Spectral Compression of Multispectral Images using Outlier Modeling and Subspace Clustering

by

Farnaz Agahian

B.Sc., Isfahan University of Technology, 2002

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

© Farnaz Agahian 2013

SIMON FRASER UNIVERSITY

Fall 2013

All rights reserved.
However, in accordance with the *Copyright Act of Canada*, this work may be reproduced, without authorization, under the conditions for "Fair Dealing." Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review and news reporting is likely to be in accordance with the law, particularly if cited appropriately.

Approval

Name: Farnaz Agahian

Degree: Master of Science

Title of Thesis: Spectral Compression of Multispectral Images using Outlier Modeling and Subspace Clustering

Examining Committee:

Chair: Dr. Binay Bhattacharya
Professor

Dr. Brian Funt
Senior Supervisor
Professor

Dr. Greg Mori
Supervisor
Associate Professor
School of Computing Science

Dr. Ghassan Hamarneh
Internal Examiner
Associate Professor
School of Computing Science

Date Defended/Approved: December 6, 2013

Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the non-exclusive, royalty-free right to include a digital copy of this thesis, project or extended essay[s] and associated supplemental files ("Work") (title[s] below) in Summit, the Institutional Research Repository at SFU. SFU may also make copies of the Work for purposes of a scholarly or research nature; for users of the SFU Library; or in response to a request from another library, or educational institution, on SFU's own behalf or for one of its users. Distribution may be in any form.

The author has further agreed that SFU may keep more than one copy of the Work for purposes of back-up and security; and that SFU may, without changing the content, translate, if technically possible, the Work to any medium or format for the purpose of preserving the Work and facilitating the exercise of SFU's rights under this licence.

It is understood that copying, publication, or public performance of the Work for commercial purposes shall not be allowed without the author's written permission.

While granting the above uses to SFU, the author retains copyright ownership and moral rights in the Work, and may deal with the copyright in the Work in any way consistent with the terms of this licence, including the right to change the Work for subsequent purposes, including editing and publishing the Work in whole or in part, and licensing the content to other parties as the author may desire.

The author represents and warrants that he/she has the right to grant the rights contained in this licence and that the Work does not, to the best of the author's knowledge, infringe upon anyone's copyright. The author has obtained written copyright permission, where required, for the use of any third-party copyrighted material contained in the Work. The author represents and warrants that the Work is his/her own original work and that he/she has not previously assigned or relinquished the rights conferred in this licence.

Simon Fraser University Library
Burnaby, British Columbia, Canada

revised Fall 2013

Abstract

This thesis introduces two different approaches for improving the performance of data reduction and reconstruction of multi-spectral images. First we introduce a new Outlier Modeling (OM) method that detects, clusters and separately models outliers with their own principal bases. In the second part of this research, a sub-space clustering strategy is used for the spectral compression of multi-spectral images. Unlike classic PCA, this approach finds clusters in different subspaces of different dimension. Consequently, instead of representing all spectra in a single low-dimensional sub-space of a fixed dimension, spectral data are assigned to multiple sub-spaces with a range of dimensions from one to eight. As a result, more resources can be allocated to those spectra that need more dimensions for accurate representation and fewer resources to those that can be modeled using fewer dimensions. This initial compression step is followed by JPEG2000 compression in order to remove the spatial redundancy in the data as well.

Keywords: Multispectral Imaging, Spectral Compression, Principal Component Analysis, Outliers, Clustering, Sub-Space Clustering.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Term	Initial components of the term
GPCA	Generalized Principal Component Analysis
MS-PCA	Multiple Sub-spaces Clustered Principal Component Analysis
MS-CPCA	Multiple Sub-spaces Principal Component Analysis
OM	Outlier Modeling
PCA	Principal Component Analysis
SC	Sub-Space Clustering

Chapter 1

Introduction

Digital imaging refers to the process of producing digital images from a real-world scene using digital imaging devices such as digital camera or scanner. Spectral imaging is defined as the "capture, processing, display, and interpretation of images with a high number of spectral channels (Fairchild et al., 2001)".The number of channels is considered as the main difference between typical color imaging and spectral imaging so as it is restricted to three in the former while can range to several hundred in later.

In this chapter, first the structure of spectral images will be described and a discussion will be done on benefits and drawbacks of multispectral images. This will be followed by a review on related work performed on spectral/multispectral image compression. The chapter will then include a brief overview on some applications of multispectral images. Finally, our motivation and the main contributions of this thesis will be stated.

1.1 Multispectral Imaging

Unlike typical digital photography, the multispectral imaging systems based on acquiring the light reflecting back at each pixel of an image provide a device-independent representation that can be rendered in the correct color under any viewing condition. The spectral reflectance defines an excellent "fingerprint" of a surface and provides the most useful and accurate information for color representation under any illuminant and for any observer. Since this type of information is completely independent of the characteristics of the acquisition

device, it can be used for precise spectral-based color reproduction under different viewing and lighting conditions (Agahian et al., 2012).

1.1.1 Structure of Multispectral Images

Unlike conventional imaging systems, multispectral cameras produce a multi-layer image in which at each layer the pixel values are non-negative numeric values corresponding to the spectral power at one narrow wavelength band. In other words, by increasing the number of channels beyond the traditional three channels of color imaging, the captured image will contain both the spatial features of the scene as well as the spectral information at each pixel. In practice, the spectrum is sampled using a large but finite number ($n >> 3$) of narrow-band spectral filters (Konig & Praefcke, 1998).

1.1.2 Multispectral Image Capture

The most common configuration for multispectral image acquisition systems is to use a monochrome CCD camera in conjunction with a set of narrow band interference filters (Imai & Berns, 1999). The position of the filters could be either between the light source and the original scene or between the scene and the camera. In the absence of fluorescence, the two configurations will lead to the same results (Konig & Praefcke, 1998). This technique of image capture suffers from some technical problems that make it unrealistic and impractical. For example, inter-reflections between the original scene and interference filters and also between filters and the camera lens are an issue that should be taken into account (Imai & Berns, 1999).

Another filtering technology to produce spectrally narrow samples is a tunable filter such as liquid crystal tunable filter (LCTF). The LCTF has the advantage of being easily controllable and reliably repeatable. In addition, the angular sensitivity discussed earlier is not a concern in the case of using a liquid crystal tunable filter (Imai et al., 2000).

Another alternative system for multispectral image capture is to use a conventional trichromatic digital camera modulated by a small number of wide-band filters (Imai et al., 2000). Finding the optimal (minimum) number of filters is a crucial issue in multispectral imaging and has received a great deal of attention in recent decades (Connah et al., 2004; Hauta-Kasari et al., 1999; Sharma et al., 1998; Vrhel & Trussell, 1994).

1.1.3 Spectral Imaging: Bane or Boon

As mentioned earlier, a typical multispectral camera uses a small number of channels to capture spectrally narrow lights reflected from the original scene, and then uses an approximation algorithm (e.g. linear models) to estimate the spectral information of each pixel from the camera responses. The number of required channels is dependent on the application and can range from 5-9 in color imaging, to a hundred channels and over a much greater wavelength range in remote sensing applications. The latter is also called hyper-spectral imaging (Fairchild et al., 2001).

Since spectral reflectance data is totally independent of the lighting, as well as characteristics of the acquisition device, it represents much better the actual scene than device-dependent colorimetric values (such as R, G, B). Multispectral images will enable us to precisely calculate and reproduce colors across different illuminants and multiple observers. In the other words, by using spectral image technology, we can produce images that are robust to changes in illumination and consequently prevent the negative effects of metamerism, which often arise in conventional trichromatic color imaging and reproduction.

System design optimization, transformation of image appearances across changes in viewing conditions, high-accuracy color printing, and optimal separation algorithms for multi-ink printing are the other advantages of using multispectral images (Burns & Berns, 1996).

Although the extra information provided by a multispectral imaging device can be very useful, the large amount of data can be a problem in terms of storage and communication requirements. Digital image compression is an important task in

image processing and provides efficient solutions for storage of a large volume of image data (Du & Fowler, 2007; Kaarna et al., 1998; Penna et al., 2007).

1.1.4 Applications of Spectral Imaging

The development of spectral imaging goes back to defense and non-defense remote sensing applications such as target detection, material mapping and material identification (Freeman et al., 1997). In the last decades, multispectral imaging has gained a growing interest in medical imaging so as the advancement of multispectral imaging technologies in medical diagnosis applications have been incredible. In fact, multispectral images leading to high fidelity color reproduction under different lighting conditions is valuable in medical applications such as dermatology, surgery video, telemedicine and pathology (Yamaguchi, 2001).

Since 1990s, the application of spectral imaging was extended to artwork conservation via making highly accurate image archives with high color accuracy. The traditional techniques of image capture used to archive artwork in the most of the museum of the world rely on the conventional photographic process. Photography has the disadvantage of dependency on the scene illuminant and suffers from poor color accuracy under different illuminants particularly when high-fidelity color reproduction is required as, for example, in the reproduction and conservation of fine arts painting. In recent decades, libraries and museums such as the National Gallery of Art in Washington, D.C and the Museum of Modern Art in New York have been trying to develop spectral imaging systems based on acquiring the spectral information of an image optimized for artwork imaging, archiving, and reproduction. This is in response to the need to build digital image databases with adequate colorimetric accuracy (Imai et al., 2001; Berns et al., 2001; Zhao et al., 2005; Maitre et al., 1996).

1.2 Spectral Compression

Spectral Analysis of both natural and man-made colorants shows that the spectral reflectances of non-fluorescent objects have smooth shapes and possess a

high degree of correlation. It means one should be able to represent a spectrum as a linear combination of a small number of principal spectra without a significant loss of information. The numbers of principal spectra for acceptable recovery of spectral data depends on the spectral dataset and the required precision.

Cohen was the first to fit a linear model to a set of surface spectral reflectance curves. He calculated the characteristic vectors of a subset of 150 spectral reflectance of 433 Munsell chips and showed that 99.18 % of the total variance of Munsell colors can be explained just using these three vectors (Cohen, 1964).

Maloney performed this experiment on 462 specimens and found that five to seven components are required for an accurate representation (Maloney 1986). However, Parkkinen's experiments on the spectral reflectance of 1257 Munsell color revealed that the spectra can be modeled precisely by using as many as eight principal components (Parkkinen et al., 1989). Hardeberg used PCA to find the effective dimension of five different databases. He showed that the effective dimension of different datasets are different and strongly depend on the statistical properties of the dataset (Hardeberg, 2002).

Laamanen et al. accomplished spectral compression of 1269 reflectance spectra of the Munsell chips and a set of 922 reflectance spectra of the samples in the Pantone Color Formula Guide using PCA and ICA (Independent Component Analysis) and showed that ICA performs slightly better but the difference was not very significant. He also stated that dimensionality might be around 20 if a general basis is used to represent every spectrum with high accuracy. Nonetheless, he emphasized that the required principal components for a given reconstruction accuracy depend on the database and the basis used (Laamanen et al., 2001).

In another effort Laamanen et al. presented a weighted compression method for spectral color information. The method is based on classic PCA, however the spectral data are weighted with a weight function before the eigenvectors are calculated. The goal was to retain color information in the compression process (Laamanen et al., 2008). Agahian et al. performed the same experiment and examined different weighting factors and found that the weighting factor based on

the square root of the principal diagonal of matrix $\boldsymbol{R}$ lead to the least colorimetric errors (Agahian et al., 2014).

Kaarna and Parkkinen developed three methods for compression of multispectral images. In the first method, PCA was preceded by clustering to remove spatial and spectral redundancy. The second method was based on the wavelet transform and the third one was a combination of PCA and wavelet for spectral and spatial compression (Kaarna & Parkkinen, 2000). Du and Fowler deployed PCA in JPEG2000 to provide spectral decorrelation as well as spectral dimensionality reduction of hyperspectral images. They showed the superiority of this coder to the classic JPEG2000. They also stated that the best results are gained when a reduced number of principal components are retained and coded (Du & Fowler, 2007). Rayat et al. proposed using Box-Cox transformation technique before applying PCA on spectral datasets to improve the efficiency of employed compression technique by increasing the degree of normality in the dataset (Rayat et al., 2012).

In summary, multispectral images are large in size and consist of a high degree of spectral as well as spatial redundancies. This can be considered as a serious problem in storage and communication. Thus, compression methods for the multispectral images must be developed. This is still an open research topic.

1.3 Motivation

Analysis of the spectral reconstruction of 1269 Matte Munsell color chips (Munsell Book of Color, 1976) indicates that some color samples, mostly in the family of purples, have a detrimental effect on the spectral reconstruction error of the whole dataset. Almost half of these samples are statistically outliers with respect to the other samples. Further investigation also shows that nearly 70% of the Munsell spectral whose reconstruction error is more than the median error of the whole dataset have a large robust Mahalanobis distance from the mean. This observation motivated us to study the effect of outlier spectra in datasets of reflectance spectra and propose that they be modeled by their own separate PCA basis. Chapter 2 will cover this idea in detail.

In the second part of the research we took advantage of the assumption that, with a high probability, a complex color image generally contains many pixels with similar colors. Color similarity is a good indicator for spectral similarity. Therefore, it seemed to us a good idea to reduce the spectral redundancy in the image by finding groups of spectra with a high degree of similarity that then can be modeled well with a very small number of principal components. This helps us devote more resources for the spectra that have a different pattern from the majority of image pixels.

This observation motivated us to examine data reduction in multiple sub-spaces with various dimensions and subsequently propose a compression strategy relying on that. This method, called Multiple Sub-spaces PCA (MS-PCA), will be described extensively in chapter 3.

1.4 Contributions of the Thesis

The main contributions of this thesis can be summarized as:

a) Introducing the Outlier Modeling (OM) method for spectral data reduction based on the following steps: (1) separate the outliers from the bulk of the data (the inliers); (2) apply k-means clustering to the outliers and refine clusters iteratively; (3) apply PCA data reduction to the clusters (both the inliers and outlier clusters) individually.

b) Describing a sub-space clustering strategy for the spectral compression of multispectral images relying on finding clusters in several subspaces of different dimension. In fact, instead of representing all spectra in a single low-dimensional sub-space of fixed dimension as classic PCA does, spectral data are assigned to multiple sub-spaces with dimensions ranging from 1 to 8. This approach is modified by a preceding preprocessing step, which groups the initial dataset into 4 distinct clusters. Increasing spectral similarity in each cluster enhances the efficiency of our proposed spectral compression approach.

Chapter 2

Outlier Modeling for Spectral Data Reduction

2.1 Introduction

It is well documented that the spectral reflectance of a non-fluorescent object is generally a smooth function of wavelength, and therefore can be modeled via dimensionality reduction techniques. Principal component analysis (PCA) is a well-known technique in multivariate data analysis (Jolliffe, 2002), which has been extensively used in the context of spectral imaging as an efficient technique for spectral decorrelation as well as spectral dimensionality reduction (Tzeng & Berns, 2005). PCA uses an orthogonal linear transformation to convert spectral data from the high-dimensional spectral space into the low-dimensional spectral subspace. Among all linear transformations it guarantees the best possible representation of the high-dimensional spectral vector in the low-dimensional subspace spanned by the small number of basis vectors. This feature has made PCA a powerful tool for spectral compression.

It should be noted that the projected data can be reconstructed back into the original space; however, the compression process will usually lead to some error in the reconstructed data. According to Laamanen et al. the number of basis vectors required for effective recovery of reflectance spectra crucially depends on the type of data involved and the basis vectors that are used. Obviously, the more correlated the input data, the better the result (in terms of reconstruction error) that is achievable when using PCA (Laamanen et al., 2001). Applying weighting factors on

individual samples (Agahian et al., 2008) and clustering the main dataset based on a predefined criterion (Garcia-Beltran et al. 1998 ; Ayala et al., 2006) are techniques that have been used to enhance the efficiency of linear models by increasing the similarity of the elements in the dataset.

One of the problems with PCA modeling is that in each dataset there are usually some elements that may be a long way from the remainder of the data or do not conform to its correlation structure. Such elements are known as *outliers* and they can have a substantial deleterious effect on the results of the dataset analysis. Therefore, it is desirable to remove or reduce the effect of such observations before applying PCA on a dataset (Jolliffe, 2002).

In this chapter, we propose a new Outlier Modeling (OM) method for spectral data reduction based on the following steps: (1) separate the outliers from the bulk of the data (the inliers); (2) apply k-means clustering to the outliers and refine clusters iteratively; (3) apply PCA data reduction to the clusters (both the inliers and outlier clusters) individually.

Fig (2.1) illustrates this procedure schematically. The effectiveness of the proposed OM method is demonstrated in tests described below using one spectral dataset and 18 multispectral images.

2.2 Mathematical Background

2.2.1 Spectral Compression and Reconstruction

Principal Component Analysis (PCA) is a well-known technique in multivariate data analysis first introduced by Pearson (Pearson, 1901), and developed by Hotelling in 1933 (Hotelling , 1933) and has found many applications in different areas such as feature extraction and data compression.

PCA is one of the most widely used techniques in compression of large spectral images and guarantees the best possible representation of the high-dimensional spectra in a low-dimensional eigenvector sub-space. This compression method gives an equal treatment to all wavelengths throughout the spectrum and tries to

minimize the squared reconstruction errors between the actual and reconstructed spectra:

$$\mathbf{e} = \sum_{\lambda=380}^{780} \left(\mathbf{R}_\lambda - \hat{\mathbf{R}}_\lambda \right)^2 \rightarrow \text{Min} \qquad (1)$$

Where $\mathbf{R}_\lambda$ is the actual and $\hat{\mathbf{R}}_\lambda$ is the reconstructed reflectance.

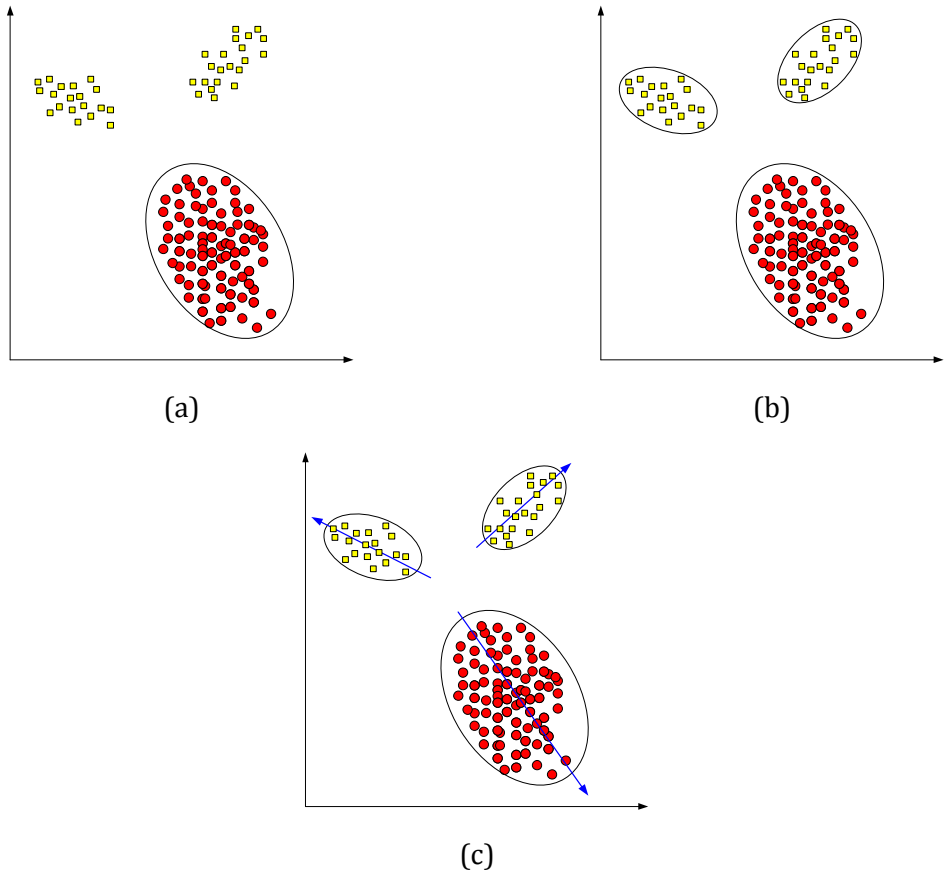


Fig 2.1 Schematic of the proposed method for spectral data reduction: (a) separate the outliers (yellow squares) from the non-outlier spectra (red circles); (b) apply k-means clustering to the outliers; (c) apply PCA data reduction to all clusters (inliers and outliers).

Here the major steps of implementation of the classic PCA in data compression are indicated.

Assume that we have N P-dimensional vectors $\mathbf{x_i} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_p}]^T$ aligned in the data matrix $\mathbf{X} \in R^{P \times N}$. In this thesis, x_i corresponds to a 31-dimensional spectral reflectance. We start with the covariance matrix calculation as follows:

$$\mathbf{C} = \frac{1}{\mathbf{N}} \tilde{\mathbf{X}} \tilde{\mathbf{X}}^T \qquad (2)$$

Where $\tilde{\mathbf{X}}$ is the zero mean dataset.

Matrix **C** is a $P \times P$ matrix with P eigenvectors and P eigenvalues, which can be diagonalized as follows:

$$\mathbf{CV=VD} \qquad (3)$$

where the columns of the matrix $\mathbf{V_{P \times P}}$ are the eigenvectors $\mathbf{v}_i$ and diagonal elements of the matrix **D** are the corresponding eigenvalues d_i.

As the first few eigenvectors (correspond to the largest eigenvalues) account for the most variability of the dataset, we can simply keep $M < P$ vectors and discard the rest. Therefore, the orthonormal matrix $\mathbf{V_{P \times M}}$ can serve as a linear transformation matrix projecting data from high-dimensional spectral space to a low-dimensional subspace and vice versa:

$$\mathbf{Z} = \mathbf{V}^T \tilde{\mathbf{X}} \qquad (4)$$

where **Z** is an $M \times N$ matrix including coordinates of spectral reflectance data in the low-dimensional subspace In this way, PCA removes the high degree of spectral redundancy by decorrelating the original spectra and provides a more manageable size of data for storing and transmission purposes.

2.2.2 Outlier Detection using Multivariate Statistical Methods

Detection of outlying observations is a primary step in many statistical analyses. Johnson (Johnson, 1992) defines an outlier as an observation in a dataset that appears to be inconsistent with the remainder of that set of data. This definition

represents well the importance of outlier detection before data analysis since the presence of such data could lead to an incorrect result.

The Mahalanobis distance is a measure based on the correlation between variables and has been widely used to detect multivariate outliers. For a multivariate vector $\mathbf{x_j} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_p}]^T$ from a dataset with mean $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]$ and covariance matrix **S** the Mahalanobis distance is defined as

$$\mathrm{MD}(\mathbf{x_i}) = \sqrt{(\mathbf{x_i} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x_i} - \boldsymbol{\mu})} \qquad (5)$$

Multivariate outliers can be defined as observations having a large Mahalanobis distance. A quantile of the chi-squared distribution ($\sqrt{\mathbf{X}^2_{\mathbf{P},0.975}}$) is usually considered as the cutoff value. However, this approach does not provide a reliable measure for multiple outliers because of the masking effect collectively created by them, which means that they do not necessarily have a large MD. Therefore, it helps to estimate the mean and covariance of the dataset using a robust procedure (Filzmoser, 2004; Rousseeuw & Van Driessen, 1999). There exist several robust estimators for mean and covariance. The minimum covariance determinant (MCD) (Rousseeuw & Van Driessen, 1999; Rousseeuw, 1984) is widely known in the literature as a computationally fast algorithm and is the used here.

The MCD objective is to find h observations (out of N) whose classical covariance matrix has the lowest determinant. The MCD estimate of the mean is then the average of these h points. The MCD estimate of scatter is their covariance matrix. In 1999 Rousseeuw and Driessen (Rousseeuw & Van Driessen, 1999) proposed the algorithm FAST_MCD, which is specifically tailored to the properties of the MCD estimator. This algorithm is a useful and robust tool for multivariate data analysis, exploring data structure as well as outlier detection in large datasets. A complete description of the algorithm is presented in (Rousseeuw & Van Driessen, 1999). A Matlab library for robust analysis is readily available (Libra, 2006).

In summary, the steps of OM for the spectral encoding phase are:

(1) Separate the outliers from the inliers based on each spectrum's Mahalanobis distance computed based on robust estimates of the dataset's mean and variance obtained using FAST_MCD;

(2) Cluster the outliers into a small number of clusters using k-means and then refine the clusters by using an iterative procedure to make sure each spectrum has been assigned to the best-fit cluster;

(3) Apply PCA to each of the clusters (outlier clusters as well as the original inliers) and retain only a small number (typically 3) of the original basis vectors along with each cluster's identifier;

(4) Project each original spectrum onto its cluster's (reduced) basis to obtain the principal components (weighting coefficients),

(5) Store the weights from (4) along with each spectrum's corresponding cluster's identifier.

A spectrum is reconstructed using the weights and cluster identifier. The reconstructed spectrum is simply the weighted linear combination of the basis vectors associated with the given cluster.

2.3 Testing on Munsell Spectral Dataset

We first tested outlier modeling on the 1269 reflectance spectra of the chips in the Munsell Book of Color – Matt Finish Collection (Munsell Book of Color, 1976). The spectra were measured by the color research group of Eastern Finland University with Perkin Elmer Lambda 18 spectrophotometer and the wavelength range was from 380 nm to 800 nm with 1 nm interval. In the current research, the reflectance data were fixed between 400 nm to 700 nm at 10 nm intervals. The results of testing the OM compression method on the Munsell matte collection are presented in detail below.

2.3.1 Munsell Matte Collection - Outlier Detection Results

The FAST_MCD algorithm was applied on the 1269 spectra of Munsell dataset in order to separate the outliers from the rest of the dataset. When using the MCD robust estimator on the 1269 reflectances, 533 spectra were detected as outliers. Fig (2.2) compares the result of using MD_{MCD} (using FAST_MCD) to $MD_{classic}$ (i.e., MD

as defined in Eq. 5) from which it is clear that they lead to very different sets of outliers. The red line represents the quantile cutoff value of $\sqrt{\chi^2_{31,0.975}}$=6.94 for classification as an outlier. Based on this criterion, 533 of the 1269 spectra are classified as outliers by MD_{MCD} versus only 244 by $MD_{classic}$. It is worth noting that a multivariate outlier that is not an extreme value for any of the original variables (i.e., wavelengths) can still be an outlier if it is inconsistent with the correlation structure of the remainder of the data (Jolliffe, 2002). The dataset is divided into outliers and non-outliers for the subsequent processing steps, which involve outlier clustering and then applying PCA to each created cluster as well as to the inlier spectra to reduce data dimensionality from 31 to 3.

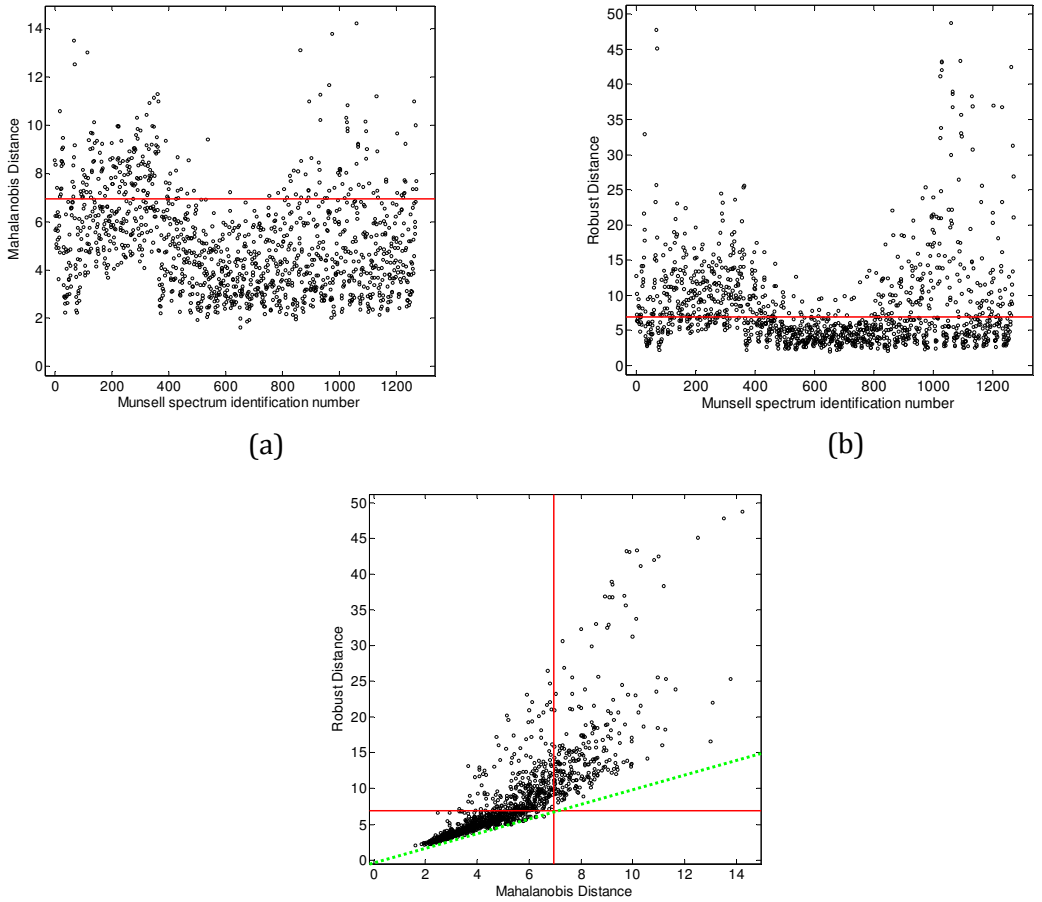


Fig 2.2 Distance measures for the 1269 Munsell spectra: (a) Classic Mahalanobis distance $MD_{classic}$ versus Munsell identifier; (b) Robust distance MD_{MCD} versus Munsell identifier; and (c) MD_{MCD} versus $MD_{classic}$ for each Munsell identifier. The horizontal and vertical lines represent the quantile cutoff value.

2.3.2 Munsell Matte Collection Outlier Clustering Results

The outlier spectra are a part of the original dataset so they cannot simply be ignored. To represent the outliers, PCA is applied separately to the set of outliers as well as to the set of inliers. By having separate PCA-derived bases for the inliers and the outlier clusters a better representation of the entire dataset is obtained than if only a single basis were to be used.

To represent the outliers we group them into several clusters based on a similarity measure such that the spectra in each group are very similar. However, determining the appropriate number of clusters to form is an issue in itself. For this step, subtractive clustering as implemented in Matlab's subclust function is used to determine the minimum number of potential clusters. This is done by gradually decreasing the number of clusters and calculating the corresponding mean normalized RMS (NRMS) error in spectral reconstruction (NRMS definition will be given in section 2.3.3). As Fig (2.3) shows, data clustering has a significant effect on the reconstruction error as the number of clusters increases from 1 to 5. Beyond 5, it reaches a plateau. Based on this analysis, outlier spectra are partitioned into 5 groups as a trade-off between reconstruction accuracy and data redundancy (the more clusters, the more basis vectors that must be included in the data to be stored/transmitted).

For the Munsell matte collection, using 5 clusters of outliers works well. The clustering is done using k-means clustering (kmeans from the statistics toolbox of Matlab (Matlab (a), 2013)) with the cosine distance as the distance parameter. The cosine distance between two spectra is the cosine of the angle between them viewed as vectors.

In order to improve the performance of clustering, K-means was followed by an iterative refinement procedure. The details of this method are given in the following section.

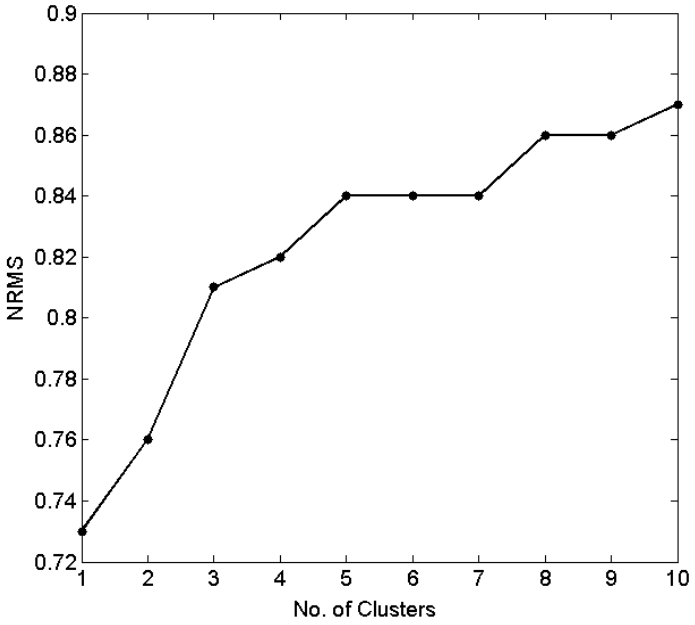


Fig. 2.3 Mean reconstruction error for the outlier spectra (excluding the inliers) as a function of the number of clusters used.

Iterative refinement method

Given an initial clustering, we can fit a low-dimensional sub-space to each cluster using classic PCA. Then given a linear PCA model for each cluster, we can assign each spectra to its closest subspace based on NRMS error and then re-estimate the sub-spaces. The convergence of the algorithm is guaranteed by the fact that there are finite (say N) number assignments of spectra to subspaces. Thus, the optimum can be found in at most N iterations (Vidal, 2011).

In this research, the number of iterations was set equal 20 based on a pre-experiment performed on some sample datasets. After each iteration, the mean of normalized root mean squares error was calculated. As Fig (2.4) shows, for the first few iterations, the amount of error decreases dramatically (i.e., NRMS increases). In the case of the current spectral dataset, the error after about 9 iterations levels off.

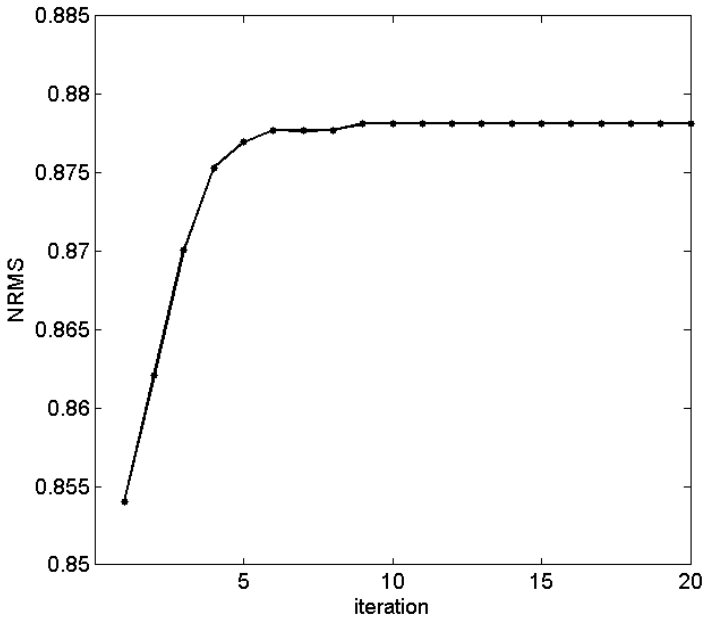


Fig. 2.4 Mean of reconstruction error for the outlier spectra versus the number of iterations.

The output of the iterative procedure is five refined clusters such that the reconstruction error of each spectral reflectance in its own subspace is less than the others, i.e. each data point falls in the best-fit sub-space.

2.3.3 Munsell Matte Collection - Spectral Reconstruction Results

For each cluster, PCA is used to reduce its dimension from 31 to 3. In total, the full dataset is partitioned into 6 clusters (counting the inlier set as a single cluster). As a result, 18 basis vectors (6 clusters × 3 bases) are required in order to the data from each cluster to be projected into its own 3D spectral sub-space. The spectra are then reconstructed using each cluster's corresponding basis. The reconstruction error is tabulated in Table (2.1) in terms of spectral accuracy. Two spectral measures, normalized root mean squares error NRMS and Goodness of Fit Coefficient, GFC are used to evaluate the goodness of the mathematical reconstruction.

Normalized RMS error is calculated using the following equation:

$$\mathrm{NRMS} = 1 - \frac{\left\|\mathbf{R}_{\mathrm{m}} - \mathbf{R}_{\mathrm{e}}\right\|}{\left\|\mathbf{R}_{\mathrm{m}} - \mathrm{mean}(\mathbf{R}_{\mathrm{e}})\right\|} \qquad (6)$$

where, || indicates the 2-norm of a vector. $\mathbf{R}_{\mathrm{m}}$ and $\mathbf{R}_{\mathrm{e}}$ are the measured and estimated spectral data, respectively.

NRMS costs vary between -∞ (bad fit) to1 (perfect fit). This function is one of built-in functions available in Matlab (Matlab (b), 2013). In this thesis, in addition to the average NRMS, the percentages of pixels with NRMS larger than 0.95 and NRMS smaller than zero are both calculated and reported. These two are considered a measure of *excellent* and *very poor recovery*, respectively.

The GFC has been proposed by Hernandez-Andres and is based in Schwartz inequality (Hernandez-Andres et al., 1998). It is described by the Eq. (7):

$$\mathrm{GFC} = \frac{\left|\sum_{\lambda} \mathbf{R}_{\mathbf{m}}(\lambda)\mathbf{R}_{\mathbf{e}}(\lambda)\right|}{\sqrt{\left|\sum_{\lambda}[\mathbf{R}_{\mathbf{m}}(\lambda)]^2\right|}\sqrt{\left|\sum_{\lambda}[\mathbf{R}_{\mathbf{e}}(\lambda)]^2\right|}} \qquad (7)$$

The value of GFC range from 0 to 1, where 1 indicates a perfect spectral match. If GFC>0.995, the reconstruction quality is judged acceptable while very good and excellent (almost exact fit) reconstruction require GFC>0.999 and GFC>0.9999.

MD_{MCD} detects and removes 533 spectra as outliers from the main dataset during the first step. The spectral dimension of the remainder of the data labeled *Inlier Cluster* is reduced to 3 via PCA. The inlier cluster benefits from the fact that the outliers have been removed, so the remaining, highly correlated data is efficiently and quite accurately represented using only a 3-dimensional basis. The outlier spectra are partitioned and reduced to 3 dimensions in the second step. As the clustering is based on a similarity measure, the spectra assigned to each cluster are again highly correlated leading to an efficient PCA-based 3-dimensional representation. The two rows of Table (2.1) highlighted in gray show that outlier modeling improves the spectral reconstruction in terms of both the normalized RMS and GFC measures. In other words, the compression ratio (CR) of OM is comparable to the classic PCA method when the number of principal components used to

represent data is equal to 4. It should be noted that the compression ratio is the ratio of the size of the original dataset (uncompressed) to the size of compressed dataset. In the case of classic PCA, the size of the compressed dataset is calculated by adding the required bits for encoding the eigenvectors, the principal component coefficients and the mean vector. However, in the case of OM, the required bits for encoding cluster identifier matrix need to be considered as well. It should be noted that, as the first eigenvector represents the majority of the variance, its weight needs more accuracy (i.e., more bits) than the 2nd weight and so on. Hence the number of resources needed for encoding different principal components vary, with more resources assigned to the first principal components and fewer to the last ones. In the current research the required number of bits for encoding each principal component was determined experimentally. In this experiment, image reconstruction was performed using principal components encoded with different numbers of bits varying from 4 to 8. Then the variation of average NRMS over all pixels versus the number of bits was assessed to find a point where the change of error is insignificant. Fig (2.5) shows this scheme for an arbitrary multispectral image. As a result, 8, 7 and 5 bits are assigned to the first to third components, respectively.

Each eigenvector as well as the mean vector was encoded using 4 bits per component. In fact, encoding eigenvectors with more than 4 bits did not lead to significant reduction of the reconstruction error. As Table (2-1) shows, classic PCA data compression in a 4-dimensional sub-space leads to a compression ratio of 9.36 while the corresponding compression ratio for OM is 9.78.

As another way of comparing PCA to OM, the residuals between the actual and reconstructed reflectance spectra at 31 wavelengths were calculated for all 1269 samples as shown in Fig (2.6). The residuals are much smaller for OM in comparison to classic PCA. If the reconstruction were done perfectly, the error would be a straight line equal to zero.

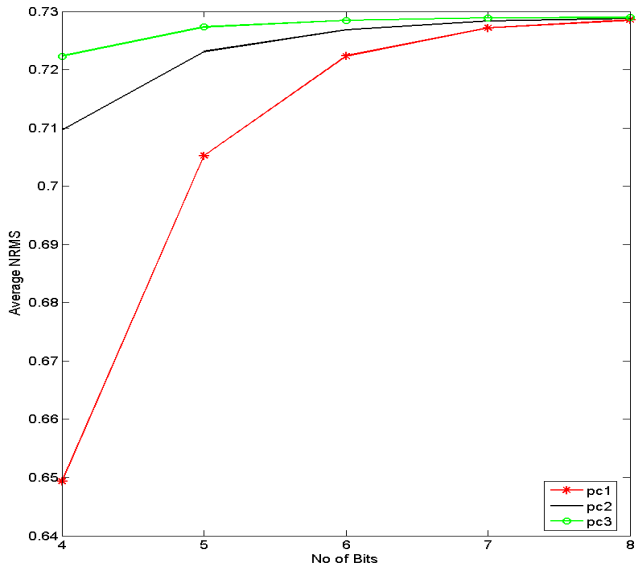
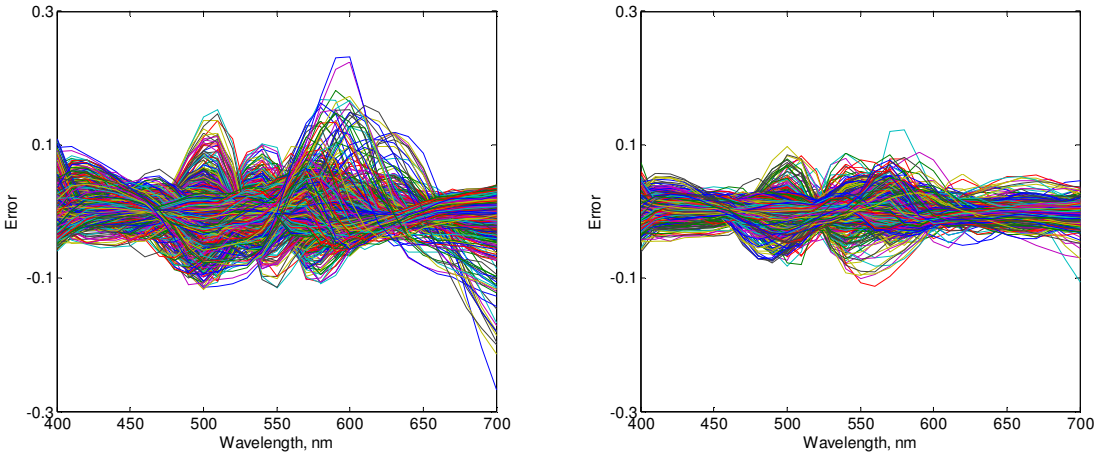


Fig 2.5 The variation of the average NRMS when the principal components of a compressed multispectral image are encoded using different numbers of bits.



(a) (b)

Fig 2.6 Spectral reconstruction error ($R_{original,\lambda} - \hat{R}_{\lambda}$) versus wavelength using for 1269 Munsell spectra: (a) classical PCA; (b) Outlier Modeling.

Table 2.1 Spectral accuracy of reflectance reconstruction of the 1269 Munsell spectra using classic PCA versus Outlier Modeling. The reconstruction error at each step of OM is listed for each cluster separately. The reconstruction errors for the entire dataset for both classic PCA and OM are found in the grey-shaded rows.

	#Spectra	CR	Normalized RMS			GFC	
			Mean	% of samples		% of samples	
				>0.95	<0	>0.995	>0.999
Classic PCA	1269						
3D		11.59	0.66	0	1.34	74.23	26.95
4D		9.36	0.77	0.23	0	88.49	55.32
5D		7.85	0.82	1.97	0	95.82	69.89
Outlier Modeling Step 1							
Inlier Cluster	736		0.76	0	0.17	89.36	53.26
Outliers	533		0.73	0	0.56	71.29	23.64
Outlier Modeling Step 2							
Outlier (Cluster 1)	71		0.85	12.68	0	85.92	60.56
Outlier (Cluster 2)	82		0.90	13.41	0	100	81.71
Outlier (Cluster 3)	162		0.89	16.05	0	99.38	77.16
Outlier (Cluster 4)	120		0.86	2.50	0	100	93.33
Outlier (Cluster 5)	98		0.88	16.33	0	87.76	47.96
Mean	533		0.88	12.19	0	95.68	73.92
Outlier Modeling for the full dataset	1269	9.78	0.81	5.12	0.09	92.01	61.93

2.4 Testing on Multispectral Images

Multispectral images are another source of large numbers of spectra. PCA and OM are compared on 10 multispectral images from the database of Hordley et al. (Hordley et al., 2004), 7 images from the Columbia University multispectral image database (Yasuma, 2008), and the "Fruits and Flowers" image, which is a 120×160

pixel image from the University of Eastern Finland spectral image database[1]. These images are displayed in Fig (2.7) to (2.9). All spectra are sampled at 10 nm intervals over the range 400 nm to 700 nm. Multispectral images taken from the Hordley database were captured in a VeriVide viewing booth with a black cloth background under CIE illuminant D75 using a Spectracube camera (Hordley et al., 2004). The images' borders were removed before analysis, so the number of spectra listed in Table (2.4) is slightly different from the actual number of pixels of the images given in (Hordley et al., 2004).

The multispectral images in the Columbia University were captured using a Cooled CCD camera (Apogee Alta U260) with a resolution of 512 x 512 pixels under CIE Standard Illuminant D65. These multispectral images represent the reflectances of the materials in the scene and were computed from the measured multispectral image using the illuminant spectrum and camera spectral response. Therefore, the spectral reflectance of each pixel is a close approximation of the true reflectance of the scene (Yasuma et al. 2008).

The distance measures MD_{MCD} and $MD_{classic}$ for the 19,200 Fruits and Flowers spectra are compared in Fig (2.10). It is evident that the MD_{MCD} distances differ substantially from those of $MD_{classic}$ (Eq. 5), and this results in two different sets of outliers. Based on the quantile cutoff value of $\sqrt{X^2_{31,0.975}}$, 7741 out of the 19,200 spectra are classified as outliers by MD_{MCD} in comparison to only 3358 by $MD_{classic}$.

[1] Eastern Finland Spectral Image Database, University of Eastern Finland, Spectral Color Research Group, https://www.uef.fi/spectral/spectral-database.

Fig 2.7 Multispectral Images from the database of Hordley et al.

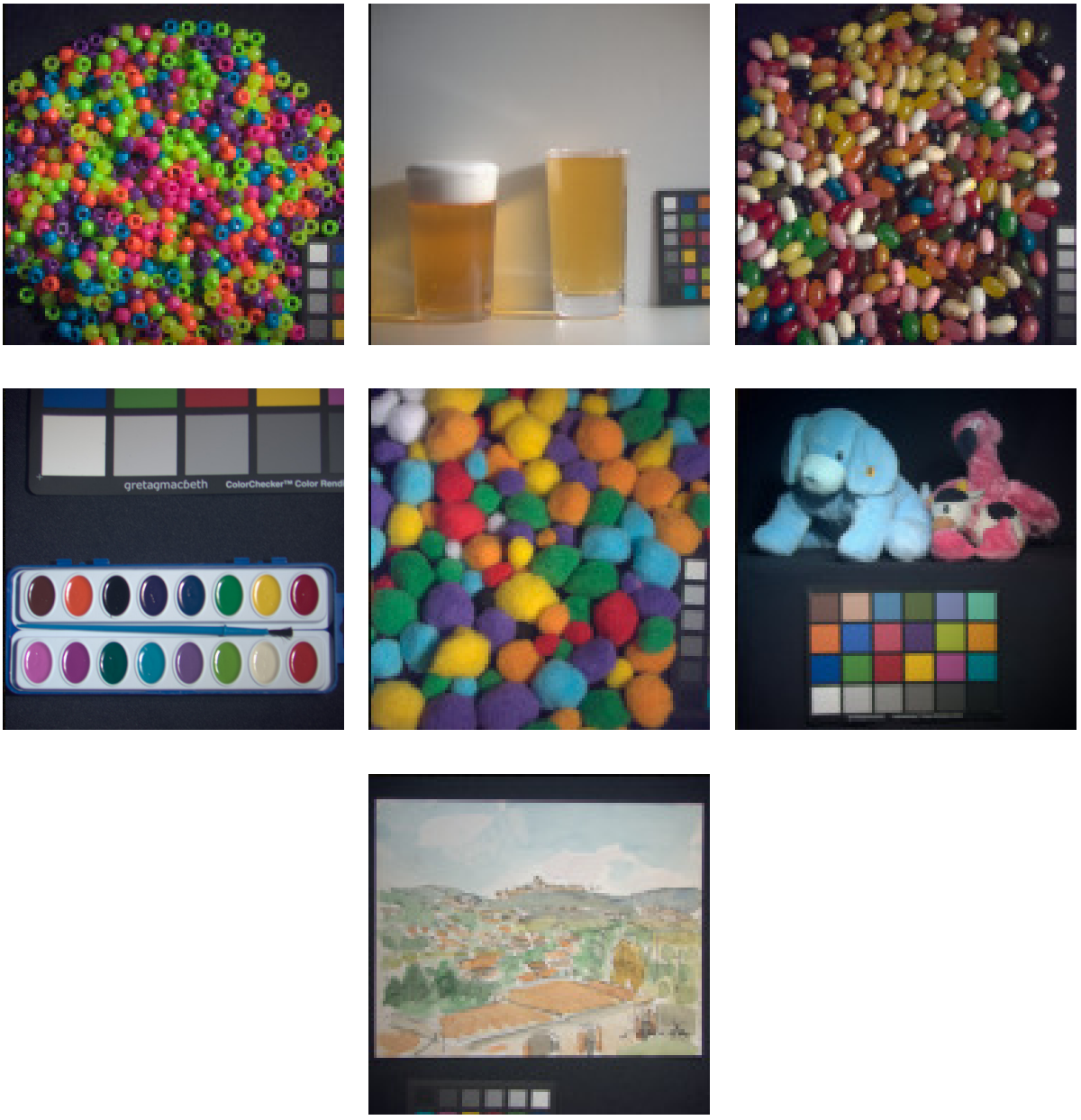
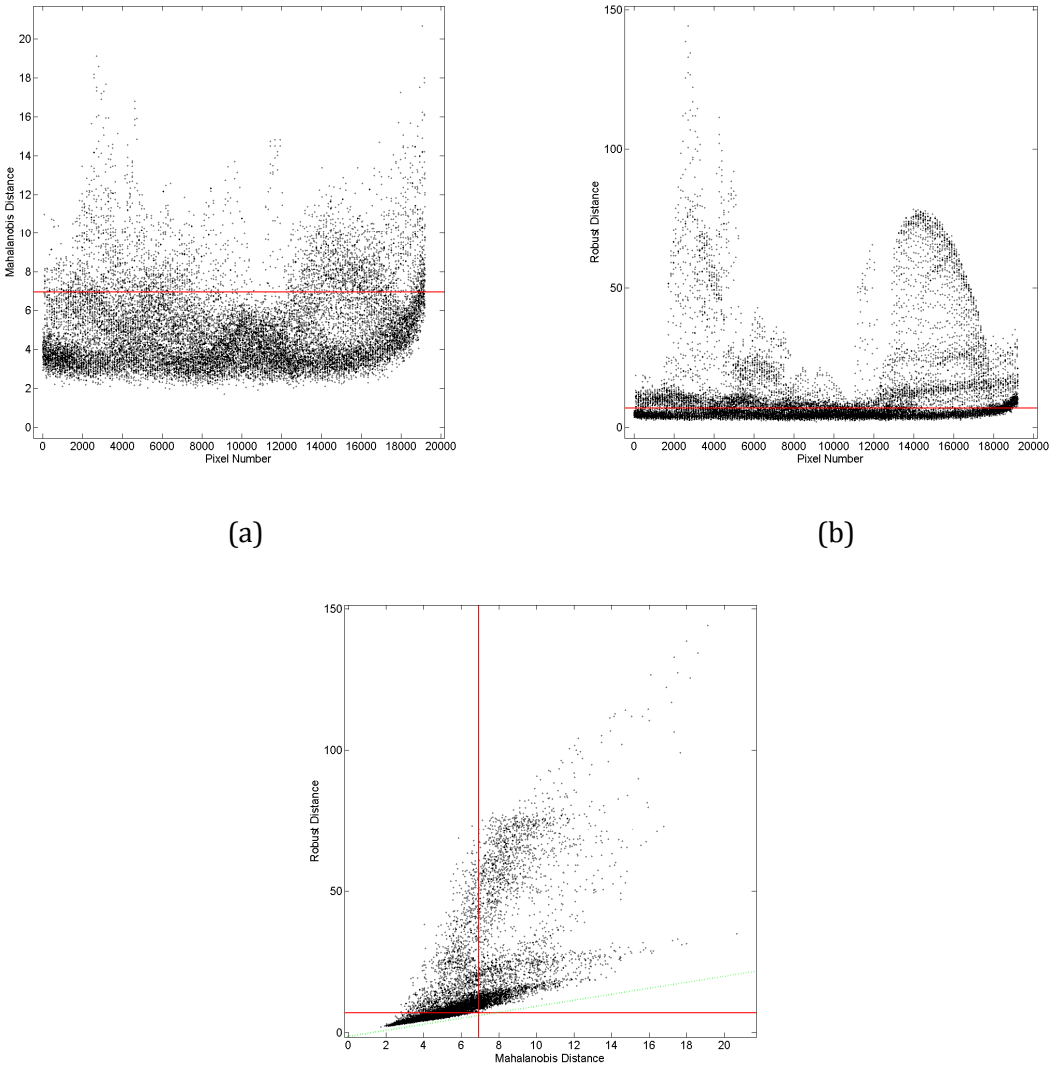


Fig 2.8 Multispectral images from the Columbia University database.

Fig 2.9 "Fruits and Flowers" image from the Eastern Finland University spectral image database.

The resulting set of outliers is partitioned into 5 clusters using K-means. The number of outlier clusters was fixed at 5 for all multispectral images based on our experimentation with subclust on a few sample datasets. In order to improve the performance of clustering, K-means was followed by an iterative refinement as explained in section (2.3.2).

As for the case of the spectral datasets of the previous section, the same analysis was performed for the 19,200 spectra of the "Fruits and Flowers" image. The spectra are divided into 6 clusters consisting of the inliers (1 cluster) and outliers (5 clusters). Applying PCA to each of these clusters and retaining only the first 3 basis vectors leads to a total of 18 basis vectors. The spectra are reconstructed using the appropriate cluster's basis. Table (2.2) lists the spectral reconstruction error in terms of normalized RMS and GFC in detail.



(a) (b)

(c)

Fig 2.10 Comparison of the classic Mahalanobis distance ($MD_{classic}$) and robust distance (MD_{MCD}) for the 19,200 Fruit and Flowers spectra: (a) Classic Mahalanobis distance versus sample number; (b) Robust distance versus sample number; and (c) MD_{MCD} versus $MD_{classic}$. The horizontal red lines represent the quantile cutoffs defining the inlier/outlier boundary.

Table 2.2 Spectral accuracy of reflectance reconstruction for the Fruits and Flowers image using standard PCA versus Outlier Modeling. The reconstruction error for each cluster of spectra is listed separately.

	#Spectra	CR	Normalized RMS			GFC	
			Mean	% of samples		% of samples	
				>0.95	<0	>0.995	>0.999
Classic PCA	19200						
3D		11.8	0.68	2.48	0.71	45.70	22.13
4D		9.52	0.77	7.00	0.40	68.47	31.24
5D		8.00	0.83	11.09	0.02	82.84	37.64
Outlier Modeling Step 1							
Inlier Cluster	11459		0.77	5.35	0	56.42	9.25
Outliers	7741		0.67	5.51	0.94	66.50	39.73
Outlier Modeling Step 2							
Outlier (Cluster 1)	963		0.90	19.94	0	92.63	45.48
Outlier (Cluster 2)	2932		0.82	32.20	0	97.95	84.86
Outlier (Cluster 3)	844		0.84	0	0	91.11	19.67
Outlier (Cluster 4)	2190		0.76	0.55	1.19	96.89	70.55
Outlier (Cluster 5)	812		0.91	47.04	0	94.58	79.06
Mean			0.82	19.76	0.37	95.89	68.19
Outlier Modeling of the full dataset	19200	10.70	0.80	11.16	0.15	72.34	33.01

As is clear from Table (2.2), clustering the 7741 outlier spectra into 5 refined clusters improves the spectral recovery significantly. As mentioned earlier, this is due to the increased similarity in each group, which improves the efficiency of linear models in representing the data points.

A comparison between the results obtained for OM (last row of the table) and those obtained for classic PCA indicates that the size of the compressed data for OM is more than PCA at 3D but less than 4D or 5D. However, OM reduces the reconstruction errors and improves the mean accuracy from 0.77 (in the case of 4D-PCA) to 0.80. The results also reveal that the compression of the given image using OM increases the percentage of samples with excellent and very good recovery

(NRMS>0.95 and GFC>0.999) in comparison to classic PCA at 3D or 4D. The reduction of the worst–case errors (NRMS<0) is also obvious when using OM.

The results on the estimation of the 10 multispectral images from the Hordley et al. database using classic PCA and OM are given in Tables (2.3) and (2.4), respectively. As the size of these multispectral images is almost the same, the compression ratio for all of these images is approximately 11.8, 9.5 and 8 for data compression using 3-, 4- and 5-dimensional classic PCA, respectively. This ratio is about 10.7 when compression is performed using the proposed OM method.

A comparison between the spectral accuracy of classic PCA and OM shows that OM performs as well as 4D-PCA in 7 spectral images, while in 2 cases 4D-classic PCA leads to less error. Nonetheless, with regards to the compression ratio and resulting accuracy, it seems reasonable to compress these two images using OM instead of 3D or 4D classic PCA. In one image (Kellogg's), OM performs better than even 5D-PCA. In order to be able to do a better comparison between classic PCA and our proposed approach, we took an average over all images and the results are shown in form of a bar graph in Fig (2.11).

Table 2.3 Accuracy of reflectance reconstruction for 10 multispectral images from the Hordley database using classic PCA with 3, 4 and 5 eigenvectors.

	Normalized RMS									GFC					
	# Eigenvectors									# Eigenvectors					
	3			4			5			3		4		5	
	Mean	% of samples		Mean	% of samples		Mean	% of samples		% of samples		% of samples		% of samples	
		>0.95	<0		>0.95	<0		>0.95	<0	>0.995	>0.999	>0.995	>0.999	>0.995	>0.999
Daz	0.21	2.08	29.50	0.45	4.10	17.72	0.62	7.10	6.56	72.66	15.33	78.02	28.43	82.93	48.11
Persilnonbio	0.65	0.49	4.65	0.71	1.73	2.30	0.77	3.27	0.98	87.33	18.63	92.07	34.85	94.61	56.23
Goaheadbars	0.61	4.20	8.61	0.69	6.71	4.83	0.75	10.66	2.72	74.20	25.07	78.82	34.86	84.31	47.42
Couscous	0.72	0.05	1.09	0.79	1.64	0.86	0.82	2.36	0.36	54.83	5.59	80.43	16.00	87.53	26.5
Elastoplast	0.72	0.02	2.33	0.78	2.77	2.23	0.81	5.33	0.77	64.16	7.65	85.79	24.20	92.03	39.10
Kellogg's	0.26	0.46	27.84	0.35	1.62	24.50	0.54	3.37	13.85	63.09	12.36	69.65	18.48	78.91	29.46
Freeform	0.74	3.18	5.26	0.80	6.39	2.84	0.85	12.09	0.89	90.91	39.24	93.85	54.21	95.59	69.09
Mulligatawny	0.74	0.00	2.46	0.78	0.09	1.21	0.81	0.35	0.42	82.51	17.59	86.91	26.57	90.54	37.59
Vanish	0.48	1.77	17.22	0.58	4.38	8.59	0.66	9.23	4.16	73.70	23.19	79.80	37.00	84.07	51.95
Fairy	0.49	0.11	11.29	0.58	0.40	6.95	0.65	1.40	3.07	77.63	18.87	84.57	29.57	87.87	46.53
MEAN	0.56	1.23	11.02	0.65	2.98	7.20	0.73	5.51	3.38	74.10	18.35	83.00	30.41	87.84	45.20

Table 2.4 Accuracy of reflectance reconstruction for 10 multispectral images from the Hordley database using OM.

	Image Size	Normalized RMS			GFC	
		Mean	% of samples		% of samples	
			>0.95	<0	>0.995	>0.999
Daz	295×241	0.52	2.25	12.06	76.73	29.07
Persilnonbio	276×279	0.70	1.57	2.40	91.75	27.12
Goaheadbars	331×286	0.71	5.41	2.41	76.35	35.67
Couscous	242×348	0.80	1.85	0.31	79.60	16.86
Elastoplast	162×218	0.78	3.57	1.16	82.68	25.12
Kellogg's	455×224	0.56	1.06	8.60	66.02	25.93
Freeform	239×285	0.75	4.13	5.37	91.63	47.40
Mulligatawny	197×276	0.80	0.39	0.82	84.58	36.49
Vanish	194×280	0.58	2.12	8.00	79.22	34.42
Fairy	294×239	0.55	0.45	7.41	80.28	28.13
MEAN		0.67	2.28	4.85	80.88	30.62

We tested OM on 7 multispectral images from the Columbia University dataset and report the results in terms of NRMS and GFC in Table (2.5). The corresponding results achieved by spectral estimation using classic PCA are given in Table (2.6). A comparison shows that the spectral accuracy achieved by OM in terms of average NRMS is the same as that obtained using classic PCA when 5 principal components are used. However, the resultant compression ratio of images compressed by OM is 10.77 versus 8 in the case of 5D-PCA. Besides, OM outperforms 5D-PCA in terms of the percentage of samples with excellent recovery as well as GFC. The percentage of samples with poor recovery is better than 3D-PCA and worse than 4D-PCA. A better comparison can be done using the bar chart illustrated in Fig (2.12).

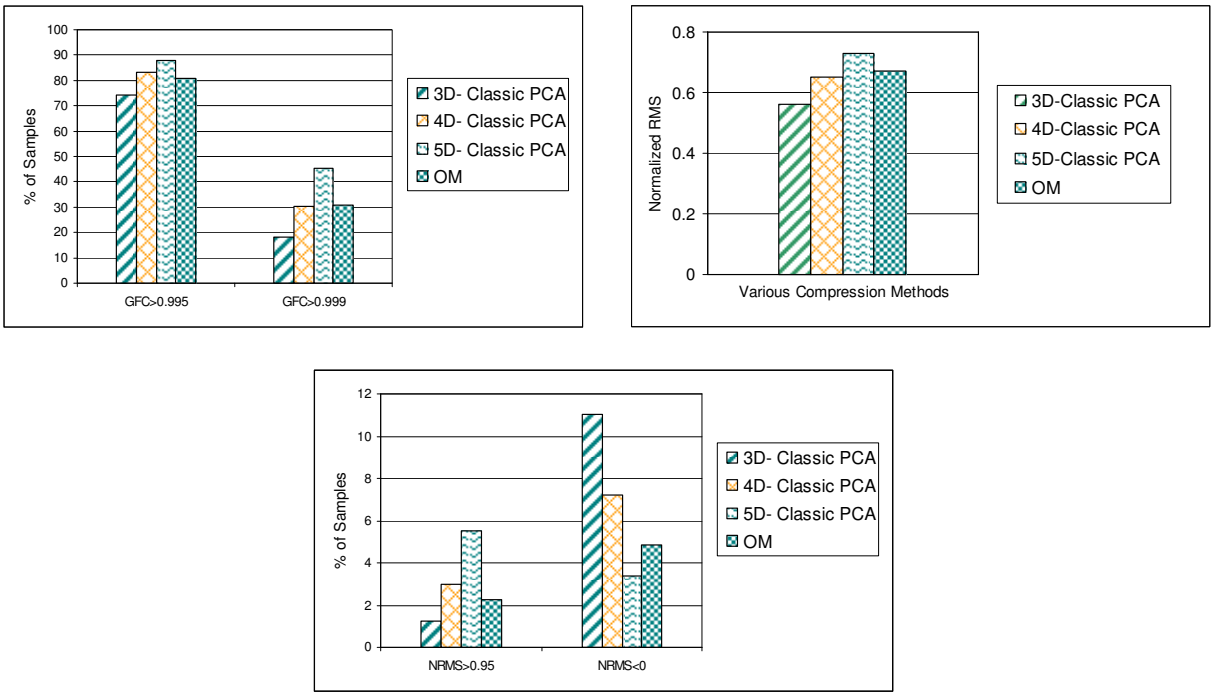


Fig 2.11 Comparison between spectral accuracy of reflectance estimation for 10 multispectral images from Hordley database using classic PCA and OM.

Table 2.5 Accuracy of reflectance reconstruction for the 7 multispectral images from the Columbia University database using OM.

	Normalized RMS			GFC	
	Mean	% of samples		% of samples	
		>0.95	<0	>0.995	>0.999
PomPom	0.86	21.42	0.32	86.03	48.50
Watercolors	0.77	3.36	1.19	81.75	67.77
Beeds	0.80	15.18	0.87	59.88	29.66
Beer	0.71	13.17	2.66	96.60	90.04
jelly_beans	0.82	5.69	0.50	72.66	33.49
stuffed_toys	0.82	9.53	1.86	87.84	41.49
paints	0.81	2.83	0.55	89.01	49.29
MEAN	0.80	10.17	1.13	81.96	51.46

Table 2.6 Accuracy of reflectance reconstruction for 7 multispectral images from the Columbia University database using classic PCA with 3, 4 and 5 eigenvectors.

	Normalized RMS									GFC					
	# Eigenvectors									# Eigenvectors					
	3			4			5			3		4		5	
	Mean	% of samples		Mean	% of samples		Mean	% of samples		% of samples		% of samples		% of samples	
		>0.95	<0		>0.95	<0		>0.95	<0	>0.995	>0.999	>0.995	>0.999	>0.995	>0.999
PomPom	0.71	0.31	0.55	0.79	0.93	0.44	0.85	15.67	0.40	49.82	8.36	71.67	18.13	86.02	44.28
Watercolors	0.71	1.46	1.87	0.77	3.55	0.57	0.81	4.62	0.52	77.21	61.47	82.28	68.01	84.29	71.74
Beeds	0.66	0.02	1.23	0.75	3.24	0.55	0.82	7.28	0.57	28.85	1.24	45.82	8.34	63.23	19.55
Beer	0.65	9.18	1.63	0.74	12.12	0.48	0.78	17.72	0.41	95.07	81.35	97.22	87.72	97.86	91.45
jelly_beans	0.65	0.37	6.19	0.80	2.36	0.64	0.85	6.67	0.51	41.72	11.47	67.22	22.72	81.47	35.41
stuffed_toys	0.56	0.41	2.44	0.65	1.63	0.96	0.73	5.02	0.79	26.71	8.99	36.80	17.16	48.50	24.67
paints	0.63	0.01	2.17	0.71	0.04	0.66	0.80	0.39	0.51	51.91	24.48	65.33	29.54	82.37	39.52
MEAN	0.65	1.68	2.29	0.74	3.41	0.61	0.80	8.19	0.53	53.04	28.19	66.62	35.94	77.67	46.66

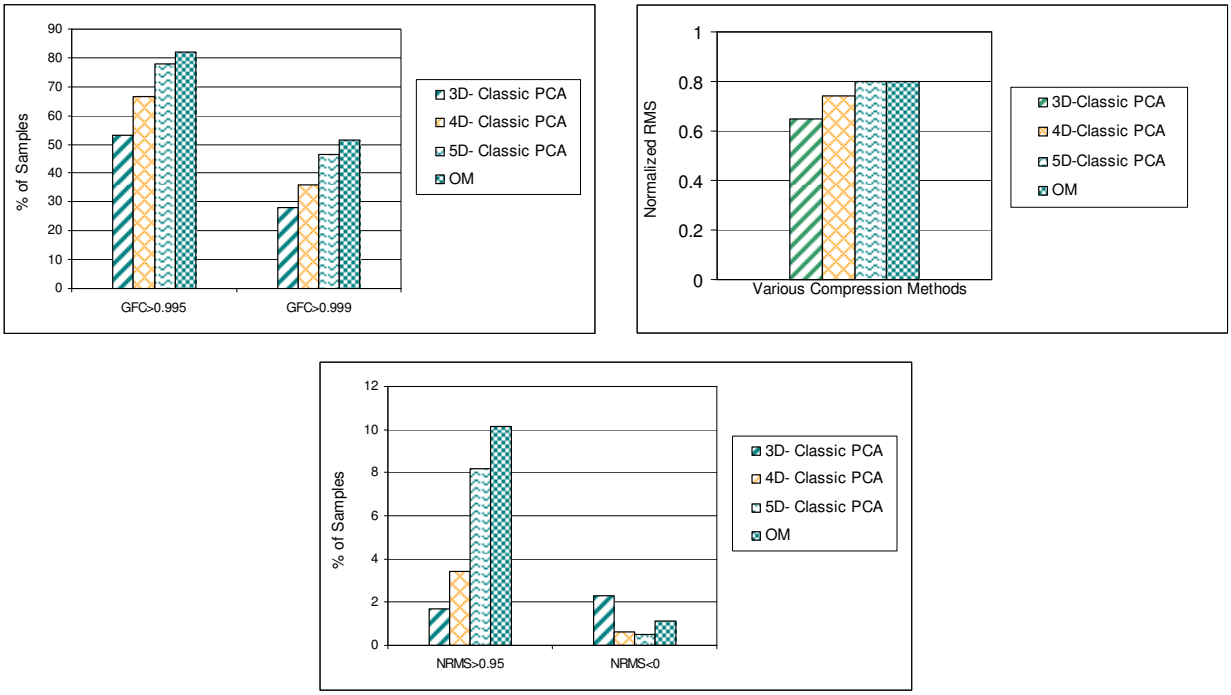


Fig 2.12 Comparison of the accuracy of the reconstruction of reflectance spectra from 7 multispectral images from the Columbia University database using classic PCA and OM.

And finally, we made a comparison between the running time of OM and 4-D classic PCA for 2 selected multispectral images, i.e. Fruits and Flowers from dataset III and Pompoms from datasets II. In fact, we chose these two images as representative of the smallest and the largest images in our datasets. The running time for different steps of OM and classic PCA implementation are given in Table 2.7.

Table 2.7 OM and Classic PCA running time (in seconds) for the smallest and largest images in the datasets.

		OM					4-D Classic PCA
	Image size	Fast MCD	K-means	Iterative refinement	PCA	Total	
Fruits and Flowers	120×160	14.24	0.16	1.06	0.29	15.75	0.72
PomPoms	512×512	20.65	1.14	26.52	6.47	54.78	9.94

Obviously, the running time of each method strongly depends on the size of the image. It should be noted that iterative refinement of clusters is an optional step and can be discarded at the expense of a small error but will speed up the OM procedure.

2.5 Summary of the Chapter

Datasets of reflectance spectra can be represented more compactly in terms of a low-dimensional basis using standard Principal Components Analysis. Most datasets, however, contain "outlier" spectra that differ markedly from the bulk of the dataset. These outliers can lead to poor reconstruction of some spectra when using a standard PCA-derived basis. The Outlier Modeling method proposed here improves upon standard PCA by separating out the outlier spectra and treating them separately. The outliers are grouped into several clusters and then a separate PCA basis is used to represent the inliers and each cluster of outliers. The outliers are identified using a robust Mahalanobis distance measure provided by the minimum covariance determinant algorithm. Tests show that outlier modeling leads to lower spectral reconstruction errors of reflectance spectra both in terms of normalized RMS and goodness of fit.

Chapter 3

Spectral Compression using Sub-Space Clustering

3.1 Introduction

Over the past few decades, the acquisition of high-dimensional data has become increasingly common in many application fields and consequently significant progress has been made to compress, store, and transmit such datasets. Many of these developments have been based on the fact that the intrinsic dimension of a high dimensional dataset is often much smaller than the dimension of the original space. Relying on this observation, multivariate statistical tools such as Principal Component Analysis (PCA) can be utilized to determine the dimension of the smallest sub-space that represents data without considerable loss of information. In classic PCA, it is assumed that that data is drawn from a single low-dimensional sub-space and therefore finding the number of dimensions of this sub-space is the only parameter to determine. Nonetheless, in practical situations the data could be drawn from various sub-spaces (Vidal, 2011). This means it seems reasonable to consider extending traditional data reduction techniques so as to discover clusters of data in several sub-spaces of the same dataset. The dimensions of these sub-spaces may be different. This strategy is called *Subspace Clustering* and has had various applications in computer vision and image processing (Yang et al., 2008; Ho et al., 2003; Wei et al., 2006).

In this chapter, a sub-space clustering strategy is used for the spectral compression of multispectral images. Unlike classic PCA, this approach finds clusters in several subspaces of different dimension.

3.2 Sub-Space Clustering for High Dimensional Data

In the case of multiple sub-spaces, at one extreme one can fit N data points using N different sub-spaces of dimension 1 (i.e., 1 sub-space per data point), or at the other extreme, using a single subspace of full dimension. Obviously, neither solution is satisfactory. The challenge is to find a small number of sub-spaces of low dimension that represents the data well. The main goal of sub-space clustering is to find the number of subspaces, their dimensions and the bases of each sub-space. When the number of sub-spaces is equal to 1, this problem reduces to classic Principal Component Analysis (Vidal, 2011). The mathematical ground of PCA data reduction technique was given in section (2-2-1).

A number of sub-space clustering algorithms have been proposed. For example, Generalized Principal Component Analysis, GPCA, is an algebraic approach to data segmentation and tries to find an analytic solution to sub-space clustering (Vidal et al., 2005). Unfortunately, the long computational time for large multispectral images prevented us to implement this method in this thesis. Vidal makes an extensive review on existing sub-space clustering algorithms and shows how they try to address different challenges with which an algorithm may be confronted (Vidal, 2011). Another comprehensive survey on various sub-space clustering algorithm has been presented in (Parsons et al., 2004).

In this research, we are motivated by the idea of sub-space clustering and propose a compression strategy relying on data reduction in multiple sub-spaces. In fact, instead of representing all spectra in a single low-dimensional sub-space of fixed dimension, spectral data are assigned to multiple sub-spaces with dimensions ranging from 1 to 8. This strategy, which is called *MS-PCA* (Multiple Subspaces – PCA) throughout this thesis, allows us to distribute spectra into different sub-spaces that best fit each cluster of

spectra. Unlike classic sub-space clustering, in this research the number of sub-spaces was considered as a constant parameter equal to 8. However, one can increase the number of sub-spaces to perform a more accurate representation of data but at the expense of increasing the amount of data stored and vice versa. The following describes the details of the proposed sub-space clustering algorithm.

Algorithm: Sub-Space Clustering

Input: *The number of sub-spaces k and a dataset*

Output: *A set of clusters that the NRMS of their members is below a given threshold*

1. *D=number of dimensions*
2. *Set D to 1.*
3. *Apply classic PCA on the initial dataset (original space) and fit a sub-space with given dimension D (D=1 in the first iteration, D=2 in the second and so on).*
4. *Assign spectra whose NRMS is less than a given threshold to this cluster and remove them from the initial dataset.*
5. *Increase the number of dimension, D, by one. Repeat steps 3 and 4 until the number of subspaces reaches k.*

In order to increase the efficiency of the proposed algorithm, a *pre-processing* step can also be performed on the initial spectral dataset. This process includes a preliminary clustering using the k-means algorithm to discover 4 clusters in the initial dataset (full-dimensional space) based on the *cosine distance measure*. This approach, denoted MS-CPCA (Multiple Subspaces – Clustered PCA), segments the initial dataset into 4 datasets whose elements are similar to each other to some extent. It is well documented that the number of dimensions required to model the majority of the variation in a given spectral dataset could be decreased if the similarity between the reflectance curves in the dataset could be increased. The idea of partitioning data space into disjoint regions and then performing PCA about each cluster was presented by

Kambhatla and Leen (Kambhatla & Leen, 1997). Sattler et al. used this knowledge to present a new geometry compression method for animations. They showed that using Clustered PCA individual clusters can be compressed more efficiently with fewer principal components compared to classic PCA (Sattler et al., 2005). This approach was later modified to alleviate the possible errors stemming from the initial random selection of the cluster centers as K-mean algorithm does (Das et al., 2010). Garcia et al. used a clustering method based on computing the hue of samples and then using hue as a measure of spectral similarity in order to improve the performance of the recovery process (Garcia-Beltran et al. 1998). As well, Ayala et al. presented a selective database method based on the similarity of samples defined in terms of the color specifications of the proposed sample (Ayala et al., 2006). Ciprian and Carbucicchio proposed a "colorimetric-spectral clustering" strategy for compression of multispectral images. In this method, the spectra are grouped based on their color difference before starting the compression procedure that is based on PCA (Ciprian & Carbucicchio, 2011).

So, it has been shown that applying clustering as a pre-processing step could enhance the efficiency of PCA. Consequently it seemed to us that this could be also helpful in boosting the performance of our proposed compression approach.

3.3 Testing the Method on Multispectral Images

To evaluate the performance of the proposed MS-PCA and MS-CPCA methods for spectral compression of multispectral images, we implemented our algorithm on 18 images taken from three different datasets and assessed the spectral accuracy of recovered spectra in terms of normalized RMS (NRMS) as well as GFC at a given compression ratio. The multispectral images used in this part of this research came from the following spectral databases:

I) 10 images from the database of Hordley et al.

II) 7 images from the Columbia University multispectral image database.

III) "Fruits and Flowers" image from the Eastern Finland University spectral image database.

In all images the spectral reflectance data was from 400nm to 700nm in 10nm steps. Descriptions of the databases were given in chapter 2. Based on our previous knowledge, we assumed that even the worst-case spectral reflectances can be represented well using 8 basis vectors. Hence, the number of sub-spaces was fixed at 8 in all experiments.

3.3.1 Spectral Recovery Results

The average spectral recovery errors for the multispectral images from each dataset for four given compression ratio are given in Tables (3.1) to (3.4) in terms of NRMS and GFC. In these tables the performance of Classic PCA, MS-PCA and MS-CPCA for spectral compression of multispectral images are compared. It should be noted that exact recovery is achieved when NRMS and GFC equal one.

Based on a preliminary experiment, we found that when the dimension of spectral reflectances is reduced from 31 to 3, 4, 5 and 6 using the classic PCA, the resultant compression ratio for all images will be *approximately* 11.8, 9.5, 8 and 7, respectively. Although the size of images from different datasets is different, this difference is not that large enough to cause significant changes in the compression ratios, so these ratios remain almost the same. In order to be able to make a comparison between the accuracy of different spectral compression methods, the value of the threshold in the sub-space clustering-based approaches was set such that the storage requirements of the output image equal those of the corresponding image compressed by classic PCA.

The Evaluation of NRMS

As the tables show, both MS-PCA and MS-CPCA outperform classic PCA in spectral compression for all three datasets in terms of mean normalized RMS. This superiority becomes more visible when the compression ratio increases, i.e. for classic PCA using a small number of components (e.g. 3). The reason can be explained well by considering the principal of the sub-space clustering approach. As mentioned in section (3.2), our strategy is based on distributing samples in different sub-spaces with different dimensions varying from 1 to 8. In this way, if a given spectrum is not well represented using the first few principal components, still it has a chance to go to another sub-space

with more dimensions. This continues until the spectrum falls into an 8D-subspace, where in most cases the spectrum will be represented accurately. Therefore, unlike classic PCA, which represents all spectra using a fixed number of principal components (e.g. just 3), our algorithm discovers bad samples (spectra with poor recovery) and pushes them towards a higher-dimensional representation. When the number of dimensions is low, it is reasonable to expect that classic PCA performs much worse than MS-PCA or MS-CPCA due to modeling all samples just using a few components, while MS approach is flexible and uses a range of dimensions to model different types of spectra. However, with an increasing number of dimensions the performance of classic PCA improves gradually such that the difference between these three strategies becomes smaller. Fig (3.1) compares the performance of these three spectral compression approaches.

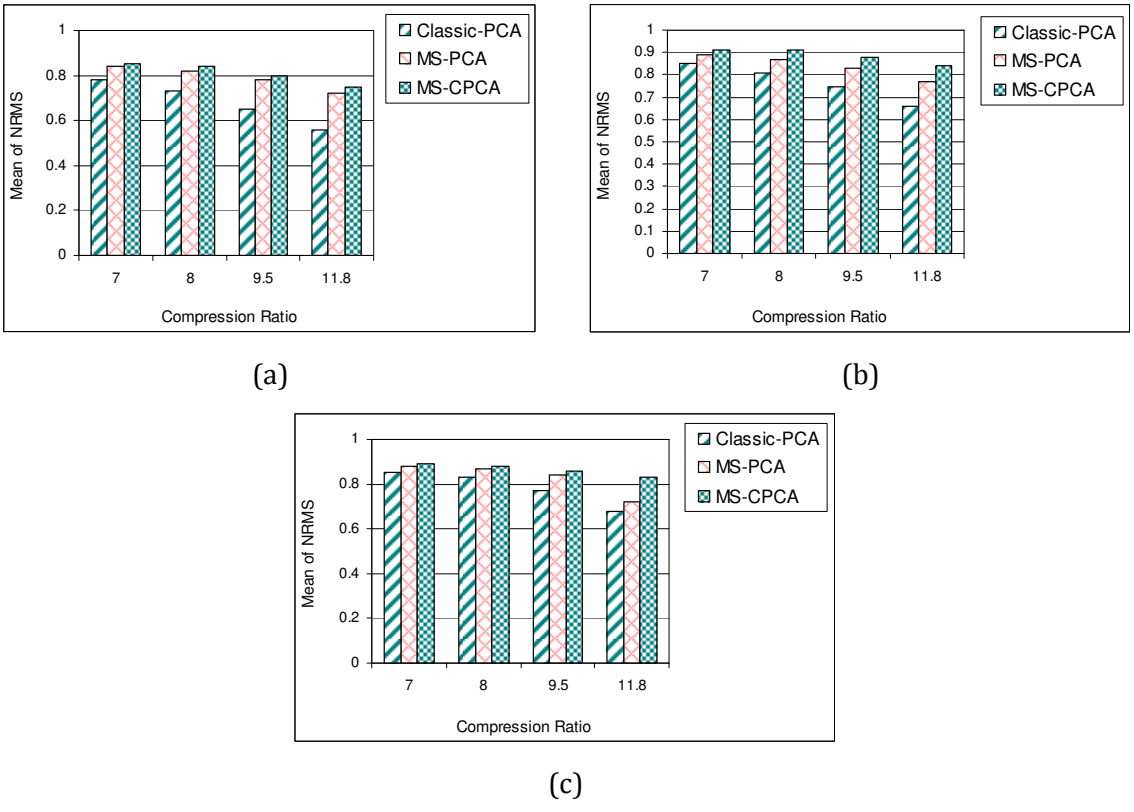


Fig 3.1 The bar chart compares the spectral accuracy of the different spectral compression methods at four given compression ratios for the three datasets I (a), II (b) and III (c).

Evaluation of the results obtained for classic PCA, MS-PCA and MS-CPCA also reveals that our proposed approaches cause a dramatic reduction in the percent of samples showing very poor recovery (NRMS<0). This reduction is an important achievement of this method, and the reduction is even more evident when a high compression ratio is used, i.e. when the classic PCA is restricted to just a small number of components (e.g. 3 components). To better compare them, refer to Fig. (3-2). This observation is exactly what would expect from the MS-PCA and MS-CPCA approaches. As mentioned earlier, in our proposed approach more resources are allocated to those spectra that need more dimensions for accurate representation. This causes a sharp decrease in the percentage of samples poorly recovered and this is reflected in the lower maximum reconstruction error.

Table 3.1 The average spectral recovery error for multispectral images taken from three different datasets, compressed by Classic PCA and multiple-subspace PCA with and without preprocessing (MS-CPCA and MS-PCA, respectively). The compression ratio for all three methods was set to 7. Classic PCA was performed using a 6-dimensional basis.

	Normalized RMS			GFC	
	Mean	% of samples		% of samples	
		>0.95	<0	>0.995	>0.999
Dataset I					
Classic PCA	0.78	10.29	1.40	92.31	60.98
MS-PCA	0.84	2.25	0.10	97.73	61.76
MS-CPCA	0.85	3.67	0.05	98.54	66.51
Dataset II					
Classic PCA	0.85	16.63	0.31	87.22	62.22
MS-PCA	0.89	8.80	0.28	97.31	70.09
MS-CPCA	0.91	24.95	0.28	99.14	88.62
Dataset III					
Classic PCA	0.85	13.18	0.015	89.04	42.17
MS-PCA	0.88	3.48	0	99.47	42.67
MS-CPCA	0.89	11.14	0	99.70	52.83

Table 3.2 The average spectral recovery error for multispectral images taken from three different datasets, compressed by Classic PCA and multiple-subspace PCA with and without preprocessing (MS-CPCA and MS-PCA, respectively). The compression ratio for all three methods was set to 8. Classic PCA was performed using a 5-dimensional basis.

	Normalized RMS			GFC	
	Mean	% of samples		% of samples	
		>0.95	<0	>0.995	>0.999
Dataset I					
Classic PCA	0.73	5.52	3.08	87.81	45.19
MS-PCA	0.82	0.96	0.08	91.82	48.80
MS-CPCA	0.84	1.85	0.05	96.39	52.96
Dataset II					
Classic PCA	0.81	8.69	0.31	79.94	50.47
MS-PCA	0.87	3.34	0.28	95.34	55.86
MS-CPCA	0.91	13.64	0.28	99.12	74.52
Dataset III					
Classic PCA	0.83	11.09	0.02	82.83	37.64
MS-PCA	0.87	3.22	0	95.16	34.59
MS-CPCA	0.88	7.22	0	99.38	40.22

Table 3.3 The average spectral recovery error for multispectral images taken from three different datasets, compressed by Classic PCA and multiple-subspace PCA with and without preprocessing (MS-CPCA and MS-PCA, respectively). The compression ratio for all three methods was set to 9.5. Classic PCA was performed using a 4-dimensional basis.

	Normalized RMS			GFC	
	Mean	% of samples		% of samples	
		>0.95	<0	>0.995	>0.999
Dataset I					
Classic PCA	0.65	2.98	7.30	83.00	30.41
MS-PCA	0.78	0.68	0.11	86.53	35.94
MS-CPCA	0.80	1.43	0.04	91.00	41.65
Dataset II					
Classic PCA	0.75	3.70	0.38	69.14	39.55
MS-PCA	0.83	1.26	0.28	81.73	44.02
MS-CPCA	0.88	7.19	0.27	96.81	62.13
Dataset III					
Classic PCA	0.77	7.00	0.40	68.47	31.24
MS-PCA	0.84	2.52	0	86.42	27.72
MS-CPCA	0.86	5.78	0	93.26	33.20

Table 3.4 The average spectral recovery error for multispectral images taken from three different datasets, compressed by Classic PCA and multiple-subspace PCA with and without preprocessing (MS-CPCA and MS-PCA, respectively). The compression ratio for all three methods was set to 11.8. Classic PCA was performed using a 3-dimensional basis.

	Normalized RMS			GFC	
	Mean	% of samples		% of samples	
		>0.95	<0	>0.995	>0.999
Dataset I					
Classic PCA	0.56	2.23	11.02	74.11	18.36
MS-PCA	0.72	0.58	0.09	72.19	25.75
MS-CPCA	0.75	0.92	0.02	79.09	31.42
Dataset II					
Classic PCA	0.66	1.91	2.01	56.23	31.88
MS-PCA	0.77	1.15	0.27	61.70	36.80
MS-CPCA	0.84	4.70	0.25	83.28	49.40
Dataset III					
Classic PCA	0.68	2.48	0.71	45.70	22.13
MS-PCA	0.72	0.49	0	38.45	16.87
MS-CPCA	0.83	4.81	0	79.25	28.19

On the other hand, the evaluation of results shows that the percent of reconstructed samples with NRMS>0.95, which indicates an *excellent fit*, is larger in the case of compression using classic PCA, particularly when the number of principal components is large (e.g. 6 components). This is due to the fact that in the MS approach some spectra, which can be modeled well with a very small number of principal components, will stay in very low-dimensional sub-spaces, e.g. one and two. Obviously, compression of these samples using, for example, 6-dimensional classic PCA will lead to much better recovery results.

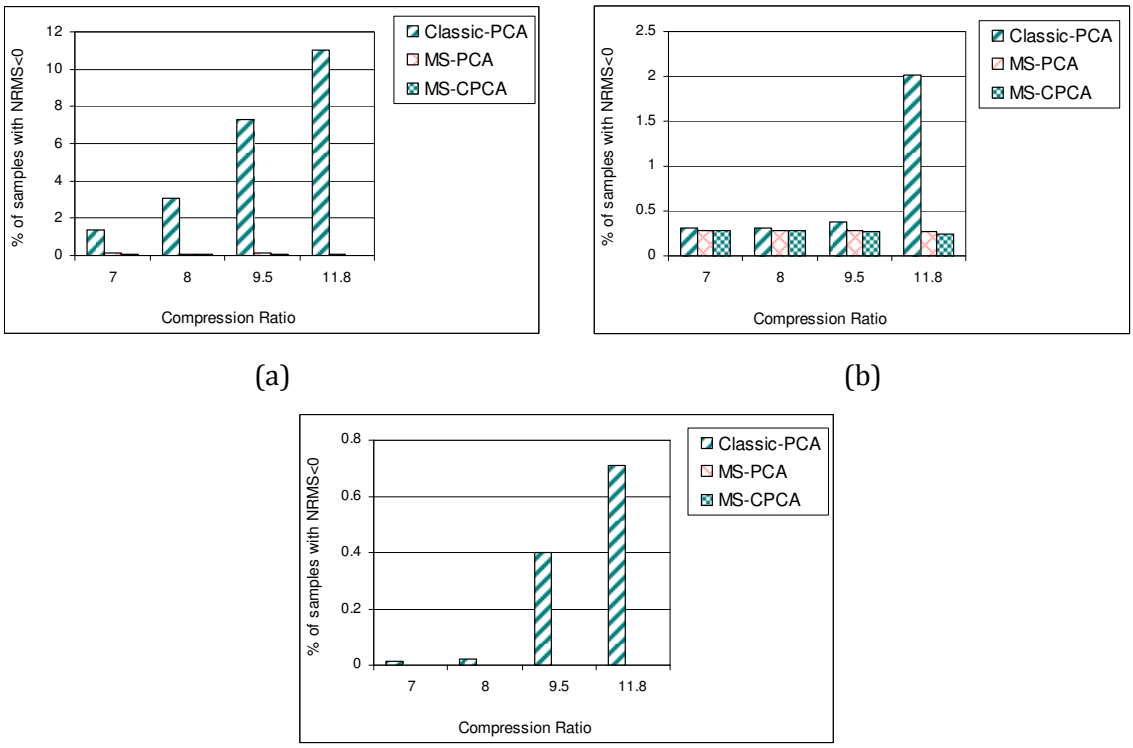


Fig 3.2 The bar chart compares the spectral accuracy of different spectral compression methods at four given compression ratios for three datasets I (a), II (b) and III (c).

The Evaluation of GFC

A comparison between the average GFC calculated over all multispectral images in each dataset shows that for all the compression ratios tested the percent of spectra that have been compressed with MS-CPCA and get recovered with good and very good accuracy (GFC>0.995 and GFC>0.999) is always higher than the percent of samples compressed with classic PCA at the corresponding compression ratio. However, in case of compression via MS-PCA the recovery accuracy measured by GFC is worse than corresponding classic PCA for a few cases, i.e., dataset I and III at CR=11.8. This could be due to the fact that in order to get such a high compression ratio, we had to assign a large number of samples to 1- or 2-dimensional sub-spaces, thereby representing them with only 1 or 2 principal components (Fig 3.7). Obviously, the accuracy of data representation using just 1 or 2 components is not as good as 3.

Comparison Between MS-PCA and MS-CPCA

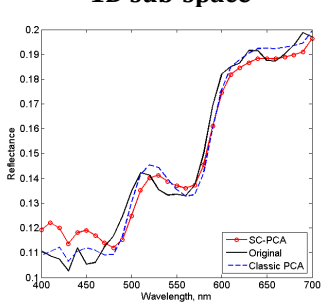
The results also reveal that the MS-CPCA approach outperforms MS-PCA in all datasets and at all given compression ratios. This superiority is very clear in the case of dataset 3 when it is compressed at CR=11.8. This significant improvement implies that applying pre-processing in the form of data clustering based on spectral similarity converts the initial dataset into four sub-datasets whose members benefit from a higher degree of similarity. Therefore, as one would have expected, these clusters are limited to sub-spaces whose dimensions are much smaller than the dimension of the initial dataset. This allows a good reconstruction of spectra even with one or two principal components.

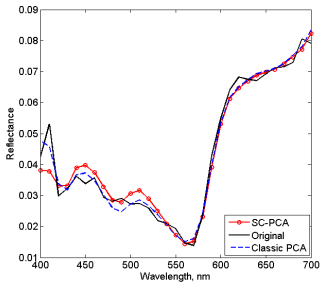
Fig (3.3) shows the results of spectral recovery of 24 randomly selected pixels from the *Kellogg's* multispectral image from dataset I. These spectra were compressed and reconstructed using classic PCA and MS-PCA approaches. Classic PCA was done using 5 basis vectors while MS-PCA used a range of sub-spaces from 1 to 8. In this figure, the first row represents the samples that fell in the 1D sub-space and were compressed using one principal component, the second row represents samples that fell in 2D sub-

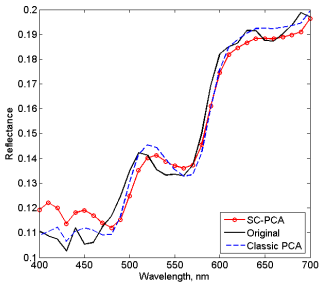
space and were compressed using two principal components and so on. It is clear that in the first few sub-spaces the degree of fit is higher when the spectrum is reconstructed using 5D, classic PCA. Nonetheless, the spectral accuracy of the spectra represented just using 1 or 2 principal components is reasonable as well.

On the other hand, MS-PCA benefits from sub-spaces with higher dimensions to represent samples that are not recovered well using 5 or fewer bases. As can be seen, in rows 5 to 8 of Fig (3.3), MS-PCA has provided a much better fit in comparison to classic PCA. Interestingly, even in the case of compression using 5 principal components, MS-PCA outperforms classic PCA particularly in the recovery of very first part of the spectrum (exact matching was achieved). In the other words, by using the same number of principal components, i.e. 5, MS-PCA provides a better approximation to the original reflectance than classic PCA. The reason could be due to the fact that classic PCA represents data in a sub-space whose bases are extracted from the complete initial data (i.e., the whole of multispectral image), which could be broad and varied. Although, these general bases can represent a wide range of spectra with an acceptable goodness of fit, they may not be able to provide a perfect match to any spectrum. On the other hand, the MS-PCA algorithm is based on removing samples from the initial dataset that it has already assigned to a sub-space. The strategy of separating out samples with different patterns makes the initial dataset gradually more specific. So, representing samples in a 5D sub-space whose bases extracted from such refined dataset appears to be more efficient.

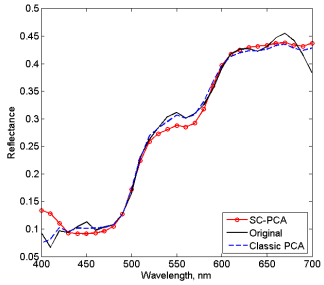
1D sub-space

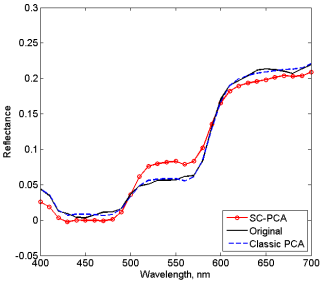


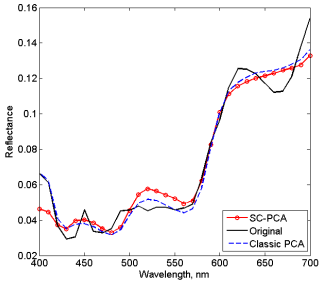




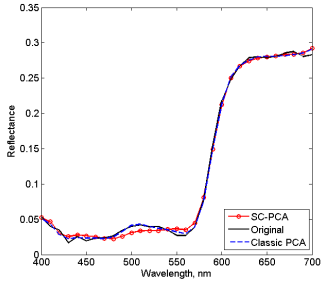
2D sub-space

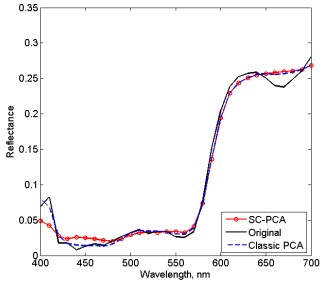


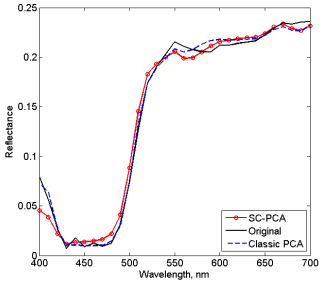




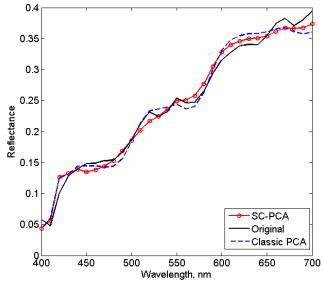
3D sub-space

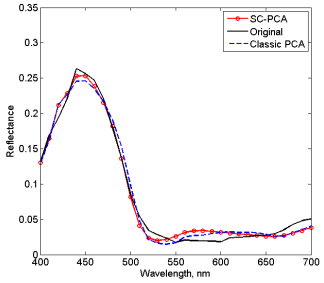


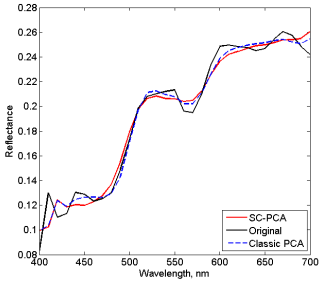




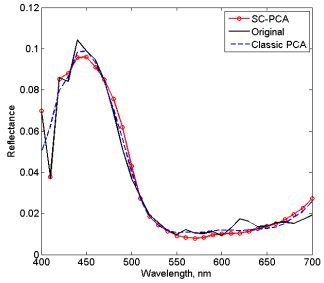
4D sub-space

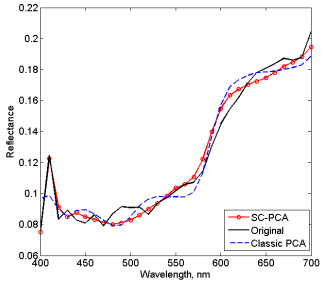


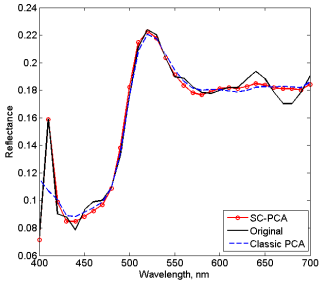




5D sub-space







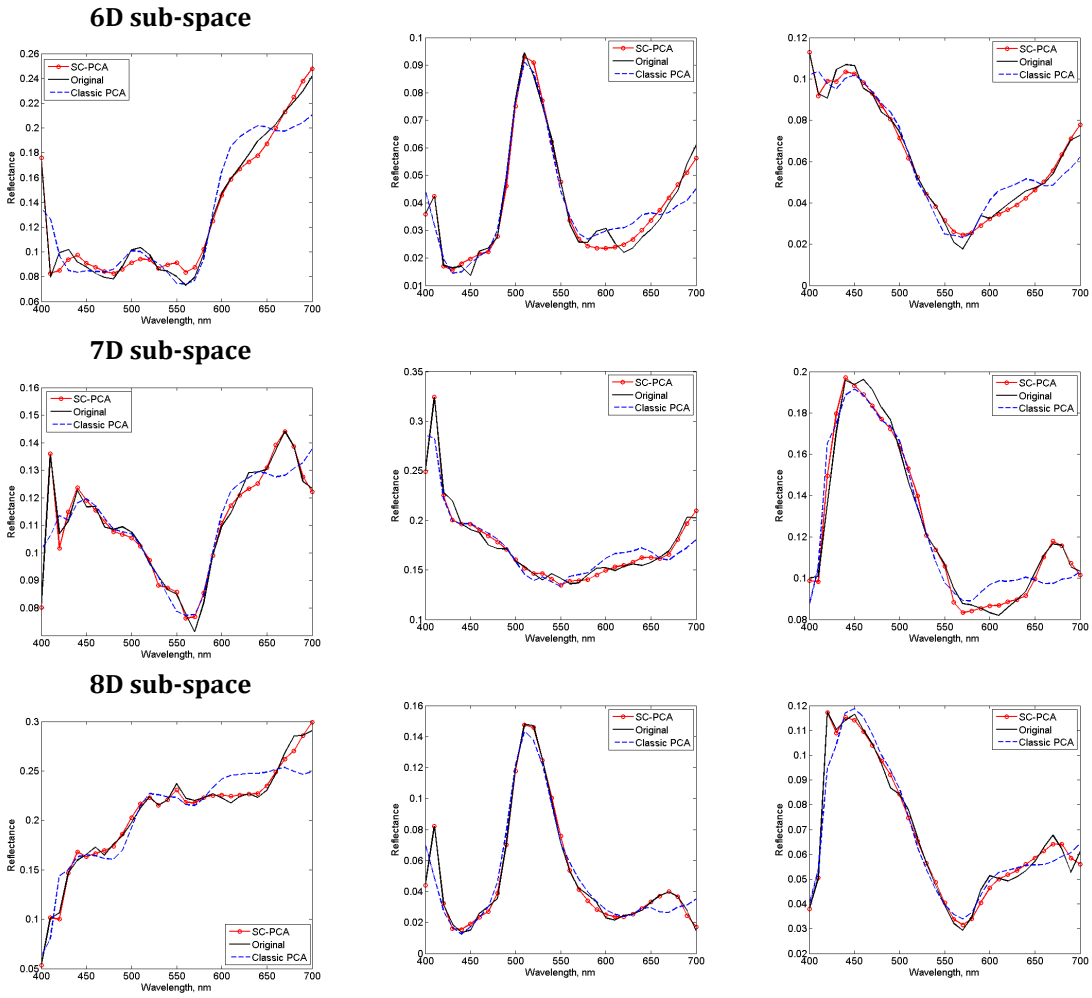


Fig 3.3 The results of spectral recovery of 24 randomly selected pixels from the "Kellogg's" multispectral image from dataset I using 5D classic PCA, and MS-PCA. Refer to the text for more information.

Fig (3.4) illustrates the frequency with which MS-PCA distributes samples into the subspaces of different dimension. For instance, Fig (3.4-a) shows the allocation of spectra to various sub-spaces so that the resultant compression ratio becomes 11.8. As mentioned earlier, this ratio is achievable via classic PCA when each spectrum is represented using 3 principal components. As can be observed, to achieve such a compression ration using MS-PCA just 35% of spectra are allocated to a 3D sub-space with around 37% assigned to 1- and 2-dimensional sub-spaces and approximately 28% to sub-spaces of dimension more than 3.

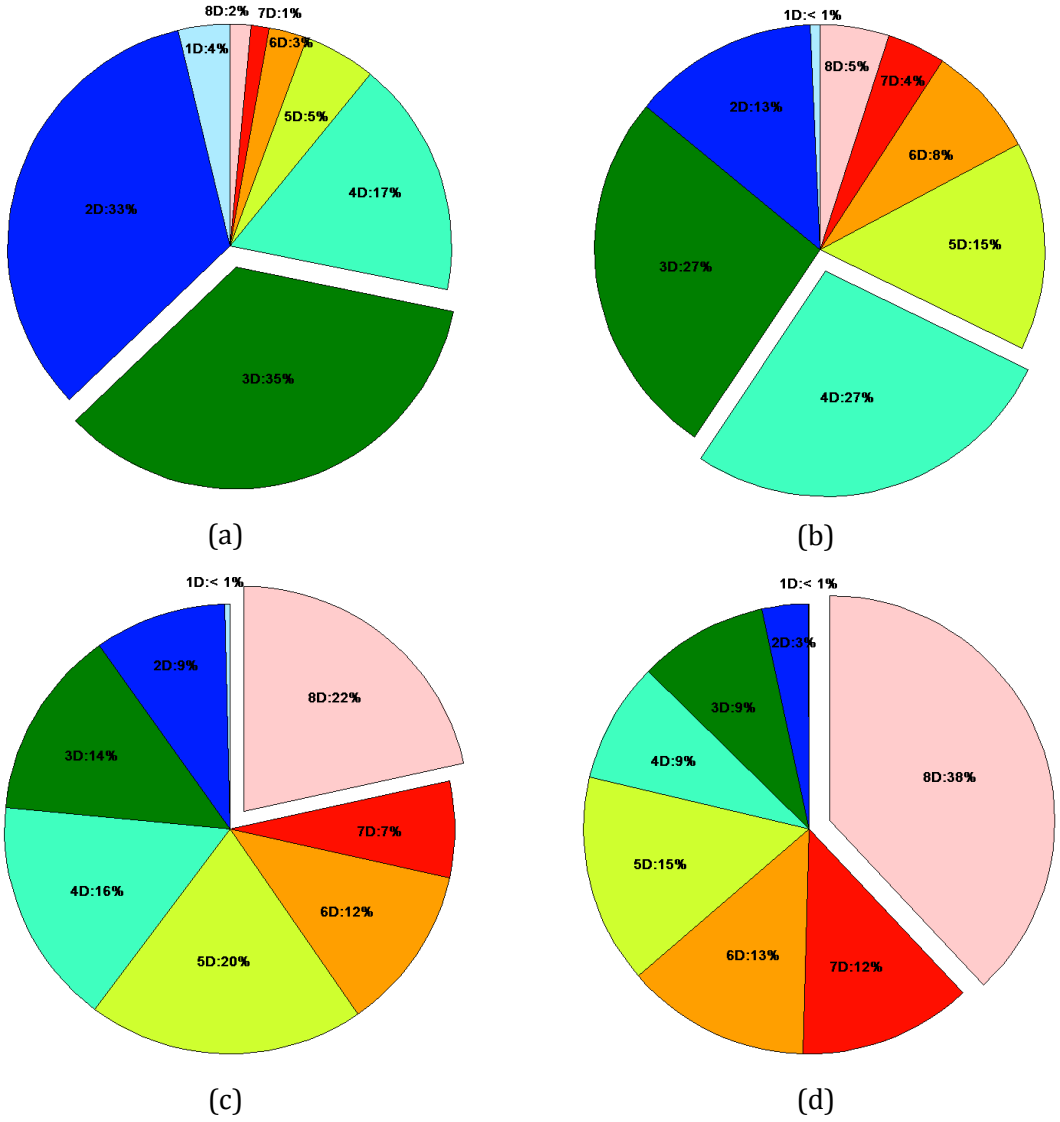


Fig 3.4 Four pie charts comparing the percent of spectra assigned to each sub-space at a given compression ratio: CR=11.8 (a), CR=9.5 (b), CR=8 (c) and CR=7 (d) using the MS-PCA approach.

Overall, it can be seen that the higher compression ratio, the greater the portion of the spectra that are assigned to the first few sub-spaces (i.e., 1D, 2D and 3D sub-spaces). For example, in the case of CR=11.8, around 72% of spectra were allocated to the first three sub-spaces whereas these sub-spaces only made up less than 13% when CR=7. In order to control this assignment, we needed to manipulate the value of thresholds such that to achieve a high compression ratio more spectra go to the first few sub-spaces and

vice versa. However, the threshold should be big enough to guarantee satisfactory reconstruction even when the spectrum is modeled using just one component.

The pie charts in Fig (3.5) show how spectra are distributed in different sub-spaces when the compression is carried out using MS-CPCA. The most obvious difference between the MS-PCA and MS-CPCA charts is the larger number of spectra assigned to the 1D sub-space using MS-CPCA. This is exactly the idea behind this scheme. In the other words, performing clustering before applying MS-PCA increases the similarity between samples within a cluster, which allows each spectrum to be represented more efficiently using fewer principal components.

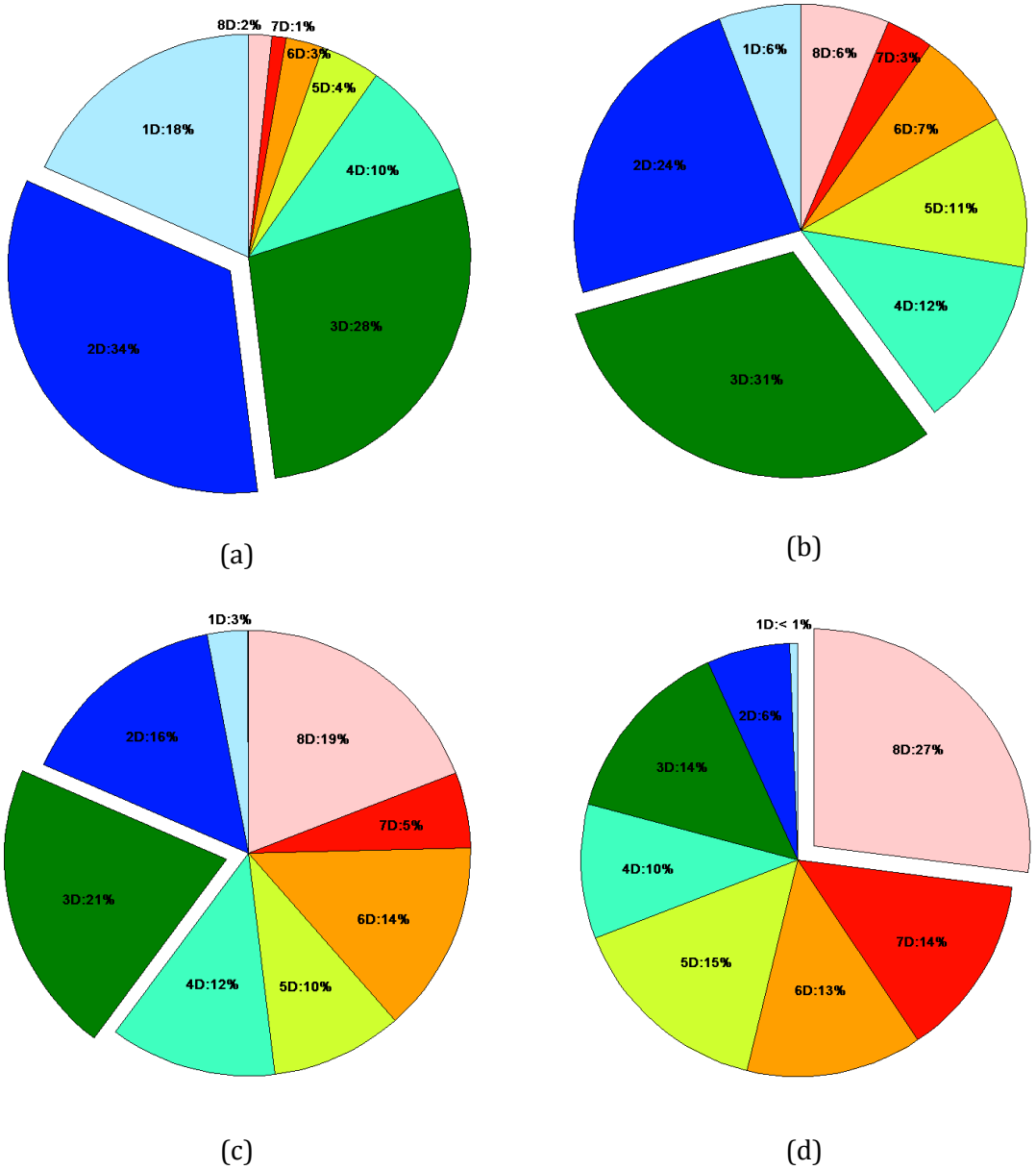


Fig 3.5 The four pie charts compare the percent of spectra assigned to each sub-space at a given compression ratio: CR=11.8 (a), CR=9.5 (b), CR=8 (c) and CR=7 (d) using MS-CPCA approach.

The same as chapter 2, in this chapter we calculated the running time of our proposed method as well as classic PCA. As mentioned in chapter 2, iterative refinement of clusters is an optional step and may be dropped to lessen computational time.

Table 3.5 MS-PCA, MS-CPCA and 5-D Classic PCA running time (in second) for the smallest and largest images in the datasets.

	Image size	MS-CPCA				MS-PCA	5-D Classic PCA
		K-means	Iterative refinement	MS-PCA	Total		
Fruits and Flowers	120×160	0.33	2.60	4.92	7.85	3.45	0.77
PomPoms	512×512	4.42	28.76	45.8	78.98	44.61	9.2

3.4 Spatial Compression using JPEG2000

In this research the main focus has been on representing datasets of spectra, and multispectral images are a good source of large numbers of spectra for testing. In addition to spectral correlation, multispectral images typically include a high degree of spatial correlation. Clearly, spectral modeling using the sub-space clustering approach can be combined with image compression techniques for further data compression.

In this part of the research, we performed *lossless* JPEG2000 compression of the 5-dimensional representation of the 18 multispectral images taken from three different datasets. To do so, rather than a classic RGB image, a 5-dimensional eigen-image that is a 5-slice image with each slice representing the principal components of the corresponding pixel at that dimension (i.e., the first slice contains the first components, the second slice contains the second components and so on) was supplied to the compression algorithm from which 5 compressed slices were generated. As a result of removing the spatial redundancy the average compression ratio for the multispectral images in datasets I and III increased from approximately 8 to 14.5, 18.5. This ratio changed differently for different images in dataset II in a range from 23.32 to 46.34.

In addition, for the 8-dimensional eigen-image of each multispectral image, provided by MS-PCA was followed by JPEG2000 for further *lossy* compression of the spatial information. This eigen-image is padded with zeroes where there are no principal components as a result of the fact that MS-PCA represents samples using a varied number of dimensions.

In order to be able to compare the accuracy of JPEG2000 combined with Classic PCA versus MS-PCA, we tried to make the compression ratio of the corresponding datasets the same. Our strategy was to compress different slices with different JPEG compression ratios. To achieve the best accuracy, we assigned a smaller compression ratio to the first few eigen-images and a higher ratio to the last. In fact, with this type of compression we tried to keep more information in the first few slices corresponding to the first few principal components.

Table (3.6) represents the spectral accuracy of 5D-classic PCA followed by lossless JPEG2000 versus MS-PCA combined with lossy JPEG2000 compression. Note that MS-PCA was followed by a lossy strategy to provide the same compression ratio as classic PCA combined with lossless compression. In addition, the original multispectral images, not reduced spectrally, are compressed using lossy JPEG2000 for comparison. The third row of table for each dataset shows these results. Nonetheless, MS-PCA still outperforms Classic-PCA particularly in terms of the mean normalized RMS as well as the percentage of samples with NRMS<0. Image compression just using JPEG2000 led to very poor recovery. It seems that a combination of techniques for subsequent reduction of spectral and spatial redundancy is the best method to improve the accuracy for a fixed compression ratio.

Table 3.6 The spectral accuracy of 5D-Classic PCA and MS-PCA combined with lossless and lossy JPEG2000 compression, respectively. Each multispectral image was also compressed using JPEG2000 by itself and the results are given in the third row for each dataset. The compression ratio of datasets I and III are 14.5 and 18.5, respectively. For dataset II, the ratio changes from 23.32 to 64.34.

	Normalized RMS			GFC	
	Mean	% of samples		% of samples	
		>0.95	<0	>0.995	>0.999
Dataset I					
PCA+Jpeg	0.72	4.55	3.48	86.94	43.40
MS-PCA+Jpeg	0.79	0.42	0.65	87.08	44.73
Jpeg	0.54	0.72	11.74	71.49	17.32
Dataset II					
PCA+Jpeg	0.78	7.16	0.43	73.33	47.84
MS-PCA+Jpeg	0.83	0.78	0.32	80.25	50.28
Jpeg	0.68	2.61	9.88	73.01	49.67
Dataset III					
PCA+Jpeg	0.81	9.34	0.03	75.94	35.16
MS-PCA+Jpeg	0.84	2.77	0.01	83.63	30.55
Jpeg	0.63	3.42	3.14	42.24	21.97

3.5 Summary of the Chapter

A new spectral data reduction strategy based on sub-space clustering was presented and examined on 18 multispectral images from three different databases. In contrast to classic PCA, which represents data in a single low-dimensional sub-space of a fixed dimension, the principal of the proposed approach relies on finding multiple sub-spaces of the initial space that differ in the number of dimensions and principal bases. This strategy allows us to distribute spectra into different sub-spaces that best fit each cluster of spectra. As a result, those spectra that can be represented well using a few principal components will stay in very low-dimensional sub-spaces; however, those spectra that need more dimensions for accurate reconstruction will pass to sub-spaces of higher dimension. In total, at a constant compression ratio, one can compress multispectral images using the proposed method to decrease the maximum reconstruction error, and improve the average error at the slight expense of decreasing the percentage of samples recovered extremely accurately.

Chapter 4

Conclusion and Future Work

This chapter includes some brief concluding remarks and discusses topics for future research.

4.1. Thesis Summary

Large multispectral datasets such as those created by multispectral images require a lot of data storage. Compression of such datasets is therefore an important problem. A common approach is to use Principal Components Analysis (PCA) as a way of reducing the data requirements as part of a lossy compression strategy. In this thesis, we first addressed the problem that outlier spectra can cause in spectral compression when using PCA. For this purpose, we employed the fast MCD (Minimum Covariance Determinant) algorithm, as a highly robust estimator of multivariate mean and covariance, in order to detect outlier spectra in a multispectral image. We then showed that by removing the outliers from the main dataset, the performance of PCA in spectral compression significantly increases. However, since outlier spectra are a part of the image, they cannot simply be ignored. Our strategy was to cluster the outliers into a small number of groups and then compress each group separately using its own cluster-specific PCA-derived bases. Overall, we showed that significantly better compression can be achieved with this approach.

In the second part of the research, we developed another spectral compression strategy to improve the conventional spectral reduction techniques. In this new strategy, whose principal is based on well-known sub-space clustering algorithms, the spectra are distributed into different sub-spaces that best fit each cluster of spectra. This strategy takes advantage of the fact that in a multispectral image the probability of quite similar spectra occurring is high. These similar spectra can be placed into sub-spaces with very low dimensions (e.g. 1), modeled just using a small number of principal components, which therefore reduce the required resources remarkably. On the other hand, those spectra that need more dimensions for accurate reconstruction will transfer to sub-spaces of higher dimension. Our error analysis has revealed that, at a constant compression ratio, the proposed approach outperforms the conventional spectral compression methods in terms of maximum and average reconstruction error. We also showed that combining this strategy with an initial clustering can improve the recovery accuracy more than before.

A comparison between the two proposed methods reveals that at a constant compression method, spectral compression using sub-space clustering based strategies performs better than OM.

JPEG2000 as a well-known compression technique was used to compress multispectral images by itself and also in combination with the classic and MS-PCA. The latter, led to the best results in all cases.

4.2. Future Work

The current research can be extended in different aspects:

- Trying other approaches for outlier detection

  To detect multivariate outliers, in this research we employed MCD algorithm to estimate multivariate mean and covariance robustly. To the best of our knowledge and based on pre-experiments performed on our data, we arrived at the conclusion that this algorithm is more successful than other existing outlier detection approaches. However, as outlier detection in multivariate datasets is an open research topic, it seems worth experimenting with other methods.

- Enhancing the existing clustering method, i.e. k-means

  As the dimension of multispectral data can be large in some applications, the conventional k-means could perform less efficiently. It is advisable to replace it with some modern multivariate clustering technique such as spectral clustering (von-Luxburg, 2007). In addition, since the measured spectral data suffers from noise, it could be wise to try clustering using the other algorithms that are more appropriate for this type of data.
- Trying other algorithms for sub-space clustering

  Throughout this research we assumed that the number of subspaces is fixed. However, there are a number of sub-space clustering algorithms that try to find the number of sub-spaces and their dimensions. Further improvement in the current proposed approach could be achieved if the number of sub-spaces were to be determined for each image separately.

References

Agahian, F., Amirshahi S. A., and Amirshahi, S. H. (2008). Reconstruction of Reflectance Spectra using Weighted Principal Component Analysis, *Col. Res. & Appl. J.,* **33**, 360.

Agahian, F., Funt B., and Amirshahi, S. H. (2012). Representing Outliers for Improved Multispectral Data Reduction. *In Proc. of CGIV*, 367.

Agahian, F., Funt B., and Amirshahi, S. H. (2014). Spectral Compression: Weighted Principal Component Analysis versus Weighted Least Squares. *In Proc. of Human Vision and Electronic Imaging Conference, SPIE* (in press).

Ayala F., Echavarri J. F., and Renet, P. (2006). Use of Three Tristimulus Values from Surface Reflectance Spectra to Calculate the Principal Components to Reconstruct These Spectra by using Only Three Eigenvector, *J. Opt. Soc. Am. A.,* **23**, 2020.

Berns, R. S. (2001). The Science of Digitizing Paintings for Color-accurate Image Archives: A Review. *J. Imaging Sci. and Technol*, **4**, 305.

Burns, P. D. and Berns, R. S. (1996). Analysis of Multispectral Image Capture. *In Proc. of Fourth Color Imaging Conference: Color Science, Systems and Applications, IS&T/SID*, 19.

Ciprian, R. and Carbucicchio, M. (2011). Colorimetric–spectral Clustering: A Tool for Multispectral Image Compression, *J. Opt.* **13**, 115402.

Cohen, J. (1964). Dependency of the Spectral Reflectance Curves of the Munsell Color Chips. *Psychon. Sci*. **1**, 369.

Connah, D., Alsam A., and Hardeberg, J. Y. (2004). Multispectral Imaging: How many Sensors do We Need?. *In Proc. of the Twelfth Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications*, *IS&T/SID*, **12**, 53.

Das S., Bora, P. K., and Gogoi, A. K. (2010). Subtractive Clustering of Vertices for CPCA Based Animation Geometry Compression, *In Proc. of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing,* 205.

Du Q. and Fowler, J. E. (2007). Hyperspectral Image Compression using JPEG2000 and Principal Component Analysis, *In Proc. of the IEEE Geoscience and Remote Sensing Letters*, **4**, 201.

Fairchild, M. D., Rosen M. R., and Johnson G. M. (2001). Spectral and Metameric Color Imaging. *Technical Report*, Munsell Color Science Laboratory.

Filzmoser, P. (2004). A Multivariate Outlier Detection Method, *In Proc. International Conference on Computer Data Analysis and Modeling*, 18.

Freeman, F. Downs, L. Marcucci, E. N. Lewis, B. Blume, and Rish, J. (1997). Multispectral and Hyperspectral Imaging: Applications for Medical and Surgical Diagnostic. *In Proc. of the 19th International Conf. IEEE/EMBS,* 700.

Garcia-Beltran, A., Nieves, J. L., Hernandez-Andres, J., Romero. J. (1998). Linear Bases for Spectral Reflectance Functions of Acrylic Paints, *Col. Res. & Appl. J.,* **23**, 39.

Hardeberg, J. Y. (2002).On the Spectral Dimensionality of Object Colours. *In Proc. of the First European Conference on Color in Graphics, Imaging and Vision (CGIV)*,
480.

Hauta-Kasari, M., Miyazawa, K., Toyooka, S., and Parkkinen, J. (1999). Spectral Vision System for Measuring Color Images, *J. Opt. Soc. Am. A.,* **16**, 2352.

Hernandez-Andres, J., Romero. J., García-Beltrán, A., and Nieves J. L. (1998). Testing Linear Models on Spectral Daylight Measurements. Appl. Opt. 37, 971.

Ho, J., Yang, M. H. Lim, J. Lee, K.C., and Kriegman, D. (2003). Clustering Appearances of Objects under Varying Illumination Conditions. *In IEEE Conf. on Computer Vision and Pattern Recognition*, 1.

Hordley, S. Finlayson, G. Morovic, P. A. (2004). Multi-Spectral Image Database and an Application to Image Rendering across Illumination, *In Proc. of the third International Conference on Image and Graphics*, 394.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, **24**, 417.

Imai, F. H. and Berns, R. S. (1999). Spectral Estimation using Trichromatic Digital Cameras. *In Proc of the International Symposium on Multispectral Imaging and Color Reproduction for Digital Archives,* 42.

Imai, F. H., Rosen M. R. and Berns, R. S. (2000). Comparison of Spectrally Narrow-band Capture versus Wide-band with a Priori Sample Analysis for Spectral Reflectance Estimation, *In Proc. of Eighth Color Imaging Conference: Color Science and Engineering, Systems, Technologies and Applications, IS&T*, 234.

Imai, F. H., Rosen M. R. and Berns, R. S. (2001). Multi-spectral Imaging of van Gogh's Self Portrait at the National Gallery of Art. *In Proc. PICS: Image Processing, Image Quality, Image Capture Systems Conference (IS&T),* 185.

Johnson, R. A., and Wichern D. W. (1992). *Applied Multivariate Statistical Analysis*, Prentice-Hall.

Jolliffe I. T. (2002) *Principal Component Analysis*, 2nd ed., Springer Series in Statistics, (Springer-Verlag).

Kaarna, A., Zemcik, P., Kalviainen, H. and Parkkinen, J. (1998). Multispectral Image Compression, *In Proc. of the 14th International Conference on Pattern Recognition,* 1264.

Kaarna, A. and Parkkinen, J. (2000). Comparison of Compression Methods for Multispectral Images, *In Proc. NORSIG Nordic Signal Process. Symp*., 2, 251.

Kambhatla, N. and Leen T. K. (1997). Dimension Reduction by Local PCA. *Neural Computation,* **9,** 1493.

Konig F. and Praefcke W. (1998). The Practice of Multispectral Image Acquisition. *In Proc. of International Symposiom on electronic capture and publishing, SPIE* **3409***.*

Laamanen, H., Jaaskelainen, T., Parkkinen, J. P. S. (2001). Comparison of PCA and ICA in Color Recognition, *In Proc. SPIE*, 367.

Laamanen, H., Jetsu, T., Jaaskelainen, T., and Parkkinen, J. P. S. (2008). Weighted Compression of Spectral Color Information. *J. Opt. Soc. Am, A*, **25**, 1383.

LIBRA: a MATLAB Library for Robust Analysis, (2006), http://users.jyu.fi/~samiayr/DM/demot/LIBRA_Contents.pdf.

Maitre, H., Schmitt, F., Crettez, J., Wu, Y., and Hardeberg, J. Y. (1996). Spectrophotometric Image Analysis of Fine Art Paintings, *In Proc. IS&T/SID Color Imaging Conf.,* 50.

Maloney, L. T. (1986). Evaluation of Linear Models of Surface Spectral Reflectance with Small Numbers of Parameters. *J. Opt. Soc. Am. A,* **3**, 1673.

MATLAB (a) and Statistics Toolbox Release 2013a, The MathWorks, Inc., Natick, Massachusetts, United States.

MATLAB (b) and System Identification Toolbox, Release 2013a, The MathWorks, Inc., Natick, Massachusetts, United States.

Munsell Book of Color - Matte Finish Collection (Munsell Color, Baltimore, Md., 1976).

Parkkinen, J. P. S., Hallikainen J., and Jaaskelainen, T. (1989). Characteristic Spectra of Munsell Colors, *J. Opt. Soc. Am. A*., **6**, 318.

Parsons L., Haque, E., Liu, H. (2004) Subspace Clustering for High Dimensional Data: A Review, ACM SIGKDD Explorations Newsletter, **6**, 90.

Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* **2**, 559–572.

Penna, B., Tillo, T., Magli, E., and Olmo, G. (2007). Hyperspectral Image Compression Employing a Model of Anomalous Pixels, *In Proc. of the Geoscience and Remote Sensing Letters,* 664.

Rayat, A., Amirshahi, S. H., Agahian, F. (2012, online published) Compression of Spectral Data using Box-Cox. *Col. Res. & Appl. J.*

Rousseeuw, P. J. (1984) Least Median of Squares Regression, *Journal of the American Statistical Association,* **79**, 871.

Rousseeuw P. J. and Van Driessen K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator, *Technometrics*, **41**, 212.

Sattler M., Sarlette R., and Klein R. (2005) Simple and Efficient Compression of Animation Sequences. *In Eurographics/ACM SIGGRAPH Symposium on Computer Animation*, 209.

Sharma, G., Trussell, H. J., and Vrhel, M. J. (1998). Optimal Nonnegative Color Scanning Filters. *IEEE Trans. Image Processing,* **7**, 129.

Tzeng D. Y., Berns, R. S. (2005) A Review of Principal Component Analysis and Its Applications to Color Technology, *Col. Res. & Appl. J.,* **30**, 84.

Vidal, R., Ma Y., and Sastry S. (2005). Generalized Principal Component Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence, **27,**1945.

Vidal, R. (2011). A Tutorial on Subspace Clustering. *IEEE Signal Processing Magazine,* **28** 52.

von Luxburg, U. (2007). A Tutorial on Spectral Clustering. *Stat. Comput*. **17**, 395.

Vrhel, M. J. and Trussell, H. J. (1994). Filter Considerations in Color Correction, *IEEE Trans. Image Processing,* **3**, 147.

Wei Hong, John Wright, Kun Huang, and Yi Ma. (2006). Multi-scale Hybrid Linear Models for Lossy Image Representation, *IEEE Trans. on Image Processing*, **15**, 3655.

Yamaguchi, M. (2001). Medical Application of a Color Reproduction System with a Multispectral Camera. *Digital Color Imaging in Biomedicine*, 33.

Yang A., Wright, J. Ma Y., and Sastry S. (2008). Unsupervised Segmentation of Natural images via Lossy Data Compression. *Computer Vision and Image Understanding*, **110**, 212.

Yasuma, F. Mitsunaga, T. Iso, D. and Nayar, S.K. (2008). Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum. *Technical Report*, Department of Computer Science, Columbia University.

Zhao, Y. Taplin, L. A., Nezamabadi, M., and Berns, R. S. (2005). Using Matrix R Method in the Multispectral Image Archives, *In Proc. AIC*, 469.