2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

Compression of Large-Scale Image Dataset using Principal Component Analysis and K-means Clustering

Rushrukh Rayan, Md. Sabir Hossain, and Asaduzzaman

Department of Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET), Chittagong-4349, Bangladesh
Email: rushrukhryan@gmail.com, sabir.cse@cuet.ac.bd, asad@cuet.ac.bd

***Abstract*— Digital images, being on the verge of its utmost popularity encompasses plenty of applications and as such are generated at an unprecedented rate. These digital form of data are often found with redundant information. Applications that require a bulk amount of images to be processed, turn out to be high regarding computational complexity. Needless to say, it leads to inefficient storage utilization. In this paper, a hybrid approach is applied to compress a large-scale image data-set by combining two popular algorithms: Principal Component Analysis (PCA) and K-means. This paper works with a view to diminishing the redundant information by implementing dimensionality reduction followed by color quantization. The PCA is used to project the data onto a lower dimensional space with retaining as maximum variance as possible. The K-means algorithm is used to restrict the distinct number of colors to represent an image by means of clustering the data together. The results obtained from the proposed method is compared with the results obtained from implementing PCA and K-means clustering algorithms independently, where the proposed method provides with a better compression ratio.**

Keywords—Dimensionality reduction; color quantization; principal component analysis; k-means clustering; unsupervised machine learning.

I. Introduction

In this modern era of growing technological development, the digital image has become a significant appliance in numerous applications such as hotspot detection in photovoltaic modules, breast cancer cell analysis, face recognition, satellite image analysis [1-4]. Often these images contain redundant features and additional distinct colors, due to which, processing and transmission yield expensive cost. In applications where a large-scale dataset is required to train a model as opposed to a single image, reducing the feature redundancy and a number of distinct colors can lead to a great deal of cost minimization of further processing along with reduced file size.

Image compression has two general classifications, namely lossless and lossy compression. In lossy compression, the resultant image permanently loses some of its data due to diminishing correlated feature and color quantization. Hence the original image cannot be retrieved [5]. Although it loses some data in the process, it is possible to retain the variance in the dataset which allows performing further processing for various applications without concerning valuable data loss. Hence, a lossy compression method can provide with a redundancy-divested data-set along with better storage efficiency.

Dimensionality reduction is basically projecting the data on to a lower dimensional space. For instance, the sample input images are first to read converting it into a data matrix where, if the number of input samples is n and each sample is of size p x q, the data matrix dimension would be n x (p x q). Each row and column represent a simple input and a feature respectively. Sometimes, the number of features can be very large, whereas a lot of them may be correlated and considered as redundant. The higher the number of features, the harder it is to obtain a visual perception from the training set and to work on it. Essentially, the data is projected onto a lower dimensional space m, where m is certainly less than (p x q).

Principal Component Analysis (PCA) is one of the most used statistical techniques, which works with the perception obtained from a large quantity of numerical data depending on the correlation amongst the variables [6]. PCA is performed with two objectives, first is dimensionality reduction for the purpose of compression of the data. Second is to have a visual perception of the data which can further provide information that originally was hidden to human eyes. We choose the number of dimensions onto which to project the data according to the optimal needs. If originally the data-set incorporates n-dimensions, PCA will reduce data to k-dimensions, with k less than n being classified.

Color quantization is a process that produces an output image with a view to preserving the resemblance as close as possible to the input image, achieved by diminishing the number of colors used to represent an image.

K-means, an efficient clustering method groups the input data together using centroids, each representing a cluster. Data are grouped together based on the concept of minimizing the squared error function [7]. It basically attempts to split a given unlabeled dataset into a fixed number of clusters incorporating two steps: Cluster assignment and Centroid movement step.

In [8], a method for compressing a single image has been devised which incorporates partitioning of the training set into k clusters and application of PCA to each of them. This approach does not allow a large-scale image dataset to be processed and does not include color quantization.

In this contribution, we incorporated PCA for dimensionality

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

reduction, along with K-means clustering to obtain color quantization, so as to compress a large-scale image dataset.

The rest of the paper is organized as follows. Section II provides the proposed methodology including the schematic diagram, PCA, K-means clustering. Section III discusses the experimental results of the proposed method. Finally, Section V concludes this paper.

II. Proposed Methodology

This section introduces our proposed methodology for compressing a large-scale image dataset. It can be divided into three basic steps as (1) pre-processing dataset with mean normalization and feature scaling, (2) dimensionality reduction using PCA, (3) color quantization using K-means clustering. Fig. 1 shows the schematic diagram of the proposed method.

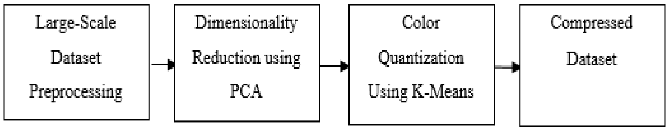
Large-Scale Dataset Preprocessing
Dimensionality Reduction using PCA
Color Quantization Using K-Means
Compressed Dataset

Fig. 1. Schematic Diagram of Proposed Method.

A. Pre-processing

Before applying Principal Component Analysis, it is ideal to have performed mean normalization on the dataset beforehand (1). As the prime motive of PCA is to retain the maximum variance in the dataset, having to mean normalization performed on the dataset leads to an efficient output; if not, it may wrongfully assume the number of components that explains all the variance in the data, thereby, perform badly.

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \qquad (1)$$

Replace each $x_j^{(i)}$ with x_j - μ_j

Where, μ_j: mean of the feature, m: number of elements in the training set, $x^{(i)}$: i^{th} element in the training set.

B. Dimensionality Reduction using PCA

As an image is being taken as an input matrix, initially, the covariance matrix is computed. A covariance matrix is a square matrix that attempts to describe the variance and covariance among the data. Each member in the covariance matrix falls under one of the three categories, such as 1) positive, if two variables under observation change in the same direction, 2) negative, if they change in the opposite direction, 3) non-existent, if one of the member changes while the other remains stationary. Computation of Covariance Matrix using (2).

$$\mathbf{\Sigma} = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T \qquad (2)$$

Eigen-decomposition of the covariance matrix is performed, where the matrix is decomposed into a set of eigenvectors and eigenvalues i.e. the diagonalization of a matrix along its eigenvectors. Eigenvectors are axes of principal force that a matrix moves the inputs along. An n x n matrix may have n eigenvectors, each representing its line of action in one dimension. Computation of Eigenvectors of matrix Σ using (3).

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = svd(\Sigma) \qquad (3)$$

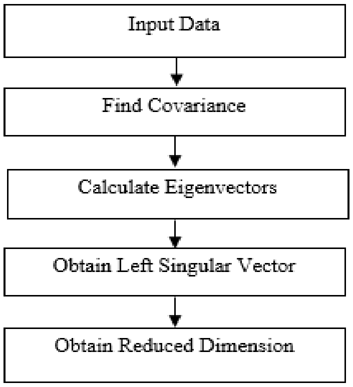

Fig. 2. Flowchart of Principal Component Analysis algorithm

where U is a unitary matrix, S is a rectangular diagonal matrix with non-negative real numbers on the diagonal and V is a unitary matrix. From the U matrix, first k vectors are selected onto which to project the data in order to reduce dimensionality. Hence, the projection of data onto a lower dimensional space is obtained. Fig. 3 shows the dimensionality reduction of an input image from the dataset.

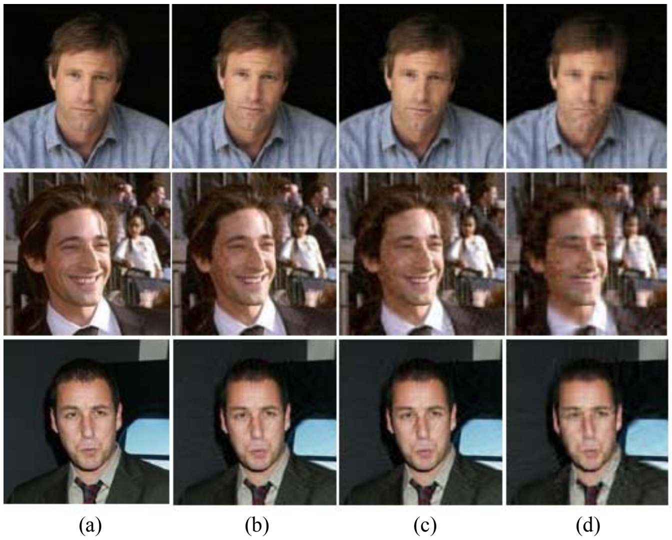

Fig. 3. Processing example of dimensionality reduction: a) input image, b) number of dimension = 35, c) number of dimension = 25 and d) number of dimension = 15

C. Color Quantization using K-means clustering

The output obtained from the dimensionality reduction using PCA step is taken as input here. The K-means clustering splits the unlabeled dataset into a fixed number of clusters. It basically incorporates two steps, namely 1) cluster assignment step and 2) move centroid step. The inputs for the K-means clustering algorithm would be (1) K: the number of clusters such that $\mu_1, \mu_2, \ldots, \mu_K \in \mathbf{R}$ and data points $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$.

After initializing clusters arbitrarily by selecting random data points as centroids, all the data points that are neighboring to a centroid are grouped together while minimizing the squared Euclidean distance as much as possible. Formally, if c_i refers to each of the centroids in a set of centroids C, then splitting the

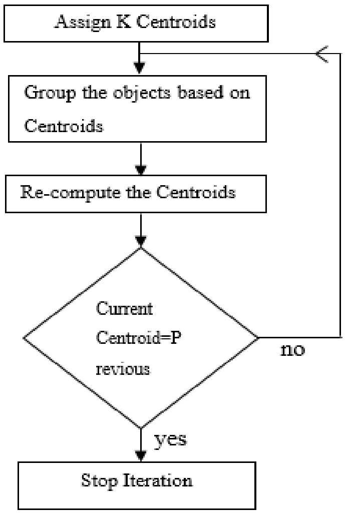
Assign K Centroids
Group the objects based on Centroids
Re-compute the Centroids
Current Centroid=P revious
no
yes
Stop Iteration

Fig. 4. Flowchart of K-means clustering algorithm

data points to the clusters is performed by (4).

$$\min_{c_i \in C} distance\ (c_i, x)^2 \tag{4}$$

where, *the distance* refers to the standard Euclidean distance. Following the cluster assignment, in the move centroid step, the cluster centroid will be moved into the next location which depends on the mean distance (5) from the data to the cluster centroid.

$$\mu_k = average\ of\ points\ assigned\ to\ cluster\ k \tag{5}$$

These two steps are repeatedly performed as long as data points are reallocated to a new cluster. The algorithm converges when no data points are reallocated to a new cluster. Fig. 5 shows the color quantization output of an input image on which dimensionality reduction has been performed with the number of dimension = 25.

Since there is no general method to obtain the right number of clusters, an empirical estimation can be retrieved from applying a method called Elbow Method. The K-means algorithm is to be run for a different number of clusters and by plotting the average within-cluster distance to centroid against the different number of clusters, the point where the rate of decrease abruptly changes can be exploited to determine the right number of K.

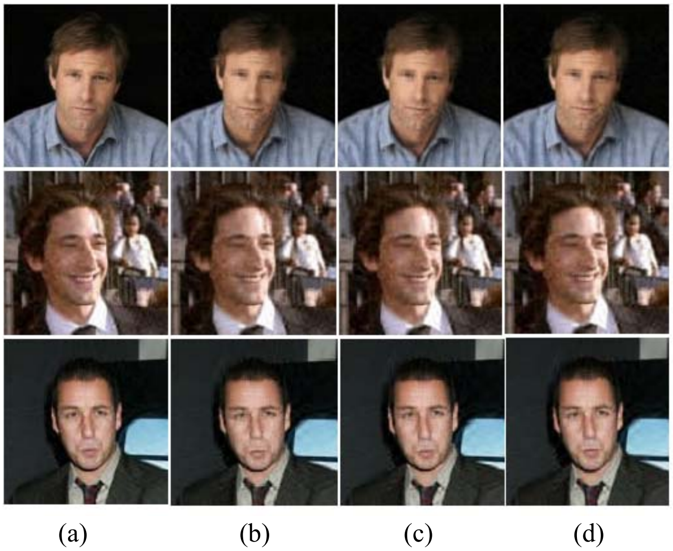

Fig. 5. Processing example of color quantization: a) input image, b) number of cluster = 128, c) number of cluster = 512 and d) number of cluster = 1024

III. Experimental Results and Discussion

The comparative analysis of the proposed method, K-means and PCA has been presented in this section assessing the benefit of the proposed method. A total of 100 images were taken into account to obtain the relevant result. To provide better insights into the proposed method, the compression ratio obtained from applying K-means with the different number of clusters and PCA with the different number of dimensions are analyzed. The comparison of compression ratio, Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR) expressed in dB among the proposed method, K-means and PCA are being presented in the graphical representation to illustrate the effectiveness of the proposed method.

In Fig. 6, the compression ratio of the proposed method against the number of clusters of K-means clustering is represented. For a different number of clusters, the proposed method provides with an increase in the compression ratio as opposed to that obtained from applying K-means clustering independently. As the number of clusters increases, the compression ratio first increases. But after a substantial amount of increase, the ratio faces a recess in the increase. Clustering of data beyond optimal level is responsible for the recess. The proposed method has been evaluated with the number of dimensions being held at 25 and 15 where the compression ratio is higher at 15.

In Fig. 7, the compression ratio of the proposed method against the dimensions of PCA is represented. For a different number of dimensions, the proposed method also provides with an increase in the compression ratio as opposed to that obtained from applying PCA independently. As the number of dimensions decreases, so does the compression ratio due to projecting the data onto a lower dimensional space. The proposed method has been evaluated with the number of clusters being held at 64 and 1024. It doesn’t encounter the sudden recess in the compression ratio, rather increases sharply with the decrease in the number of dimensions. By reducing the number of dimensions, we increasingly limit the space onto which to project the data, thereby do not encounter the recess.

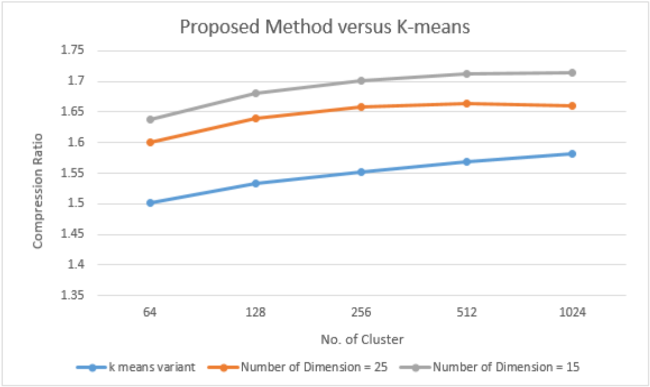

Fig. 6. Number of clusters versus compression ratio

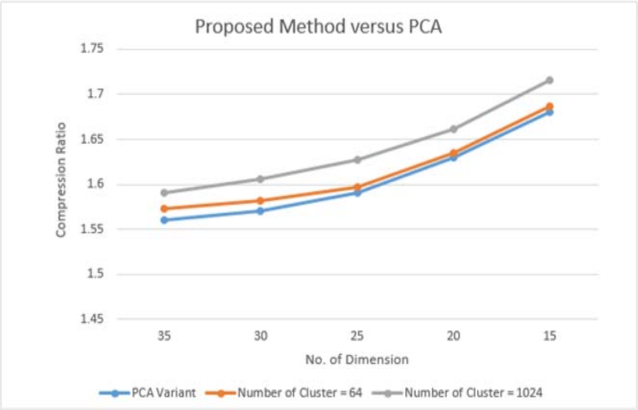

Fig. 7. Number of dimensions versus compression ratio

In Fig. 8, a comparative analysis among K-means, PCA and the proposed method has been represented. Here, the average compression ratio for a set of hundred images has been represented for K-means, PCA and the proposed method each.

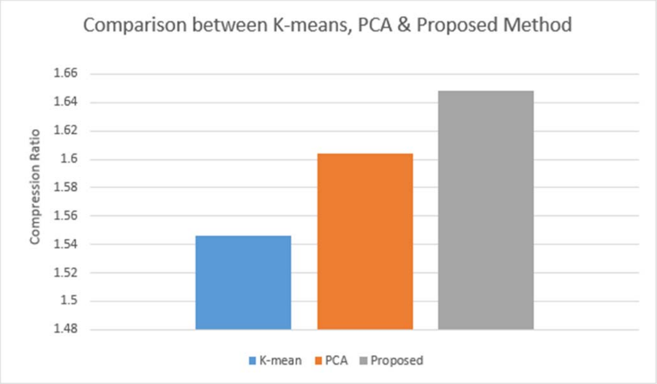

Fig. 8. Comparison between K-means, PCA and proposed method

The proposed method achieves 6.8% and 3.1% increased compression ratio than that of K-means and PCA algorithms respectively.

As this paper works with lossy image compression, the MSE calculation for each of the methods is represented in Fig. 9. The mean squared error is the cumulative squared error between the compressed and the original image. The mathematical formulae for MSE is in (6).

$$MSE = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}[\, I(x,y) - I'(x,y)]^2 \qquad (6)$$

where $I(x,y)$ is the original image, $I'(x,y)$ is the compressed image and $M,\ N$ implies that the image is of size M x N. The increase is MSE indicates increased reconstruction error in the output image. As more compression ratio is being obtained, the compressed image permanently loses some of its data, as for which, the MSE value increases with the increase in compression ratio denoting the loss in image data as well as the decrease in file size. The average MSE is calculated for the outputs obtained from K-means, PCA and the proposed method for the same set of hundred input images.

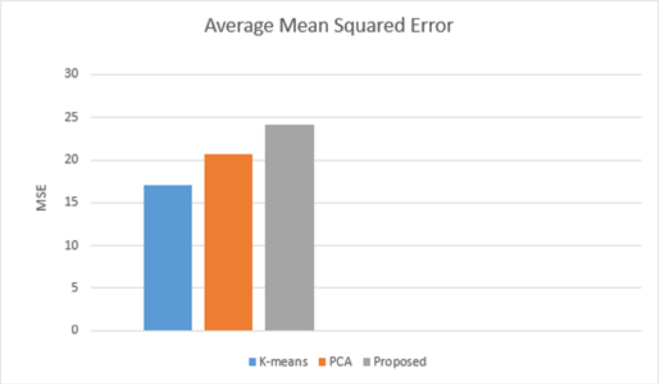

Fig. 9. Comparison of MSE of K-means, PCA and proposed method

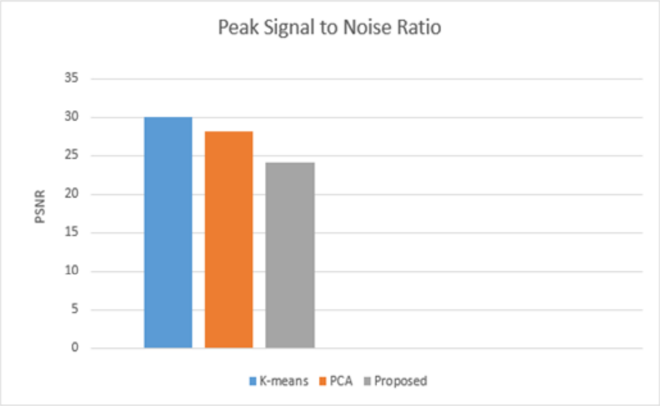

Fig. 10. Comparison of PSNR of K-means, PCA and proposed method

The PSNR provides insights into the peak error. As expected, it has an inverse relationship with MSE, a decrease in PSNR results in an output which has higher noise than that of the input. While working with lossy image compression, the output permanently loses some of its data which leads to higher noise. It intuitively indicates the increase in compression ratio [9]. The mathematical formula for PSNR is:

$$PSNR = 10\log_{10}(\frac{255^2}{MSE}) \qquad (7)$$

In Fig. 10, the comparison of K-means, PCA and the proposed method has been depicted based on PSNR value expressed in dB. As more compression ratio is obtained, the MSE increases. Hence, PSNR decreases denoting the decrease in file size.

IV. CONCLUSIONS

In this paper, an unsupervised method is developed to compress a large-scale image dataset by using PCA and K-means Clustering. It provides a better compression ratio, thereby, better storage efficiency along with a redundancy-divested dataset to be able to be implemented for further processing reducing the computational complexity.

The results have been compared to the results obtained from applying K-means clustering and PCA individually which shows that the proposed method provides with a better compression ratio. And to have a detailed perception on how the compression ratio changes along with the change in the number of clusters and number of dimensions, analysis has been performed on a various number of clusters and dimensions.

The computation time increases substantially as the number of clusters in K-means clustering increases. The clusters are being initialized randomly in this paper. A heuristic approach to initialize the cluster can decrease the cost of computation.

This method can be helpful for preprocessing a dataset for further processing in regard to human facial recognition, object detection, character and handwriting recognition.

References

[1] G. C. Ngo and E. Q. B. Macabebe, "Image Segmentation Using K-Means Color Quantization and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for Hotspot Detection in Photovoltaic Modules," In *IEEE Region 10 Conference (TENCON) — Proceedings of the International Conference*, 2016.

[2] N. Sharma and K. Saroha, "A Novel Dimensionality Reduction Method for Cancer Dataset using PCA and Feature Ranking," In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015.

[3] T. N. Palghamol and S. P. Metkar, "Constant Dimensionality Reduction for Large Databases using Localized PCA with an application to Face Recognition," In *Proceedings of the IEEE Second International Conference on Image Information Processing (ICIIP-2013),* 2013.

[4] T. Celik, "Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering", *IEEE Geoscience and Remote Sensing Letters,* Vol. 6, No. 4*,* October 2009.

[5] A. M. Rufai, G. Anbarjafari and H. Demirel, "Lossy Image Compression Using Singular Value Decomposition and Wavelet Difference Reduction," In *Digital Signal Processing,* Vol 24*,* 2014.

[6] H. Abdi and L. J. Williams, "Principal Component Analysis," In *John Wiley & Sons, Inc.,* Vol. 2, July/August 2010.

[7] C. Ding and X. He, "K-means Clustering via Principal Component Analysis," In *Proceedings of the SIAM International Conference on Data Mining*, 2004.

[8] C. W. Wang and J. H. Jeng, "Image Compression Using PCA with Clustering," In *IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2012),* November 4-7, 2012.

[9] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," in Electronics Letters, vol. 44, no. 13, pp. 800-801, 19 June 2008.