A Robust Band Compression Technique for Hyperspectral Image Classification

Qazi Sami ul Haq,Lixin Shi,Linmi Tao,Shiqiang Yang
Key Laboratory of Pervasive Computing, Ministry of Education
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

***Abstract*—Dimension reduction is the key step of hyperspectral image classification. Many techniques have been developed in the past years, but our classification experiments show that some of these techniques are not robust while others suffer from the accuracy and the effectiveness for the classification of hyperspectral data. In this paper, a novel band compression algorithm is proposed based on the fusion of segmented principle component analysis (SPCA) and linear discriminant analysis (LDA) for dimension reduction. We first select the bands independently via SPCA and LDA. Theoretical analysis shows that the selected bands have little correlation, and therefore, an iterative algorithm is adopt to adaptively co-optimizing both the parameter of merging SPCA bands and LDA bands, and the classification accuracy. Our extensive experiments on two real hyperspectral datasets (AVIRIS 1992 Indian pine image and HYDICE image of Washington DC Mall), proves that the proposed technique is not only robust but offers more classification accuracy than many conventional dimension reduction techniques over several well known classifiers.**

***Index Terms*—Hyperspectral data, fusion,image region classification,remote sensing,spectral band compression,aviris.**

I. Introduction

The advent of hyperspectral sensors has been the most notable event in the field of remote sensing. It offers much more information than any other sensor technology in remote sensing due to the high spectral resolution and because of which it is possible to discriminate a large number of materials [1]. With such high dimensional data, not only does it incur high computational cost but also Hughes phenomenon comes into play, which means large amount of training samples would be required to achieve good classification. Unfortunately, in remote sensing, usually it is extremely difficult and expensive to label the samples because of which, the training samples rarely exist in large numbers. Meanwhile, many bands of hyperspectral images are very noisy due to the reflection of atmosphere, and have little information about the surface of the earth. As a result, dimension reduction, both for reducing the computational cost and for removing the noisy bands, is the key step to construct a robust and accurate classifier.

The easiest way to do dimension reduction is to select some subset of bands based on some optimal criteria like classification accuracy, but this can only be done in case of multispectral images with few number of bands. In the case of hyperspectral images, the number of bands are very high (> 200 bands), so it becomes prohibitively computational intensive to try out all combinations of the subsets of bands.

Many dimension reduction techniques have been introduced into remote sensing, which can be grouped into three approaches: variance based approach, such as PCA [2], factor analysis [3], sparse PCA [4], segemented PCA [5]; discrimination based approach, such as LDA [6], MDA [7], MDF [8]; and random projection [9] based approach.Out of these approaches, PCA,Segmented PCA and LDA are known to provide good classification accuracy and robustness.

All these approaches are widely used in many remote sensing areas, in practice, we still need criteria for evaluating the algorithms on the purpose of selecting an algorithm of dimension reduction for the better classification. There are two criteria that should be met for this purpose i.e informativeness and redundancy. Informativeness measures the total amount of information over the compressed data. This criteria minimizes the loss of valued information in the compression. Information redundancy means to minimize the wasted information of the bands. Wasted information comes from information overlap, which means that the information is shared by most of the samples. The balance between these two criteria is a complex problem that has confused researchers in the related fields for a long time and specifically for data compression in hyperspectral imaging, one method tends to aim at only one of the criteria above.

Therefore, an obvious shortcoming of existing methods is that they don't fulfill these two criteria and are not robust for different dataset and classifiers e.g despite its usefulness in dimension reduction, PCA may not be an optimal method for dimension reduction in hyperspectral data as it may overlook subtle but useful information if directly applied to hyperspectral data [10]. Similarly LDA provides good discrimination but its more sensitive to noise. Jia and Richards [5] proposed a segmented principal component analysis in which they observed correlation matrix to identify highly correlated bands and to divide them in groups but it suffers from discrimination information. As an example, we can see in the fig. 7 that linear discriminant analysis and conventional principal component analysis are not robust for different datasets and different classifiers as the variance in their accuracies across different datasets and classifiers is larger.

To address this dual-criteria problem by searching for a new band reduction algorithm is a hard and time consuming task. This study has been carried out in our lab, and a novel

978-1-4244-6585-9/10/$26.00 ©2010 IEEE

band reduction algorithm is proposed, which is not only robust for different datasets and classifiers but offers more classification accuracy. Our extensive experiments (section IV) shows that the proposed technique is more robust and offers more classification accuracy than the existing techniques.

The rest of paper is organized as follows. Section II describes Segmented PCA and LDA. Section III describes the proposed approach. Section IV presents the experimental results and comparison. Finally the section V concludes the paper.

II. Segementd PCA and LDA

In this section, we briefly describe segmented principle component analysis and linear discriminant analysis.

A. Segmented PCA

Segmented PCA was first introduced in remote sensing by Jia and Richards [5]. Segmented PCA comes from the observation that the data is actually highly segmented, the complete set of bands is divided into number of subgroups and each subgroup contains the most correlated set of bands. PCA is then applied on each sub-group. From each sub-group some transformed dimensions are selected which covers the maximum variance and those are then combined to form the new dimension set. In order to demonstrate it, we calculate the correlation matrix which is defined as follows:

Mean vector:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Covariance matrix:

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \tag{2}$$

Correlation matrix:

$$R = (r_{ij})_{b\times b},\ where\ r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}},\ \Sigma = (\sigma_{ij})_{b\times b} \tag{3}$$

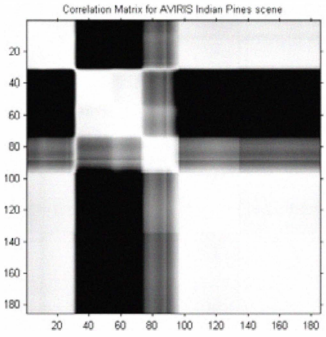

Fig. 1. The correaltion matrix of AVIRIS Indian Pines scene

B. LDA

Also knows as Fisher's linear discriminant analysis [11] [12]. By having training samples with their labels, it tries to find a linear transform that maximizes the ratio of average between class-variation over average within class-variation.

LDA computes a subspace W to perform a projection. Suppose we have calculated the mean vector (eq.1) $\mu_1, \mu_2, ..., \mu_C$ and covariance matrix $\Sigma_1, \Sigma_2, ..., \Sigma_C$. Then calculate their expectations using prior distributions:

$$\boldsymbol{\mu}_w = \mathbb{E}(\mu_i) \tag{4}$$

$$\Sigma_w = \mathbb{E}(\Sigma_i) \tag{5}$$

Define between-class scatter

$$\Sigma_b = \sum_i (\boldsymbol{\mu}_i - \boldsymbol{\mu}_w)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_w)^T$$

then we can optimizes the object function:

$$W = \text{argmax}\frac{|W^T\Sigma_b W|}{|W^T\Sigma_b W|}$$

III. Proposed Approach

In this section, we'll talk about why we combine segmented PCA and LDA together, and details for how to integrate them.

A. Fusion of Segmented PCA and LDA

From the explanation of SPCA and LDA in (section II), we can obeserve that these two methods apply to the two criteria i.e informativeness and information redundancy, respectively. For SPCA, it selects a small number of "principal" components exhibiting the highest variance, i.e., the most informativeness features; for LDA, it maximizes the rate of between-class scatter matrix over within-class scatter matrix, which means that it tries its best to eliminate the shared information and emphasizes the discriminant information.

This combination will improve the informativeness without resulting in more redundancy. Intuitively, we can see that PCA and LDA are two ends at a balance, which itself implies low redundancy. To make it more clear, fig. 2 shows the correlation matrix(eq. 3) of the dimension selected by PCA and LDA. It demonstrates evidently that these two methods don't share high overlap:

Each of SPCA and LDA has either direct or indirect constraints in the number of dimensions they can choose. For PCA, the expressiveness of small dimensions is very high, fig. 3 gives an evidence: If we use SPCA, 3 or 4 bands are often enough for each segment. If the classification algorithm have a space to increase the number of bands, it's not possible that increasing the number of bands will increase the performance under SPCA. For LDA, it has a more direct constraint: if we have c classes, then at most $c-1$ dimensions are available. Even though we will give a grouped approach to overcome this

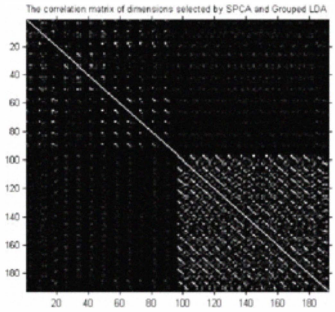

Fig. 2. This figure shows the correlation of the dimensions compressed by Fusion. Dark cells stand for lower correlation.

constraint later, it faces the same problem of expressiveness as PCA does.

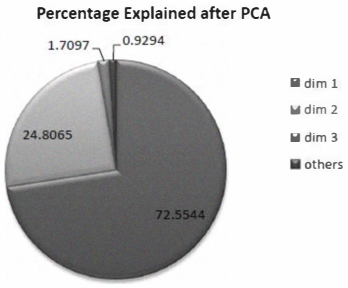

Fig. 3. Percentage explained by each dimension of $\mathcal{D}_c$ after PCA

B. How to fusion SPCA and LDA

Note that both SPCA and LDA projects the original data vector onto a new subspace. Suppose the projection matrix is W_{SPCA} and W_{LDA} respectively, we can define the new projection matrix as

$$W = \left[\begin{array}{cc} W_{SPCA} & W_{LDA} \end{array} \right]$$

The result space is simply the union of the subspace generated by SPCA and LDA. By the fig. 2 we have shown that there's little correlation between the two subspaces. Therefore, it's reasonable to fusion them together in this way.

The process of fusing SPCA and LDA is shown in fig. 4.

C. Co-optimization of the combination rate and the classification accuracy

It seems that it is a straight forward idea to put W_{SPCA} and W_{LDA} together to have the new projection matrix, but the real problem is how to determine the number of bands for each single method, or the rate of bands selected by the methods. We suppose that the total number of dimensions are fixed due to classification method and specific data set. We use an adaptive way to iteratively determine the number of bands for them. Starting from an initial division of bands, say, half for SPCA and half for LDA, the band rate is dynamically updated through the learning. The training set is divided into 2 subsets T_1 and T_2, first run data compression and classification

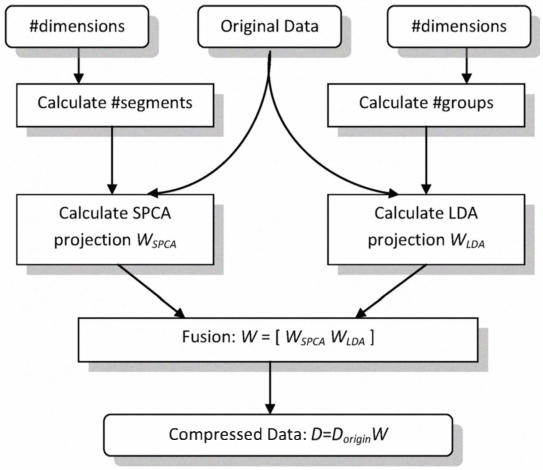

Fig. 4. The process of fusing SPCA and LDA

on T_1, test the classification accuracy on T_2, this works as a feedback information, then adaptively update the rate of dimensions in a determined way; this update is accepted only if the accuracy is improved. Finally, the updating process finished when there's no update accepted, and we can get the final rate of dimensions.

Given the number of dimensions for SPCA, we determine the number of segments by the observation from fig. 3. Actually for each segment, 3 or 4 dimensions are enough, we can let the number of bands be #bands/3 or so.

Another problem is that since LDA has a $c-1$ constraint on the number of dimensions, how can we choose arbitrary number of bands using LDA. Here we use a similar method as SPCA: In order to get rid of this constraint, we can modify the LDA to be grouped LDA. Nonetheless note that the grouping *cannot* be the same with SPCA. The reason is that LDA compress data while retaining the *discriminant* information to distinguish different classes, hence it's not a good idea to make LDA compress data which shares great similarity between dimensions, which is an assumption of adjacent bands. Actually we expect that bands in the same group are as far away as we can. Hence the grouping could be in a way that bands $i, g+i, 2g+i, \cdots$ are in group i if there are g groups totally.

IV. Experiments

In this section, the performance of proposed technique is evaluated. We used two real datasets i.e AVIRIS 1992 Indian Pine Image and HYDICE image of Washington DC Mall [1]. The aim of these experiments is to demonstrate the effectiveness and robustness of the proposed technique with different sizes of training samples and with various classifiers. The training samples have been selected randomly. For all experiments, we used maximum likelihood,support vector machine and artificial neural networks for testing the performance of the proposed technique. All results are the average of 10 runs. The comparisions with other techniques

have also been presented.

A. Real AVIRIS Image

The 1992 AVIRIS hyperspectral image from Northern Indiana was taken on June 12,1992 as shown in fig.

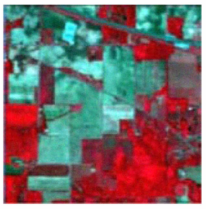

Fig. 5. 1992 AVIRIS hyperspectral image

It has 145*145 pixels and 220 spectral bands. 20 Noisy bands(104-108,150-163,220) due to water absorption have been removed, so in total 200 bands are used. The total classes available are 16 but due to insufficient training samples, 7 classes were not used. Two scenarios are used here.

1) AVIRIS SubImage: In the first scenario, we used a subset scene from the AVIRIS image i.e from columns [27-94] and rows [31-116] with size of 68*86 with four classes. Table 1,2,3 shows the results of classifcation of proposed technique using ml,svm and ann respectively, in comparison with the other techniques. The number of training samples used are 50%,40%,30% and 20% of the total samples. The rest of the samples are used for testing. We can see from the results that in every case the proposed technique performed better than the other techniques.

TABLE I
COMPARISON OF PROPOSED TECHNIQUE USING ML ON AVIRIS SUBIMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**95.10**	93.82	94.17	92.32
40%	**94.86**	93.75	93.94	91.91
30%	**94.58**	93.21	93.76	91.57
20%	**93.99**	92.88	93.28	90.67

TABLE II
COMPARISON OF PROPOSED TECHNIQUE USING SVM ON AVIRIS SUBIMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**96.56**	96.49	96.39	90.02
40%	**96.52**	96.20	96.26	89.87
30%	**95.80**	95.77	95.51	89.65
20%	**95.48**	95.30	95.31	88.71

2) AVIRIS Full Image: In the second scenario the whole 145*145 image is used with 9 classes. The number of training samples used are 50%,40%,30%,20% of the total samples. The rest of the samples are used for testing.From Table 4,5,6 the results of classification of the proposed technique in comparison with other techniques can be observed. In all cases the performance of proposed technique is better than other techniques but only when we select 20% of samples for trainining then in case of mle and svm, the results of segmented pca are fractionally better than the proposed technique.

TABLE III
COMPARISON OF PROPOSED TECHNIQUE USING ANN ON AVIRIS SUBIMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**94.11**	93.27	92.16	93.34
40%	**93.68**	93.10	90.57	93.49
30%	**92.36**	91.93	81.82	92.14
20%	**90.94**	90.14	80.96	90.21

TABLE IV
COMPARISON OF PROPOSED TECHNIQUE USING ML ON AVIRIS FULL IMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**89.17**	87.29	86.07	84.33
40%	**88.56**	87.13	85.70	83.85
30%	**87.49**	86.54	84.60	83.27
20%	**85.26**	85.43	82.98	81.47

TABLE V
COMPARISON OF PROPOSED TECHNIQUE USING SVM ON AVIRIS FULL IMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**91.91**	91.86	90.28	76.64
40%	**91.50**	91.40	89.84	75.70
30%	**91.40**	91.01	89.32	72.37
20%	**89.55**	89.60	87.68	69.12

TABLE VI
COMPARISON OF PROPOSED TECHNIQUE USING ANN ON AVIRIS FULL IMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**84.48**	83.27	81.82	83.37
40%	**83.61**	83.16	80.96	83.52
30%	**82.57**	81.78	79.98	82.04
20%	**80.66**	79.53	77.80	80.28

B. Washington DC Mall Image

This is the image of airborne hyperspectral data set of the Washington DC mall. It has 210 spectral bands from the 0.4 to 2.4 in visible and infrared spectrum regions. It contains 1208 scan lines with 307 pixels in each scan line. The total available classes are 9. The training samples used are the 50%,40%,30%, and 20% of total training samples available and the rest of samples are used for testing. Tables 7,8,9 shows the classification results by using mle,svm and ann respectively. Here the accuracy of classifiers for all techniques is very high as the data is not as much difficult as the AVIRIS

data, still the results show that the classification accuracy for proposed technique is higher in most cases and comparable in few cases.

TABLE VII
COMPARISON OF PROPOSED TECHNIQUE USING ML ON WASHINGTON DC MALL IMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**99.92**	99.84	99.82	99.90
40%	**99.89**	99.81	99.79	99.85
30%	**99.84**	99.75	99.71	99.74
20%	**99.31**	99.35	99.33	98.75

TABLE VIII
COMPARISON OF PROPOSED TECHNIQUE USING SVM ON WASHINGTON DC MALL IMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**99.73**	99.74	99.72	99.91
40%	**99.76**	99.64	99.68	99.93
30%	**99.61**	99.60	99.60	99.93
20%	**99.54**	99.46	99.48	99.84

TABLE IX
COMPARISON OF PROPOSED TECHNIQUE USING ANN ON WASHINGTON DC MALL IMAGE

Train samples	Fusion	SPCA	PCA	LDA
50%	**99.71**	99.74	99.31	99.43
40%	**99.69**	99.64	99.16	99.27
30%	**99.62**	99.60	98.89	99.01
20%	**99.46**	99.46	98.41	98.67

We can see from the fig. 6 and fig. 7 that the proposed technique not only provides better accuracy but also is more robust than the existing techniques across different datasets and classifiers. From fig. 7, we can observe that the overall variance of proposed technique across all datasets and classifiers is far less than the other existing techniques and therefore is more robust.

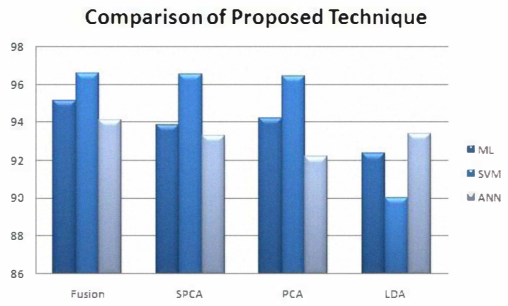

Fig. 6. Comparison of classification accuracy of Fusion based technique

V. CONCLUSION

In this paper, we proposed dual-criteria of dimension reduction for better hyperspectral data classification i.e informative-

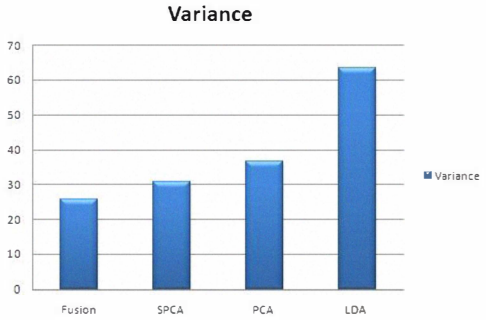

Fig. 7. Comparison of variance in accuracies of all techniques.The variance is calculated over all datasets and classifiers

ness and redundancy. We developed a novel co-optimization method to adjusting the parameter of merging SPCA bands and LDA bands for dimension reduction, and to increase the classification accuracy iteratively. The extensive tests on two real hyperspectral datasets, three typical classification algorithms and comparisons with the other techniques, prove the robustness and the effectivness of the proposed technique.

ACKNOWLEDGMENT

This research was supported in part by the National Natural Science Foundation of China under Grant Nos. 60873266 and 90820304.

REFERENCES

[1] Landgrebe, D. A., [Signal Theory Methods in Multispectral Remote Sensing], John Wiley, Hoboken, NJ (2003).
[2] E. Garcia (2009) PCA & SPCA Tutorial.Http://www.miislita.com/information-retrieval-tutorial/pca-spca-tutorial.pdf (July 04, 2009)
[3] Foundations of Factor Analysis, Second Edition,Stanley A Mulaik, CRC Press, September 25, 2009.
[4] Zou, H., Hastie, T., & Tibshirani, R. (2004). Sparse principal component analysis (Technical Report). Statistics Department, Stanford University.
[5] Jia, X. and Richards, J. A. (1999) Segmented principal components transformation for efficient hyperspectral remote sensing image display and classification. IEEE Transactions on Geoscience and Remote Sensing, 37, pp. 538-542.
[6] Bandos, T. V., Bruzzone, L., & Camps-Valls, G. (2009). Classification of hyperspectral images with regularized linear discriminant analysis. IEEE Transactions on Geoscience and Remote Sensing, 47(3), 862-873.
[7] Zhu Jingbo, Ye Na, Chang Xinzhi, Chen Wenliang and Benjamin K Tsou. 2005. Using Multiple Discriminant Analysis Approach for Linear Text Segmentation. In Proceedings of the Second International Joint Conference on Natural Language Processing, pp. 292-301
[8] Pavei, N., Ribari, S., Grad, B., Comparison of PCA -, MDF -, and RD-LDA - based Feature Extraction: Approaches for Hand-based Personal Recognition, Proceedings of the 2007 International Conference on Computer Systems and Technologies, 2007
[9] J. Lin and D. Gunopulos. Dimensionality reduction by random projection and latent semantic indexing. In Proceedings of the Text Mining Workshop, at the 3rd SIAM International Conference on Data Mining, May 2003
[10] Tsai, F., Lin, E.K. and Yoshino, K. 2007: Spectrally segmented principal component analysis of hyperspectral imagery for mapping invasive plant species. International Journal of Remote Sensing 28, 102339.
[11] R.Duda, P.Hart, D.Stork: .Pattern Classification . Second Edition., John Wiley & Sons, 2001.
[12] A.Jain, R.Bolle, S.Pankanti Eds.: .Biometrics . Personal Identification in Networked Society., Kluwer Academic Publishers, 1999.