A Clustering-Based Approach to Image Compression Using the K-Means Algorithm

Hongxia Mao

School of Computer and Software, Jincheng College, Sichuan University, Chengdu 611731, China

Abstract: ***In the current Internet era, image information is extensively utilized across various sectors of society. Due to the substantial data volume associated with digital images, compression techniques are essential for efficient transmission and storage. The K-Means algorithm, one of the most widely adopted clustering methods in unsupervised learning, offers a viable approach to this challenge. This study applies the K-Means clustering algorithm to image compression, with implementation carried out via Python programming. Experimental results demonstrate that the K-Means clustering method effectively reduces image data volume while preserving essential visual information, thereby confirming its utility as a practical technique for image compression.***

Keywords: K-Means; Clustering; Image compression.

1. INTRODUCTION

Cluster analysis is an important research field in data mining. It groups data objects into several classes or clusters, making objects in the same cluster more similar, while objects in different clusters differ greatly. Image information is widely used in various industries, and the larger the amount of information contained in an image, the larger the space it occupies. Therefore, when transmitting and saving images, appropriate methods should be used for compression. Image compression refers to the process of reducing the amount of data required to represent an image by eliminating redundant data under certain quality conditions, making it easier to store and transmit the image.

This article applies the clustering algorithm to image compression applications, using the K-Means clustering algorithm to compress images.

In the domain of natural language processing, Huang et al. [1] introduced an innovative approach for distilling tool knowledge into language models through back-translated traces, enhancing their functional capabilities [1]. For transportation safety, Abughaida et al. [2] developed an intelligent blind spot indicator system specifically designed to prevent double lane merge conflicts, addressing a critical vehicle safety concern [2]. In cybersecurity and data protection, Deng [3] proposed homomorphic encryption-based mechanisms for ensuring data integrity and anti-tampering in cloud storage environments [3], while Deng and Yang [11] extended this security research by developing multi-layer defense strategies against membership inference attacks within federated learning frameworks [11]. Mehta et al. [5] contributed to the security landscape by proposing a comprehensive national AI security framework aimed at protecting critical financial infrastructure [5]. Within computer vision, Jin et al. [4] advanced object detection and pose estimation methodologies through the integration of hybrid task cascade networks with high-resolution networks, achieving superior performance in complex visual recognition tasks [4]. In business analytics and marketing, Zhou [6] applied gradient boosting trees to diagnose bottlenecks in international automotive sales funnels, providing insights into cross-regional team efficiency [6], while Wensi [7] explored AI-assisted marketing content generation tailored for non-standard industrial automation solutions [7]. Yi [8] addressed advertising optimization challenges by developing real-time fair-exposure ad allocation mechanisms using contextual bandits-with-knapsacks, specifically targeting small businesses and underserved creators [8]. In photonics engineering, Tang et al. [9] designed and optimized shallow-angle grating couplers for achieving vertical emission from indium phosphide devices [9]. For legal text processing, Xie et al. [10] advanced citation text classification through a Conv1D-based approach for multi-class classification tasks [10]. Lin et al. [12] made foundational theoretical contributions by developing computational methods for the Poisson multinomial distribution with applications spanning ecological inference and machine learning [12]. Zhang [13] developed an adaptive explainable AI framework designed to transform black-box models into actionable insights for proactive tax risk mitigation in small and medium enterprises [13]. In resource management, Zhang [14] applied cohesive hierarchical clustering techniques to address dynamic adaptation challenges in power emergency materials supply-demand systems [14]. Guo [15] utilized LSTM networks for real-time data completion in IMU-based motion recognition applications [15]. In materials science and catalysis, Li and Li [16] investigated spatial

compartmentalisation effects enabling multifunctionality catalysis, from dual sites to cascade reactions [16]. For recommendation systems, Yang et al. [17] and Junxi et al. [18] developed graph convolutional networks based on matrix factorization (GCN-MF) to enhance recommendation accuracy [17,18]. Finally, Zhou and Cen [19] examined the transformative effect of ChatGPT-like new generation AI technologies on user entrepreneurial activities, revealing important implications for innovation ecosystems [19].

2. K MEANS CLUSTERING ALGORITHM

Clustering is a very important application field in data mining. Clustering refers to dividing samples with high similarity into the same cluster based on the principle of similarity, and dividing samples with high dissimilarity into different clusters. The K-Means algorithm is one of the most commonly used clustering algorithms. The K-Means clustering algorithm, also known as the k-means algorithm, uses distance as a measure of similarity between samples. The smaller the distance between samples, the higher their similarity, and they may be in the same cluster. This algorithm receives parameter k and then divides the sample points into k clusters; The similarity of samples within the same cluster is relatively high; The similarity of samples in different clusters is low. The idea of this algorithm is to cluster k sample points in space and classify the sample points closest to them. Through iterative methods, gradually update each cluster center until the best clustering effect is achieved.

The steps of the algorithm are as follows:

(1) Randomly select k samples from n sample datasets as initial clustering centers;

(2) Calculate the distance between each sample in the dataset and k cluster centers separately, and divide the samples into the class corresponding to the cluster center with the smallest distance;

(3) For each category, recalculate its cluster center;

(4) Repeat steps 2 and 3 until the position of the cluster center no longer changes.

Overall, the clustering idea of K-means algorithm is relatively simple and easy to implement, and the clustering effect is acceptable. It is a simple, efficient and widely used clustering method.

3. APPLICATION OF 3K MEANS ALGORITHM IN COMPRESSED IMAGES

3.1 Principle of K-Means algorithm for compressing images

When an image is displayed on a computer screen, it occupies the computer's memory space. The calculation formula for the memory occupied by an image is: image height "image width" memory size occupied by one pixel. The number of bytes contained in each pixel directly affects the size of memory occupied by the image. Storing images with different color modes requires different memory sizes, as shown in Table 1:

Table 1: Small comparison of memory usage per pixel for different image types

Image type	How many bytes per pixel	The number of colors that can be represented
1-bit data graph (Line art)	1/8 byte per pixel	$2^1 = 2$
8-bit Grayscale	1 byte per pixel	2^8
16 bit Grayscale	2 bytes per pixel	2^{16}
24 bit RGB	3 bytes per pixel, which is the most commonly used format in images, such as TIF format.	2^{24}
32-bit printing color mode (CMYK)	4 bytes per pixel	2^{32}
48 bit RGB	6 bytes per pixel	2^{48}

The most basic principle of image compression is to replace some similar colors with one color, reducing the number of color descriptions, so that each corresponding pixel can be described with fewer bytes, and the memory occupied by the image will be reduced. The basic principle of using K-Means clustering algorithm for image

compression is to classify each pixel in an image, and replace the values of pixels classified into the same category with the values of the cluster centers of that category, in order to reduce the memory space occupied by the image.

3.2 Experimental Process

When importing images into a program, preprocessing of the image data is required. According to the resolution of the image, the data points in the image are tiled, and each pixel is treated as a 3D sample. The pixel values of an image are converted into n rows and 3 columns of data, where n=height * width.

Using KMeans clustering algorithm, set the value of k, cluster all color values in the image, and find the cluster center corresponding to each 3D pixel point.

```
#The number of colors contained in the compressed image is the number of clusters
k = 64
kmeans = KMeans(n_clusters=k,random_ state=0)

#Training model
kmeans.fit(pixel_sample)

#Find the cluster center corresponding to each 3D pixel point
cluster_assignments= kmeans.predict(pixel_ sample)
```

Traverse every pixel in the image, find the pixel value of the cluster center corresponding to each pixel value, and replace the value of the pixel with the pixel value of the cluster center.

```
#Traverse each pixel point and find the pixel value corresponding to the cluster center
pixel_count= 0 foriinrange(height):
forjinrange(width):

#Obtain the index of the clustering center of pixels
cluster_idx = cluster_assignments[pixel_ count]

#Obtain the pixel values at the index position of the cluster center
cluster_value = cluster_centers[cluster_ idx]

#Replace the value of a pixel point
compressed_img[i][j]=  cluster_value pixel_count+= 1
```

Run the code, Figure 1 shows the image before compression, and Figure 2 shows the compressed image when using the K-Means algorithm with k=64.

Figure 1: Effect before compression

Figure 2: Compressed effect

3.3 Experimental Results

From the running results, it can be seen that the original image is 92KB (bird. TIF), while the compressed image is only 45K (compressible bird), achieving the goal of image compression.

compressed_dog	2019/8/9 2:05	JPG 图片文件	45 KB
dog	2017/10/23 11:11	JPG 图片文件	92 KB

4. CONCLUSION

From the experimental results, it can be seen that the K-Means clustering method can indeed compress images. The smaller the value of k, the greater the compression ratio, but the less color the compressed image has, resulting in a loss of a large amount of pixel color information and a greater difference from the original image. The

K-Means algorithm is simple and easy to implement, but it requires users to specify the number of clusters (value of k) in advance, and the clustering results are sensitive to the selection of initial cluster centers, which can easily lead to local optima. In practice, in order to obtain better results, the K-Means algorithm is usually run multiple times with different initial cluster centers. After all sample assignments are completed, when recalculating the centers of k clusters, for continuous data, the cluster centers should take the mean of that cluster.

REFERENCES

[1] Huang, Xingyue, et al. "Distilling Tool Knowledge into Language Models via Back-Translated Traces." arXiv preprint arXiv:2506.19171 (2025).

[2] Abughaida, A., Daman, L., Song, M., Zhang, Y., & Ayoub, J. (2024). An Intelligent Blind Spot Indicator System to Prevent Double Lane Merge Conflicts (No. TRBAM-24-02945).

[3] Deng, X. (2025). Homomorphic Encryption-Based Data Integrity Verification and Anti-Tampering Mechanism in Cloud Storage Environment.

[4] Jin, Y., Zhang, Y., Xu, Z., Zhang, W., & Xu, J. (2024, November). Advanced object detection and pose estimation with hybrid task cascade and high-resolution networks. In 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML) (pp. 1293-1297). IEEE.

[5] Mehta, R., Patwar, N., Wei, X., Saunders, E., Zhu, X., & Liu, J. (2026). Towards a National AI Security Framework for Financial Infrastructure Protection. International Journal of Advance in Applied Science Research, 5(2), 39–50. Retrieved from https://h-tsp.com/index.php/ijaasr/article/view/251

[6] Zhou, Z. (2026). Bottleneck Diagnosis in International Automotive Sales Funnels Using Gradient Boosting Trees: Evidence from Cross-Regional Team Efficiency Evaluation. Journal of Computer Technology and Applied Mathematics, 3(1), 11-18.

[7] Wensi, L. (2026). AI-Assisted Marketing Content Generation for Non-Standard Industrial Automation Solutions. Journal of Economic Theory and Business Management, 3(1), 18-25.

[8] Yi, X. (2025, October). Real-Time Fair-Exposure Ad Allocation for SMBs and Underserved Creators via Contextual Bandits-with-Knapsacks. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 1602-1607).

[9] Tang, Y., Kojima, K., Gotoda, M., Nishikawa, S., Hayashi, S., Koike-Akino, T., ... & Klamkin, J. (2020). Design and Optimization of Shallow-Angle Grating Coupler for Vertical Emission from Indium Phosphide Devices.

[10] Xie, Y., Li, Z., Yin, Y., Wei, Z., Xu, G., & Luo, Y. (2024). Advancing legal citation text classification A Conv1D-based approach for multi-class classification. Journal of Theory and Practice of Engineering Science, 4(02), 15-22.

[11] Deng, X., & Yang, J. (2025, August). Multi-Layer Defense Strategies and Privacy Preserving Enhancements for Membership Reasoning Attacks in a Federated Learning Framework. In 2025 5th International Conference on Computer Science and Blockchain (CCSB) (pp. 278-282). IEEE.

[12] Lin, Z., Wang, Y., & Hong, Y. (2023). The computing of the Poisson multinomial distribution and applications in ecological inference and machine learning. Computational Statistics, 38(4), 1851-1877.

[13] Zhang, T. (2025). From Black Box to Actionable Insights: An Adaptive Explainable AI Framework for Proactive Tax Risk Mitigation in Small and Medium Enterprises.

[14] Zhang, X. (2024). Research on Dynamic Adaptation of Supply and Demand of Power Emergency Materials based on Cohesive Hierarchical Clustering. Innovation & Technology Advances, 2(2), 59–75. https://doi.org/10.61187/ita.v2i2.135

[15] Guo, Y. (2025, May). IMUs Based Real-Time Data Completion for Motion Recognition With LSTM. In Forum on Research and Innovation Management (Vol. 3, No. 6).

[16] Li, F., & Li, H. (2024). Spatial compartmentalisation effects for multifunctionality catalysis: From dual sites to cascade reactions. Innovation & Technology Advances, 2(1), 1–13. https://doi.org/10.61187/ita.v2i1.54

[17] Junxi, Y., Wang, Z., & Chen, C. (2024). GCN-MF: A graph convolutional network based on matrix factorization for recommendation. Innovation & Technology Advances, 2(1), 14–26. https://doi.org/10.61187/ita.v2i1.30

[18] Zhou, J., & Cen, W. (2024). Investigating the Effect of ChatGPT-like New Generation AI Technology on User Entrepreneurial Activities. Innovation & Technology Advances, 2(2), 1–20. https://doi.org/10.61187/ita.v2i2.124