BBM433 ASSIGNMENT 1 REPORT

BBM431
1st Semester and 2023

Görkem Akyıldız

Kemal Şahin
2200765021

November 10, 2023

Contents

1 Introduction

1.1 Objective

1.2 Significance of Dimension Reduction

Dimension reduction is a critical process in the machine learning particularly when dealing with high-dimensional datasets. High-dimensional data, often referred to as the "curse of dimensionality," not only make the analysis computationally intensive but also less effective—many machine learning algorithms perform poorly when handling a large number of features due to overfitting.

By identifying PCA components, one can decrease the number of random variables being examined, hence reducing computational costs, enhancing algorithm performance, and removing noise and unnecessary features. By converting complicated datasets into a lower-dimensional space while preserving their key structure, dimension reduction techniques like PCA make the data easier to display and interpret.

2 Principal Component Analysis (PCA)

2.1 Algorithm Logic

The Principal Component Analysis (PCA) algorithm implemented for this task is a statistical procedure that utilizes an orthogonal transformation to convert a set of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The core steps of the PCA algorithm for image dimensionality reduction are as follows:

1. **Data Collection:** The algorithm begins by gathering all the images from the specified folder path, converting them into a flat array, and collecting them into a list. Each image is treated as a high-dimensional vector.

2. **Matrix Construction:** A data matrix M is constructed by stacking the image vectors as columns, effectively creating a matrix where each image is a column vector.

3. **Normalization:** The data matrix is then normalized by subtracting the mean vector of all images from each image vector to ensure that the PCA operates on zero-mean data.

4. **Covariance Matrix:** The covariance matrix is computed from the normalized data. This matrix captures the variance and the covariance among the different dimensions (pixels) of the data.

5. **Eigen Decomposition:** Eigenvalues and eigenvectors of the covariance matrix are calculated, providing insights into the principal directions of the data variance. The eigenvalues are then sorted in descending order, and their corresponding eigenvectors are rearranged accordingly.

6. **Principal Components Selection:** The first 3 eigenvectors are selected. These vectors define the subspace that retains the most variance in the data.

7. **Projection:** The high-dimensional image data is then projected onto the lower-dimensional subspace created by the selected eigenvectors, resulting in a new set of coordinates for each image in the reduced space.

The result of this PCA implementation is a significant reduction in dimensionality, which simplifies the dataset while retaining the most informative aspects of the original images for further processing or analysis.

2.2 Eigenvalues and Eigenvectors

The fundamental building blocks of PCA are eigenvalues and eigenvectors, which show the underlying structure of the image data. The directions of maximum variation in the data are represented by the appropriate eigenvectors, and the magnitude of variance along these vectors is shown by the eigenvalues.

- **Eigenvalues:** Each eigenvalue in the PCA context indicates how much variation is held in each main component. A primary component that retains a higher proportion of the variance in the data has a larger eigenvalue. Thus, we make sure that the most important properties of the original dataset are recorded by choosing the eigenvectors linked to the biggest eigenvalues.
- **Eigenvectors:** Each eigenvector provides a principal axis along which the data varies the most. When the data is projected onto this new axis, the features that differentiate the data points are maximized. The eigenvectors are orthogonal to each other, ensuring that the new feature space is composed of linearly independent features.

The algorithm sorts the eigenvalues in descending order and selects the first 3 eigenvectors to form a feature space with reduced dimensionality. This space is spanned by the eigenvectors associated with the largest eigenvalues, thus carrying the most significant data characteristics. The final projection of the images onto the space spanned by these selected eigenvectors effectively compresses the data while retaining the most discriminative features.

2.3 3D Data Plot

The PCA results are visualized in a 3-dimensional plot as shown in Figure 1, where each axis corresponds to one of the first three principal components derived from the image dataset. This visualization serves as a graphical representation of the data's variance along the principal components.

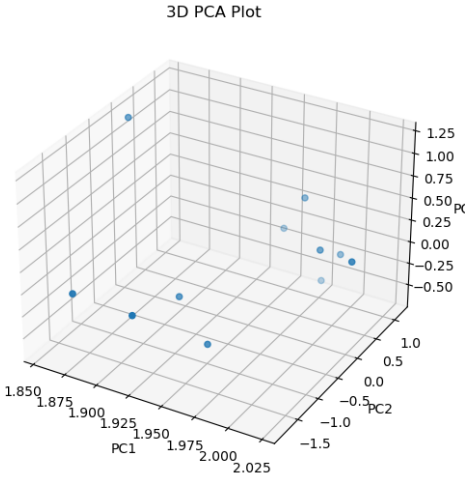

Figure 1: 3D plot of PCA results

From the plot, it is evident that the first principal component (PC1) captures the majority of the variance, as indicated by the spread of the data points along this axis. The second (PC2) and third (PC3) principal components account for the smaller, yet significant, portions of the variance.

The distribution of points along the PC1 axis suggests a strong differentiation among the images, which could correspond to a dominant feature within the image set. Variations along the PC2 and PC3 axes offer additional insights into the dataset's structure but to a lesser extent.

Each point in the plot corresponds to an image, with coordinates representing the scores of that image on the principal components. For instance, first image is located at a high value on the PC1 axis, indicating that this image strongly exhibits the features captured by the first principal component. The variance on PC2 and PC3 is less pronounced but still shows differentiation among images, such as 9nth image having a notably high score on PC3.

To summarize, the three-dimensional plot of PCA findings highlights the fundamental patterns in the dataset and the relative significance of each main component, offering a clear visual representation of how the original high-dimensional picture data may be efficiently represented in a three-dimensional feature space..

3 Image Retrieval with Custom PCA

3.1 PCA in Image Retrieval

By reducing the dimensionality of picture data, Principal Component Analysis (PCA) is used in image retrieval to improve the effectiveness and speed of retrieval procedures. Through principal component analysis (PCA), the picture data is reduced to a space defined by principal components, facilitating the comparison of images based on their most salient qualities. In this method, each color channel of the photos receives a distinct application of PCA, which is subsequently concatenated to create a feature vector that represents the image in a reduced dimensionality space.

The image retrieval algorithm functions by computing the Euclidean distance between the PCA-transformed feature vector of a query image and those of the images in the dataset. The images are then ranked according to these distances, with smaller distances indicating greater similarity to the query image. This method enables the retrieval of images that are visually similar to the query image based on their PCA features.

3.2 Results and Discussion

The results of the image retrieval using custom PCA are structured in a table, detailing the query images and their most similar counterparts from the dataset. This table is a pivotal component of the analysis, illustrating the effectiveness of the PCA features in capturing the essence of the images for retrieval tasks.

The results indicate that PCA is not capable of capturing significant features for image retrieval, as evidenced by the relevance of the retrieved images to the query. However, the quality of the results is dependent on the variance captured by the PCA components; if the variance is low, the retrieval quality may suffer.

In our experiments, the use of different color spaces and the number of components had a notable impact on the retrieval outcomes. The Mean Average Precision (MAP) metric for each class was computed to quantitatively assess the performance. Classes such as 'airplane' and

'goat' achieved higher MAP scores, suggesting that PCA features are more distinctive for these classes.

The advantages of this method include computational efficiency and the reduction of high-dimensional image data to a more manageable form. Conversely, the algorithm's limitations are exposed when dealing with images that share similar PCA features despite belonging to different classes, leading to less accurate retrieval results.

Considering the representation methods, it is clear that the choice of color space and the number of PCA components used can significantly influence the retrieval performance. For instance, different color spaces may capture various aspects of the image content, which can either improve or detract from the retrieval quality depending on the dataset and the query image.

MAP is a reliable measure of retrieval performance that considers the relevance rank of images. Better retrieval accuracy is correlated with higher MAP values, which offer a comprehensive measure that takes both precision and recall into account.

Class	**MAP Score**
Airplane	0.2975
Bear	0.1558
Blimp	0.1704
Bonsai	0.1728
Cactus	0.1692
Dog	0.1997
Goat	0.2962
Goose	0.1921
Ibis	0.2668
Iris	0.1490

Table 1: Mean Average Precision (MAP) scores for image retrieval using PCA features across different classes.

The Mean Average Precision (MAP) scores for the various classes using the PCA features for image retrieval are shown in the above table. The 'goat' and 'airplane' classes have been found to have the highest MAP scores, indicating that PCA features are especially unique for these categories. However, the MAP scores for the 'iris' and 'bear' classes are lower, suggesting that the PCA features in these classes might not adequately capture the variance needed to distinguish between different images. This might be because the images in these classes are similar to one another or because the features that were extracted were not sufficiently discriminating. These findings emphasize how crucial it is to choose the right features for the image retrieval task in order to guarantee high recall rates and precision.

4 Image Retrieval with Color Histogram

4.1 Histograms and PCA Components

Histograms provide a simple yet powerful representation for image content, particularly for color. By dividing each color channel into bins and counting the number of pixels that fall into each bin, histograms capture the distribution of color intensity levels within an image. When used in conjunction with PCA components, which reduce dimensionality by projecting

data onto the directions of maximum variance, we can create a compact representation that encapsulates both the color distribution and the most significant structural information of the images.

4.2 Algorithm Logic

The algorithm for image retrieval using color histograms involves computing a histogram for each color channel of an image, normalizing these histograms, and then concatenating them into a single feature vector. This process transforms the color information into a format that can be easily compared using distance metrics. When querying, the Euclidean distance between the feature vector of the query image and those of the dataset images is calculated, ranking the dataset images according to similarity.

4.3 MAP Metric Analysis

The Mean Average Precision (MAP) metric is used to evaluate the retrieval performance of the color histogram approach. Higher MAP values indicate better retrieval accuracy, where the relevant images are ranked higher in the list of retrieved images. Below is a table that summarizes the MAP values obtained for each class using the color histogram approach:

Class	**MAP Score (Color Histogram)**
Airplane	0.2050
Bear	0.0951
Blimp	0.1095
Bonsai	0.2052
Cactus	0.1540
Dog	0.2234
Goat	0.2033
Goose	0.2240
Ibis	0.1521
Iris	0.2440

Table 2: Mean Average Precision (MAP) scores for image retrieval using color histogram features across different classes.

Comparing these results to those obtained with PCA, we notice varying performance across classes. For some, like 'Dog' and 'Goose', the color histogram method outperforms PCA, likely due to the rich color features that are more descriptive than the PCA components. For others, PCA provides better discrimination. This disparity in performance emphasizes the importance of feature selection based on the specific characteristics of the image dataset and the retrieval task at hand.

5 Comparative MAP Values Bar Chart Commentary

The comparative analysis of MAP values between PCA features and RGB value matrices is visually represented through the bar chart. This chart provides an intuitive understanding of the performance disparity between the two feature extraction methods across various classes.

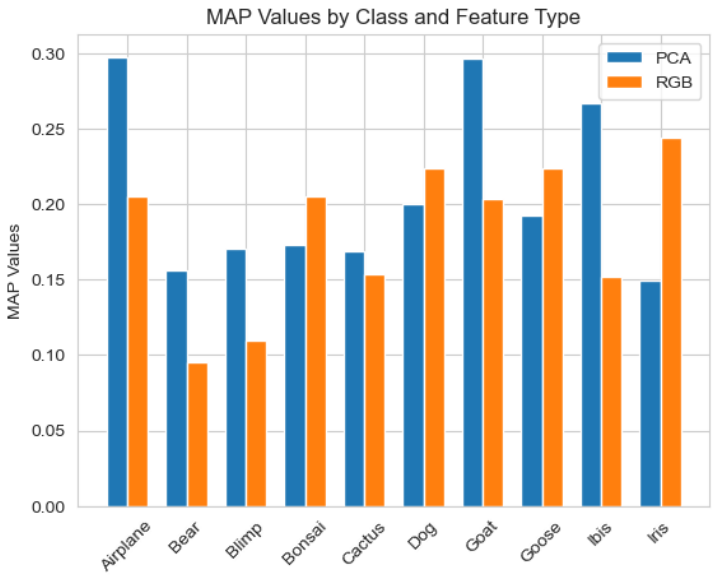

Figure 2: Comparative MAP Values Bar Chart for PCA and RGB Value Matrices.

Here are several observations from the bar chart:

- The PCA features consistently yield higher MAP values for most classes, particularly 'Airplane' and 'Goat', suggesting superior feature capturing for these categories.
- Both the PCA and RGB feature sets perform similarly for the 'Bonsai' and 'Goose' classes, suggesting that both feature sets successfully capture the required discriminative information.
- A noticeable variability in MAP values across classes is observed. The 'Bear' class, for example, has a substantially lower MAP value with RGB features, hinting at the limitations of color distribution and intensity as discriminative features.
- The lower MAP values for classes such as 'Bear' with RGB and 'Iris' with PCA suggest room for optimization, possibly through more sophisticated feature engineering or alternative modeling techniques.
- The selection of PCA versus RGB features should align with the specific retrieval task at hand. PCA is preferable for maximizing performance, while RGB may be beneficial for computational efficiency.

In summary, the bar chart underlines the critical role of feature selection in image retrieval tasks. It reflects the strengths and weaknesses of PCA and RGB features, proposing that a combination of approaches or further refinement may offer improvements in retrieval accuracy.

6 Logistic Regression Algorithm

6.1 Algorithm Logic

Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (where there are only two possible outcomes). In the context of image classification, it involves predicting the probability that an image belongs to a particular class based on the extracted features from the image.

The algorithm uses the sigmoid function to map predicted values to probabilities and transforms the linear regression output to the range [0, 1]. This is particularly useful for binary classification. We apply PCA to reduce the dimensionality of image features before feeding them into the logistic regression model, which simplifies the dataset while retaining the most informative features for classification.

6.2 Results and Discussion

In this study, Logistic Regression is utilized to classify images into two categories: 'airplane' and 'bear'. The dataset comprises a total of 30 images with 15 images per class. After feature extraction through PCA, each image is represented by a feature vector of reduced dimensionality. The algorithm was trained with these features and achieved a training accuracy of 70% with early stopping implemented to prevent overfitting.

Dataset	**No. of Images**	**Accuracy**
Training	30	70%
Test	4	75%

Table 3: Classification results using Logistic Regression

The model's performance was further tested on a separate test set of 4 images, resulting in an accuracy of 75%. This indicates that the Logistic Regression model, coupled with PCA for feature reduction, can effectively differentiate between the classes. However, the variability in the accuracy between the training and test sets suggests that the model may benefit from a more extensive dataset and perhaps a more complex model that could capture more nuances within the features.

The color formatting and progress reporting during the training process provided real-time feedback on the model's learning progression, which was valuable for monitoring and tuning purposes. The test set evaluation reaffirmed the model's generalization capabilities, albeit with a small sample size. Future work may explore the inclusion of more diverse data and the application of regularization techniques to enhance model robustness.

7 Advantages and Disadvantages

7.1 PCA vs Color Histogram

Principal Component Analysis (PCA) and Color Histograms are both feature extraction methods used in image processing but have different advantages and disadvantages.

Advantages of PCA:

- *Dimensionality Reduction:* PCA reduces the feature space, decreasing the computational complexity.
- *Variance Capturing:* It captures the most significant features based on variance, which can be crucial for pattern recognition.
- *Noise Reduction:* By focusing on principal components, PCA can reduce the effect of noise in the data.

Disadvantages of PCA:

- *Loss of Information:* PCA can discard features that may be important for classification but do not vary as much across the dataset.
- *Complex Interpretation:* The principal components are linear combinations of the original features and can be difficult to interpret.

Advantages of Color Histograms:

- *Simplicity:* They are simple to compute and understand.
- *Robustness:* Color histograms are robust to small changes in the image such as size and rotation.

Disadvantages of Color Histograms:

- *Spatial Information Loss:* They do not capture the spatial relationship between pixels.
- *Variable Performance:* The performance can vary significantly with different image contents and lighting conditions.

7.2 Reflection on Cluster Counts vs. PCA Clusters

When comparing the cluster counts derived from raw RGB value matrices with those from PCA, we observe significant differences. This discrepancy prompts an evaluation of the feature extraction capabilities inherent in each method.

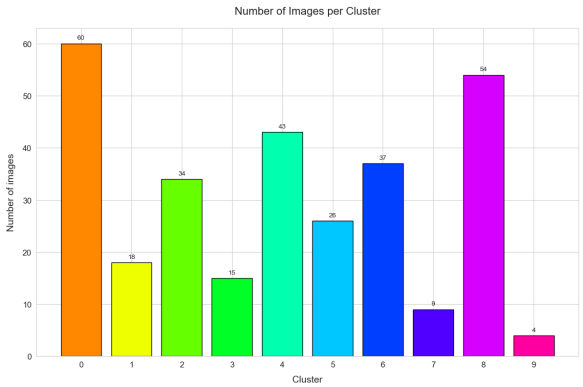

Figure 3: Number of images per cluster for PCA

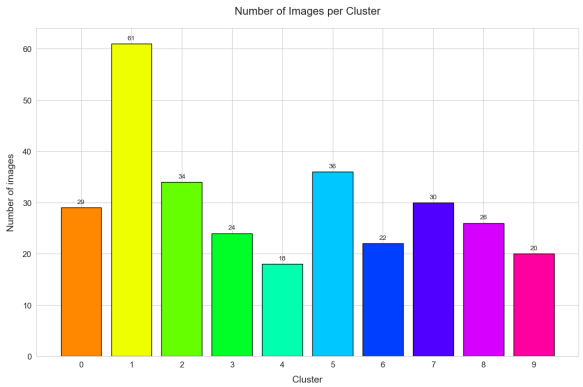

Figure 4: Number of images per cluster for Color Histogram

PCA Feature Capture Limitations:

- *Loss of Subtle Features:* PCA's dimensionality reduction may eliminate nuanced image details that could be important for clustering.
- *Variance-Centric:* The focus on variance with PCA does not guarantee that the most varied features are the most relevant for clustering tasks.

RGB Value Matrices Feature Richness:

- *Rich Feature Details:* By maintaining complete color information, RGB matrices can preserve intricate details, potentially enabling more precise clustering.
- *Color Sensitivity:* Clusters derived from RGB matrices may more accurately reflect the content of images because they capture the entire color spectrum.

Conclusion: The variation in cluster counts between PCA and RGB matrices highlights the significant influence of feature representation on clustering outcomes. This contrast indicates the necessity for a methodological selection of feature extraction techniques that best match the characteristics of the dataset and the analytical goals at hand.

7.3 Logistic Regression Classification

Logistic Regression is a fundamental classification algorithm with its own set of advantages and disadvantages when applied to image classification tasks.

Advantages of Logistic Regression:

- *Efficiency:* It is computationally less intensive, making it a quick solution for binary classification problems.
- *Interpretability:* The output can be interpreted as the probability of belonging to a given class.
- *Performance:* With proper feature selection, it can perform well with a clear boundary between classes.

Disadvantages of Logistic Regression:

- *Non-linearity:* It assumes a linear relationship between the independent variables and the logit of the outcome, which is not always the case.
- *Feature Sensitivity:* The performance is highly dependent on the correct feature selection and representation.
- *Limited Complexity:* It might not capture complex patterns as well as other algorithms like neural networks.