Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 111 (2017) 113–119

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Conference on Advances in Information Technology, IAIT2016, 19-22 December 2016, Macau, China

Principal component analysis to reduce dimension on digital image

S. C. Ng*

School of Information Technology, SEGi University, PJU 5, Kota Damansara, Selangor 47810, Malaysia

Abstract

High-resolution image is referred as high-dimensional data space as each image data is organized into two-dimensional pixel values in which each pixel consists of its respective RGB bits value. The representation of image data poses a challenge to sharing image files over Internet. The lengthy image uploading and downloading time has always been a major issue for Internet users. Apart from data transmission problem, high-resolution image consumes greater storage space. Principal Component Analysis (PCA) is a mathematical technique to reduce the dimensionality of data. It works on the principal of factoring matrices to extract the principal pattern of a linear system. This paper aims to evaluate the application of PCA on digital image feature reduction and compare the quality of the feature reduced images with difference variance values. As a result of summarizing the preliminary literature, dimension reduction process by PCA generally consists of four major steps: (1) normalize image data (2) calculate covariance matrix from the image data (3) perform Single Value Decomposition (SVD) (4) find the projection of image data to the new basis with reduced features. Experimental results showed that PCA technique effectively reduces the dimension of image data while still maintaining the principal properties of the original image. This technique achieved 35.3% for the file size reduction for the best feature reduced quality. The transmission time of image file over Internet has achieved significant improvement especially for the download activity via mobile devices.

Peer-review under responsibility of the organizing committee of the 8th International Conference on Advances in Information Technology

Keywords: principal component analysis (PCA); digital image compression; single value decomposition (SVD)

* Corresponding author. Tel.: +60123721643.
E-mail address: ashleyng@segi.edu.my

Peer-review under responsibility of the organizing committee of the 8th International Conference on Advances in Information Technology
10.1016/j.procs.2017.06.017

1. Introduction

In this advanced information era, efficient transmission of digital images has become a major challenge in information technology[1]. Vast volumes of digital images are shared over Internet every day. In the meantime, the number of people who are using mobile devices to access to Internet is increasing exponentially. High resolution images need greater bits for storage and transmission. It would take considerable time to fully render a high resolution image on the browser screen. Considering the bandwidth consumption over multimedia communication, having an effective compression coding scheme with low bitrates is crucial to enable the effective and fast transmission of image data over the network. The underlying basis of image compression is reducing the amount of data (bits) used to represent the image. The compression process removes redundant data (details) from the image contents. There are two types of compression methods: *lossy* and *lossless*. The formal method discards redundant data from the image content to reduce the image file size. This method will cause the information loss during the compression process. Thus, the resulting image will have a tolerable degree of degradation in the visual quality. Lossy compression has been widely applied on multimedia file where the quality loss of the resulting digital media is not noticeable but the file size can be reduced significantly. On the other hand, lossless compression maintains the quality of the original file when the file is decompressed. The application of lossless compression can be found in [2-6]. This study applies the statistical approach to discard the redundant features of digital images that consume the storage space. Generally, high resolution image can be perceived as high-dimensional data. Image data is represented in the form of two-dimensional matrix pixel in a problem space. Each pixel element contains bits value for RGB to represents a pixel's colour. Principal Component Analysis (PCA)[7,8] is a technique to reduce the dimension space presented by mathematic functions. This technique extracts the fundamental properties of a linear system by single value decomposition method[9]. PCA has been widely applied to discard noise data in digital signal processing, image recognition as well as solving classification problems[10-14]. Ref.[15] applied PCA to compress digital images used in medicine. Other PCA-based digital image compression literature can be found in[16-20]. This paper consolidates the PCA digital image compression technique into four major steps and compares the quality of the feature reduced images with difference variance values. The transmission time of the original and the resulting image will be compared in two different modes of Internet access which is via desktop and mobile devices. This study assumes that image data is represented by a linear function. Data normalization is conducted on image data prior to the compression process to ensure that all the image data fall into the range specified.

2. Related Work

State-of-art well established compression coding schemes are Discrete Wavelet Transform (DWT), artificial neural network, Fast Fourier Transform (FFT) and Discrete Cosine Transform (DCT) which has been applied in JPEG standard[21-25]. To compress an image, both FFT and DCT transform a two-dimensional matrix of pixel values into an equivalent matrix of two-dimensional waveform in frequency domain. Redundant frequency components are then removed in the frequency domain. In Fourier series analysis, sinusoids are chosen as the basis function. On the other hand, wavelet analysis is based on a decomposition of the data using an orthonormal basis functions[26]. DWT first analyze the image data by decomposing the image data into multiple levels of frequency components based on a specific wavelet. Most of the image data information is contained in low frequency components. Thus, low frequency components are normally associated with the pattern in the image. Details in the image are provided by high frequency component. De-noising is then conducted in high frequency components in which the image data with the value smaller than the specified threshold will be discarded[27-29]. The filtered image data is then synthesized to become an image with lower dimension. Compared to other methods, DWT compression method is sophisticated provided that a proper threshold value is selected in the de-noising process. Another advantage of DWT is that wavelet transform is able to process data in different scales and translation. Owing to this flexibility, wavelet transform consists of infinite set of possible basis function. However, wavelet transform cannot yield a finite resolution in the time and frequency domain simultaneously and it requires significant amount of computation time to process the image data.

3. Principal Component Analysis

PCA is a statistical approach to find the principal features of a distributed dataset based on the total variance[30]. Given a set of multivariate distributed data in X-Y coordinate system (Fig. 1), PCA first finds the maximum variations of the original datasets. These data points are then projected onto a new axis called U-V coordinate system. The direction of U and V-axis is known as principal components. The principal direction in which the data

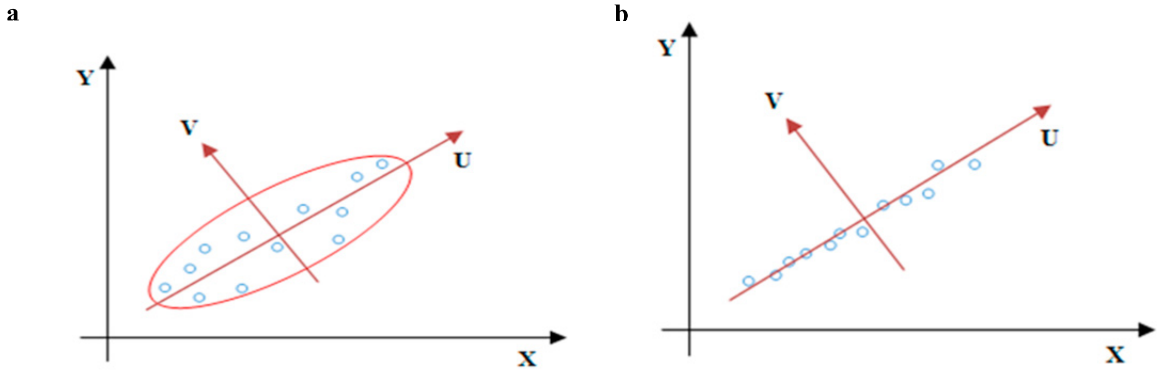

Fig. 1. PCA for dimension reduction (a) adjusted axis system; (b) variable v is discarded

varies is shown by U-axis following by its orthogonal direction, V-axis. In the case where all the data points on V-axis are very close to zero as shown in Figure 1(b), the data set can be represented by only one variable U and the variable V can be discarded.

4. Dimension Reduction Process on Digital Image

A digital image can be represented by a group of pixels arranged in the form of two-dimensional matrix as shown in (1)[15]. These pixels are in floating point values for coloured image and they are in discrete values when it is a gray scaled image.

$$F(x,y) = \begin{bmatrix} f(0,0) & \cdots & f(0,m-1) \\ \vdots & \ddots & \vdots \\ f(n-1,0) & \cdots & f(n-1,m-1) \end{bmatrix} \quad (1)$$

where x and y are the coordinates of the pixels in the image and $f(x,y)$ is the corresponding colour or gray level depending on its value type. In general, the image dimension reduction technique by PCA[15] can be consolidated into four major steps: image normalization, finding the covariance matrix of the image data, calculate the eigenvectors and eigenvalues of the covariance matrix and lastly transforming image data into new basis.

4.1. Image Normalization

As the pre-processing phase of PCA, normalization is performed on the image dataset. The process of subtracting the image dataset with their mean value (2) can improve the signal-to-noise ratio (SNR) and discrimination power (DP) of the dataset.

$$F_{normalized}(x,y) = \begin{bmatrix} f(0,0) & \cdots & f(0,m-1) \\ \vdots & \ddots & \vdots \\ f(n-1,0) & \cdots & f(n-1,m-1) \end{bmatrix} - [\bar{f}(0,0) \quad \ldots \quad \bar{f}(0,m-1)] \quad (2)$$

where $[\bar{f}(0,0) \quad \cdots \quad \bar{f}(0, m-1)]$ is the column vector containing the mean value for y_1 to y_m

4.2. Covariance Matrix of Image Data

Covariance matrix is calculated to locate the greatest variance value of the dataset. The covariance matrix of the image, $\mathrm{Cov}(X, Y)$ is obtained by the following equation:

$$Cov(X,Y) = \frac{F_{normalized}(x,y)*F_{normalized}(x,y)^T}{m-1} \tag{3}$$

where *m* is the number of element *y*.

4.3. Eigenvectors and Eigenvalues of the Covariance Matrix

Eigenvector with the largest eigenvalue is the direction of greatest variation in the image data set. The eigenvector matrix represents the principal features of the image data. Eigenvectors and eigenvalues of covariance matrix, AA^T, can be obtained by the following SVD equation.

$$AA^T = Cov(X,Y) = UD^2U^T \tag{4}$$

where *U* are the eigenvectors of $A\,A^T$, and the square of the singular values in *D* are the eigenvalues of $A\,A^T$.

4.4. Transforming Image Data into New Basis

This step transforms the original image dataset into a new axis with reduced dimension. Equation (5) is used to conduct the projection of the original image with the eigenvectors obtained from (4):

$$F_{transformed}(x,y) = U^T F_{normalized}(x,y) \tag{5}$$

where U^T is the transpose of the eigenvectors matrix and $F_{normalized}(x,y)$ is the adjusted original image dataset.

5. Results and Discussion

Fig. 2 shows the comparison of the original image with the feature-reduced image after applying PCA. Figure 2(b) and 2(c) are the feature-reduced image with the variance 0.97 and 0.95 respectively. It can be observed that after PCA process the less significant features of the image data has been discarded but the image still maintains its principal characteristics. Instead, the file size of Figure 2(b) and (c) has reduced 39% and 40% respectively compared to the original image. Fig. 3 shows the comparison of the original image with its feature-reduced image (variance = 0.99). After PCA process, the dimension of the original image has been reduced without losing much information of the image. The file size of the feature-reduced image has achieved 35.3% reduction.

a b c

Fig. 2. (a) Original image (size = 175KB); (b) feature-reduced image (size = 107KB, variance = 0.97); (c) feature-reduced image (size = 101KB, variance = 0.95)

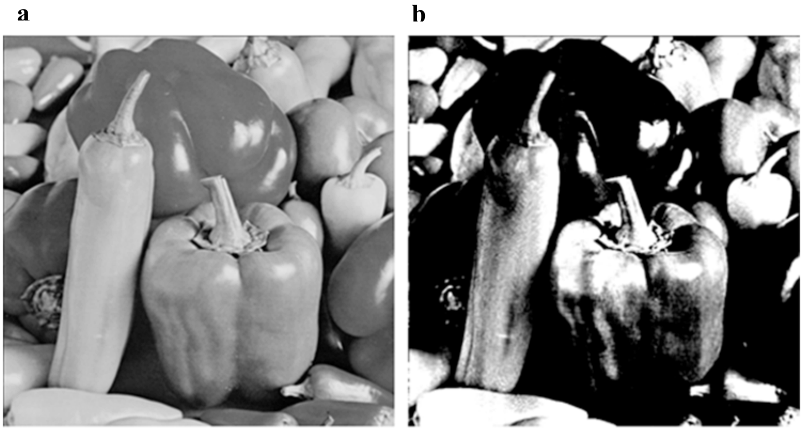

Fig. 3. (a) Original image (size = 156KB); (b) feature-reduced image (size = 101KB, variance = 0.99)

Fig. 4 shows the comparison of the elapsed time to upload and download an image over Internet via desktop and mobile devices. Experimental results show that the uploading time has improved 13.3% when the image has been compressed. The downloading time via desktop and mobile devices has respectively decreased 9.29% and 28.9%. The results show the significant improvement of time taken to download an image for mobile device users.

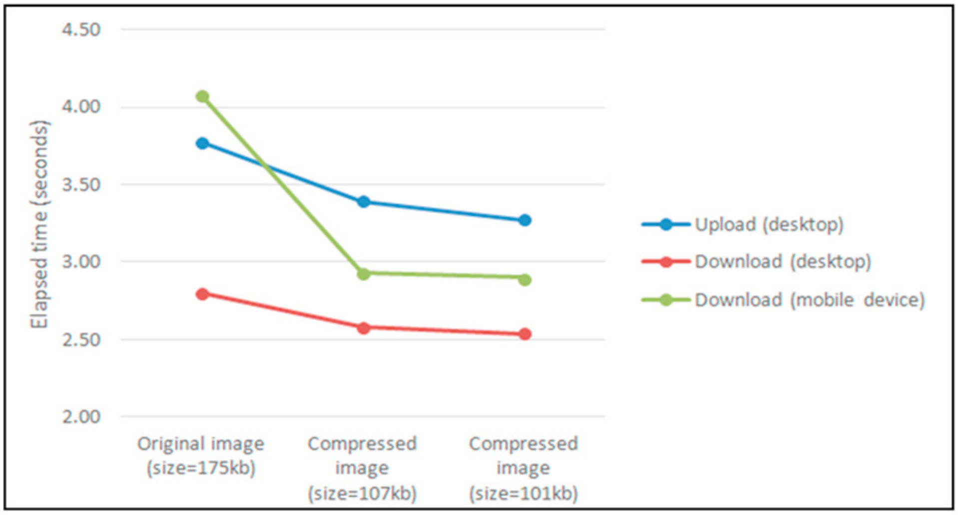

Fig. 4. Transmission time of image file over the Internet

6. Conclusion

PCA reduces dimension of problem space by locating the greatest variance on distributed datasets. This study evaluates the application of PCA on feature reduction in image processing. By calculating the eigenvectors and eigenvalues of the covariance matrix, principal components of the dataset can be obtained. These principle components hold the feature factor information of the original image. The application of PCA on data reduction in image is a lossy compression method. To obtain an effective digital image sharing over the Internet environment where the network bandwidth is part of the factor concerned, the trade-off for the human visual quality of the image to its relative file size compressed has to achieve a balance. The decision on the variance value is important as it dictates the quality and file size of the resulting image. Smaller file size image can save the transmission time (uploading and downloading) but certain degree of visual quality will be compromised. Experimental results show that this approach contributes significant saving of storage and transmission time for the image files while still manage to maintain the integrity of the image. Nevertheless, this approach is limited to linear systems with uncorrelated variables. Further research will be investigating an alternative dimension reduction approach for solving non-linear problem space with correlated variables.

Acknowledgements

This work is supported by the SEGi Internal Research Fund under the project no. SEGi/2015-02/SoIT-3/85.

References

1. Y. E. Gelogo, and T. Kim, "Compressed images transmission issues and solutions," *International Journal of Computer Graphics*, vol. 5, no. 1, pp. 1-8, 2013.
2. V. A. Kokovin, S. S. Uvaysova and S. U. Uvaysov, "Lossless compression algorithm for use in telecommunication systems," *2016 International Siberian Conference on Control and Communications (SIBCON)*, Moscow, 2016, pp. 1-4.
3. P. C. Lin, Y. H. Pai, Y. H. Chiu, S. Y. Fang and C. C. P. Chen, "Lossless compression algorithm based on dictionary coding for multiple e-beam direct write system," *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, 2016, pp. 285-288.
4. A. Dzhagaryan and A. Milenkovic, "On effectiveness of lossless compression in transferring mHealth data files," *2015 17th International Conference on E-health Networking, Application & Services (HealthCom)*, Boston, MA, 2015, pp. 665-668.
5. S. Wandelt; X. Sun; Y. Zhu, "Lossless Compression of Public Transit Schedules," *IEEE Transactions on Intelligent Transportation Systems*, vol.PP, no.99, pp.1-12.
6. G. Vallathan and K. Jayanthi, "Lossless compression based on hierarchical extrapolation for biomedical imaging applications," *2015 International Conference on Microwave, Optical and Communication Engineering (ICMOCE)*, Bhubaneswar, 2015, pp. 146-149.
7. B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Transactions on Automatic Control*, vol. 26, no. 1, 17-32, 1981.
8. I. Jolliffe, *Principal Component Analysis*, Springer, 2002.
9. J. Tzeng, "Split-and-combine singular value decomposition for large-scale matrix," *Journal of Applied Mathematics*, 2013.
10. R. Petrolis, and A. Krisciukaitis, "Multi stage Principal component analysis based method for detection of fetal heart beats in abdominal ECGs," *Computing in Cardiology Conference 2013*, Zaragoza, 2013, pp. 301 - 304.
11. H. Shi, "Application of principal component analysis to general contracting risk assessment," *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, Sanya, 2009, pp. 53-56.
12. G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Principal component analysis of image gradient orientations for face recognition," *2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, Santa Barbara, CA, 2011, pp. 553-558.
13. Li, H., "Accurate and efficient classification based on common principal components analysis for multivariate time series," *Neurocomputing*, vol. 171, pp. 744-753, 2016.
14. P. Kamencay, T. Trnovszky, M. Benco, R. Hudec, P. Sykora and A. Satnik, "Accurate wild animal recognition using PCA, LDA and LBPH," *2016 ELEKTRO*, Strbske Pleso, 2016, pp. 62-67.
15. R. D. Santo, "Principal component analysis applied to digital image compression," *Einstein*, vol. 10, no. 2, pp. 135-139, 2012.
16. Q. Du and James E. Fowler, "Low-complexity principal component analysis for hyperspectral image compression," *International Journal of High Performance Computing Applications*, vol. 22, no. 4, pp. 438 - 448, 2002.
17. C. Lee, S. Youn, T. Jeong, E. Lee and J. Serra-Sagristà, "Hybrid Compression of Hyperspectral Images Based on PCA With Pre-Encoding Discriminant Information," in *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 7, pp. 1491-1495, July 2015.
18. C. W. Wang and J. H. Jeng, "Image compression using PCA with clustering," *2012 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS),* New Taipei, 2012, pp. 458-462.
19. A. Vaish and M. Kumar, "WDR coding based image compression technique using PCA," *2015 International Conference on Signal Processing and Communication (ICSC),* Noida, 2015, pp. 360-365.
20. S. T. Lim, D. F. W. Yap and N. A. Manap, "Medical image compression using block-based PCA algorithm," *Computer, Communications, and Control Technology (I4CT), 2014 International Conference on*, Langkawi, 2014, pp. 171-175.
21. M.-S. Wu, "Genetic algorithm based on discrete wavelet transformation for fractal image compression," *Journal of Visual Communication and Image Representation*, vol. 25, no. 8, pp.1835-1841, 2014.
22. A. J. Hussain, D. Al-Jumeily, N. Radi, and P. Lisboa, "Hybrid neural network predictive-wavelet image compression system," *Neurocomputing*, vol. 151, no. 3, pp. 975-984, 2015.
23. K. M. M. Prabhu, K. Sridhar, M. Mischi, and H. N. Bharath, "3-D warped discrete cosine transform for MRI image compression," *Biomedical Signal Processing and Control*, vol. 8, no.1, pp. 50-58, 2013.

24. AnjanaB and Shreeja R, "Image compression using artificial neural network," *Communication and Computing (ARTCom2012), Fourth International Conference on Advances in Recent Technologies in*, Bangalore, India, 2012, pp. 288-290.
25. J. Quijas and O. Fuentes, "Removing JPEG blocking artifacts using machine learning," *Image Analysis and Interpretation (SSIAI), 2014 IEEE Southwest Symposium on*, San Diego, CA, 2014, pp. 77-80.
26. D. Datta, "Wavelet analysis based estimation of probability density function of wind data," *International Journal of Energy, Information and Communications*, vol. 5 no. 3, pp. 23-34, 2014.
27. Z. Zhou et al., "Improvement of the signal to noise ratio of Lidar echo signal based on wavelet de-noising technique," *Optics and Lasers in Engineering*, vol. 51, no. 8, pp. 961-966, 2013.
28. M. S. Ibraheem, S. Z. Ahmed, K. Hachicha, S. Hochberg and P. Garda, "Medical images compression with clinical diagnostic quality using logarithmic DWT," *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Las Vegas, NV, 2016, pp. 402-405.
29. A. Baviskar, J. Baviskar, A. Mulla, N. Jain and P. Dave, "Coded sub band replacement DWT based space image compression," *2016 IEEE Aerospace Conference*, Big Sky, MT, 2016, pp. 1-8.
30. B.-K. Bao, G. Liu, C. Xu, S. Yan, "Inductive robust principal component analysis," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3794-3800, 2012.