Comparative Analysis of Distance Metrics and PCA in k-Nearest Neighbors for Face Recognition

By: Saad Rasheed

1. Introduction

This report explores face recognition using the K Nearest Neighbours (k-NN) algorithm and Principal Component Analysis (PCA). The goal is to develop an effective face recognition system and evaluate its performance on the CMU Pose, Illumination, and Expression (PIE) database [1]. The dataset consists of 10 subjects with 170 resized face images each, and we preprocess it by normalizing the face image vectors. We then randomly split the dataset into training and testing sets, conducting experiments with the k-NN classifier by varying distance measures, k values, and the number of classes involved. Additionally, we apply PCA for dimensionality reduction and assess its impact on computation times and classification performance. The report aims to identify the best configurations for accurate face recognition and provide insights into the utilization of k-NN and PCA techniques.

2. Methodology

The implemented methodology is highlighted in the following paragraphs:

1. Dataset Preprocessing: The first step in our methodology is to preprocess the dataset by normalizing each face image vector to unit length. This normalization process involves dividing each vector by its magnitude. By performing this normalization, we ensure that all the vectors have the same scale and improve the performance of the subsequent classification algorithms.
2. Dataset Splits: To evaluate the performance of our face recognition system, we need to divide the dataset into training and testing sets. For each of the 10 subjects in the dataset, we randomly select 150 images for training purposes, while the remaining 20 images are reserved for testing. This random split process is repeated five times to obtain multiple sets of training and testing data, allowing us to assess the robustness of our algorithms across different splits.
3. k-NN Classifier with Different Distance Measures: Our first experiment focuses on implementing the k-NN classifier using the training set. We explore different distance measures, such as Euclidean distance, Mahalanobis distance, and cosine similarity. By

calculating the distances between test samples and their k nearest neighbors in the training set, we assign class labels based on majority voting. We vary the value of k, representing the number of nearest neighbors considered, to examine its influence on the accuracy of the face recognition system.

4. Fewer Training Images: To assess the robustness of the k-NN classifier under limited training data scenarios, we reduce the number of training images. Instead of using the full 150 training images per subject, we limit it to 100 training images, while keeping 70 test images per category. This experiment allows us to evaluate the system's performance when faced with a smaller training set, providing insights into its generalization capabilities.
5. PCA for Dimensionality Reduction: In this experiment, we incorporate Principal Component Analysis (PCA) as a dimensionality reduction technique. PCA is an unsupervised method that extracts the most informative features from high-dimensional data. We train PCA on all the training images to capture the essential facial characteristics necessary for accurate recognition. By reducing the dimensionality of the feature space, we aim to improve computation times and classification performance. We vary the number of principal components used to identify the optimal configuration that maximizes accuracy.
6. Evaluation Metrics and Analysis: To evaluate the performance of our experiments, we employ several metrics. We calculate the average accuracy and standard deviation over the five random splits for each experiment, providing insights into the consistency and reliability of the results. Additionally, we measure the computation times for each experiment, assessing the efficiency of the algorithms. By analyzing the results and comparing the performance across different experiments, we aim to identify the most effective distance measures, k values, and PCA configurations for achieving accurate and efficient face recognition.

By following this methodology, we systematically explore the performance of the k-NN classifier and PCA technique in the context of face recognition. The experiments conducted provide valuable insights into the effectiveness of different techniques, parameter settings, and dataset characteristics, enabling us to identify the optimal configurations for achieving accurate and efficient face recognition.

3. Results and Analysis

In this study, we conducted several experiments to explore the performance of the face recognition system using the k-NN algorithm and PCA. Here are the key findings from each experiment:

3.1. Distance Metrics: We experimented with three distance measures: Euclidean distance, cosine similarity, and Mahalanobis distance. Among the three distance measures, cosine similarity consistently yielded the best results in terms of accuracy for face recognition. Euclidean distance and cosine similarity performed comparably, while Mahalanobis distance showed relatively lower accuracy.

3.2. Varying k Values: We tested different values of k (number of nearest neighbors) ranging from 3 to 15 with increments of 2. The results consistently showed that using k=3 achieved the highest accuracy across all distance measures. This suggests that considering a smaller number of nearest neighbors provides better discriminative power for face recognition.

Distance Metrics	k=3	k=7	k=11	k=15
Euclidean Distance	0.9650	0.9130	0.8880	0.8580
Cosine Similarity	0.9790	0.9490	0.9380	0.9190
Mahalanobis Distance	0.8260	0.6560	0.4840	0.3850

3.3. Training and Testing Set Sizes: Two variations were explored: using 100 images per class for training and 70 images per class for testing, versus using 150 images per class for training and 20 images per class for testing. The results demonstrated that the 150/20 training and testing set configuration consistently yielded better accuracy compared to the 100/70 configuration. Having a larger training set allows for better model generalization and improved recognition performance.

Dataset Splits	Euclidean Dist. (k=3)	Cosine Similarity (k=3)	Mahalanobis Dist. (k=3)
150-20	0.9650	0.9790	0.8260
100-70	0.9069	0.9466	0.1009

3.4. PCA with Varying Principal Components: PCA was integrated into the pipeline for dimensionality reduction. We experimented with different numbers of principal components, ranging from 1024 to 64 with decrements of 64. For Euclidean distance and cosine similarity, the choice of the number of principal components did not significantly affect the average accuracy. However, for Mahalanobis distance, reducing the number of principal components led to higher average accuracy.

Distance Metrics	n=64	n=256	n=512	n=1024
Euclidean Distance	0.9630	0.9640	0.9620	0.9620
Cosine Similarity	0.9640	0.9630	0.9630	0.9630
Mahalanobis Distance	0.9700	0.9510	0.9000	0.8000

3.5. Computation Times: The computation times for making predictions varied among the distance measures. Euclidean distance took the least amount of time, followed closely by cosine similarity. In contrast, Mahalanobis' distance required considerably more time, approximately 20 times longer than cosine similarity.

These results highlight the importance of selecting the appropriate distance measure and parameter settings for achieving accurate face recognition. Cosine similarity consistently outperformed Euclidean and Mahalanobis distances, while using a smaller k value (k=3) yielded better accuracy. Additionally, the results emphasize the benefits of a larger training set (150/20) and demonstrate the potential of PCA in improving accuracy, particularly in conjunction with Mahalanobis distance. Considering computation times, Euclidean distance demonstrated the shortest prediction times, making it a more efficient choice for real-time applications.

4. Conclusion

In this study, we investigated the application of the k-NN algorithm and PCA for face recognition using the CMU PIE dataset. Through a series of experiments, we examined the impact of different factors on the system's performance, including distance measures, k values, training and testing set sizes, and the incorporation of PCA.

Our findings revealed that cosine similarity consistently outperformed Euclidean and Mahalanobis distances, demonstrating its effectiveness in capturing the similarity between face images. Furthermore, selecting a smaller number of nearest neighbors (k=3) yielded higher accuracy, indicating the importance of considering local information for accurate face recognition.

We observed that a larger training set (150 images per class) coupled with a smaller testing set (20 images per class) resulted in improved accuracy, highlighting the significance of having sufficient training data for model generalization. Additionally, the integration of PCA for dimensionality reduction showcased promising results, particularly in combination with Mahalanobis distance, where reducing the number of principal components enhanced accuracy.

Computation times varied among the distance measures, with Euclidean distance demonstrating the shortest prediction times, followed closely by cosine similarity. However, Mahalanobis distance required considerably more time for predictions, indicating its higher computational complexity.

Overall, our study provides insights into the factors influencing the performance of the face recognition system. The results underscore the importance of selecting appropriate distance measures, optimizing the choice of k value, and considering the size of the training and testing sets. The incorporation of PCA can further improve accuracy, particularly in scenarios where Mahalanobis distance is employed. These findings contribute to the advancement of face recognition techniques and provide valuable guidance for designing more accurate and efficient systems.

Future research can explore other distance measures, feature extraction techniques, and classification algorithms to enhance face recognition performance. Additionally, investigating the impact of larger and more diverse datasets can help assess the scalability and robustness of the proposed approaches. By continuously refining and expanding our understanding of face recognition methods, we can pave the way for broader applications in various domains, including security systems, surveillance, and human-computer interaction.

5. References

[1] CMU Pose, Illumination, and Expression (PIE) Database. Available at: http://ieeexplore.ieee.org/abstract/document/1004130/