# PCA Dimensionality Reduction and Face Recognition

Qiu MingRui
1024041125
*Cybersecurity Major*
*Nanjing University of Posts and Telecommunications*
Nanjing, China

***Abstract*—This study explores the impact of PCA-based dimensionality reduction on image reconstruction and face recognition accuracy using various methods, including CNN, Euclidean distance, and random forest. The experiments reveal that as the dimensionality and the number of training images increase, both the quality of reconstructed images and the accuracy of face recognition improve.**

***Index Terms*—PCA,CNN,Euclidean distance,random forest.**

## I. Introduction

In the era of big data applications, numerous fields, particularly image processing and face recognition, face the complex challenge of managing and processing high-dimensional data. Although this data is rich in valuable information, it also brings significant issues, such as increased computational demands and substantial storage requirements. The effective management and processing of high-dimensional data have therefore become critical research topics.

Dimensionality reduction techniques address these challenges by projecting high-dimensional data into a lower-dimensional space, thereby reducing feature redundancy, decreasing computational complexity, and improving algorithmic efficiency. Principal Component Analysis (PCA), as a classical and widely adopted dimensionality reduction method, has found extensive applications in image processing. PCA is capable of substantially reducing the dimensionality of the data while retaining the most significant features, which is particularly crucial in image data analysis. By reducing the dimensionality of the data, PCA not only enhances computational efficiency but also mitigates the effects of noise interference, thereby improving the generalization ability of models.

To validate the reliability of the PCA technique, this paper will also conduct experiments involving image reconstruction and face recognition following dimensionality reduction, further assessing its effectiveness in practical applications.

## II. Related Work

This section will introduce some of the technical principles used in this paper, including PCA dimensionality reduction and various face recognition methods.

### A. PCA

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique aimed at projecting high-dimensional data into a lower-dimensional space by identifying the most representative directions within the data. This reduces redundant information and improves processing efficiency. PCA works by preserving the largest variance in the data, ensuring that the main features are retained after the transformation.

*1) Basic Principle of PCA:* The goal of PCA is to transform the original high-dimensional data into a new lower-dimensional space through orthogonal transformations, where each dimension (principal component) is a linear combination of the original features.

*a) Data Standardization:* The first step of PCA is to standardize the data. The goal of standardization is to ensure that each feature has a mean of 0 and a variance of 1, which avoids the influence of differing scales across features. If the dataset $X$ consists of $n$ samples and $m$ features, the standardization process is as follows:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

where:

- $\mu$ is the mean of each feature.
- $\sigma$ is the standard deviation of each feature.
- $X_{\text{std}}$ is the standardized data matrix.

*b) Covariance Matrix Calculation:* The covariance matrix is computed from the standardized data to capture the linear relationships between the features. The covariance matrix $\Sigma$ can be calculated as:

$$\Sigma = \frac{1}{n-1} X_{\text{std}}^T X_{\text{std}}$$

where:

- $X_{\text{std}}$ is the standardized data matrix.
- $\Sigma$ is the $m \times m$ covariance matrix, reflecting the correlation between features.

*c) Eigenvalue and Eigenvector Calculation:* Eigenvalue decomposition (or Singular Value Decomposition, SVD) is performed on the covariance matrix to obtain the eigenvalues

$\lambda_i$ and the corresponding eigenvectors $v_i$. The eigenvalue equation is as follows:

$$\Sigma v_i = \lambda_i v_i$$

where:

- $\lambda_i$ is the $i$-th eigenvalue, indicating the "importance" of the principal component.
- $v_i$ is the $i$-th eigenvector, representing the direction of the principal component.

*d) Selecting Principal Components:* The eigenvalues are sorted in descending order, and the top $k$ eigenvectors corresponding to the largest eigenvalues are selected as the principal components. These eigenvectors represent the most important directions in the data. The principal components with the largest eigenvalues explain the greatest variance in the data.

*e) Projection onto the New Space:* The original data is projected onto the selected $k$ principal components to obtain the reduced-dimensional data. The reduced-dimensional data $X_{\text{PCA}}$ can be computed as:

$$X_{\text{PCA}} = X_{\text{std}} V_k$$

where:

- $V_k$ is the matrix formed by the top $k$ eigenvectors.
- $X_{\text{PCA}}$ is the reduced-dimensional data.

*2) Advantages of PCA:*

- **Reduction of Redundant Information**: PCA effectively reduces redundancy in the data by preserving the largest variance and eliminating correlations between features, which results in lower dimensionality.
- **Noise Reduction**: PCA can remove noise in the data by selecting larger eigenvalues and ignoring smaller ones, which is especially effective in image processing where variations in lighting and viewpoint may introduce noise.
- **Improved Computational Efficiency**: The dimensionality reduction achieved by PCA significantly reduces computational cost, making subsequent machine learning and pattern recognition algorithms faster and more efficient.

### *B. Face Recognition Algorithms*

The main goal of face recognition algorithms is to extract features from images and perform identity verification or recognition. This paper selects three face recognition methods for experimentation: Convolutional Neural Networks (CNN), Euclidean Distance, and Random Forest.

*a) Convolutional Neural Networks (CNN):* Convolutional Neural Networks (CNNs) are an important neural network structure in deep learning, particularly suited for processing image data. CNNs automatically extract features from images through multiple convolutional layers, avoiding the complexity of manually designing feature extraction algorithms as in traditional methods. The basic structure of a CNN consists of convolutional layers, pooling layers, and fully connected layers. The convolutional layers are used to extract local features from the input image, the pooling layers reduce the feature dimensions through downsampling to enhance computational efficiency, and the fully connected layers perform classification to output the final recognition result.

In face recognition tasks, CNNs can automatically learn useful facial features from raw images through end-to-end learning. Compared to traditional methods, CNNs exhibit greater robustness under complex conditions such as lighting changes, angle variations, and facial expression changes. During training, CNNs adjust the network parameters continuously through optimization algorithms and loss functions (such as cross-entropy loss and triplet loss) to improve recognition accuracy. During the feature extraction process, CNNs can automatically identify key facial information, such as eyes, nose, and mouth, thus enhancing the performance of face recognition.

Furthermore, training CNNs on large-scale datasets further improves their performance in face recognition. Common deep network structures, such as AlexNet, VGGNet, and ResNet, have been successfully applied to face recognition tasks. With the continuous development of network structures, deep learning models can handle more complex and detailed image features, thus providing higher accuracy and efficiency in applications such as identity verification and security monitoring.

*b) Euclidean Distance:* Euclidean Distance face recognition is a method that performs identity recognition by calculating the Euclidean distance between image feature vectors. The method first converts face images into high-dimensional feature vectors using feature extraction techniques such as PCA or CNN. Then, it uses Euclidean distance to measure the similarity between different face features. A smaller Euclidean distance indicates higher similarity, meaning the two faces are similar. Ultimately, the identity with the smallest distance is selected for recognition.

In the recognition process, each person's face feature vector is first extracted and stored in a database. During recognition, the system converts the face image to be identified into a feature vector and then calculates the Euclidean distance between this feature vector and all the feature vectors in the database. Based on the smallest distance, the system determines whether the face matches, with a smaller Euclidean distance indicating higher similarity, leading the system to recognize them as the same person.

The advantage of Euclidean distance face recognition is its simplicity and intuition, as it does not require a complex training process, making it suitable for recognition tasks with small-scale databases. Its computation is fast and especially well-suited for processing single, clear images. However, when handling high-dimensional data, the method may face the problem of "the curse of dimensionality," where the effectiveness of Euclidean distance decreases as the feature dimensions increase. Additionally, under complex conditions such as variations in lighting, pose, or partial occlusion, the recognition accuracy of Euclidean distance may be affected. Therefore, it is often combined with other methods in practical applications to improve recognition accuracy and robustness.

*c) Random Forest:* Random Forest is an ensemble learning algorithm widely used for classification and regression tasks. It makes predictions by constructing multiple decision trees and aggregating their outputs through voting or averaging, thus improving the model's accuracy and robustness. In facial recognition tasks, Random Forest serves as an efficient classifier capable of identifying facial features in complex visual data.

The basic idea of Random Forest is to build multiple decision trees, each trained on different subsets of the training data and using random feature subsets for decision-making. This random selection helps prevent overfitting and enhances the model's generalization ability. In the end, the prediction results of all decision trees are combined through a voting mechanism to produce the final recognition outcome. In facial recognition, Random Forest learns the facial features in the training dataset and identifies faces in input images.

Compared to traditional single decision trees, Random Forest demonstrates greater stability and accuracy in facial recognition. It can efficiently handle a large number of features and is highly tolerant to noise and missing values in the data. Due to its ensemble nature, Random Forest makes decisions from multiple perspectives, making it robust against variations in facial expressions, lighting, and angles.

## III. Experimental Design

Based on PCA dimensionality reduction, four experiments were conducted in total: one reconstruction experiment and three face recognition experiments. The reconstruction experiment and the face recognition experiment using Random Forest are detailed below. The designs for the three face recognition experiments are similar.

### A. Reconstruction Experiment

In the experiment, the PCA algorithm was used to reduce the dimensionality of the face images.The reconstruction experiment performs an inverse transformation on the dimensionality-reduced data, converting the reduced image data back into the original image shape. Different dimensionality-reduced images are observed and compared in terms of image quality and the degree of detail retention.

### B. Face Recognition with Random Forest

The design for the face recognition experiment using Random Forest is as follows:

1) **Dataset Preparation:** The ORL face database, which contains 40 individuals with 10 images each, is used. For each individual, a number $k$ of images were selected randomly to form the training set, and the remaining images were used as the testing set.
2) **Image Vectorization:** Each image is read and converted to grayscale, followed by vectorization. This process transforms each image from a 2D matrix into a 1D vector, which is used as input to the PCA algorithm.
3) **PCA Dimensionality Reduction:** The PCA algorithm is applied to the training data to reduce the dimensionality from the original size (112x92) to a smaller size, such as 10, 20, 30, or 40 dimensions. The first $r$ principal components are selected for dimensionality reduction.
4) **Training the Classifier:** A Random Forest classifier is trained using the dimensionality-reduced data. The classifier is then used to predict the labels of the test set images. The accuracy of the classification is evaluated.
5) **Performance Evaluation:** The classification accuracy is calculated for different values of $k$ (the number of images used for training per person) and different dimensionality reduction sizes (ranging from 10 to 40 dimensions).
6) **Result Visualization:** The accuracy results are plotted as a function of $k$ (the number of training images per person) for each dimensionality reduction setting. A comparison of accuracies for 10, 20, 30, and 40 dimensions is made to analyze the effect of dimensionality reduction on the classification performance.

### C. Repetition of Face Recognition Experiments

The other two face recognition experiments follow the same design as described above, with variations in the number of principal components used for dimensionality reduction (10, 20, 30, and 40 dimensions) and the number of training images used (from $k = 1$ to $k = 9$).

### D. Expected Results

The expected outcome of these experiments is to identify the optimal number of principal components for dimensionality reduction that retains sufficient information for accurate face recognition. Additionally, the experiments will show how the number of training images $k$ affects the performance of the Random Forest classifier in terms of classification accuracy.

## IV. Result and Evaluation

This section will present the experimental results, followed by data evaluation and result analysis.

### A. Reconstruction

The reconstruction experiment has completed the dimensionality reduction and reconstruction of the dataset, including experimental results for dimensions ranging from 10 to 50, as shown in the figure below.

When performing image reconstruction on data reduced to different dimensions using PCA, it can be observed that the image quality and detail retention gradually improve as the dimensionality increases. In low-dimensional reconstructions, many details in the images are lost, resulting in blurriness and distortion. However, as the dimensionality increases, the reconstructed images progressively recover more details, and the visual quality significantly improves.

This indicates that selecting an appropriate dimensionality is crucial in the dimensionality reduction process to preserve the critical information of the original images. Too low a dimensionality may lead to excessive information loss, while too high a dimensionality could reduce the efficiency of the reduction.

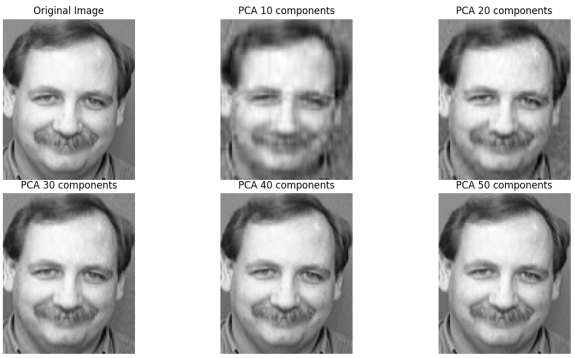


Fig. 1. Reconstruction figure

### B. CNN

The CNN face recognition results are shown in Table 1, and the corresponding line chart is shown in Figure 2.

| Samples | 10 Dim | 20 Dim | 30 Dim | 40 Dim |
|---|---|---|---|---|
| 1 | 17.78% | 28.06% | 18.33% | 27.78% |
| 2 | 68.12% | 46.56% | 64.38% | 70.94% |
| 3 | 63.21% | 77.50% | 76.43% | 66.07% |
| 4 | 82.08% | 80.83% | 73.33% | 84.58% |
| 5 | 88.00% | 86.50% | 82.50% | 88.50% |
| 6 | 94.38% | 90.62% | 91.87% | 91.25% |
| 7 | 88.33% | 93.33% | 95.83% | 96.67% |
| 8 | 93.75% | 98.75% | 96.25% | 91.25% |
| 9 | 95.00% | 92.50% | 90.00% | 95.00% |

TABLE I
CNN FACE RECOGNITION RESULTS

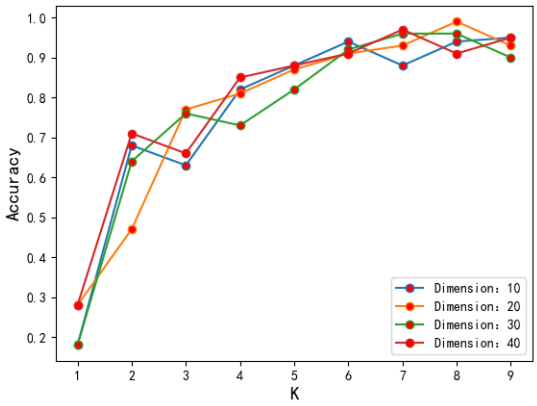


Fig. 2. CNN figure

### C. Euclidean Distance

The results of Euclidean Distance face recognition are as follows in Table 2, and the corresponding line chart is shown in Figure 3.

### D. Random Forest

The results of Random Forest face recognition are in Table 3, and the corresponding line chart is shown in Figure 4.

| Samples | 10 Dim | 20 Dim | 30 Dim | 40 Dim |
|---|---|---|---|---|
| 1 | 60.28% | 66.67% | 64.17% | 64.17% |
| 2 | 76.88% | 74.38% | 83.44% | 81.56% |
| 3 | 86.07% | 84.64% | 86.43% | 88.57% |
| 4 | 85.83% | 89.58% | 89.58% | 93.75% |
| 5 | 88.00% | 90.50% | 92.50% | 97.50% |
| 6 | 91.25% | 92.50% | 95.00% | 97.50% |
| 7 | 98.33% | 94.17% | 95.00% | 95.00% |
| 8 | 95.00% | 97.50% | 96.25% | 98.75% |
| 9 | 92.50% | 97.50% | 100.00% | 100.00% |

TABLE II
EUCLIDEAN DISTANCE FACE RECOGNITION RESULTS

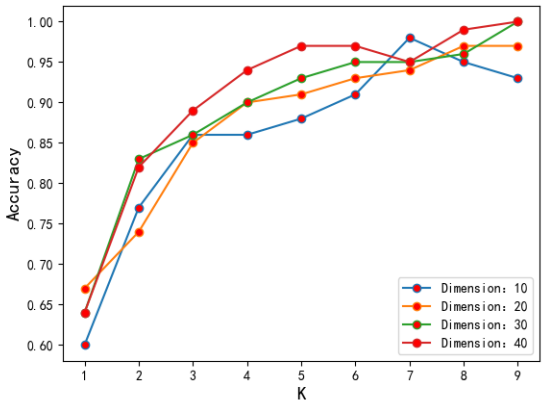


Fig. 3. Euclidean Distance figure

| Samples | 10 Dim | 20 Dim | 30 Dim | 40 Dim |
|---|---|---|---|---|
| 1 | 41.67% | 48.61% | 37.22% | 35.56% |
| 2 | 64.38% | 66.25% | 70.62% | 57.19% |
| 3 | 81.07% | 77.50% | 75.36% | 81.43% |
| 4 | 85.42% | 88.33% | 89.58% | 84.17% |
| 5 | 85.50% | 88.00% | 90.00% | 92.50% |
| 6 | 89.38% | 89.38% | 89.38% | 91.25% |
| 7 | 90.83% | 92.50% | 88.33% | 94.17% |
| 8 | 91.25% | 93.75% | 97.50% | 95.00% |
| 9 | 90.00% | 95.00% | 95.00% | 87.50% |

TABLE III
RANDOM FOREST RECOGNITION RESULTS

For the three face recognition methods, instead of comparing their differences, we focus on their commonalities. It can be observed that as the number of training images and the dimensionality of the reduced data increase, the recognition accuracy gradually improves. At lower dimensions, the recognition accuracy decreases due to the loss of critical information during dimensionality reduction. However, as the dimensions increase, more features are retained, and the recognition accuracy improves, indicating that higher dimensions help capture more details in the images, thereby enhancing recognition performance.

Additionally, as the number of training images increases, the model's learning and generalization capabilities are enhanced, leading to higher recognition accuracy. This trend suggests that

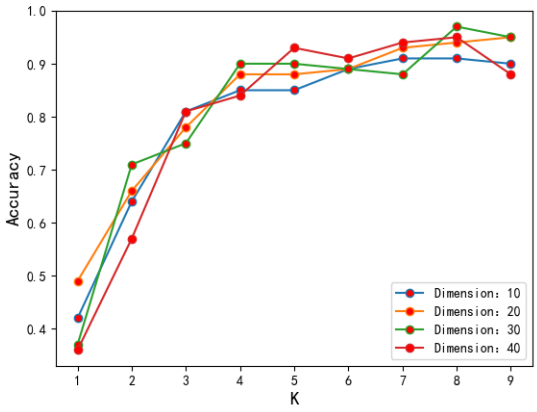


Fig. 4. Random Forest figure

selecting an appropriate dimensionality and providing sufficient training data are crucial for optimizing the performance of random forests in face recognition tasks.

## V. Conclusion

The experimental results indicate that both the selection of dimensions and the number of training images significantly affect the performance outcomes in image reconstruction and face recognition experiments based on PCA dimensionality reduction.

In the image reconstruction experiment, the quality and detail retention of the images improved progressively with the increase in dimensions. At lower dimensions, the reconstructed images exhibited blurriness and distortion, whereas at higher dimensions, the image quality significantly improved, highlighting the importance of appropriate dimension selection to retain critical information in the images.

In the face recognition experiments, regardless of whether CNN, Euclidean distance, or random forest methods were used, the recognition accuracy increased with the rise in dimensions and the number of training images. This suggests that higher dimensions and ample training data effectively enhance the model's recognition performance. Particularly in the case of the random forest method, increasing both dimensions and training samples significantly improved the recognition accuracy, further underscoring the critical role of dimensions and data volume in recognition tasks.

Overall, the experimental results support a balanced approach in dimensionality reduction tasks: reducing dimensions to improve computational efficiency while ensuring sufficient dimensions to retain key features of the data. Additionally, providing adequate training data is crucial for the model's learning and generalization capabilities, thereby achieving better recognition results in practical applications.

## References

[1] D. Bertsimas and D. L. Kitane, "Sparse pca: a geometric approach," *J. Mach. Learn. Res.*, vol. 24, no. 1, Mar. 2024.

[2] L. Hai and H. Guo, "Face detection with improved face r-cnn training method," in *Proceedings of the 3rd International Conference on Control and Computer Vision*, ser. ICCCV '20. New York, NY, USA: Association for Computing Machinery, 2021, p. 22–25. [Online]. Available: https://doi.org/10.1145/3425577.3425582

[3] C. Ping, H. Da-Peng, and L. Zu-Ying, "Automatic attendance face recognition for real classroom environments," in *Proceedings of the 2018 2nd International Conference on Big Data and Internet of Things*, ser. BDIOT '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 65–70. [Online]. Available: https://doi.org/10.1145/3289430.3289436

[4] L. Tong, S. Zhang, and X. Yue, "Research and application of intelligent education management platform based on convolutional neural network face recognition," in *Proceedings of the 2024 9th International Conference on Intelligent Information Processing*, ser. ICIIP '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 287–291. [Online]. Available: https://doi.org/10.1145/3696952.3696990

[5] N. M. Stausholm, "Improved differentially private euclidean distance approximation," in *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, ser. PODS'21. New York, NY, USA: Association for Computing Machinery, 2021, p. 42–56. [Online]. Available: https://doi.org/10.1145/3452021.3458328

[6] Z. Liu, H. Li, R. Li, Y. Zeng, and J. Ma, "Graph embedding based on euclidean distance matrix and its applications," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 1140–1149. [Online]. Available: https://doi.org/10.1145/3459637.3482261

[7] G. Feng, D. Desai, S. Pasquali, and D. Mehta, "Open set recognition for random forest," in *Proceedings of the 5th ACM International Conference on AI in Finance*, ser. ICAIF '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 45–53. [Online]. Available: https://doi.org/10.1145/3677052.3698631

[8] S. Yucer, F. Tektas, N. Al Moubayed, and T. Breckon, "Racial bias within face recognition: A survey," *ACM Comput. Surv.*, vol. 57, no. 4, Dec. 2024. [Online]. Available: https://doi.org/10.1145/3705295

[9] F. Dietz, L. Mecke, D. Riesner, and F. Alt, "Delusio - plausible deniability for face recognition," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. MHCI, Sep. 2024. [Online]. Available: https://doi.org/10.1145/3676494

[10] F. Zhou, Q. Zhou, B. Yin, H. Zheng, X. Lu, L. Ma, and H. Ling, "Rethinking impersonation and dodging attacks on face recognition systems," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 2487–2496. [Online]. Available: https://doi.org/10.1145/3664647.3681440