International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 3– No.7, August 2012 – www.ijais.org

Eigen Value based K-means Clustering for Image Compression

K. Somasundaram
Gandhigram Rural Institute-Deemed University
Gandhigram

M. Mary Shanthi Rani
Gandhigram Rural Institute-Deemed University
Gandhigram

ABSTRACT

In this paper, a new method has been proposed to enhance the performance of K-means clustering using the significance of Eigen values in spectral decomposition. Experimental results with standard images show that the proposed method shows faster convergence and reduced bit rate than standard K-means without compromise in the quality of the reconstructed images measured in terms of Peak Signal to Noise Ratio(PSNR).

KEYWORDS

Codebook, Covariance, Spectral Decomposition, Eigen value

1. INTRODUCTION

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data and it is an unsupervised learning technique. Clustering algorithms partition the given dataset into subsets (clusters) based on some distance measure, such that one attribute within a cluster is more similar to each other than that of other clusters [1]-[3]. Generally, clustering techniques can be classified as hierarchical and partitioning based on the method adopted in the formation of clusters. Partitioning algorithms determine all clusters at once whereas hierarchical algorithms build clusters hierarchically. The outcome of partitioning methods is commonly a set of K clusters and every cluster is characterized by a centroid which represents the summary description of all the objects in that cluster.

K-means [3] is one of the most common partitioning techniques which has large number of applications in the fields of image and video compression [4],[5], image segmentation[6], pattern recognition[7] and data mining [8]. It was also graded as one of the top ten algorithms in data mining [9]. It is an iterative method which generates a codebook from the training data using a distortion measure appropriate for the given application [3], [10].

It is very simple, effective and easy to implement. The main factor that determines the performance of K-means is its convergence time which mostly depends on the amount of training data, codebook size, code vector dimension, and distortion measure. The two major steps in K-means clustering process are cluster initialization and classification based on computational and convergence criteria.

The algorithm starts by choosing K initial seeds or centroids from the training data, either at random or based on some heuristic measure. Next, it constructs a new partition by assigning each point to its closest initial centroid based on distance measure and the centroid of each set is recalculated. The algorithm is repeated by alternate application of these two steps until convergence is reached where the points no longer switch clusters or the overall squared error is less than the convergence threshold . The overall quality of clustering is the average distance from each data point to its associated cluster centroid. Typically, the K-means algorithm uses squared euclidean distance measure to determine the distance between an object and its cluster centroid.

Though the K-means algorithm always converge, it does not guarantee to yield the most optimal clustering as it is significantly sensitive to the randomly selected initial cluster centroids. A better approach to minimize this problem is to make multiple runs of the algorithm with different K initial seed centroids and choose the best one for a given problem. The two main crucial issues in a K-means clustering model are finding the optimal number of clusters (K) to create and the initial centroid of each cluster. The first issue is application specific as it depends on the problem domain whereas the second issue plays a significant role in the performance of K-means algorithm . Usually, the input data points are scattered and don't fall exactly into easily recognizable groups which consequently increases the time for convergence. This poses a serious concern which initiated the development of robust strategies for fast convergence of K-means.

A modified K-means algorithm (KMOD) was proposed by Lee et al. [11] which converges faster than the conventional K-means by using a scaled updating scheme along the direction of the local gradient by a step-size larger than that used by the centroid update of the conventional K-means algorithm. While the use of a scaled-update can accelerate the convergence, use of a "fixed" scaling for the entire range of iterations results in the use of larger step sizes even at iterations closer to convergence. Consequently, this increases the number of iterations required for convergence and undesirably high perturbations of the codevectors as well, to a poorer local optimum. Kuldip and Ramasubramanian (NEW) [12] proposed the use of a variable scale factor which is a function of the iteration number. It offers faster convergence than the KMOD algorithm with a fixed scale factor, without affecting the optimality of the codebook.

However, despite the efficiency of K-means, bad choice of initial seeds may yield poor results in addition to increase in computation time for convergence. This has stimulated the researchers to focus their attention on the initialization

step and many novel initialization methods have been proposed.

The earliest method of initializing the K means algorithm was done by Forgy in 1965 [13] where the seed points are chosen randomly. A pioneering work on seed initialization was proposed by Ball and Hall (BH) [14]. A similar approach named as Simple Cluster Seeking (SCS) was proposed by Tou and Gonzales [15]. The SCS method chooses the first input vector as the first seed and the rest of the seeds are selected, provided they are 'd' distance apart from all selected seeds. The SCS and BH methods are sensitive to the parameter d and the order of the inputs.

A novel initialization method was proposed by Bradley and Fayyad (BF) [16] for large datasets. The core idea of their algorithm is to select 'm' subsamples from the dataset, apply the K-means on each subsample independently, keep the final N centers from each subsample and produce a set that contains mN points. A new approach for optimizing K-means clustering in aspects of accuracy and computation time has been proposed by Ali and Kiyoki [17]. This algorithm designates the initial centroids' positions in the farthest accumulated distance between the data points in the vector space analogous to choosing pillars' locations as far as possible from each other within the pressure distribution of the roof structures for a stable building. Data points with maximum accumulated distance among the data distribution are chosen as the initial centroids.. Yanfeng Zhang *et al* [18] proposed an Agglomerative Fuzzy K-means clustering method for learning optimal number of clusters that provides significant clustering results. High density areas are detected based on which the initial cluster centers with a neighbour sharing selection approach is determined. Huang and Harris [19] proposed the Direct Search Binary Splitting (DSBS) method. This method is similar to the Binary Splitting algorithm except that the splitting step is enhanced through the use of Principle Component Analysis (PCA).

In this paper, an innovative approach based on spectral decomposition has been proposed to construct the initial codebook to achieve fast convergence, eventually reducing the computational time of K-means clustering technique for compressing images. The rest of the paper is organized as follows: Section 2 briefly describes the proposed method, Section 3 presents the results and performance analysis of the proposed method and Section 4 concludes our work.

2. PROPOSED METHOD

The efficiency of any clustering algorithm depends upon the knowledge of the affinity between the points in euclidean space. Clustering techniques employ spectral decomposition as a method for determining the block structure of the affinity matrix. Spectral techniques which make use of information obtained from the eigenvectors and eigenvalues of a matrix, have attracted the researchers' attention with respect to clustering recently. The significance of eigen value has found a great number of applications in science and engineering such as solution to linear and non-linear differential equations, boundary value problem, markov chain, network analysis and population growth model etc. Principal Component Analysis (PCA) also known as (Karhunen-Loeve or Hotelling transform), is a linear transform which is one of the eigen value based spectral techniques[20]-[21]. This method serves as a powerful tool for data analysis and pattern recognition in the fields of signal and image processing .

Several research work on the application of eigen values have reported that the eigen vector with the highest eigen value formed out of the covariance matrix of the given image is the principal component of an image[20]. Eigen vectors with eigen values from the highest to lowest represent the principal components of an image in decreasing order of significance. This property has been well exploited for dimension reduction in image data thereby yielding compression. In the proposed method, we make use of the eigen value for initialization process in K-means Clustering.

The proposed method divides the given image into 4x4 pixel blocks and the eigen value is calculated for every image block. The method proceeds by partitioning the input image blocks into high and low eigen valued blocks. This classification is done based on a threshold which is the average eigen value of all the image blocks. Blocks with high eigen values contain all the principal components of the image whereas blocks with low eigen values consist of less detailed image components which can be subjected to high compression.

Next, clustering is performed on each partition with K1 and K2 blocks as initial seeds where K1 and K2 are the number of blocks having high eigen values in their respective partitions. The codebooks resulting from the high eigen and low eigen partitions are merged to form the global initial codebook. Finally, a last run of K-means with global codebook as the initial codebook is done to construct the final global codebook.

The following steps describe the algorithm of the proposed method.

1. Divide the input image into 4x4 blocks.
2. Form the covariance matrix of each block.
3. Find the eigen value of each block.
4. Compute the average eigen value of all the blocks and set it as the threshold value T.
5. Classify the input image blocks with eigen value greater than the threshold T as blocks with high eigen value (BHE) and remaining as blocks with low eigen value (BLE).
6. Convert each image block into 16-element vector.
7. Set K1 = N1/ δ where N1 is the number of vectors in the BHE and "δ" is the rate of compression.
8. Form K1 initial seeds for BHE.
9. Set K2 = N2/ (2 * δ) where N2 is the number of vectors in the BLE.

10. Form K2 initial seeds for BLE.

11. Perform K-means clustering with the corresponding initial seeds in BHE and BLE and generate codebooks CBH and CBL.

12. Merge the two codebooks CBH and CBL to form a global codebook of size K (=K1+K2)

13. Perform a final run of K-means on image data with the global codebook as the initial codebook to construct the final global codebook**.**

4. RESULTS AND DISCUSSION

To evaluate the performance of the proposed method, experiments are conducted using the proposed method on several test images of size 256×256 pixels with initial block size of 4×4 on a machine with core2 duo processor at 2 GHZ using MATLAB 6.5.
The quality of the reconstructed image is measured in terms of Peak Signal to Noise Ratio (PSNR) defined by

$$PSNR = 20 \log_{10} \left(\frac{255}{\sqrt{MSE}} \right) \quad (1)$$

where, *MSE* is the mean-square error measuring the deviation of the reconstructed image from the original image. Compression is measured in bits per pixel (BPP). The computed values for bpp and PSNR for test images Lena, Boats and Barbara using the proposed method and K-means are given in Table 1. From Table 1, we observe that the proposed method accomplishes enhanced bit rate and fast convergence at comparable picture quality than K-means with random initialization of seeds. The execution speed of the proposed method is almost double of that of standard K-means.

The parameter 'δ' is used to tune the rate of compression depending on the type of application. Compression rate increases with increase in the value of 'δ'. Therefore, a high value of 'δ' achieves high compression ratio. Furthermore, the codebook size K1 of BHE is set to be greater than K2 of BLE as the high eigen partition represents the principal components of the given image which attribute to better quality of the reconstructed image. It has been found out from our experiments that good rate-distortion performance is achieved if the value of 'δ' is set to 32. Thus the proposed method works adaptively by setting different bit rates for high and low eigen valued blocks thereby achieving higher compression rate without compromising the picture quality. Figures 1-3 show the reconstructed images of Lena, Barbara and Boats using K-means with random initialization and the proposed methods for visual comparison.

Table 1. Performance Comparison of the Proposed Method

Test Image	Method	Execution time in sec.	Bit rate(bpp)	PSNR
Lena	Proposed Method	22.36	0.19	29.01
	K-means	65	0.25	29.16
Boats	Proposed Method	26.87	0.21	27.48
	K-means	135	0.25	27.90
Barbara	Proposed Method	35.1	0.20	25.03
	K-means	77.5	0.25	25.29

(a) (b) (c)
Figure 1: a) Original Lena Image; Reconstructed images using b) K-means c) Proposed method

(a) (b) (c)
Figure 2: a) Original Barbara Image; Reconstructed image using b) K-means c) Proposed method

(a) (b) (c)
Figure 3: a) Original Boats Image; Reconstructed image using b) K-means c) Proposed method

5. CONCLUSIONS

In this paper, we have proposed a new approach to generate codebook used in Vector Quantization for image compression. The proposed method subdivides the given image into sub-image blocks. Eigen values for these blocks are computed and a principal component analysis (PCA) is done. This PCA process divides the blocks into blocks of high and low eigen valued blocks. Codebooks of these two categories are merged to form the global codebook. The proposed method shows better performance in terms of bit rate and significant decrease in computation time than the standard K-means method at comparable picture quality.

6. REFERENCES

[1] Ankerst, M., M. Breunig, H.P. Kriegel and J. Sander, "OPTICS: Ordering points to identify the clustering structure", Proceedings of ACM SIGMOD International Conference on Management of Data Mining, ACM Press, Philadelphia, Pennsylvania, United States, pp. 49-60, June 1999.

[2] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is 'nearest neighbor' meaningful?", Proceedings of the International Conference on Database Theory, Jerusalem, pp. 217–235, January 1999.

[3] Linde Y., Buzo A., and Gray R.M., "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communication, Vol. 28, pp. 84–95, 1980.

[4] N. Venkateswaran, and Y. V. Ramana Rao, "K-Means Clustering Based Image Compression in Wavelet Domain", Information Technology Journal , Vol.6, pp. 148–153, 2007.

[5] Gersho A. , and Gray R.M. "Vector Quantization and Signal compression", Kluwer Academic Publishers, New York, pp. 761, 1992.

[6] H.P. Ng, S.H. Ong, K.W.C. Foong, P.S. Goh, and W.L. Nowinski, "Medical image segmentation using k-means clustering and improved watershed algorithm", IEEE Southwest Symposium on Image Analysis and Interpretation, Denver, pp.61-65, 2006.

[7] Duda, R.O. and P.E. Hart, Pattern Classification and Scene Analysis. John Wiley Sons, New York, pp. 482, 1973.

[8] Jiang, D.J. Pei and A. Zhang, "An interactive approach to mining gene expression data", IEEE Transactions on Knowledge and Data Engineering, Vol.17, pp. 1363-1380, 2005.

[9] XindongWu , Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang ,Hiroshi Motoda ,Geoffrey J. McLachlan ,Angus Ng , Bing Liu ,Philip S. Yu , Zhi-Hua Zhou , Michael Steinbach ,David J. Hand , Dan Steinberg, "Top 10 algorithms in data mining", Knowledge and Information Systems Journal, Vol.14, pp. 1-37, 2008.

[10] MacQueen, J.B., Some Method for Classification and Analysis of Multivariate Observations, Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, (MSP'67), Berkeley, University of California Press, pp: 281-297, 1967.

[11] D. Lee, S. Baek, and K. Sung, "Modified k-means algorithm for vector quantizer design", IEEE Signal Processing Letters, Vol. 4, pp. 2–4, 1997.

[12] Kuldip K. Paliwal and V. Ramasubramanian, Comments on "Modified K-means Algorithm for Vector Quantizer Design", IEEE Transactions on Image Processing, Vol. 9 , No. 11, pp.1964-1967, 2000.

[13] E. Forgy, "Cluster analysis of multivariate data: efficiency vs interpretability of classification", Biometrics, Vol. 21, pp.768-769, 1965.

[14] Ball G.H. and Hall D.J., "PROMENADE-an Online Pattern Recognition System", Stanford Research Institute Memo, Stanford University, 1967.

[15] Tou, J. and R. Gonzales, Pattern Recognition Principles, Addision-Wesley, Reading, MA., pp: 377, 1977

[16] Bradley, P.S. and U.M. Fayyad, " Refining initial points for K-means clustering", Proceedings of the 15th International Conference on Machine Learning (ICML'98), ACM Press, Morgan Kaufmann, San Francisco, pp. 91-99, 1998.

[17] Ali Ridho Barakbah and Yasushi Kiyoki, "A New Approach for Image Segmentation using Pillar-Kmeans Algorithm ", World Academy of Science, Engineering and Technology Journal , Vol.59, pp.23-28, 2009.

[18] Yanfeng Zhang, Xiaofei Xu and Yunming Ye, "An Agglomerative Fuzzy K-means clustering method with automatic selection of cluster number" ,2nd International Conference on Advanced Computer Control (ICACC), Vol. 2, pp. 32-38, 2010

[19] C. Huang and R. Harris, "A comparison of several codebook generation approaches", IEEE Transactions on Image Processing, Vol. 2 (1), pp. 108–112, 1993.

[20] H. H. Barret. *Foundations of Image Science*. John Wiley & Sons, New Jersey, U.K., third edition, 2004.

[21] R. C. Gonzales and R. E. Woods. *Digital Image Processing*. Prentice Hall, second edition, 2002.