Representing Outliers for Improved Multi-Spectral Data Reduction

Farnaz Agahian and Brian Funt; Simon Fraser University; Vancouver, B.C., Canada

Seyed Hossein Amirshahi, Amirkabir University of Technology (Tehran Polytechnic); Tehran, Iran

Abstract

Large multi-spectral datasets such as those created by multi-spectral images require a lot of data storage. Compression of these data 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 paper, we employ the fast MCD (Minimum Covariance Determinant) algorithm, as a highly robust estimator of multivariate mean and covariance, to detect outlier spectra in a multi-spectral image. We then show 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 is 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 show that significantly better compression can be achieved with this approach.

Introduction

Conventional 3-channel image color imaging devices capture limited spectral information about each scene location. RGB images are device-dependent in that they depend on the spectral sensitivity functions, which may differ from one device to another. In addition, the RGB color information depends on the scene illuminant. A change in illuminant leads to the problems of metamerism. The limitations of 3-channel color imagery, especially when high-fidelity color reproduction is required as, for example, in the reproduction and conservation of fine arts painting, are frequently overcome by moving to multi-spectral image capture [1-4].

The spectral reflectance defines an excellent "fingerprint" of a surface and provides the most useful information for color specification under any illuminant and for any observer. In the last decade, multi-spectral imaging has gained a growing interest in several applications such as color reproduction [4-5], medical imaging [6], art conservation science and digital image archives with high color accuracy [1-4]. Unlike typical digital photography, the multi-spectral imaging systems based on acquiring the spectral reflectance at each pixel of an image provide a device-independent representation that can be rendered as a correct color under any viewing condition.

Although the extra information provided by a multi-spectral 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 [7-9].

It is well documented that the spectral reflectance of a non-fluorescent objects is generally a smooth function of wavelength, and therefore can be modeled via dimensionality reduction techniques. In the other words, the smooth spectral reflectances are usually highly correlated and can be represented as a linear combination of a few basis vectors. Principal component analysis (PCA) is a well-known technique [10] in multivariate data analysis that has been extensively used in the context of spectral imaging as an efficient technique for spectral decorrelation as well as spectral dimensionality reduction [11]. PCA determines a linear transformation from the high-dimensional spectral space to the low-dimensional spectral subspace, which among all linear transformations guarantees the best possible representation of the high-dimensional spectral vector in the low-dimensional subspace, spanned by the a few numbers of basis vectors. This feature has made PCA a powerful tool for spectral compression.

It should be noted that the projected data can reconstructed to the original space; however, the compression process will usually lead to some error in the reconstructed data. According to Laamanen et al. [12], the number of basis vectors required for effective recovery of reflectance totally 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 by using PCA. Applying weighting factors on individual samples [13] and clustering of the main dataset based on a predefined criterion [14-15] are techniques that have been used to enhance the efficiency of linear models by increasing the similarity of the elements in the dataset.

It is worth noting that in each dataset there are 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 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 [10].

Analysis of the spectral reconstruction of 1269 Matte Munsell color chips [16] indicates that some color samples, mostly in the family of purples, have a detrimental effect on the spectral and colorimetric 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 (in terms of RMS) is more than the median error of the whole dataset also have a large robust Mahalanobis distance from the mean. If we omit purples from Munsell dataset and then extract eigenvectors and use these eigenvectors for reconstruction of all 1269 samples, the error is less than reconstruction with bases extracted from all samples (including purples). This observation motivated us to study the effect of outlier spectra in a large datasets of reflectance spectra, including those derived from multi-spectral images, and then propose a new method of compressing spectra based on the following steps: (1) separate the outliers from the non-outliers; (2) use standard PCA data reduction on the non-outliers; (3) apply k-means clustering to the outliers; (4) apply PCA data reduction to the clusters individually.

Outlier Detection in a Spectral Dataset

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 $\mathbf{x}_j = [x_1, x_2, \ldots, x_p]^T$ from a dataset with mean $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]$ and covariance matrix $\mathbf{S}$ the Mahalanobis distance is defined as

$$\mathrm{MD}(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})} \quad (1)$$

Multivariate outliers can be defined as observations having a large Mahalanobis distance. A quantile of the chi-squared distribution ($\sqrt{X^2_{P,0.975}}$) 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 [17-18]. There exist several robust estimators for mean and covariance. The minimum covariance determinant (MCD) [18-19] is widely known in the literature as a computationally fast algorithm and is the one we employ 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. A complete description of the algorithm is presented in [18-19]. A Matlab library for robust analysis is readily available [20].

In this study we used one multispectral image entitled "Fruits and Flowers" from the Joensuu spectral image database [16] and four multi-spectral images available from the database of Hordley et al. [21]. "Fruits and Flowers" is a 120×160 pixel image containing 19,200 spectral reflectances sampled at 10 nm intervals over the range 400 nm to 700 nm. Another four multispectral images have also been measured in the same wavelength band with the same sampling rate. The number of spectra in each image is reported in Table II. It should be noted that the border of these images was removed before analysis, so the reported number of spectra in Table II is slightly different from the actual size of the images in [21]. In this paper, we show the steps of our method on "Fruits and Flowers" and report only the final results for the other images in Table II.

The result of using the "Fast MCD" algorithm [18] in conjunction with the MD distance (denoted $\mathrm{MD}_{\mathrm{MCD}}$) on the 19,200 Fruits and Flowers spectra is shown in Fig. 1. As can be seen, there is a substantial difference in the distances as measured by $\mathrm{MD}_{\mathrm{MCD}}$ as compared to $\mathrm{MD}_{\mathrm{classic}}$ (i.e., MD as defined in Eq. 1), and this leads to very different sets of outliers. The red line represents the quantile cutoff value of $\sqrt{X^2_{31,0.975}}$ =6.94 for the classification as an outlier. Based on this criterion, 7741 out of the 19,200 spectra were recognized as outliers by $\mathrm{MD}_{\mathrm{MCD}}$ in comparison to only 3358 by MDclassic. 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 [10]. The dataset is divided into outliers and non-outliers for the next processing steps, which involve applying PCA to the non-outliers and clustering of the outliers.

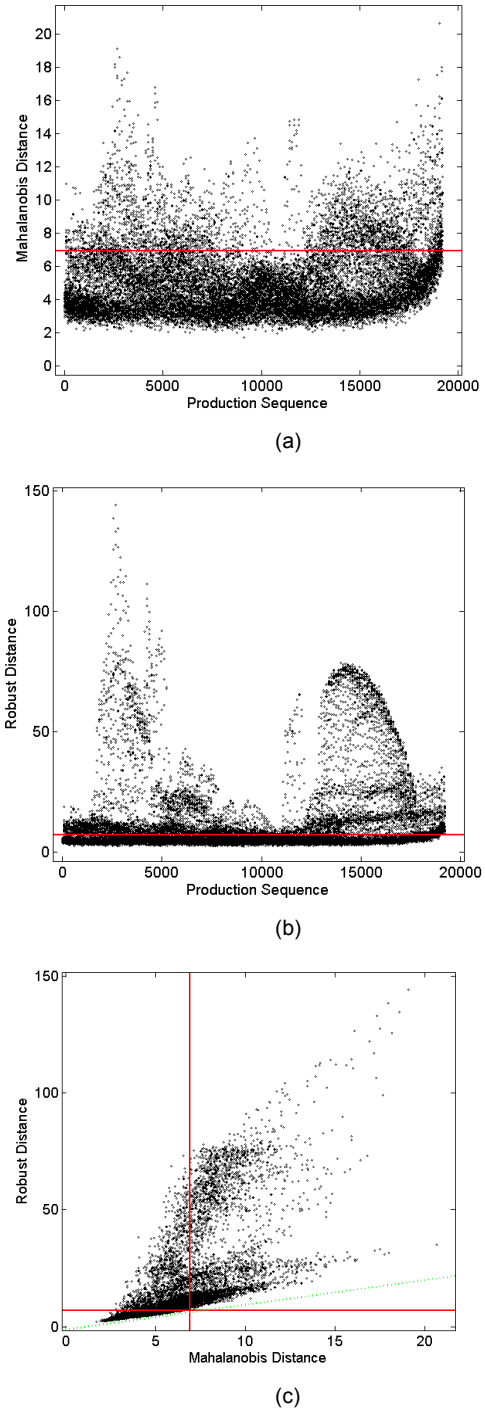

Figure 1. *Distance measures for the 19,200 Fruit and Flowers spectra: (a) Classic Mahalanobis distance $MD_{classic}$ versus pixel number; (b) Robust distance MD_{MCD} versus pixel number; and (c) MD_{MCD} versus $MD_{classic}$. The horizontal and vertical lines represent the quantile cutoff value.*

Clustering the Outliers

The outlier spectra are a part of the original dataset so we cannot simply ignore them. However, we can apply PCA to the non-outlier set and thereby get a better representation of it than of the entire dataset, and then represent the outliers separately. To represent the outliers, we group them into several clusters

Table I. Spectral and colorimetric accuracy of reflectance reconstruction of the Fruits and Flowers image using classic PCA and the proposed method. The errors of each step of this method are given separately. The final results obtained both with classic PCA and the proposed method are in the grey rows.

	#Spectra	RMS		GFC	ΔE^*_{00}	
		Mean	Max		D65	A
Classic PCA						
Original Dataset	19200	0.0114	0.0744	0.9892	3.88	3.99
Proposed Method 1st Step						
Central Cluster	11459	0.0058	0.0209	0.9938	2.29	2.42
Outliers	7741	0.0166	0.068	0.9905	2.99	2.86
Proposed Method 2nd Step						
Outlier (Cluster 1)	661	0.0050	0.0179	0.9964	1.31	1.50
Outlier (Cluster 2)	3276	0.0100	0.0428	0.9977	1.83	1.40
Outlier (Cluster 3)	2701	0.0092	0.0448	0.9988	1.18	1.15
Outlier (Cluster 4)	1103	0.0075	0.0387	0.9980	0.84	0.71
Proposed Method 3rd Step	19200	0.0079	0.0448	0.9970	1.49	1.43

based on a similarity measure such that the spectra in each group are highly correlated. However, finding the number of appropriate number of clusters is an issue in itself. For this step, we used subtractive clustering as implement in Matlab's subclust function to help determine the minimum number of potential clusters. This is done by gradually decreasing the number of clusters and calculating the corresponding mean RMS error in spectral reconstruction. As Fig. 2 shows, data clustering has a significant impact on the reconstruction error when the number of clusters is increased from 1 to 4. Beyond 4, the error goes down slightly until the number of clusters reaches 12, which is the optimal number of clusters as determined by *subclust*. Based on this analysis, we partitioned outlier spectra into 4 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 Flowers and Fruits spectra, using 4 clusters of outliers worked well. The final clustering was done using k-means clustering (*kmeans* from the statistics toolbox of Matlab [22]) with the number of clusters set to 4 and the *cosine* distance was chosen as the distance parameter. The *cosine* distance between two spectra is the *cosine* of the angle between them viewed as vectors.

Spectral Compression

For each cluster, PCA was used to reduce the dimension of the spectra from 31 to 3. In total, the whole dataset is partitioned into 5 clusters (counting the non-outlier set as a cluster). As a result, 15 basis vectors (5 clusters × 3 bases) are required so the data from each cluster can be projected into its own 3D spectral subspace. The spectra are then reconstructed using the appropriate cluster's basis. The reconstruction error is tabulated in Table I both in terms of spectral and color accuracy. Two spectral measures (RMS and Goodness of Fit Coefficient, GFC [23]) and two colorimetric measures (CIEDE2000 color differences under illuminants A and D65) were used.

A comparison between the two rows of Table I highlighted in gray shows a considerable improvement in the spectral reconstruction using the proposed method in terms of both reduced spectral and colorimetric errors. As can be seen, 7741 spectra detected by MD_{MCD} as outliers are removed from the main dataset during the first step. The spectral dimension of the remainder of the data labeled Central Cluster is reduced to 3 via PCA. The central cluster benefits from the fact that the outliers have been removed so the remaining highly correlated data is efficiently and accurately represented using only 3 dimensions.

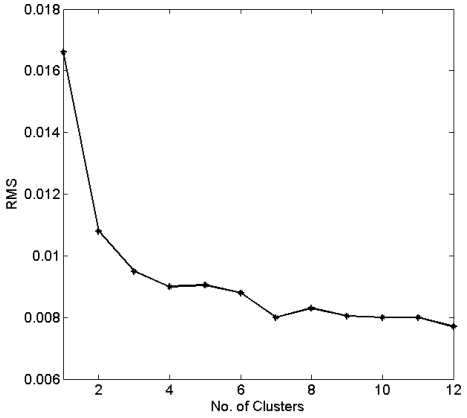

***Figure 2**. Mean of reconstruction error for the outlier spectra as a function of the number of clusters used.*

Table II. Spectral and colorimetric accuracy of reflectance reconstruction of four spectral images taken from the multi-spectral database [21] using classic PCA and the proposed method.

	#Spectra	RMS		GFC	ΔE^*_{00}	
		Mean	Max		D65	A
Persilnonbio Image						
Classic PCA	77004	0.0137	0.0589	0.9969	1.48	1.21
Proposed Method		0.0108	0.0317	0.9989	1.10	1.01
Goaheadbars Image						
Classic PCA	94666	0.0089	0.0365	0.9944	1.54	1.41
Proposed Method		0.0054	0.0204	0.9980	1.01	1.19
Couscous Image						
Classic PCA	84216	0.111	0.0611	0.9934	2.08	1.64
Proposed Method		0.009	0.0218	0.9981	1.43	1.31
Elastoplast Image						
Classic PCA	35316	0.0145	0.0640	0.9937	2.07	1.82
Proposed Method		0.0104	0.0280	0.9982	1.53	1.48

The outlier spectra are partitioned and reduced to 3 dimensions in the second step. As the clustering was performed based on a similarity measure, the spectra assigned to each cluster are again highly correlated leading to an efficient PCA-based 3-dimensional representation.

As Zhao et al. [3] point out, while high spectral reconstruction accuracy is important, in some applications such as digital image archives for museums, high colorimetric accuracy under various lighting conditions can also be important. The ΔE^*_{00} errors in Table I decrease by more than a factor of 2 when the proposed approach is used in place of standard PCA.

As Table II shows, using the proposed method in place of classic PCA improves the reconstruction of the four multi-spectral images from the Hordley et al. database [21] in terms of both spectral and colorimetric accuracy.

Although multi-spectral image compression is beyond the scope of the present paper, it should be noted that, in addition to spectral redundancy, multi-spectral images typically include a high degree of spatial correlation. The proposed 3-step technique can be combined with image compression techniques for further data compression. For example, lossless JPEG2000 compression of the 3-dimensional representation of the Fruits and Flowers spectra reduces the amount of space required by a factor of two without increasing the reconstruction error.

Conclusion

Large spectral datasets such as those provided by the spectra from a multi-spectral image can be represented in lower dimensions using traditional PCA [24]; however, often the datasets include spectra that differ markedly from the bulk of the dataset and this can lead to poor spectral reconstruction. The proposed method improves the efficiency of a model of any fixed dimension by separating out the outlier spectra and treating them separately. A separate PCA basis is used to represent each cluster of outliers. The outliers are identified using a robust Mahalanobis distance measure provided by the minimum covariance determinant. Tests show that the proposed 3-step method leads to lower reconstruction errors both in terms of spectral and color differences.

References

1. F. H. Imai, M. R. Rosen, and R. S. Berns, Multi-spectral imaging of van Gogh's self portrait at the National Gallery of Art, Proc. PICS: Image Processing, Image Quality, Image Capture Systems Conference (IS&T), 185 (2001).
2. R. S. Berns, The science of digitizing paintings for color-accurate image archives: a review, J. Imaging Sci. and Technol, 4, 305 (2001).
3. Y. Zhao, L. A. Taplin, M. Nezamabadi and R. S. Berns, Using matrix R method in the multispectral image archives, Proc. AIC, 469 (2005).
4. H. Maitre, F. Schmitt, J Crettez, Y. Wu and J. Y. Hardeberg, Spectrophotometric image analysis of fine art paintings, Proc. IS&T/SID Color Imaging Conf., 50 (1996).
5. M. Yamaguchi, T. Teraji, K. Ohsawa, T. Uchiyama, H. Motomura, Y. Murakami, and N. Ohyama, Color image reproduction based on the multispectral and multiprimary imaging: experimental evaluation, Proc. SPIE, 15 (2002).
6. M. Okuyama, N. Tsumura, and Y. Miyake, Evaluating a multi-spectral imaging system for mapping pigments in human skin, Opt. Rev., 10, 580 (2003).
7. A. Kaarna, P. Zemcik, H. Kalviainen, and J. Parkkinen, Multispectral image compression, Proc. of the 14th International Conference on Pattern Recognition, 1264 (1998).
8. Q. Du, J. E. Fowler, Hyperspectral image compression using JPEG2000 and principal

component analysis, Proc. of the Geoscience and Remote Sensing Letters, 201. (2007).
9. B Penna, T Tillo, E Magli, and G Olmo, Hyperspectral Image Compression Employing a Model of Anomalous Pixels, Proc. of the Geoscience and Remote Sensing Letters, 664 (2007).
10. I. T. Jolliffe, Principal Component Analysis, 2nd ed., Springer Series in Statistics, (Springer-Verlag, NY, 2002).
11. D. Y. Tzeng, R. S. Berns, A review of principal component analysis and its applications to color technology, Color Res. Appl., 30, 84 (2005).
12. H. Laamanen, T. Jaaskelainen, J. P. S Parkkinen, Comparison of PCA and ICA in color recognition, Proc. SPIE, 367. (2001).
13. F. Agahian, S. A. Amirshahi and S.H. Amirshahi, Reconstruction of Reflectance Spectra Using Weighted Principal Component Analysis, Col. Res. & Appl. J., 33, 360 (2008).
14. A. Garcia-Beltran, J. L. Nieves, J. Hernandez-Andres, J. Romero. Linear bases for spectral reflectance functions of acrylic paints, Color Res Appl, 23, 39 (1998).
15. F. Ayala, J. F. Echavarri and P. Renet, 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 (2006).
16. Spectral Database, University of Joensuu Color Group, http://spectral.joensuu.fi/
17. P. Filzmoser, A multivariate outlier detection method, Proc. International Conference on Computer Data Analysis and Modeling, 18. (2004).
18. P. J. Rousseeuw and K. Van Driessen, A Fast Algorithm for the Minimum Covariance Determinant Estimator, Technometrics, 41, 212 (1999).
19. P. J. Rousseeuw, Least Median of Squares Regression, Journal of the American Statistical Association, 79, 871 (1984).
20. http://wis.kuleuven.be/stat/robust/LIBRA.html
21. S. Hordley, G. Finlayson, P. Morovic, A Multi-Spectral Image Database and an Application to Image Rendering Across Illumination, Proc. Thrid International Conference on Image and Graphics. (2004).
22. The MathWorks Inc. MATLAB R2010b, Version 7.1.
23. F. H. Imai, M. R. Rosen and R. S. Berns, Comparative Study of Metrics for Spectral Match Quality, Proc. CGIV, 492. (2002).
24. J. P. S. Parkinnen, J. Hallikainen, and T. Jaaskelainen, Characteristic spectra of Munsell colors, J. Opt. Soc. Am. A., 6, 318 (1989).

Author Biography

Farnaz Agahian is a doctoral student in the department of Computing Science, Simon Fraser University, having already received the PhD degree in Color Science from Tehran Polytechnic in 2009. Her research areas include color reproduction, reconstruction of spectral data and metamerism.

Brian Funt is Professor of Computing Science at Simon Fraser University where he has been since 1980. He obtained his Ph.D. from the University of British Columbia in 1976. His research focus is on computational approaches to modeling and understanding color.

Seyed Hossein Amirshahi received the Ph.D. degree from the University of New South Wales, Australia in 1994. Currently, he is professor of Color Science at Tehran Polytechnic and does research in classic color physics, recovery of spectral data from colorimetric values and digital color imaging and printing.