[[1]](#footnote-1)

Unsupervised Learning Techniques for Image Clustering and Compression(January 2024)

Amr Zakaria, Omar Hisham, Youssef Hany, and Saed Torbany, IEEE

*Abstract*—This paper investigates the use of KMeans clustering and Principal Component Analysis (PCA) in image compression. The purpose of the study is to find out how well these techniques work to reduce the size of image files without sacrificing visual authenticity. The notebook compares several models to find the one that provides the best trade-off between compression rate and image quality. The study sheds light on the effectiveness of PCA and KMeans clustering approaches in the context of image processing and data reduction by concentrating on the smallest image size that can be achieved with each model. The results of this study may have a big impact on fields like digital media, online content distribution, and remote sensing that need effective image transmission and storage.

# INTRODUCTION

I

n the era of digital technology, effective image transmission and storage are critical. In order to lower the bandwidth and storage needs of digital photographs, image compression techniques are essential. Principal Component Analysis (PCA) and KMeans clustering are used in this study to present a novel method of image reduction. The capacity to convert data into a new coordinate system that emphasizes their most important qualities makes PCA a well-known technique in data processing and dimensionality reduction. On the other hand, pattern identification and data clustering commonly employ the KMeans clustering technique. The study attempts to assess how well these two approaches work together to compress photographs to the smallest size without sacrificing significant quality. Through a series of experiments and comparative analyses, the notebook documents the process of applying these techniques to image data, assessing their performance, and determining the best model based on the accuracy and compression ratio achieved. The results are expected to contribute to the ongoing development of more efficient and effective image compression methods.

# Data Acquisition

For this research project, we collected our own dataset of Egyptian images from the internet to infuse a personal and cultural touch into our research. The dataset consists of two key images that capture the essence of Egypt.

## Camel

Egyptian Camel Image: To represent the cultural heritage of Egypt and its iconic deserts, we manually searched the web for an image of an Egyptian camel. The selected image is available at the following URL: [Egyptian Camel Image](https://media-cdn.tripadvisor.com/media/attractions-splice-spp-720x480/07/92/67/cb.jpg).

## Cairo

Cairo, Egypt Image: As Cairo is one of Egypt's vibrant and bustling cities, we sought to include an image that showcases the urban landscape and colors of the capital. After careful manual selection, we obtained an image of Cairo, Egypt, which is accessible at this URL: [Cairo, Egypt Image](https://alittlenomad.com/wp-content/uploads/2019/10/panorama-of-cairo.jpg).

## Collection Methods:

The images were gathered by manually searching the internet for photographs that encapsulate the heart of Egypt. Special attention was given to selecting images that were rich in color and represented different aspects of Egypt's culture and landscape.

# Exploratory Data Analysis

In this phase of our study, we conducted a thorough exploration and preprocessing of the Egyptian image dataset, focusing on achieving optimal conditions for applying PCA and KMeans clustering techniques. The key steps and challenges encountered in this process are outlined below:

## Data Cleaning and Preprocessing:

The dataset did not require extensive data cleaning since it was manually curated and consisted of high-quality images. Standardizing Image Shapes: We resized the images to a uniform dimension of 300x300 pixels, maintaining the aspect ratio. This resizing was essential for uniformity across the dataset and to reduce computational complexity for analysis.

## The Visualization and Shape Analysis:

We extensively visualized the images at various stages of preprocessing. Initially, we displayed the original images to understand their composition and color distribution. The images were then resized, and their reshaped versions were visualized. This step was crucial to confirm that the resizing process preserved the essential features of the images.

We further visualized the reconstructed images from the PCA components to assess the impact of dimensionality reduction on the image quality. The visualization also included plotting the original and reduced color spaces using 3D scatter plots. This was key to understanding how the color distribution was altered after applying PCA and KMeans clustering.

## Challenges Encountered and Solutions:

Dimensionality Reduction: One significant challenge was balancing the number of principal components in PCA. This balance was crucial for maintaining a good compression rate without significant loss of image quality. We conducted multiple trials to find the optimal number of components. Computational Constraints: Given the computations required for PCA and KMeans, we optimized our code for efficiency. Working with resized images initially helped in managing these computational demands.

# Model Selection

In our research, we have focused on the utilization of unsupervised learning algorithms, specifically Principal Component Analysis (PCA) and K-Means clustering, for the task of image clustering and compression. The justification for choosing these models, as well as an overview of their suitability for our task, is based on the detailed analysis and experiments conducted in our notebook.

## Justification for Using Unsupervised Learning Algorithms:

The primary goal of our project is to compress images by reducing their complexity while maintaining as much of the original information as possible. Unsupervised learning is ideal for this purpose as it does not require labeled data. Instead, it identifies patterns and structures within the data itself.

Image compression inherently involves finding patterns and reducing redundancy in the image data, which aligns well with the capabilities of unsupervised learning methods.

## Principal Component Analysis (PCA):

PCA is a statistical technique used for dimensionality reduction. It transforms the data into a new coordinate system, reducing the number of dimensions without significant loss of information.

In our notebook, PCA was employed to reduce the dimensionality of the image data. This reduction is crucial in compressing the image by retaining only the most significant features (principal components) of the images, thereby reducing the file size while preserving the essential visual information.

The effectiveness of PCA in our project was demonstrated through the reconstruction of images from the retained principal components, which showed a good balance between compression and quality retention.

Image compression inherently involves finding patterns and reducing redundancy in the image data, which aligns well with the capabilities of unsupervised learning methods.

## K-Means Clustering:

K-Means is a popular clustering algorithm that partitions data into K distinct clusters based on feature similarity. In the context of image processing, it's used for color quantization – reducing the number of colors in an image.

In the notebook, we applied K-Means clustering to group similar pixels together, significantly reducing the number of distinct colors in each image. This step is pivotal in the compression process as it leads to a smaller file size by limiting the color palette used in the image.

The application of K-Means was complemented with visualizations of the color spaces before and after clustering, highlighting the algorithm's efficiency in reducing color variety while maintaining the overall appearance of the image.

# Model Training

In our research, we have implemented and trained Principal Component Analysis (PCA) and K-Means clustering models for image compression. This section outlines the specific steps taken in the training process, including hyperparameter tuning and customization, along with the technical details pertinent to PCA and K-Means as demonstrated in our notebook.

## Principal Component Analysis (PCA) Implementation and Training:

Initialization: PCA was initialized with parameters aimed at maximizing variance retention while reducing dimensionality. The number of components was a key hyperparameter, which we tuned based on the variance ratio to balance between image detail retention and compression rate.

Training Process: The training involved fitting the PCA model to the reshaped image data. Each image, converted into a 2D array where rows represented pixels and columns the color channels, was subjected to PCA to identify principal components.

Reconstruction: Post PCA application, we reconstructed images from the reduced feature set to visualize and assess the quality of the compressed image. This step was crucial to evaluate the performance of PCA in retaining essential information.

## K-Means Clustering Implementation and Training:

Initialization: K-Means was initialized with the number of clusters as the primary hyperparameter. The selection of the number of clusters (colors) was critical and was done based on the Elbow Method, which helped in determining the optimal balance between compression and quality.

Training Process: The K-Means model was trained on the pixel data of the images. The aim was to cluster pixels into groups representing the most prominent colors. This process effectively reduced the color palette of the image to the specified number of clusters.

Cluster Centers for Color Quantization: After training, the cluster centers determined by K-Means were used to create a new, compressed image. Each pixel in the original image was replaced with the nearest cluster center, thereby reducing the overall color variety in the image.

## Technical Details and Customizations

PCA Technicalities: The PCA process involved handling high-dimensional data and transforming it to a lower-dimensional space. The computation of eigenvalues and eigenvectors was key to this transformation, and we ensured that the computational load was manageable by resizing images prior to PCA application.

K-Means Technicalities: The K-Means algorithm's efficiency depended heavily on the initial placement of centroids. To address this, we experimented with different initialization strategies, including random initialization and k-means++ for better centroid placement.

Hyperparameter Tuning: For both PCA and K-Means, hyperparameter tuning was an iterative process. We continuously adjusted parameters like the number of components in PCA and the number of clusters in K-Means, assessing the impact on image compression and quality.

# Evaluation

In the evaluation phase of our research, we thoroughly assessed the performance of the unsupervised learning models, namely Principal Component Analysis (PCA) and K-Means clustering, used for image clustering and compression. The evaluation involved both quantitative and qualitative analysis, focusing on metrics such as PCA variance, file size reduction (in MB and KB), the number of colors used, and the effectiveness of color space transformations.

## PCA Variance Retention:

The PCA model's performance was primarily evaluated based on the variance it retained from the original images. This metric was crucial as it indicated how much of the original information was preserved after dimensionality reduction.

We quantitatively measured the variance ratio for the selected principal components. A higher variance ratio indicated better information retention, which is desirable for maintaining image quality during compression.

## File Size Reduction:

A key metric for assessing the effectiveness of our image compression approach was the reduction in file size, measured in both megabytes (MB) and kilobytes (KB).

We compared the sizes of the original, PCA-reduced, and K-Means processed images. The reduction in size was quantified to demonstrate the efficiency of our models in compressing the images. Significant reductions in file size, while maintaining image integrity, were indicative of successful compression.

## Number of Colors and Color Space Analysis:

For K-Means clustering, the reduction in the number of colors used in the compressed images was a vital metric. The fewer colors used, the more significant the compression, as long as the visual quality was not overly compromised.

We also performed an analysis of the color spaces before and after applying our models. This qualitative assessment involved visualizing the color distributions and observing how the color spaces were altered by PCA and K-Means. The effectiveness of color space transformations in enhancing compression efficiency while retaining visual quality was a key aspect of our evaluation.

## D. Qualitative Analysis:

The PCA model's performance was primarily evaluated based on the variance it retained from the original images. This metric was crucial as it indicated how much of the original information was preserved after dimensionality reduction. Apart from the quantitative metrics, we also conducted a qualitative analysis. This involved visually inspecting the compressed images to assess any loss of quality or significant changes in color accuracy.

The visual comparisons between the original and compressed images provided insights into the perceptual impact of our compression techniques.

# Interpretation

The evaluation of our image compression project using Principal Component Analysis (PCA) and K-Means clustering has yielded insightful findings. Here, we interpret these results, discussing the broader implications of image clustering and compression for various applications and analyzing the strengths and limitations of the techniques employed.

## Interpreting Findings from the Evaluation:

The high variance retention by PCA indicates that the technique was successful in reducing the dimensionality of the image data while preserving significant information. This is crucial in maintaining the quality of the compressed images.

The reduction in file size (in both MB and KB) demonstrates the effectiveness of our approach in compressing images, making them more manageable for storage and transmission.

The reduction in the number of colors, as achieved by K-Means clustering, contributed significantly to file size reduction. This was done without drastically altering the perceptual quality of the images, as evidenced by our qualitative analysis.

The results show a significant reduction in image size when using both PCA and KMeans for compression. The original image size is 33.55 MB, which is substantially reduced to 2.06 MB with PCA, achieving a reduction ratio of 0.06. This indicates that the PCA-compressed image is only 6% of the size of the original image, demonstrating a high level of compression.

For KMeans compression, the size reduction is also notable, especially considering the number of colors used. With only 2 colors, the image size is reduced to 0.09 MB, which is even smaller than the PCA-compressed size. This trend continues as the number of colors increases (8, 64, 256), with the image size remaining around 0.09 MB. Remarkably, the reduction ratio compared to PCA is consistently 0.04 across all color levels, indicating that KMeans is more efficient than PCA in terms of size reduction for this particular image and settings.

## Implications for Various Applications:

Digital Media and Web Applications: In areas where bandwidth and storage are limited, such as online content delivery, our approach could facilitate faster loading times and lower data usage.

Remote Sensing and Satellite Imagery: The techniques could be invaluable for compressing satellite images, where transmitting large amounts of data is often costly and impractical.

Medical Imaging: In medical fields, where the integrity of the image is paramount, the ability to compress without significant loss of quality is particularly beneficial.

## Strengths of the Chosen Techniques:

PCA: Its strength lies in its ability to reduce data dimensionality while retaining the most critical information. This makes it highly effective for initial data reduction in image compression.

K-Means Clustering: The algorithm excels in simplifying the color palette of images, which is a key factor in reducing file size. It's particularly effective when a balance between compression rate and image quality is needed.

## Limitations and areas for improvement:

PCA Limitations: While PCA is effective in reducing dimensionality, it may sometimes overlook finer details in images, which can be crucial in certain applications like medical imaging.

K-Means Limitations: The reliance on the number of clusters as a hyperparameter requires careful tuning and may not always capture the most optimal color representation, especially in images with a vast color spectrum.

Computational Intensity: Both PCA and K-Means are computationally intensive, especially for high-resolution images, which could be a limiting factor in real-time applications.

# Conclusion

Our research project, exploring image clustering and compression through Principal Component Analysis (PCA) and K-Means clustering, has yielded significant insights into the potential of unsupervised learning techniques in image processing. Key findings include the effective dimensionality reduction by PCA, which preserved crucial visual information, and the substantial file size reduction achieved by K-Means clustering through color palette simplification, while maintaining image quality. These methods collectively present a balanced approach to image compression, addressing the need for reduced storage and transmission resources without compromising the integrity of the images. This study is particularly relevant in fields such as digital media, online content delivery, remote sensing, and medical imaging, where efficient image management is crucial. Future directions could include incorporating advanced machine learning algorithms for further enhancement of the compression process, automating hyperparameter selection for optimization, and exploring real-time compression techniques. The significance of this study lies in its practical approach to handling image data, providing a cost-effective solution for managing image databases, and offering insights that extend both academic understanding and real-world applications of image compression. This research marks a vital contribution to the digital world, demonstrating the importance of unsupervised learning in the important field of image processing and data management.

References

1. Analytics Vidhya (2018). Unsupervised Deep Learning Algorithms for Computer Vision. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/06/unsupervised-deep-learning-computer-vision/.
2. Edeza, T. (2021). Image Processing with Python — Unsupervised Learning for Image Segmentation. [online] Medium. Available at: <https://towardsdatascience.com/image-processing-with-python-unsupervised-learning-for-image-segmentation-90ebd23d91a4>.
3. Kaloyanova, E. (2020). How to Combine PCA and K-means Clustering in Python? [online] 365 Data Science. Available at: <https://365datascience.com/tutorials/python-tutorials/pca-k-means/>.
4. Liu, F., Hernandez-Cabronero, M., Sanchez, V., Marcellin, M. and Bilgin, A. (2017). The Current Role of Image Compression Standards in Medical Imaging. Information, 8(4), p.131. doi:https://doi.org/10.3390/info8040131.
5. Reddaiah, B. (2020). A Study on Image Compression and its Applications. International Journal of Computer Applications, 177(38), pp.33–36. doi:https://doi.org/10.5120/ijca2020919887.
6. Riswanto, U. (2023). Image Compression Techniques: A Closer Look at Principal Component Analysis. [online] Medium. Available at: https://ujangriswanto08.medium.com/image-compression-techniques-a-closer-look-at-principal-component-analysis-67cf7a29fdb9#:~:text=In%20image%20processing%2C%20PCA%20is.
7. Sharma, N. (2023). Image Compression Using Different Machine Learning Techniques. [online] Comet. Available at: <https://www.comet.com/site/blog/image-compression-using-different-machine-learning-techniques/#:~:text=Listed%20below%20are%20some%20of>.

1. . [↑](#footnote-ref-1)