# UNSUPERVISED FEATURE EXTRACTION BASED ON A MUTUAL INFORMATION MEASURE FOR HYPERSPECTRAL IMAGE…

**Conference Paper** · July 2011

CITATIONS
4

READS
8

**3 authors:**

Dr Md Ali Hossain
UNSW Australia
**10** PUBLICATIONS **23** CITATIONS
SEE PROFILE

Mark R. Pickering
UNSW Canberra
**218** PUBLICATIONS **1,036** CITATIONS
SEE PROFILE

Xiuping Jia
UNSW Australia
**42** PUBLICATIONS **613** CITATIONS
SEE PROFILE

# UNSUPERVISED FEATURE EXTRACTION BASED ON A MUTUAL INFORMATION MEASURE FOR HYPERSPECTRAL IMAGE CLASSIFICATION

*Md. Ali Hossain, Mark Pickering, Xiuping Jia*

School of Engineering and Information Technology
The University of New South Wales, Australian Defence Force Academy, Canberra, ACT-2600

## ABSTRACT

Finding the most informative features from high dimensional space for reliable class data modeling is one of the most challenging problems in hyperspectral image classification. The problem can be address using two basic techniques: feature selection and feature extraction. One of the most popular feature extraction methods is Principal Component Analysis (PCA), however its components are not always suitable for classification. In this paper, we present a feature reduction method (MI-PCA) which uses a nonparametric mutual information (MI) measure on the components obtained via PCA. Supervised classification results using a hyperspectral data set confirm that the new MI-PCA technique provides better classification accuracy by selecting more relevant features than when using either PCA or MI on the original data.

***Index Terms***— Hyperspectral image, mutual information, nonparametric feature extraction, principal component analysis, small sample size

## 1. INTRODUCTION

Hyperspectral sensors simultaneously measure hundreds of continuous spectral bands with a fine resolution, e.g $0.001\mu m$. For instance, the AVIRIS hyperspectral sensor has 224 spectral bands ranging from $0.4\mu m$ to $2.5\mu m$. This high resolution means that hyperspectral images have proven to be beneficial in many different applications as they provide more accurate and detailed information for classification [1]. On the other hand, this large number of spectral bands has a direct impact on the required computational cost for classification. This cost is also related to the complexity of the adopted classifier (linear, quadratic etc.). Also, as the feature space dimension increases, if the size of the training data does not grow correspondingly, a reduction in the classification accuracy of the testing data is observed due to poor parameter estimation of the supervised classifier. This effect is known as the Hughes phenomenon [2, 3, 4]. Two approaches are available to overcome this small-sample-size (SSS) problem. One is to apply a feature reduction method to reduce the dimensionality of the input data. The other is to modify the classifier design to be suitable for the SSS problem [5, 6, 7, 8]. The main purpose of the feature reduction is to mitigate the Hughes effect or the curse of dimensionality.

Feature reduction can be addressed using two basic approaches. One is the selection of the most informative feature subset in the original feature space. This approach is known as feature selection. An alternative approach is the extraction of a limited number of features after a mapping to a new feature space based on some criteria like variance [9, 10]. One well known linear feature extraction technique is Principal Component Analysis (PCA). The linear transform in PCA is derived from eigenvectors corresponding to the largest eigenvalues of the covariance matrix of the data.

PCA seeks to optimally represent the data in terms of maximal variance or minimal mean-square-error between the new representation and the original data. Although this approach can reduce the data with a significant volume it does not emphasize individual spectral classes or signatures of interest [11]. The main limitation of PCA is that it does not consider the class separability since it does not take into account the class label when generating the eigenvectors. PCA simply performs a coordinate rotation that aligns the transformed axes with the direction of maximum variance; there is no guarantee that the direction of maximum variance will contain good features for classification [12]. Another limitation of PCA is that it only considers the second order statistics of the data. There may exist some small objects which are small and do not contribute to the overall variance and hence are not included in the first few principal components.

To address this limitation, this paper propose a feature reduction method, which combines feature extraction using PCA and feature selection from the resulting principle components (PCs) using a mutual information (MI) measure. In our experiment we found that the features obtain by this new method provided better classification accuracy than using either PCA or MI on the original data. The proposed MI-PCA algorithm selects the PC transformed features with higher values of mutual information as this provides a better way to measure the relevance of the components. Our results show that the proposed method provides better class discrimination as well as a significant increase in classification accuracy with reduced dimensionality. The rest of this paper is organized as follows, section 2 contains a brief introduction to PCA and

mutual information, section 3 outlines the proposed MI-PCA feature reduction method and section 4 presents a description of experimental procedure and results. Section 5 contains the conclusions drawn from our results.

## 2. RELATED WORKS

### 2.1. Principal component analysis

The principal component transform produces a new set of uncorrelated images that are ordered by variance. This analysis mainly depends on the eigenvalue decomposition of the covariance matrix of the input hyperspectral data. The new feature space ***Y*** is given by

$$Y = E^T X \tag{1}$$

Where ***X*** is the input image, ***Y*** is the transformed output image and $E^T$ is the matrix of normalized eigenvectors of the input hyperspectral image covariance matrix ordered by the respective singular eigenvalues. The first few components are often selected as extracted features.

### 2.2. Mutual information

According to Shannon's information theory, entropy measures the information in terms of uncertainty. For a given random variable *A* with marginal probability *P(A)* the entropy is defined by:

$$H(A) = -\sum_{A} P(A) log P(A) \tag{2}$$

This concept can be extended to two random variables, where one is the measured spectral image itself and the other is the target reference image that is related to the classification objective. Measuring the mutual information (MI) between these two images provides an ideal framework to measure the similarity between them. If we consider another variable *B*, then the mutual information between these two variables is defined as follows:

$$I(A,B) = \sum_{A}\sum_{B} P(A,B) \frac{log P(A,B)}{P(A)P(B)} \tag{3}$$

Mutual information can also be described using entropy as:

$$I(A,B) = H(A) + H(B) - H(A,B) \tag{4}$$

Where *H(A)* and *H(B)* are the entropies of *A* and *B* and *H(A,B)* is their joint entropy.

## 3. COMBINING MI AND PCA FOR FEATURE REDUCTION

In feature selection, the relevant features have important information required to produce the classification output, whereas the irrelevant features contain little information required for classification [1, 13]. The aim of feature selection is to find a subset of $k < n$ features that is maximally informative. The success of any feature selection algorithm depends critically on how much relevant information is contained in the selected features [14]. According to Shannon's theory the problem of selecting relevant input features can be solved by computing the mutual information. In order to find informative features from the transformed features obtained by unsupervised PCA, nonparametric mutual information (MI) is proposed as the criterion since it is sensitive to the class structure and suitable for multi class problems and non-gaussian data. Consider a set of PC transformed features ***Y***, $y_i \epsilon R^d$ obtained from Eq. (1) and the reference image $f_j$. The mutual information between $y_i$ and $f_j$ is given by:

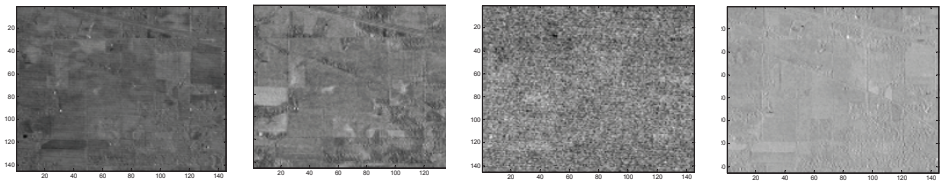
$$I(y_i, f_j) = \sum_{i}\sum_{j} P(y_i, f_j) \frac{log P(y_i, f_j)}{P(y_i)P(f_j)} \tag{5}$$

where $P(y_i), p(f_j)$ are marginal probabilities and $P(y_i, f_j)$ stands for the join probability of $y_i$ and $f_j$. If feature $f_j$ is believed to be the most informative image then $I(y_i, f_j)$ can be used as an indicator to quantify how informative $y_i$ is. To use this concept for hyperspectral feature selection, the reference feature needs to be defined first. We proposed to use the mean channel over all the spectral bands. The rest of the features are then re-ranked based on the MI value between each of them and the reference feature in order of decreasing magnitude. The advantage of using the mean channel is that, averaging the channels has the power to increase the signal to noise ratio (SNR) and also increase the bit depth of the image beyond what would be possible with a single image [15, 16].

It is straight forward to perform feature selection based on the MI between the original band and their mean features (we call this method MI-Org). However the components obtained by MI-Org are very correlated and hence provide poor classification accuracy. This is why we propose to apply the feature selection process on the transformed space to improve the effectiveness of our proposed method. In this way we obtain the reduced feature set which is uncorrelated, informative in general, can avoid the Hughes problem and improve the classification accuracy for both training and test data compared to the PCA and MI-Org techniques.

## 4. EXPERIMENTAL RESULTS

The hyperspectral image used in our experiments was the "Indian Pine" data set. This data set consists of a 145x145 pixels of an AVIRIS [17, 18] image acquired over Northwestern Indian Pine in June 1992 [9]. All the 220 original bands are employed as an input to a principal component analysis for feature transformation. The output is then used as the input to the MI-PCA method. The new and older ranked components based on two different estimation techniques of variance and



a. PC6 b. PC9 c. PC7 d. PC12

**Fig. 1**. Selected principal components for the Indian Pines data set

**Table 1**. Top ten ranked features obtained by PCA, MI-Org and MI-PCA

| Ranking Method | Ranked Features | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $1^{st}$ | $2^{nd}$ | $3^{rd}$ | $4^{th}$ | $5^{th}$ | $6^{th}$ | $7^{th}$ | $8^{th}$ | $9^{th}$ | $10^{th}$ |
| PCA | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 | PC10 |
| MI-Org | B35 | B99 | B100 | B85 | B84 | B94 | B98 | B87 | B86 | B88 |
| MI-PCA | PC1 | PC2 | PC3 | PC4 | PC5 | PC9 | PC12 | PC6 | PC16 | PC15 |

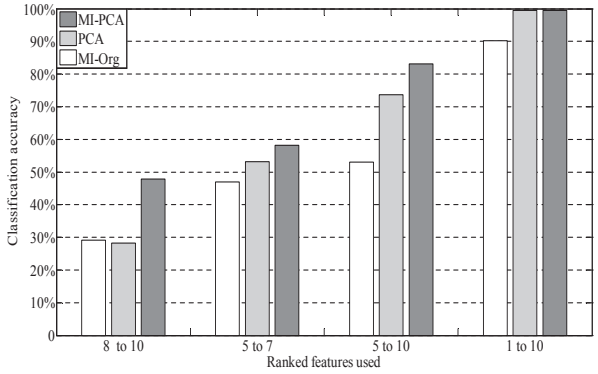



**Fig. 2**. Classification accuracy of training data

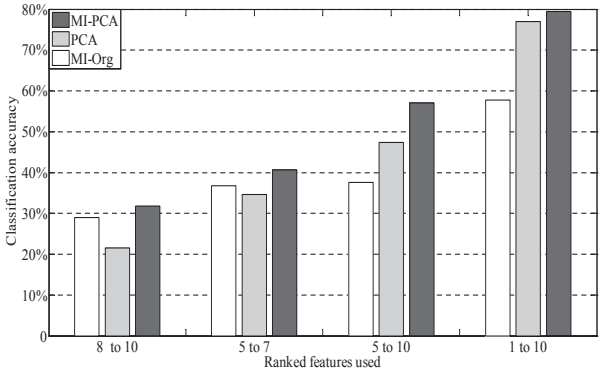



**Fig. 3**. Classification accuracy of test data

MI value are used as the input to a supervised maximum likelihood classifier (MLC) individually and their classification accuracy is measured with respect to 10 different classes. All the feature extraction methods are evaluated in terms of classification accuracy and are listed in Table-1 and Fig-2 and 3. We can see that the new ranked features obtain by MI-PCA provide better discrimination than the order obtained by traditional PCA. It can also be seen that MI-PCA performs better in 7 cases out of 8 cases for both the training and test data. This is due to the fact that, mutual information orders the features according to their relevancy, whereas the PCA orders the data based on the global variance. Fig-1 shows a visual representation of the benefits of using mutual information. Notice that even though PC7 has higher variance it contains just noise where PC12 contains spatial information that can be used in the classification process. These images highlight the main limitation of PCA, i.e. maximum variance does not guarantee maximum information. From the above experimental result it can be said that, features selected using a combination of maximum spatial similarity and variance can provide better discrimination and a corresponding increase in classification accuracy.

**Table 2**. No of training and testing samples in each class

| Class no | Class name | Pixels training set | Pixels of testing set |
|---|---|---|---|
| 1 | Hay-windrowed | 143 | 144 |
| 2 | Soybean-notill | 182 | 72 |
| 3 | woods | 252 | 63 |
| 4 | Wheat | 78 | 54 |
| 5 | Grass-trees | 147 | 96 |
| 6 | corn-notill | 224 | 68 |
| 7 | Corn-min | 77 | 88 |
| 8 | Soybean-min | 132 | 198 |
| 9 | Grass-pasture | 88 | 45 |
| 10 | Stone stell tower | 48 | 24 |

**Table 3**. Feature extraction algorithms for comparison

| Name | Feature extraction algorithm |
|---|---|
| PCA | Principal component analysis |
| MI-Org | Mutual information between original data and their mean channel |
| MI-PCA | Mutual information based principal component analysis |

## 5. CONCLUSION

In this research we have demonstrated the advantages of combining nonparametric mutual information with unsupervised Principal Component Analysis and its effects on classification accuracy. The selection criterion of this method is based on maximizing MI, which has natural applicability to multi-class problems and non-Gaussian data. The proposed method can identify features that can obtain 80%(best among three) classification accuracy with only 10 features for test data.

## 6. REFERENCES


[1] R. Archibald and G. Fann, "Feature selection and classification of hyperspectral images with support vector machines," *IEEE Geoscience and Remote Sensing Letters*, vol. 4 no. 4, pp. 674–677, October 2007.

[2] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. IT-14 no. 1, pp. 55–63, January 1968.

[3] AVIRIS, "Airborne visible/infrared imaging spectrometer," *[Online]*, vol. Available: http://aviris.jpl.nasa.gov/.

[4] B. Guo, R. Gunn, R.I. Damper, and J.D.B. Nelson, "https://engineering.purdue.edu/ multispec/hyperspectral.html," .

[5] J. Yang, P. Yu, and B. Kuo, "A nonparametric feature extraction and its application to nearest neighbour classification for hyperspectral image data," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 48 no. 3, pp. 1279–1293, March 2010.

[6] S. Tajudin and D.A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 2113–2118, July 1999.

[7] C. Lee and D.A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 15 no. 4, pp. 388–400, April 1993.

[8] L.O. Jimenez and D.A. Landgrebe, "Supervised classification in high dimensional space: Geometrical, statistical, and asymptotical properties of multivariate data," *IEEE Transaction on Systems, Man, Cybernetics-Part c*, vol. 28 no. 1, pp. 39–54, February 1998.

[9] Landgrebe D.A., *Signal Theory Methods in Multispectral Remote Sensing*, A John Wiley and Sons Publication, Hoboken, New Jersy, 2003.

[10] Richards J.A. and X. Jia, *Remote Sensing Digital Image Analysis, $4^{th}$ Edition*, Springer-verlag Berlin Heidelberg, Germany, 2006.

[11] S. Subramanian, N. Gat, M. Sheffield, J. Barhen, and N. Toomarian, "Methodology for hyperspectral image classification using novel neural network," *Algorithms for Multispectral and Hyperspectral Imagery III, SPIE*, vol. 3071–Orlando, FL, pp. 1–9, 1997.

[12] Borges J.S. and A.R.S. Marcal, *Evaluation of Feature Extraction and Reduction Methods for Hyperspectral Images, New Developments and Challenges in Remote Sensing*, Millpress, Rotterdam, ISSBN 978-90-5966-053-3, 2007.

[13] N. Yao, Z. Lin, and J. Zhang, "Feature selection based on mutual information and its application in hyperspectral image classification," *KSEM 2010 Proceedings of the $4^{th}$ International Conference on Knowledge Science, Engineering and Management*, pp. 561–566, 2010.

[14] N. Kwak and C. H. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24 no. 12, pp. 1065–1076, December 2002.

[15] Gobzalez R.C. and R.E. Woods, *Digital Image Processing, Third Edition*, Addison-Wesley Publishing Company, 2008.

[16] D.F. Vitale, G. Lauria, N. Pelaggi, G. Gerundo, C. Bordini, D. Leosco, C. Rengo, and F. Rengo, "Optimal number of averaged frames for noise reduction of ultrasound images," *Computers in Cardiology, Proceedings*, pp. 639 –641, September 1993.

[17] J.A. Benediktsson, J.R. Sveinsson, and K. Arnason, "Classification and feature extraction of aviris data," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 33 no. 5, pp. 1194–1205, September 1995.

[18] S.B. Serpico and G. Moser, "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 45 no. 2, pp. 484–495, February 2007.