Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 3, June 2020, pp. 1183~1188
ISSN: 2302-9285, DOI: 10.11591/eei.v9i3.1985

Quality and size assessment of quantized images using K-Means++ clustering

Davin Ongkadinata, Farica Perdana Putri
Department of Informatics, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia

Article Info

Article history:

Received: Jul 29, 2019
Revised Sep 30, 2019
Accepted Jan 16, 2020

Keywords:

Color quantization
Image compression
K-Means++
Machine learning
True color image

ABSTRACT

In this paper, an amended K-Means algorithm called K-Means++ is implemented for color quantization. K-Means++ is an improvement to the K-Means algorithm in order to surmount the random selection of the initial centroids. The main advantage of K-Means++ is the centroids chosen are distributed over the data such that it reduces the sum of squared errors (SSE). K-Means++ algorithm is used to analyze the color distribution of an image and create the color palette for transforming to a better quantized image compared to the standard K-Means algorithm. The tests were conducted on several popular true color images with different numbers of K value: 32, 64, 128, and 256. The results show that K-Means++ clustering algorithm yields higher PSNR values and lower file size compared to K-Means algorithm; 2.58% and 1.05%. It is envisaged that this clustering algorithm will benefit in many applications such as document clustering, market segmentation, image compression and image segmentation because it produces accurate and stable results.

Corresponding Author:

Davin Ongkadinata,
Department of Informatics,
Faculty of Engineering and Informatics,
Universitas Multimedia Nusantara,
Scientia Boulevard, Gading Serpong, Tangerang, 15227, Indonesia.
Email: davin.ongkadinata@gmail.com

1. INTRODUCTION

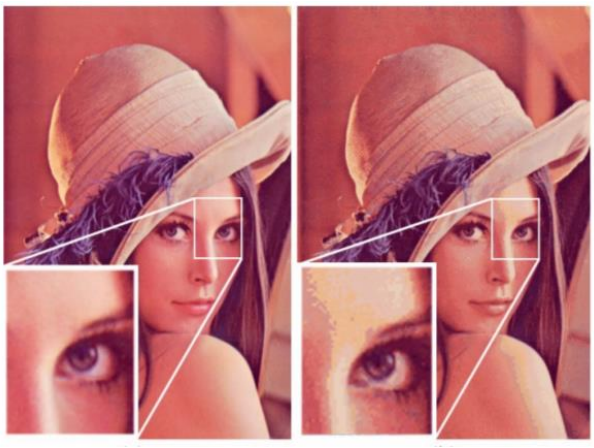
The development of information and communication technologies favors an increasing need for image and video data, which require large storage space and high transmission bandwidth. Image processing technology has undergone significant growth. Therefore, to save up existing storage devices and transmission channels, methods which are able to compress data are widely investigated [1, 2]. A true-color image consists of 24-bit of storage, where 8 bits for red, 8 for green, and 8 for blue [3]. This kind of image has a big number of pixel data which can represent up to 16,777,216 colors and makes its display, processing, transmission, and storage problematic. As a result, color quantization is used to solve for many image processing and graphics problems. In the past, color quantization was needed due to the limitations of graphics capabilities of display hardware. Color quantization still maintains its practical value even though 24-bit display hardware has become common [4-6].

Color quantization is a preprocessing technique that is used to reduce the number of colors in images such that the reconstructed image should be visually close to the original image. It is able to eliminate unnecessary information from images. For instance, unnecessary color information in topographic maps needs to be eliminated so as to accurately construct digital evaluation model (DEM). Color quantization plays a critical role in many other digital applications such as segmentation,

color texture analysis, content-based retrieval, watermarking, text detection, and non-photorealistic rendering, and. In general, this technique is done by performing two steps. The first is to select the palette design from the colors in the original image. The second step is pixel-mapping, it is done by changing each color with the color in the palette. Color quantization is an implementation of lossy image compression [7-10]. Figure 1 shows the representation between a 24-bits original image and 64 colors of quantized image.

Currently, there are two algorithms for palette design, clustering-based and splitting-based. Some well-known splitting algorithms are the median-cut [11], center-cut [12], Octree [13], variance-based [14], RWM-cut [15], and binary splitting [16]. The splitting-based algorithms generally split the color space of the color image into two disjoint groups according to their splitting criteria. Then, the splitting is iterated until the wanted number of groups is achieved. Finally, the cluster center of each group is chosen as the palette color [17]. Clustering-based algorithms include Hierarchical Clustering, K-Means, and K-Medoids [18]. The performance of clustering-algorithm-based techniques varies greatly depending on how the K representative colors are chosen. These techniques have to make a tradeoff between their computational efficiency and minimization of the distortion measure. For example, the standard K-Means algorithm can minimize the quantization error efficiently if enough number of iterations are allowed [19, 20]. K-Means clustering algorithm [21] was originally proposed for pattern recognition. It is commonly recognized as a sub-optimal quantization technique that can also be applied to color vision, image segmentation, and vector quantization. For image vector quantization, the generalized Lloyd algorithm, also called the LBG algorithm, is identical to the standard K-means algorithm [22].

K-Means++ was proposed by David Arthur and Sergei Vassilvitski in 2007. It is an improvement to the standard K-Means in order to surmount the random selection of the initial cluster centers. This algorithm uses the squared distance weighting method to select the next center. This algorithm also reduces the fluctuation happening in K-Means and provide better and stable clustering results [23]. Authors in [24] found that the K-Means++ algorithm has proved to be more accurate than the standard K-Means in crime document clustering. Therefore, we investigate the PSNR and file size of quantized images using K-Means++ compared to the standard K-Means.



(a) (b)

Figure 1. The different representation between, (a) An original image in 24-bits RGB color, (b) Quantized image is reduced to a palette of 64 colors

The rest of the paper is organized as follows; Section 2 describes the research method, including the K-means++ algorithm, and PSNR as the quality assessment method. Section 3 presents the results and analysis. Finally, Section 4 explains the conclusions and future work.

2. RESEARCH METHOD

In this experiment, the image is converted to a collection of RGB color values as *X*. Identify the number of clusters (K value) that is desired to represent the image. The numbers of clusters proposed are 32, 64, 128, and 256. The number of clusters must be less than the actual number of colors that exist in

the image. The next step is to choose K colors to represent the original image. The chosen colors would act as the centroids to classify all the colors in the original image. Then proceed the following K-Means++ Clustering algorithm [25]:

- Select a center c_1, which is chosen uniformly at random from X.
- Select a new center c_i, by choosing $x \in X$ with probability $\frac{{D_{(x)}}^2}{\sum_{x \in X} {D_{(x)}}^2}$.

  $D(x)$ denotes the shortest distance from a data point to the closest center we have already chosen.

$$ {D_{(x)}}^2 = (x_i - c_{i1})^2 + (x_i - c_{i2})^2 + \ldots + (x_i - c_{ir})^2 \quad (1) $$

- Repeat Step 2, k centers have been taken altogether.
- For each $i \in \{1,\ldots,k\}$, set the cluster C_i to be the set of points in X that are closer to c_i than they are to c_j for all $j \neq i$.
- For each $i \in \{1,\ldots,k\}$, set c_i to be the center of mass of all points in C_i: $c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$.
- Repeat Step 4 and 5, until C no longer changes.

After the algorithm is done, each color is looped through and replaced with the center color that has the closest distance to it. This step is known as color remapping in color quantization. Then, the remapping process would reproduce an image that is visually similar to the original image, but only with 'K' colors exist in it.

Color quantization algorithms efficiency is measured by PSNR (Peak-to-Signal Noise Ratio), where a higher PSNR means the quantized image is closer to the original visually [26].

$$ PSNR = 10\, log_{10}\left(\frac{M^2}{MSE}\right) \quad (2) $$

where M is the original image's maximum value. The typical value of PSNR for lossy compressed images are between 30-50 dB [27]. MSE measures the distortion or deviation between the original image and its reconstructed image (quantized image) [28].

$$ MSE = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} \left(S_{xy} - C_{xy}\right)^2 \quad (3) $$

3. RESULTS AND DISCUSSION

To evaluate the PSNR value and file size of K-Means++ clustering algorithm, 4 (four) popular images from USC-SIPI University Image Database (http://sipi.usc.edu/database/); "Baboon", "Fruits", "Lena", and "Pepper" are used. The images are 512x512 pixels sized. The original image file sizes are as follows 624 KB, 464 KB, 464 KB, and 528 KB. Figure 2 shows the tested images.

(a) (b) (c) (d)

Figure 2. The test images, (a) Baboon, (b) Fruits, (c) Lena, (d) Peppers

Experiments are implemented using Java programming language on an i7 7700HQ 2.8GHz processor, 8GB RAM, and NVIDIA GeForce GTX 1050Ti 4GB VGA. K-Means++ is compared to the standard K-Means clustering algorithm. The tests were done 3 times for each 'K' value. Table 1 depicts

the average PSNR values and file sizes of the tested images. Figure 3 shows the quantization results of "Fruits". Figure 4 shows the comparison between K value and PSNR, while Figure 5 shows the comparison between K value and file size.

Table 1. The average PSNR and file sizes

No.	Image	Number of colors	Size (KB)	K	K-Means++		K-Means	
					PSNR (dB)	Size (KB)	PSNR (dB)	Size (KB)
1	Baboon	230427	622	32	27.07	129	26.31	130.67
				64	28.95	169.67	28.39	170
				128	30.89	215	29.93	218
				256	32.67	273.33	30.80	275.33
2	Fruits	49451	461	32	30.25	61.67	29.90	62
				64	32.85	82	32.53	83.33
				128	35.25	110.33	34.14	110.67
				256	37.06	137.67	36.86	138.67
3	Lena	148279	462	32	32.08	93.67	31.00	95
				64	33.81	129.67	32.26	130
				128	34.88	161	32.68	163.33
				256	35.08	175.67	35.05	176
4	Peppers	183525	526	32	28.56	82.67	28.15	84
				64	30.95	117.67	30.06	120
				128	33.40	159.33	32.34	163.33
				256	34.90	207	34.87	208.67

(a) (b) (c) (d)

Figure 3. Quantized images of "Fruits", (a) K=32, (b) K=64, (c) K=128, (d) K=256

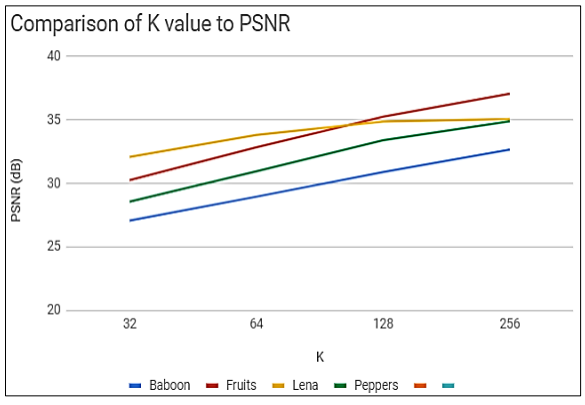


Figure 4. The comparison between K value and PSNR

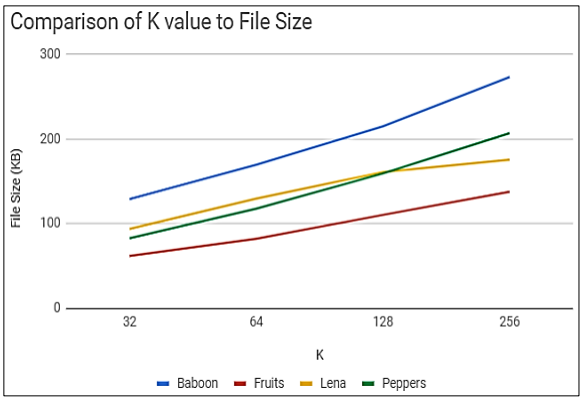


Figure 5. The comparison between K value and file size

Based on the results depicted in Table 1, the quantized images altogether have the average PSNR more than 30 dB on K value of 128 and 256. The smallest average file size produced is 61.67 KB on Fruits image with K value of 32 and the biggest is 273.33 KB on Baboon image with K value of 256. The average file sizes have a scale of 2 to 7 times smaller than the original size. The PSNR and file sizes of K-Means++

clustering are better than K-means clustering with PSNR and file size percentage decreases for 32, 64, 128, and 256 clusters are 2.19% and 1.2%, 2.59% and 1%, 3.93% and 1.39%, 1.61% and 0.61%. The standard K-Means algorithm initializes the cluster centers uniformly at random, whereas K-Means++ tends to initialize near the center of the squares. Therefore, the results given by the standard K-Means are different in every run. In this case, the colors chosen as the centers could have adjacent RGB values. So, the PSNR and file sizes obtained by the standard K-Means are worse than the K-Means++. Based on Figure 4 and 5, more clusters (K) produce a better PSNR and file size, while the file sizes are still much lower than the original size. If the number of clusters gets bigger, then the mass of pixels that have similarities with the cluster centers produce quantized image that is closer to the original.

4. CONCLUSION

By using K-Means++ algorithm, the quantized images altogether have the average PSNR more than 30 dB on K value of 128 and 256. The average file sizes have a scale of 2 to 7 times smaller compared to the original size. The PSNR and file size increase as the K value gets bigger. In addition, the average PSNR and file size produced of each quantized image achieves better results compared to the original K-Means algorithm; 2.58% and 1.05%. K-Means++ algorithm has been proved to be better and more accurse that K-Means, in color quantization. The reason is because the fact that K-Means algorithm initializes the cluster centers uniformly at random, whereas K-Means++ tends to initialize near the center of the squares. Based on the research conducted, different distance metrics can be used to make a representation of the images so that it can be compared in terms of weighting. It is also recommended that the extension of this work can be done for images with bigger resolutions.

ACKNOWLEDGEMENTS

We would like to thank Universitas Multimedia Nusantara for providing us an opportunity to publish this research paper.

REFERENCES

[1] G. Ramella and G. S. D. Baja, "A New Method for Color Quantization," *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Naples, 2016, pp. 1-6.
[2] P. Pangestu, D. Gunawan, and S. Hansun, "Histogram equalization implementation in the preprocessing phase on optical character recognition," *Int. J. Technol.*, vol. 8, no. 5, pp. 947–956, 2017.
[3] Z. Ye, H. Mohamadian and H. Majlesein, "Adaptive Enhancement of Gray Level and True Color Images with Quantitative Measurement Using Entropy and Relative Entropy," *2008 40th Southeastern Symposium on System Theory (SSST)*, New Orleans, LA, 2008, pp. 127-131.
[4] S. Erwin and M. I. Prasetiyowati, "Identification of object shapes on two dimensional digital image using normalized cross correlation," in *CONMEDIA*, 2015.
[5] H. J. Park, K. B. Kim, and E. Y. Cha, "An effective color quantization method using color importance-based self-organizing maps," *Neural Netw. World*, vol. 25, no. 2, pp. 121–137, 2015.
[6] M. E. Celebi, "An effective color quantization method based on the competitive learning paradigm," *International Conference on Image Processing, Computer Vision, and Pattern Recognition*, vol. 2, pp. 876–880, 2009.
[7] C. H. Lee, H. Y. Lu, and J. H. Horng, "Color quantization by hierarchical octa-partition in RGB color space," *IEEE International Conference on Applied System Invention (ICASI)*, pp. 147–150, 2018.
[8] D. Barman, A. Hasnat, S. Sarkar, and M. A. Rahaman, "Color image quantization using Gaussian particle swarm optimization(CIQ-GPSO)," *International Conference on Inventive Computation Technologies*, pp. 1-4, 2016.
[9] R. Samet and E. Hancer, "A new approach to the reconstruction of contour lines extracted from topographic maps," *J. Vis. Commun. Image Represent.*, vol. 23, no. 4, pp. 642–647, 2012.
[10] C. Ozturk, E. Hancer, and D. Karaboga, "Color image quantization: A short review and an application with artificial bee colony algorithm," *Inform.*, vol. 25, no. 3, pp. 485–503, 2014.
[11] K. L. Shelley and D. P. Greenberg, "Path specification and path coherence," *9th Annual Conference on Computer Graphics and Interactive Techniques*, vol. 16, no. 3, pp. 157–166, 1982.
[12] G. Joy and Z. Xiang, "Center cut for color image quantization," *Vis. Comput.*, vol. 10, no. 1, pp. 62–66, 1993.
[13] M. Gervautz and W. Purgathofer, "A simple method for color quantization: Octree quantization," *New Trends in Computer Graphics*, pp. 219-231, 1988.
[14] S. J. Wan, P. Prusinkiewicz, and S. K. M. Wong, "Variance-based color image quantization for frame buffer display," *Color Res. Appl.*, vol. 15, no. 1, pp. 52–58, 1990.
[15] C. Y. Yang and J. C. Lin, "RWM-cut for color image quantization," *Comput. Graph.*, vol. 20, no. 4, pp. 577–588, 1996.
[16] M. T. Orchard and C. A. Bouman, "Color quantization of images," *IEEE Trans. SIGNAL Process.*, vol. 39, no. 12, pp. 2677–2690, 1991.

[17] M.-G. Lee, "K-means-based color palette design scheme with the use of stable flags," *J. Electron. Imaging*, vol. 16, no. 3, pp. 1-11, 2007.
[18] K. M. Archana Patel and P. Thakral, "The best clustering algorithms in data mining," *International Conference on Communication and Signal Processing*, no. 2, pp. 2042–2046, 2016.
[19] O. Verevka, "The local K-means algorithm for colour image quantization," *Graphics Interface*, pp. 128–135, 1995.
[20] M. Frąckiewicz and H. Palus, "Clustering with K-Harmonic means applied to colour image quantization," *8th IEEE International Symposium on Signal Processing and Information Technology*, pp. 52–57, 2008.
[21] J. T. Tou and R. C. González, Pattern recognition principles," *Addison-Wesley Pub. Co*., 1974.
[22] A. Buzo, R. M. Ray, and Y. Linde, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, 1980.
[23] Z. Min and D. Kai-Fei, "Improved research to K-means initial cluster centers," *9th International Conference on Frontier of Computer Science and Technology*, pp. 349–353, 2015.
[24] B. Aubaidan, M. Mohd, and M. Albared, "Comparative study of k-means and k-means++ clustering algorithms on crime domain*," Jour. of Comput. Sci.*, vol. 10, no. 7, pp. 1197–1206, 2014
25] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," *18th Annu. ACM-SIAM Symp. Discret. Algorithms*, pp. 1027–1035, 2007.
[26] K. H. Thung and P. Raveendran, "A survey of image quality measures," *International Conference for Technical Postgraduates*, pp. 1-4, 2009.
[27] M. A. Hassan, "Joint color quantization and dithering techniques," *Commun. Appl. Electron.*, vol. 3, no. 7, pp. 1–8, 2015.
[28] S. Ilic, M. Petrovic, B. Jaksic, P. Spalevic, L. Lazic, and M. Milosevic, "Experimental analysis of picture quality after compression by different methods," *Prz. Elektrotechniczny*, vol. 89, no. 11, pp. 190–194, 2013.

BIOGRAPHIES OF AUTHORS

Davin Ongkadinata is a graduate student of Technology Management program at Universitas Multimedia Nusantara. He received his bachelor's degree in Computer Science from the same university in 2019. He currently works as a frontend engineer in a local startup.

Farica Perdana Putri received her bachelor's degree in Computer Science from Universitas Multimedia Nusantara. She then continued her study in Computer Science at National Taipei University. She has been a lecturer and researcher at Universitas Multimedia Nusantara since 2017. She is also a member of APTIKOM and actively publishes many publications. Her research interests mainly focus in semantic analysis on natural language processing, digital image processing, and machine learning where she has successfully been granted some research grants from the government and Universitas Multimedia Nusantara.