整幅图像的表现能力都在增强，而优选法产生的样本集展现出了较好的稳定性；图 1(c)对比了大样本数的随机法样本集(样本数 50 000)和小样本数的优选法样本集(样本数 500)，可以看出，随机法只有在样本数较大的前提下才能相对稳定地反映整幅图像的统计特性，而优选法用较小的样本数就可以做到这一点并且稳定性很好.

用优选法样本集(150 个样本)和随机法样本集(5 000 个样本)生成核矩阵 $\boldsymbol{K}$，并对高光谱图像 Cuprite 做 KPCA 操作，选用高斯核函数. 表 1 列出了前 7 个特征值以及对应的方差和累计方差，为了对比，PCA 的结果也在表中列出.

表 1 中的数据首先表明 KPCA 算法相对于

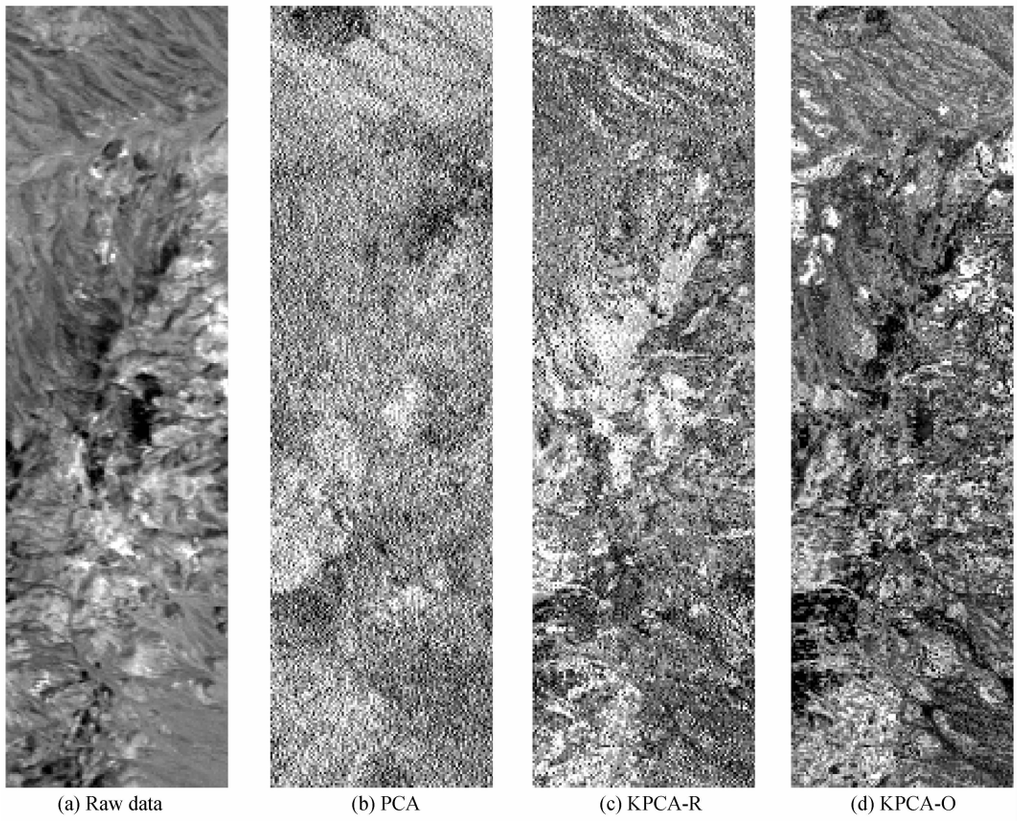
PCA 的优越性，在以往大量对高光谱图像进行的实验中，PCA 中数值大的特征值一般只有 3 个左右，极少数的特征值就包含了整幅图像将近 95% 的方差. 这是由于高光谱图像光谱之间的相关性比较复杂并且多为非线性，而 PCA 只是简单的依据二阶统计特性将原始数据做线性投影造成的现象. PCA 用于高光谱图像降维有很大的局限性，经过 PCA 降维后的数据能量过于集中导致大量细节的丢失，对以后的分类和小目标提取等工作都会产生不利的结果. KPCA 挖掘了高光谱数据间的非线性相关，克服了传统 PCA 算法固有的缺陷，从表 1 中的数据可以看出，KPCA 算法能够更“平缓”地分布原始数据的方差，可以用较多且有意义的主成分来描述原始数据. 表 1 中列出 KPCA(样本集不同)算法产生的前 7 个特征值所占的累积方差为 60% 和 90% 左右，而累积方差接近 100% 时则需要 50 和 20 左右个特征值. 比较采用不同样本集的 KPCA 算法结果可以得出结论，采用优选法小样本集的 KPCA 算法显著节省了计算时间和空间，结果无论从单个特征值所占方差还是累积方差分布的角度来看，都是可以接受的.

表 1 PCA、KPCA(随机法样本集)和优选法样本集的特征值、方差和累积方差

Table 1 Eigenvalues, variance and cumulative variance by the PCA, the KPCA with randomly selected sample set and the KPCA with optimized sample set

	PCA			KPCA with randomly selected sample set			KPCA with optimized sample set		
PCs	Eigenvalues	Variance	Cumulative variance	Eigenvalues	Variance	Cumulative variance	Eigenvalues	Variance	Cumulative variance
1	58 665.078 1	89.998 4	89.998 4	26.155 9	22.525 2	22.525 2	18.542 6	37.980 1	37.980 1
2	4 138.907 7	6.349 5	96.347 9	15.191 0	13.082 4	35.607 6	9.574 9	19.612 0	57.592 1
3	1 407.473 3	2.159 2	98.507 1	8.024 4	6.910 5	42.518 1	5.579 6	11.428 6	69.020 7
4	336.103 5	0.515 6	99.022 7	5.776 0	4.974 2	47.492 3	3.634 0	7.443 3	76.464 0
5	108.726 1	0.166 8	99.189 5	4.554 2	3.922 0	51.414 3	2.614 3	5.354 8	81.818 8
6	103.666 5	0.159 0	99.348 5	3.626 3	3.123 0	54.537 3	2.529 3	5.180 7	86.999 5
7	57.461 7	0.088 2	99.436 7	2.912 7	2.508 4	57.045 7	1.491 1	3.054 1	90.053 6

用表 1 中的三种算法对高光谱图像 Cuprite 做降维操作，降维后的维数设为 20. 其中第 10 维的灰度图如图 2. 图中截取了图像细节比较丰富的部分用于显示，通过和原图比较不难发现，PCA输出的



(a) Raw data (b) PCA (c) KPCA-R (d) KPCA-O

图 2 PCA、KPCA(随机法样本集)和 KPCA(优选法样本集)对原图降维后的第 10 维数据

Fig. 2 The 10th dimension data for the PCA, the KPCA with randomly selected sample set and the KPCA with optimized sample set

第 10 维基本由噪音构成，而两种 KPCA（使用随机样本集的 KPCA 输出简称 KPCA-R；使用优选样本集的 KPCA 输出简称 KPCA-O）算法的输出则包含了大量有意义的信息，为后续的各种应用提供了保障. 使用优选样本集 KPCA 算法在大大减轻计算量的前提下，输出结果令人满意.

4　结论

本文针对目前使用较广的 KPCA 在高光谱遥感图像降维方面的应用，提出了一种从原图像优选样本的方法. 相对于传统的随机法选择样本集，本文的优选法充分考虑了高光谱遥感图像的特性，包括光谱相关和能量分布，并以样本集和整幅图像之间的能量比作为衡量样本集对 KPCA 算法适用度的考量标准. 传统的随机法并没有考虑图像本身，因此只有在样本数很大的情况下才能产生较稳定的输出，运算量巨大. 本文提出的优选法结合原图像来选择样本集，确保由少数样本组成的样本集与整幅图像之间的能量关系维持在一个比较稳定的水平，所以由优选法选择的样本集更适合作为 KPCA 算法的输入来对高光谱遥感图像进行降维，实验数据证明了这一点.

参考文献

[1] SHAW G, MANOLAKIS D. Signal processing for hyperspectral image exploitation [J]. *IEEE Signal Processing, Magazine*, 2002, **19**(1):12-16.

[2] LANDGREBE D. Hyperspectral image analysis[J]. *IEEE Signal Processing Magazine*, 2002, **19**(1):17-28.

[3] JIA X, RICHARDS J A. Segmented principal components transformation for efficient hyperspectral remote sensing image display and classification [J]. *IEEE Transactions on Geosscience and Remote Sensing*, 1999, **37**(1):538-542.

[4] STEIN D W J, BEAVEN S J, HOFF L E, *et al*. Anomaly detection from hyperspectral imagery [J]. *IEEE Signal Processing Magazine*, 2002, **19**(1):58-69.

[5] FAUVEL M. Decision fusion for hyperspectral classification in hyperspectral data exploitation: theory and applications[M]. New Jersey: John Wiley & Sons, 2007.

[6] HUGHES G. On the mean accuracy of statistical pattern recog-nizers[J]. *IEEE Transactions on Information Theory*, 1968, **14**(1):55-63.

[7] SCHÖLKOPF B, SMOLA A, MÜLLER K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. *Neural Computa-tion*, 1998, **10**(5):1299-1319.

[8] JOLLIFFE I. Principal component analysis[M]. New York: Springer-Verlag, 1986.

[9] TAYLOR J S. Kernel methods for pattern analysis [M]. cristianini N. Cambridge: Cambridge University Press, 2004.

[10] FAUVEL M. Spectral and spatial methods for the classification of urban remote sensing data[D]. Reykjavik: Institute National Polytechnique de Grenoble, 2007.

[11] LANDGREBE D. Signal theory methods in multispectral remote sensing[M]. New Jersey: John Wiley & Sons, 2003.

[12] SCHÖLKOPF B, SMOLA A, MÜLLER K R. Kernel principal component analysis [J]. *Neural Computation*, 1999, **24**(10):1299-1319.

[13] RUIZ A, PEDRO E L T. Nonlinear kernel-based statistical pattern analysis [J]. *IEEE Transactions on Neural Networks*, 2001, **12**(1): 16-32.

A Dimensionality Reduction Method Based on KPCA with Optimized Sample Set for Hyperspectral Image

WANG Ying, GUO Lei, LIANG Nan

(*Institute of Automatic, Northwest Polytechnical University, Xi'an* 710129, *China*)

Abstract: Dimensionality reduction is a common preprocessing for hyperspectral image, and Kernel Principal Components Analysis(KPCA), as a common feature extraction method, makes use of nonlinear mapping to capture higher-order statistics of data. An optimization sample set algorithm, which is used in KPCA for dimensionality reduction of hyperspectral image was proposed. This algorithm picks sample set used in KPCA taking the statistics of the whole hyperspectral image into account simultaneously, and the minimum sample set with similar energy distribution of the full image is the final selection. The algorithm was implemented in IDL7. 0 and tested by using the real hyperspectral image from Cuprite. The experiment results show that the new algorithm is able to save computing time significantly and perform better than conventional KPCA in dimensionality reduction.

Key words: Hyperspectral image; Kernel Principal Components Analysis(KPCA); Nonlinear mapping; Trace; Dimensionality reduction