Dimensional Reduction of Hyperspectral Image Data Using Band Clustering and Selection Based on Statistical Characteristics of Band Images

Muhammad Sohaib, Ihsan-Ul-Haq, and Qaiser Mushtaq

***Abstract*—In this paper an approach for the dimensionality reduction of the hyperspectral image data using the method of band selection based on the statistical measures is introduced. The spread hyperspectral image data is measured in each band and the calculated bands are clustered using the K-means clustering technique. The K-means clustering of bands is performed in such a way that the intra-cluster variance is kept minimize and the inter-cluster variance maximum. The optimal number of band selection is done using the concept of Virtual Dimensionality (VD). The endmember or targets are extracted through Vertex Component Analysis (VCA). The experimental results are compared with other unsupervised band selection techniques to show the effectiveness of the proposed technique.**

***Index Terms*—Dimesionaity reduction, k-means clustering, VD, VCA.**

I. Introduction

Hyperspectral sensors- used for hyperspectral imagery collect information as a set of images represented by different bands. A remotely sensed image is an image in a cubic form with the third dimension specified by spectral wavelengths. The collected image data by hyperspectral remote sensors is simultaneously in hundreds of narrow, adjacent spectral bands over the wavelengths that can range from the near ultraviolet through the thermal infrared at 5nm of fine resolutions. Each pixel contains a hyperspectral signature that represents different materials. As a result of high spectral resolution, hyperspectral systems produce a massive amount of data. These measurements make it possible to derive a continuous spectrum for an image data [1]. Hyperspectral data helps the analyst in detection of more materials, objects and regions with enhanced accuracy.

Hyperspectral images provide a vast amount of information about a scene, but most of that information is redundant as the bands are highly correlated. For computational and data compression reasons, it is desired to reduce the dimensional of the data set [2] while maintaining good performance in image analysis tasks. There are some of the challenges we face during the analysis of hyperspectral images, first is due to huge data volume, we face data storage and transmission problem. Second is redundancy of information because redundancy in data can cause convergence instability. Third is high processing time. As a result, the imposition of requirements for storage space, computational load and communication bandwidth are against the real time applications and it is difficult to visualize or to classify such a huge amount of data. Dimensionality reduction is a good choice to overcome these challenges. The reduction of dimensionality is necessary for high accuracy in unmixing of the pixels, classification and detection.

There are several methods of dimensionality reduction which can be further categorized into two groups; feature extraction and feature or band selection. Feature selection is preferable for dimensionality reduction because feature extraction need most of the original data representation for extraction of features [3]. Secondly due to transformation in feature extraction the critical information may have been distorted. Compare to feature extraction, feature selection preserve the relevant original information. There are many band selection techniques used in the past [4]. Search base Methods [4], Transform based Methods [5], ICA-based band selection Method [6] and information based Methods [7].

Hierarchical structure for clustering was used in [12] and the bands were clustered based on the similarity calculated by city block, Euclidean distance and cosine distance. The hierarchical structure was clustered by using three linkage methods i.e. single link, average and ward. For each of the metric these three linkage strategies were used for clustering bands [12].

In this paper we have used the clustering approach for band selection and use three different statistical methods to measure data. We have used Standard Deviation, MAD and Variance in our proposed work. K-means clustering is used to cluster the bands using two distance metrics, city block and Square Euclidean. Bands are clustered through K-means and those bands which have maximum value are selected. Virtual Dimensionality VD [8] is used for the bands estimation. Minimum number of bands is selected through VD and the maximum information is preserved. Vertex Component Analysis (VCA) [9] is used for the unmixing and detection of endmembers. The results obtained are compared to the Constrained Based Selection (CBS) [10] methods i.e. The Linear Constrained Minimum Variance (LCMV) and Constrained Energy Minimization (CEM) and also with Minimum Variance Principal Component Analysis (MVPCA) [11].

In this paper Section 2 explains the Band Clustering using K-means and Selection of Bands. The experimental results and comparison are presented in Section 3 and conclusion in Section 4.

Manuscript received August 2, 2012; revised October 1, 2012.

The authors are with the faculty of engineering & technology, International Islamic University, Islamabad, Pakistan.

II. Band Clustering and Selection

Band Clustering and Selection are two steps used in our work. Clustering of band images keeps the intra-cluster variance minimum and the inter-cluster variance maximum. The method in which dimensionality is reduced by selecting a subset of the original dimensions are known as band/ feature selection. The hyperspectral data is spread in some direction. This data can be measured by using different statistical methods which include MAD (Mean Absolute Deviation), moment, variance, mean, geometric mean and standard deviation. We have used MAD, Standard Deviation and Variance in our proposed work. Suppose that we have $\{B_l\}_{l=1}^{L}$ band images in our hyperspectral image data cube where L is the total number on bands, if each band image is of size M×N and $\bar{B}_l$ the mean of the l^{th} band image. The statistical characteristics we use for data are given below.

MAD for the l^{th} band is

$$d_l = \frac{1}{MN}\sum_{i=1}^{MN}|b_i - \bar{B}_l| \quad (1)$$

Standard Deviation for the l^{th} band image is

$$d_l = \left(\frac{1}{MN}\sum_{i=1}^{MN}(b_i - \bar{B}_l)^2\right)^{\frac{1}{2}} \quad (2)$$

Variance for the l^{th} band image is

$$d_l = \frac{1}{MN}\sum_{i=1}^{MN}(b_i - \bar{B}_l)^2 \quad (3)$$

The result from the above statistical methods for L band images is given by:

$$d = \{d_l\}_{l=1}^{L}$$

III. K-Means Clustring

For clustering the bands (band images) K-means clustering technique is used. For K-means clustering city block and Square Euclidean distance metrics are used. K-means clustering is one of the simplest unsupervised algorithms and is well-known for solving the problem of clustering. The flowchart of K-means clustering is shown in figure 1. K-means follows a simple and easy way to classify a given data set through clusters; the number of clusters is fixed and is given a prior. The number of centroids i.e. K are defined for each cluster and which are placed far away from each other as possible. The points which belong to the given data set are taken and are associated to the nearest centroid which results in K number of groups. Again K new centroids are recalculated for new centers of the cluster and a new binding has to be done between the same data set points and the nearest new centroid. A loop is run for the K centroids to change their location step by step until there is no change and the centroids are fixed. The centroids of the clusters are calculated by minimizing the sum of squared errors. The K means algorithm performs three steps until convergence.

1) Determine the centroid coordinate
2) Determine the distance of each object to the centroids
3) Group the object based on minimum distance

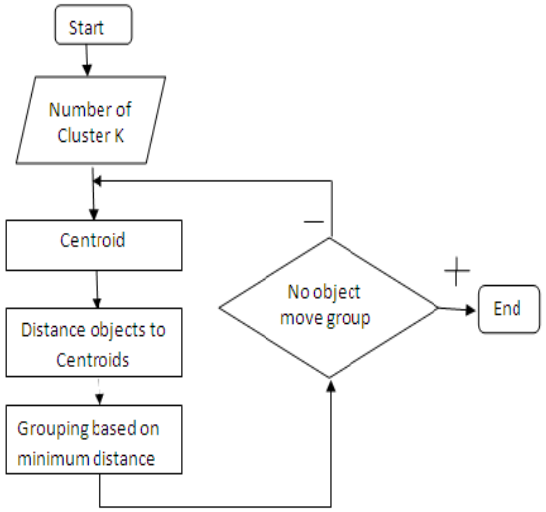


Fig. 1. K-means clustering flow chart.

For the observations $X = (x_1, x_2, x_3 \ldots x_n)$, the K-means clustering method divides the n observations into k sets (k < n), $K = \{S_1, S_2, S_3 \ldots S_k\}$, minimizing the sum of squares with-in clusters i.e.

$$min_s \sum_{i=1}^{k}\sum_{x_j \in S_i}\|x_j - \mu_i\|^2$$

where μ_i is the mean of points in clusters C_k

K-means computes centroid clusters differently for the different supported distance measures. We have used the following distance metrics in K-means clustering.

A. *Square Euclidean*

For an m-by-n data matrix $X = (x_1, x_2, x_3 \ldots x_m)$ the distance between the vector xr and xs is given as:

$$d_{rs}^2 = (x_r - x_s)\, D^{-1}\, (x_r - x_s) \quad (4)$$

where D is the diagonal matrix

B. *City Block metric*

For an m-by-n data matrix $X = (x_1, x_2, x_3 \ldots x_m)$ the distance between the vector xr and xs is defined as

$$d_{rs} = \sum_{j=1}^{n}|x_{rj} - x_{sj}| \quad (5)$$

Bands are clustered based on their statistical characteristics i.e. Variance, MAD (Mean Absolute Deviation) and Standard Deviation by K-means clustering technique. After that a band is selected from each cluster (group) which has maximum variance with in the cluster. The proposed technique using Variance with city block as distance metric is abbreviated as VAR-CB. The proposed technique using Variance with Square Euclidean as distance metric is abbreviated as VAR-SE. The proposed technique using standard deviation with city block as distance metric is abbreviated as STD-CB and similarly for Standard Deviation with Square Euclidean as STD-SE. The proposed technique using MAD with city block as distance metric is abbreviated as MAD-CB and the technique using MAD with Sq. Euclidean is abbreviated as MAD-SE.

IV. Proposed Algorithm

Following are the steps of the proposed algorithm to summarize the band clustering and selection:

1) Calculate the number of bands i.e. VD.
2) Calculate or measure the data of each band image using VAR, MAD and STD.
3) Band clustering using K-means clustering and using distances among the measured values to examine the proximity of band images to each other.
4) According to VD, clusters are created which contain all the measured values.
5) From each cluster, one band having maximum value is picked.

Now the question is how many bands need to be selected preserving the necessary information. This problem can be solved by using the new concept of Virtual Dimensionality (VD) [8] to estimate the minimum number of bands and preserve the maximum useful information. The selected bands are analyzed for the endmember detection. VCA [9] is then used for the unmixing process of the hyperspectral image and the results are compared.

V. Experimental Results

We have used a well known Airborne Visible/ Infrared Imaging Spectrometer [13] for our research work. The Cuprite image is used to compare and evaluate the proposed research work. The image scene is shown in Fig. 2. And it is available at website [14]. It was collected by 224 spectral bands with 10 nm spectral resolutions over the Cuprite mining site, Nevada in 1997, where Cuprite is a mining area in the south of Nevada with minerals and little vegetation. The geologic summary and mineral map can be found in [15]. Cuprite has been widely used for experiments in remote sensing and has become a standard test site to compare different techniques of hyperspectral image analysis. In our research work, a sub image of size 350×350 with 224 bands of a data set taken on the AVIRIS flight of June 19, 1997. The instrument of AVIRIS covers 0.41 – 2.45 μm regions in 224 bands with a 10 nm bandwidth and flying at an altitude of 20 km, it has an Instantaneous Field Of View (IFOV) of 20 m and views a swath over 10 km wide. Prior to the analysis of AVIRIS Cuprite image data, low SNR bands 1 – 3, 105 – 115 and 150 – 170 have been removed and the remaining 189 bands are used for experiments. The ground truth of spatial positions of four pure pixels corresponding to four mineral alunite (A), buddingtonite (B) , calcite (C) and kaolinite (K) are labeled and encircled by “A”, “B”, “C”, and “K” respectively. Endmembers extracted by an endmember algorithm are verified by using these labels of spatial positions. The USGS signatures of “A”, “B”, “C” and “K” are also shown in Fig. 3.

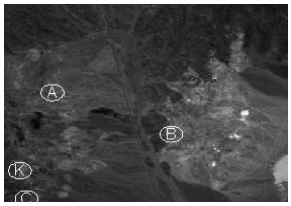


Fig. 2. Ground truth of spatial positions of four pure pixels corresponding to following minerals: Alunite (A), Buddingtonite (B), Calcite (C), and Kaolinite.

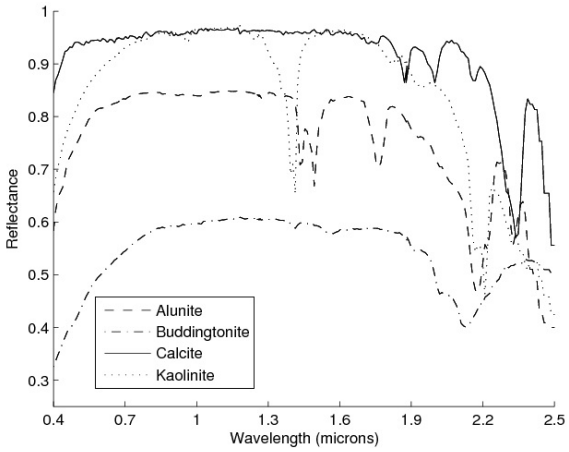


Fig. 3. USGS spectral signatures of Alunite (A), Buddingtonite (B), Calcite (C), and Kaolinite (K)

TABLE I: Selection of Bands Using Different Techniques

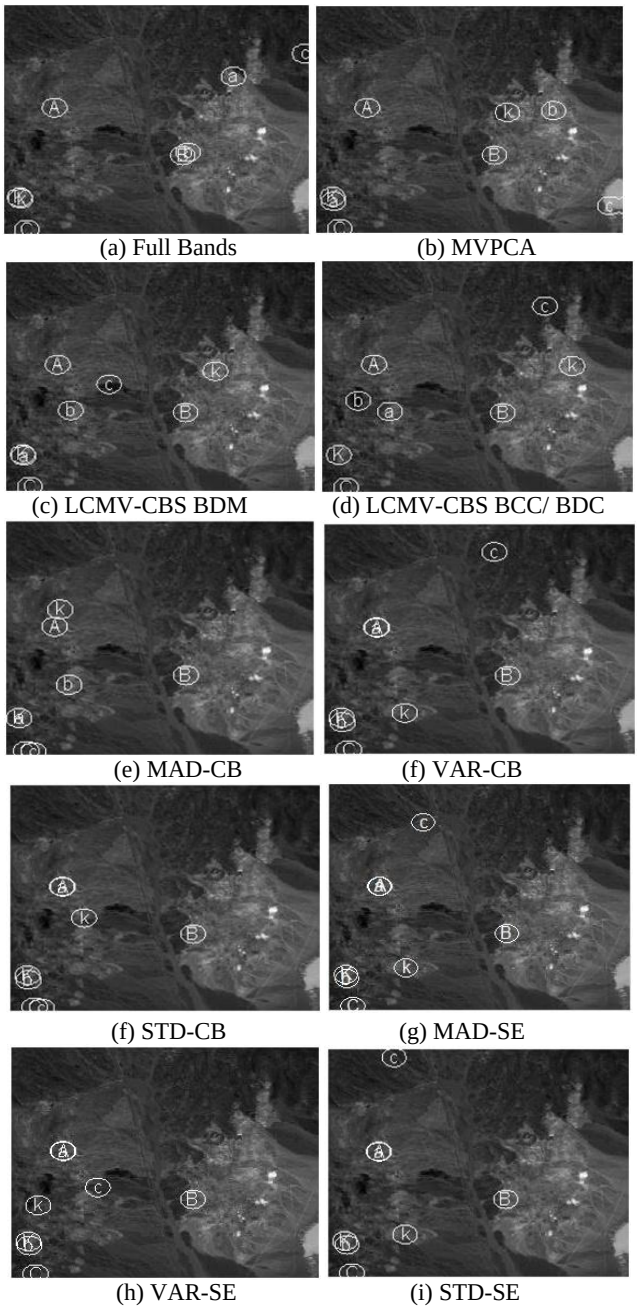
Criteria	Selected bands
MAD-CB	92, 77, 12, 57, 99, 65, 129, 102, 35, 18, 87, 5, 119, 155, 69, 131, 11, 149, 105, 178, 116, 163
VAR-CB	39, 100, 89, 154, 13, 51, 18, 177, 148, 22, 168, 102, 64, 4, 183, 61, 67, 87, 70, 32, 81, 35
STD-CB	5, 85, 13, 170, 51, 89, 151, 33, 73, 177, 18, 81, 64, 147, 67, 155, 168, 87, 9, 176, 23, 61
MAD-SE	53, 19, 102, 87, 17, 170, 168, 35, 155, 69, 183, 39, 178, 149, 112, 4, 24, 48, 94, 107, 65, 164
VAR-SE	82, 107, 39, 26, 163, 128, 57, 51, 21, 177, 5, 147, 89, 99, 14, 151, 78, 34, 72, 18, 87, 67
STD-SE	91, 61, 36, 117, 74, 177, 125, 130, 21, 67, 135, 57, 141, 14, 87, 5, 51, 131, 83, 26, 99, 70
LCMV-CBS CM/BDM	26, 117, 48, 37, 189, 64, 1, 185, 10, 172, 47, 4, 60, 28, 165, 17, 5, 2, 151, 158, 3, 94
LCMV-CBS CC/BDC	185, 37, 2, 3, 5, 64, 8, 9, 6, 7, 10, 165, 4, 11, 12, 14, 151, 13, 28, 15, 16, 153
MVPCA	87, 85, 88, 86, 89, 84, 91, 80, 78, 90, 92, 83, 82, 79, 93, 81, 98, 99, 97, 189, 77, 76

Preserving the maximum information, the number of bands required and estimated VD are 22. In our research work we have tabulated 22 bands. These bands are selected by LCMV-CBS, MVPCA and our proposed technique of clustering according to 22 VD. Fig. 4 shows the extraction of four end members and also the extracted endmembers by VCA using the 22 selected bands given in table I, the detected endmember/ targets are labeled with “a”, “b”, “c”, “k”. the detected endmembers are compared with the ground truth endmember pixels which are labeled as “A”, “B”,”C”,”K”. In addition the measurement of the spectral similarity between the endmember pixels “a”,”b”,”c”,”k” and the ground truth endmember pixels “A”,”B”,”C”,”K”, we have calculated the Spectral Angle Mapper (SAM), the results of which are tabulated in table II. The location of the “A”,”B”,”C”,”K” and “a”,”b”,”c”,”k” in the image scene are also included in the form of coordinates in brackets. The coordinates for both the target endmembers and the ground truth endmembers in brackets shows the location in the image scene, included in the table II. The result obtained

from the simulations shows that the performance of the clustering-based band selection techniques, using K-means clustering are better than the techniques of LCMV-CBS and MVPCA and Full bands. The values of the Spectral Angle Mapper (SAM) among the same target/ endmembers minerals are highlighted which shows a good similarity. The detection of the endmember pixel using the selected bands and K-means clustering, by VCA gives better results compare to the LCMV-CBS and MVPCA, therefore the detected endmember pixel have high spectral similarities

TABLE II: SPECTRAL SIMILARITY MEAUREMENTS OF GROUND TRUTH TARGETS AND FOUND TARGETS

	A (61,161)	B (209,234)	C (22,298)	K (22,298)
Full Band				
a (267,113)	0.0822	0.2146	0.2578	0.1136
b (215,229)	0.1330	0.0685	0.1089	0.1378
c (349,78)	0.2172	0.1141	0.0818	0.2408
k (23,300)	0.1043	0.1734	0.2165	0.0341
LCMV-CBS BCC/BDC				
a (23,305)	0.06230	0.1959	0.2354	0.1092
b (277,165)	0.1247	0.0921	0.1477	0.1403
c (342,312)	0.1680	0.1017	0.0796	0.1975
k (224,168)	0.0888	0.1834	0.2283	0.0340
LCMV-CBS BCM/BDM				
a (23,300)	0.0388	0.1353	0.1807	0.1075
b (77,231)	0.2046	0.0908	0.1027	0.2071
c (121,191)	0.1839	0.1027	0.0871	0.2012
k (243,171)	0.1043	0.1734	0..2165	0.0341
MVPCA				
a (80,232)	0.0606	0.1727	0.2202	0.0913
b (44,216)	0.1763	0.0777	0.1076	0.1626
c (257,72)	0.1944	0.0748	0.0513	0.2067
k (288,163)	0.0684	0.1702	0.2097	0.0668
MAD-CB				
a (22,298)	0.0335	0.1413	0.189	0.0969
b (788,248)	0.1643	0.0726	0.1035	0.1928
c (38,349)	0.1871	0.0839	0.0477	0.192
k (68,135)	0.0961	0.1733	0.2114	0
VAR-CB				
a (60,161)	0.0165	0.1654	0.2125	0.0933
b (23,305)	0.1329	0.0633	0.0856	0.1468
c (194,45)	0.2202	0.1002	0.055	0.2393
k (93,291)	0.0889	0.1834	0.2283	0.0341
STD-CB				
a (61,160)	0.0165	0.1654	0.2125	0.0933
b (23,305)	0.202	0.0795	0.0925	0.2179
c (38,349)	0.1871	0.0839	0.0477	0.192
k (86,309)	0.0889	0.1834	0.2283	0.0341
MAD-SE				
a (61,161)	0.0172	0.1645	0.2115	0.0962
b (23,305)	0.1412	0.0583	0.0827	0.1541
c (112,62)	0.1962	0.0995	0.0563	0.209
k (92,288)	0.0889	0.1834	0.2283	0.0341
VAR-SE				
a (61,160)	0.0165	0.1654	0.2125	0.0933
b (23,304)	0.1966	0.0787	0.0889	0.1996
c (101,216)	0.2021	0.0813	0.0597	0.1989
k (33,243)	0.104	0.1787	0.2215	0.0348
STD-SE				
a (61,160)	0.0165	0.1654	0.2125	0.0933
b (24,304)	0.1412	0.0583	0.0827	0.1541
c (79,16)	0.2465	0.1204	0.056	0.2528
k (92,288)	0.0938	0.1766	0.2207	0.0294



(a) Full Bands (b) MVPCA (c) LCMV-CBS BDM (d) LCMV-CBS BCC/ BDC (e) MAD-CB (f) VAR-CB (f) STD-CB (g) MAD-SE (h) VAR-SE (i) STD-SE

Fig. 4. Extracted endmembers by VCA using full bands and the selected bands given in table I.

VI. CONCLUSION

The proposed techniques for dimensional reduction and target detection give better results as compare to LCMV-CBS and MVPCA. The results shows that if the dimensions of Hyperspectral data are reduced by clustering the band images using their statistical parameters, then it gives better results of unmixing and detection than other techniques like LCMV-CBS and MVPCA etc. In proposed technique of band clustering and selection using K-means method, band from each cluster is selected such that intra-cluster variance is kept maximum and inter-cluster variance is minimum. Furthermore from our proposed technique, STD-SE is better than others. The proposed technique is simple to implement and computes the result very fast. The computation takes seconds for band clustering and selection. All the endmember/ targets are detected well and have high spectral

similarities. Therefore it is concluded from the results of experiments that the proposed clustering techniques are promising and authentic techniques for band clustering and band selection.

REFERENCES

[1] C.-I. Chang, "Hyperspectral imaging: Techniques for spectral detection and classification," *Chapter 2, Kluwer Academic/Plenum Publishers, New York*, 2003

[2] R. Smith, "Introduction to hyperspectral imaging with tmips," *MicroImages Tutorial Web site*, July 2006

[3] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Transactions on Geosciences And Remote Sensing*, vol. 44, no. 6, pp. 1575-1585, 2006

[4] G. Petrie, P. Heasler, and T. Warner, "Optimal band selection strategies for hyperspectral data sets," *Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS '98. 1998 IEEE International*, vol. 3, pp. 1582-1584, July 1998.

[5] P. Groves and P. Bajcsy, "Methodology for hyperspectral band and classification model selection," *IEEE Workshop on Advances in Techniques for Analysis of Re- motely Sensed Data*, pp. 120-128, 2003.

[6] H. Du, H. Qi, X. Wang, R. Ramanath, and W. E. Snyder, "Band selection using independent component analysis for hyperspectral image processing," in *Applied Imagery Pattern Recognition Workshop, IEEE Computer Society*, Los Alamitos, CA, USA, pp. 93-98, 2003.

[7] A. M. U. Uso, F. Pla, J. M. Sotoca, and P. G. Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 45, pp.158-4171, Dec. 2007

[8] C.-I. Chang. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York: Plenum, 2003

[9] J. M. P. Nascimento and J. M. B. Diasb, "Independent component analysis applied to unmixing hyperspectral data," in *Image and Signal Processing for Remote Sensing (L. Bruzzone, ed.)*, Spie, Bellingham, WA, vol. 5238, pp. 306-315, 2004.

[10] C.-I. Chang and S. Wang, "Constrained band selection for hyperspectral imagery," *IEEE Transactions on Geosciences And Remote Sensing*, vol. 44, no. 6, pp. 1575-1585, 2006

[11] C.-I. Chang, Q. Du, T.-L. Sun, and M. L. G. Althouse, "A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 2631-2641, Nov. 1999.

[12] I. U. Haq and X. Xu "A new approach to band clustering and selection for hyperspectral imagery," *IEEE ICSP Proceedings* 2008.

[13] G. Vane, R. Green, T. Chrien, H. Enmark, E. Hansen, and W. Porter, "The airborne visi-ble/infrared imaging spectrometer (aviris)," *Remote Sensing of the Environment*, no. 44, pp. 127-143, 1993.

[14] USGS Spectroscopy Lab – Cuprite. [Online]. Available: http://speclab.cr.usgs.gov/cuprite.html.

[15] G. Swayze, S. S. R. N. Clark, and A. Gallagher, "Ground-truthing aviris mineral mapping at cuprite, nevada," in *Third Annual JPL Airborne Geosciences Workshop*, pp. 47–49, 1992