# ESE 2180 Project 3 Writeup

Chris Brusie
*Deptartment of Electrical Engineering*
*Washington University in St. Louis*
St. Louis, United States
brusie@wustl.edu

Sebastian Theiler
*Deptartment of Electrical Engineering*
*Washington University in St. Louis*
St. Louis, United States
s.k.theiler@wustl.edu

## I. Introduction

Image processing and compression is a promising use-case of applied linear algebra. The goal of this project was to explore the use of Principle Component Analysis (PCA) for dimensionality reduction and compression of the Extended Yale B ("Eigenfaces") dataset, which consists of images of various people. Singular Value Decomposition (SVD) is applied to identify principal components, compress data, and approximate images. The report discusses methods to quantify information retention, storage reduction, and reconstruction accuracy. Results demonstrate the trade-off between feature retention and compression, providing insights into PCA's effectiveness for image analysis.

## II. Background

PCA transforms data into a new coordinate system such that variance is maximized along orthogonal axes. Image analysis is a special case of SVD where an image with $M$ pixels is represented as a flattened vector $x_i \in \mathbb{R}^M$. The dataset contains $N$ images and is therefore represented as a matrix $X \in \mathbb{R}^{M \times N}$. PCA uses SVD to decompose $X$ into $U\Sigma V^T$, where $\Sigma$ contains singular values along its diagonals. These values quantify the variance captured by each principal component. Retaining fewer components (i.e., taking a submatrix of $\Sigma$) reduces dimensionality, enabling compression and approximation, at the cost of reduced reconstruction accuracy.

## III. Training the PCA Compression Model

### A. Dataset

The Extended Yale B, "Eigenfaces," dataset contains images of 15 individuals. 13 subjects were used for training set and two were reserved for testing. Images were resized for uniformity and flattened into a vector to construct the matrix $X$, where columns are 11368-vectors representing images. $X$ was then demeaned such that each column was centered on zero. This is important so that SVD can capture the variance in the data itself, rather than the shift from a non-zero mean.

```
file_list = os.listdir(folder_path)
for i, filename in enumerate(
    file_list[:num_imgs_total]
):
    img_path = os.path.join(
        folder_path,
        filename
    )
    img = cv2.imread(img_path)
    img_gray_flat = cv2.cvtColor(
        img,
        cv2.COLOR_BGR2GRAY
    ).flatten()
    img_matrix[:, i] = img_gray_flat
```

### B. SVD

SVD was performed on $X$ to extract singular values and principal components. This was done with Numpy's built-in SVD function: `U, s, Vh = np.linalg.svd(X, full_matrices=False)`. The cumulative variance explained by the singular values was analyzed to determine the number of components required for 70%, 80%, 90%, and 95% data retention (

Images were reconstructed using varying numbers of principal components $d$ ($d = 20, 50, 70, 100$), and reconstruction error was calculated as:

$$\text{Error} = \frac{|X - \hat{X}|_F}{|X|_F}$$

where $\hat{X}$ is the approximation of $X$ using $d$ components. As seen in Figure TODO, as the number of features retained increases, the reconstructed image gradually becomes more similar to the original.

## IV. Compressing Test Data

Finally, using 50 principal components from the training data, test images from two subjects were approximated.

```
d = 100
U_50 = U[:, :d]
projections = U_50.T @ X_test
test_reconstruct = U_50 @ projections + np.outer(
    np.mean(img_matrix_test, axis = 1),
    np.ones(img_matrix_test.shape[1])
)
```

Reconstruction error was measured for these test images and a rotated image ('subject15rotated.jpeg').

## V. Discussion and Conclusion

This study demonstrates PCA's capability in reducing the dimensionality of high-dimensional image data while preserving significant information. The results indicate that a relatively small number of principal components can capture a large percentage of the variance, enabling substantial storage reduction. The approximation errors remain low even with significant compression, validating PCA's effectiveness for image compression and feature extraction.

However, PCA's performance is contingent on the training data's representativeness. The increased error observed with rotated images highlights PCA's limitation in handling variations not present during training. Future work could explore incorporating additional transformations in the training set or employing more robust dimensionality reduction techniques to enhance invariance.

## References


[1] Cleveland clinic, 2021, "Spanish Flu: What Is It, Causes, Symptoms & Pandemic," Cleveland Clinic, Sep. 21, 2021. Available: https://my.clevelandclinic.org/health/diseases/21777-spanish-flu

[2] World Health Organization, 2021, "The True Death Toll of COVID-19: Estimating Global Excess Mortality," World Health Organization, May. Available: https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality

[3] United States Census Bureau, 2021, "Census Regions and Divisions of the United States," United States Census Bureau. Available: https://www.census.gov/geographies/reference-maps/2020/geo/division.html

[4] Walmart, 2024, "Mezorrison KN95 Face Masks, 50-Pack, Black," Walmart. Available: https://www.walmart.com/ip/Mezorrison-KN95-Face-Masks-50-Pack-Black/518596184?classType=VARIANT