

PCA-Based Eigenface Analysis and Face Reconstruction

Harshit Kumar Moore
Kunal Sati
Syed Ali
Uday Raj Verma

1 Introduction

Principal Component Analysis (PCA) forms the mathematical foundation of the well-known *Eigenfaces* method for face representation and recognition. In this practical, we use the Olivetti Faces dataset to construct a low-dimensional facial subspace that captures the most significant variance across a population of aligned human faces. Each image is resized to 32×32 pixels and reshaped into a 1024-dimensional vector. PCA is then applied to identify the principal components that describe at least 95% of the total variance.

A selfie image, preprocessed in the same way, is then projected onto this PCA subspace. Using the top eigenfaces, we reconstruct an approximation of the selfie and compare it with the original image. The difference in reconstruction quality reflects how well the selfie conforms to the statistical structure learned from the dataset.

2 Mathematical Background

2.1 Image Representation

Each face image of size 32×32 pixels is reshaped into a column vector

$$x_i \in \mathbb{R}^{1024}.$$

Given N such images, we construct a data matrix

$$X = [x_1, x_2, \dots, x_N]^T \in \mathbb{R}^{N \times 1024}.$$

Before applying PCA, we compute the mean face:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i.$$

Each image is mean-centered:

$$\tilde{x}_i = x_i - \mu.$$

2.2 Covariance Matrix and Eigenvectors

The covariance matrix of the dataset is:

$$C = \frac{1}{N} \tilde{X}^T \tilde{X},$$

where \tilde{X} is the centered data matrix.

PCA solves the eigenvalue problem:

$$Cv_k = \lambda_k v_k,$$

where v_k are eigenvectors (principal components) and λ_k their corresponding eigenvalues. Each eigenvector v_k can be reshaped into a 32×32 image, known as an *eigenface*.

The eigenvectors are sorted in decreasing order of eigenvalues, so that the first components capture the most variance.

2.3 Variance Retention

Let the eigenvalues be $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{1024}$. The proportion of variance explained by the first k components is:

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^{1024} \lambda_i}.$$

We select the minimum k such that:

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^{1024} \lambda_i} \geq 0.95.$$

3 Eigenfaces

The first few eigenfaces correspond to the directions along which the dataset varies most. These typically highlight global structures such as:

- overall face shape
- illumination gradients
- eye and eyebrow regions
- nose and cheek contours

Figure 1 illustrates the top five eigenfaces reshaped into 32×32 images.

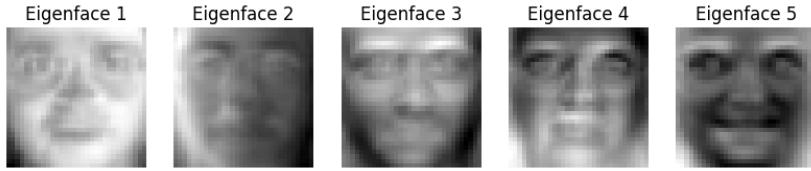


Figure 1: Top 5 eigenfaces obtained from PCA (placeholder).

4 Projection and Reconstruction

4.1 Projection

Given a new preprocessed face image x , its coordinates in the PCA subspace (also called the *eigenface coefficients*) are computed as:

$$y = W(x - \mu),$$

where $W \in \mathbb{R}^{k \times 1024}$ is the matrix whose rows are the top k eigenvectors.

4.2 Reconstruction

Using the k retained components, the reconstructed image is:

$$\hat{x} = \mu + W^T y.$$

If k is much smaller than 1024, the reconstruction is an approximation of the original face. It captures only the patterns represented by the dataset’s principal components.

Figure 2 shows an example comparison between the original selfie and its PCA reconstruction.

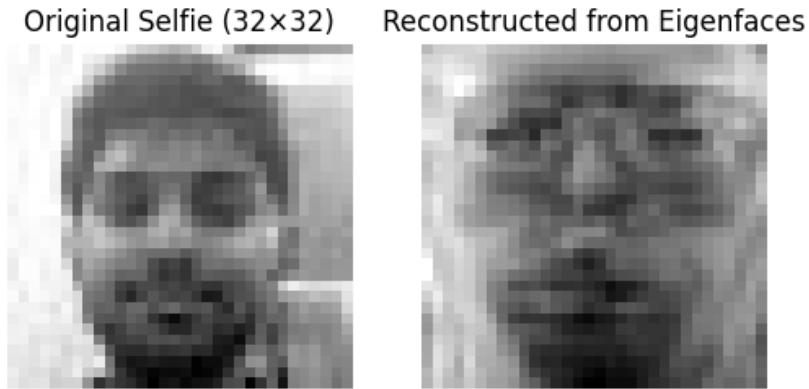


Figure 2: Original selfie vs. PCA reconstruction (placeholder).

5 Why the Selfie Reconstruction is Poor Compared to Dataset Faces

The Olivetti dataset consists of studio-like, well-aligned, grayscale, low-noise facial images. PCA therefore learns a very specific statistical model of faces:

1. **Pose Consistency** – All dataset faces are centered, similarly scaled, and frontal. A selfie usually differs in rotation, scale, or cropping.
2. **Lighting Distribution** – PCA learns the lighting patterns present in the dataset. A selfie may have smartphone lighting, shadows, or highlights that deviate from the dataset distribution.
3. **Camera Differences** – The dataset images were captured using the same camera; a selfie uses a different sensor, lens, and noise characteristics.
4. **Feature Mismatch** – Facial hair, glasses, or unique features may not exist in the dataset, so PCA cannot represent them well.

5. **Out-of-Distribution Input** – Mathematically, the selfie x lies outside the linear subspace spanned by the dataset eigenfaces. Thus the projection

$$y = W(x - \mu)$$

forces the selfie into a space that cannot fully represent it.

These factors lead to noticeable loss of detail, smoothing, and distortion in the reconstructed selfie. In contrast, faces from the Olivetti dataset are well-represented because the PCA model was trained directly on them.

6 Conclusion

This practical demonstrates how PCA can be used to construct an eigenface basis that captures the most important modes of variation in a set of aligned facial images. By projecting images into this lower-dimensional subspace, we can reconstruct approximations of the original faces.

While dataset images reconstruct well due to their conformity to the learned model, external images such as selfies suffer significant degradation. This highlights the importance of data distribution, alignment, and lighting conditions in statistical face modeling.