Classical Face Recognition Under Real-World Variations

Ayoub EL KBADI Fotios KAPOTOS Jean-Vincent MARTINI
CentraleSupélec

Abstract

Face recognition is a classical problem in computer vision that predates modern deep learning approaches and relies on explicit feature extraction and linear subspace modeling. This report presents a comparative study of three well-established appearance-based face recognition methods: Eigenfaces (PCA) [10], Fisherfaces (LDA) [2], and Local Binary Pattern Histograms (LBPH) [1]. Using the ORL (AT&T) dataset [7] and the Yale Face Dataset [4] we evaluate these methods under controlled and realistic acquisition conditions. The evaluation focuses on robustness to common image degradations, including illumination changes, additive noise, blur, and partial occlusions, as well as sensitivity to limited training data. Quantitative results are complemented by qualitative analyses such as subspace visualizations, confusion matrices, and representative success and failure cases, providing insight into the strengths and limitations of global versus local visual representations. Overall, the study highlights the interpretability, failure modes, and practical relevance of non-deep-learning face recognition techniques. The full code is available at `https://github.com/fotisk07/VIC-Project`.

1. Introduction and Motivation

Face recognition is a fundamental problem in computer vision with applications ranging from access control, security, identity verification, to human - computer interaction. The goal of face recognition is to automatically identify or verify a person based on visual information captured by devices such as cameras. We will not focus on deep learning-based methods, which have become dominant in recent years, but rather on classical appearance-based techniques which are of practical interest due to their interpretability, low computational requirements, and minimal training data requirements.

This project focues on three well-established classical face recognition methods: Eigenfaces, based on Principal Component Analysis (PCA) [10]; Fisherfaces, based on Linear Discriminant Analysis (LDA) [2]; and Local Binary Pattern Histograms (LBPH) [1], which encode local texture information. These methods embody fundamentally different modeling philosophies, ranging from global linear subspace representations to local, texture-based descriptors.

We aim to provide a comprehensive comparative evaluation of these methods under varying image acquisition conditions and degradations, such as illumination changes, noise, blur, and partial occlusions. The goal is to determine how robust are these classical techniques to real-world variations and how their performance degrades under those conditions. We will also analyze their sensitivity to limited training data, which is a common practical constraint.

Understanding these questions is important not only for appreciating the evolution of face recognition techniques but also for informing the design of lightweight, interpretable systems in constrained environments or low-data scenarios. Potential applications may include embedded systems, mobile devices, or baseline models for benchmarking and analysis.

2. Problem Definition

We consider a supervised face recognition problem defined over a labeled dataset of face images. Let

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

denote a dataset of N grayscale face images, where $x_i \in \mathbb{R}^{H \times W}$ represents the i-th image of height H and width W, and $y_i \in \{1, 2, \ldots, C\}$ is the corresponding identity label indicating one of C distinct individuals. Each image is assumed to be preprocessed to a standard size and aligned such that facial features are approximately centered, but may vary in terms of illumination, pose, expression, noise level, or partial occlusions.

Given a training subset $\mathcal{D}_{train} \subset \mathcal{D}$, the task is to learn a mapping function

$$f : \mathbb{R}^{H \times W} \to \{1, 2, \ldots, C\}$$

that predicts the identity label of previously unseen test images.

To evaluate robustness, we consider a family of image degradation operators

$$\delta_\alpha : \mathbb{R}^{H \times W} \to \mathbb{R}^{H \times W}$$

parameterized by a severity level α, modeling effects like illumination changes, additive Gaussian noise, blur, rotation, etc. The primarily objective is always to maximize recognition accuracy

$$\text{Accuracy} = \frac{1}{|\mathcal{D}_{test}|} \sum_{(x_j,y_j)\in\mathcal{D}_{test}} \mathbb{I}(f(\delta_\alpha(x_j)) = y_j)$$

on the test set $\mathcal{D}_{test}$ and to analyze how this accuracy degrades with respect to the severity α of the applied degradations.

3. Related Work

The Eigenfaces method introduced by Turk and Pentland applies Principal Component Analysis (PCA) to project face images into a low-dimensional subspace that captures the main modes of variation [10]. While effective in controlled settings, Eigenfaces remain highly sensitive to illumination changes.

Fisherfaces extend this approach by using Linear Discriminant Analysis (LDA) to maximize class separability, leading to improved robustness under varying lighting conditions, provided that sufficient training samples are available [2]. Unlike PCA, which minimizes total reconstruction error, LDA explicitly discriminates between classes, making it more suitable for classification tasks where lighting variation exceeds identity variation.

Local Binary Pattern Histograms (LBPH) represent faces using local texture descriptors computed over small neighborhoods [1]. This local representation has been shown to be more robust to lighting variations and partial occlusions compared to holistic methods, albeit at the expense of increased sensitivity to noise.

Significant advancements in face recognition have recently been driven by deep convolutional neural networks (CNNs), which learn hierarchical feature representations directly from raw pixels. A pivotal shift occurred with the introduction of DeepFace, which approached human-level performance by utilizing a deep CNN trained on a massive dataset to learn generic face representations [9]. This was followed by FaceNet, which introduced the triplet loss function to map face images directly into a compact Euclidean space where distances correspond to face similarity, unifying representation and verification [8]. More recently, margin-based loss functions, such as those used in ArcFace, have further enhanced discrimination by enforcing angular margins between classes in the embedding space, significantly improving performance on unconstrained large-scale benchmarks [3].

4. Methodology

4.1. Face Recognition Algorithms

4.1.1 Eigenfaces

The Eigenfaces method, introduced by Turk and Pentland [10], is based on Principal Component Analysis (PCA). The core idea is to represent face images in a low-dimensional linear subspace that captures the dominant modes of variation present in the training data.

Each grayscale face image $x_i \in \mathbb{R}^{H\times W}$ is vectorized into a column vector $\mathbf{x}_i \in \mathbb{R}^D$, where $D = H \cdot W$. Given a training set of N images, the mean face is computed as

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i.$$

The centered data vectors are then defined as $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}$ and stacked into a data matrix

$$X = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_N] \in \mathbb{R}^{D\times N}.$$

PCA aims to find an orthonormal basis that maximizes the variance of the projected data. This is achieved by computing the eigenvectors of the covariance matrix

$$C = \frac{1}{N} XX^\top.$$

Since the image dimensionality D is typically much larger than the number of training samples N, the eigendecomposition is efficiently performed using the smaller matrix $X^\top X$. The leading K eigenvectors of C, corresponding to the largest eigenvalues, define the principal subspace. When reshaped back to image form, these eigenvectors are referred to as *eigenfaces*.

A face image $\mathbf{x}$ is projected onto the eigenface subspace by

$$\mathbf{z} = U_K^\top(\mathbf{x} - \boldsymbol{\mu}),$$

where $U_K \in \mathbb{R}^{D\times K}$ contains the top K eigenvectors. Face recognition is then performed by comparing projected feature vectors using a distance metric such as Euclidean distance, typically within a nearest-neighbor classification framework.

4.1.2 Fisherfaces

Fisherfaces [2] is a discriminative subspace method that combines Principal Component Analysis (PCA) with Linear Discriminant Analysis (LDA). It learns a linear projection that maximizes class separability by maximizing the ratio of between-class to within-class scatter.

Let each training image be vectorized as $\mathbf{x} \in \mathbb{R}^n$ and let there be c subjects. Denote by $\boldsymbol{\mu}_i$ the mean of class i, and

by $\boldsymbol{\mu}$ the global mean. The scatter matrices are

$$S_B = \sum_{i=1}^{c} N_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \tag{1}$$

$$S_W = \sum_{i=1}^{c} \sum_{\mathbf{x}_k \in X_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T. \tag{2}$$

The LDA directions solve the generalized eigenproblem $S_B\mathbf{w} = \lambda S_W\mathbf{w}$, with at most $c-1$ discriminant directions.

In face recognition, $n \gg N$, making S_W singular in the original image space. Fisherfaces therefore first projects the data with PCA to a subspace of dimension $k_{\text{pca}} \leq N - c$, then applies LDA in that PCA space to obtain up to $k_{\text{lda}} \leq c-1$ discriminant components. The final projection is $W = W_{\text{pca}}W_{\text{lda}}$.

We use 1-nearest neighbor (1-NN) with Euclidean distance in Fisherface space.

In experiments, we report both a paper-aligned setting ($k_{\text{pca}} = N - c$, $k_{\text{lda}} = c - 1$ when feasible) and stability-oriented variants, such as capping k_{pca} and adding a small ridge term to the within-class scatter in PCA space.

4.1.3 Local Binary Pattern Histograms (LBPH)

LBPH is a texture-based face recognition method that encodes local appearance information while preserving spatial structure. This approach was introduced by Ahonen et al. [1], motivated by the observation that face images can be viewed as compositions of local texture patterns, such as edges, spots, and flat areas.

The core component of LBPH is the Local Binary Pattern (LBP) operator introduced by Ojala et al. [5]. Applied to a grayscale image, the LBP operator is computed at each pixel location (x, y) by thresholding the pixel's neighborhood against its center value. In the basic 3×3 case, the operator compares the center pixel intensity g_c with its 8 surrounding neighbors $\{g_p\}_{p=0}^{7}$, producing a binary code:

$$\text{LBP}(x,y) = \sum_{p=0}^{7} s(g_p - g_c)2^p, \quad s(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{3}$$

This formulation was later generalized to circular neighborhoods P sampling points on a circle of radius R [6], allowing for multi-scale texture analysis. Another important extension to the original LBP operator is the concept of uniform patterns [6], which reduces the number of possible LBP codes by focusing on patterns with at most two bitwise transitions. This significantly decreases the dimensionality of the resulting histograms while retaining discriminative power.

While a global LBP histogram captures texture information, it discards spatial layout, which is crucial for face recognition. To address this, the LBPH method divides the face image into m distinct rectangular regions $\{R_j\}_{j=0}^{m-1}$. For each region, an LBP histogram $H_{i,j}$ is computed:

$$H_{i,j} = \sum_{(x,y)} \mathbb{I}\{\text{LBP}(x,y) = i\}\mathbb{I}\{(x,y) \in R_j\} \tag{4}$$

where i indexes the LBP labels. The final feature vector for the face image is obtained by concatenating all regional histograms into a single feature vector. This representation encodes texture information at three different levels: pixel-level (via LBP codes), region-level (via histograms), and global-level (via concatenation).

Face recognition is then performed using a nearest-neighbor classifier in the histogram feature space. Given two feature vectors S and M, many distance metrics can be used to measure similarity, such as histogram intersection, log-likelihood, and χ^2 distance. We ultimately use the χ^2 distance, due its better performance in practice [1]:

$$\chi^2(S, M) = \sum_{i,j} \frac{(S_{i,j} - M_{i,j})^2}{S_{i,j} + M_{i,j}} \tag{5}$$

Moreover, the χ^2 formulation naturally extends to a weighted version, allowing for differential emphasis on certain regions if desired:

$$\chi^2_w(S, M) = \sum_{i,j} w_j \frac{(S_{i,j} - M_{i,j})^2}{S_{i,j} + M_{i,j}} \tag{6}$$

As explained, LPBH involves several hyperparameters that can be tuned to optimize performance for a given dataset and application. These include the number of sampling points P and radius R for the LBP operator, the number of regions m to divide the face image into, etc. Ahonen et al. show that LBPH performance is relatively insensitive to moderate changes in these parameters, offering a favorable trade-off between recognition accuracy and feature dimensionality. Despite its robustness to illumination changes, facial expressions, and moderate misalignment, LBPH has notable limitations. The concatenated histograms can become high-dimensional, leading to increased memory usage and slower matching for large datasets. Furthermore, the method relies on handcrafted features and nearest-neighbor classification, limiting its ability to model complex intra-class variations.

4.2. Datasets

We conduct our experiments on two standard face recognition datasets that are widely used in the evaluation of classical appearance-based methods.

ORL (AT&T) Face Dataset. The ORL face dataset [7] contains images of 40 individuals, with 10 grayscale images per subject. The images exhibit moderate variations in facial expression, pose, and the presence of accessories such as glasses. Due to its controlled nature and limited variability, this dataset is well suited for analyzing baseline performance and the impact of training set size on recognition accuracy.

Yale Face Dataset. The Yale face dataset [4] includes frontal face images captured under systematically varying illumination conditions. This dataset is particularly challenging for global appearance-based methods and is therefore well suited for studying robustness to lighting changes. It provides a classical benchmark for evaluating the effectiveness of discriminative subspace methods and local texture-based representations.

5. Evaluation

5.1. Global Comparison of Methods

We now provide a global comparison between the three classical face recognition methods. For both methods, we evaluate recognition accuracy as a function of the number of training images per subject, using a fixed test protocol and multiple random splits. Reported values correspond to mean accuracies over repeated experiments.

ORL dataset. Figure 1 reports the accuracy obtained on the ORL dataset as the number of training images per subject increases from 1 to 7.

Eigenfaces exhibit a strong dependency on the amount of available training data. With very limited supervision (1-2 images per subject), performance is relatively low and unstable, reflecting the difficulty of estimating a meaningful global PCA subspace from few samples. Accuracy improves steadily as more images are available, eventually reaching a plateau around 92–93% for 6-7 training images per subject.

In contrast, LBPH consistently outperforms Eigenfaces across all training regimes. Even in the low-data setting, LBPH achieves higher accuracy, and its performance saturates earlier, exceeding 97% accuracy from 5 training images per subject onward. This behavior highlights the robustness of local texture-based representations when only limited data is available.

Yale dataset. Figure 2 presents the same comparison on the Yale dataset, which is more challenging due to stronger illumination variations and facial expression changes.

Overall accuracies are lower than on ORL for both methods. Eigenfaces show clear saturation around 82%, even as

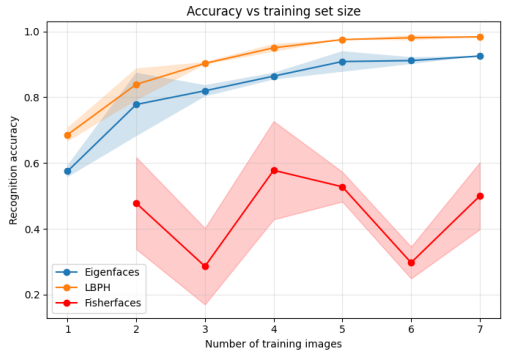


Figure 1: Recognition accuracy on the ORL dataset as a function of the number of training images per subject. Eigenfaces and LBPH are compared using identical evaluation protocols.

the number of training images increases, indicating a limited capacity to model the variability present in the dataset. Increasing the training set beyond 6 images per subject brings little to no improvement.

LBPH again demonstrates superior robustness. While the gap between the two methods is smaller than on ORL, LBPH consistently achieves higher accuracy, reaching approximately 84–85% at saturation. This confirms that local descriptors are less sensitive to global appearance changes such as illumination, which strongly affect PCA-based representations.

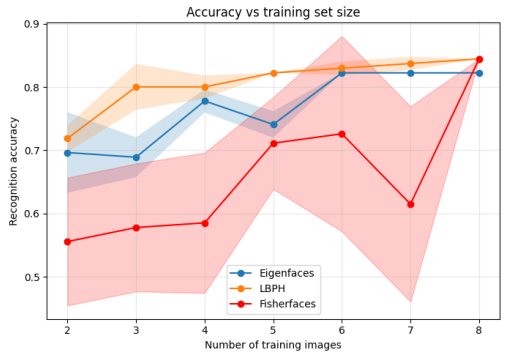


Figure 2: Recognition accuracy on the Yale dataset as a function of the number of training images per subject. The dataset exhibits stronger illumination and expression variability, resulting in lower overall performance.

Discussion. Across both datasets, LBPH dominates Eigenfaces in all training regimes. The advantage is particularly pronounced in low-data settings and on datasets with strong appearance variability. Eigenfaces rely on a global linear subspace and precise pixel-level alignment, making

them sensitive to limited data and distribution shifts. LBPH, by contrast, encodes local texture patterns and is therefore more robust to such variations.

5.2. Robustness Analysis by Method

5.2.1 Eigenfaces

Influence of the number of principal components. We first analyze the sensitivity of Eigenfaces to the number of retained principal components K. The results can be seen in 3. This parameter directly controls the dimensionality of the PCA subspace.

Across both ORL and Yale datasets, we observe a rapid improvement in recognition accuracy as K increases from very small values. However, this gain quickly saturates: beyond approximately $K = 50$, increasing the number of components yields marginal or no improvement. In some cases, performance even slightly fluctuates due to overfitting to dataset-specific variations. This behavior indicates that most discriminative identity information is captured by the leading eigenvectors, and that retaining additional components mainly encodes noise, illumination changes, or other non-discriminative factors.

As a result, all subsequent Eigenfaces experiments are conducted using values of K within this saturation regime, ensuring that observed failures cannot be attributed to an under-parameterized representation.

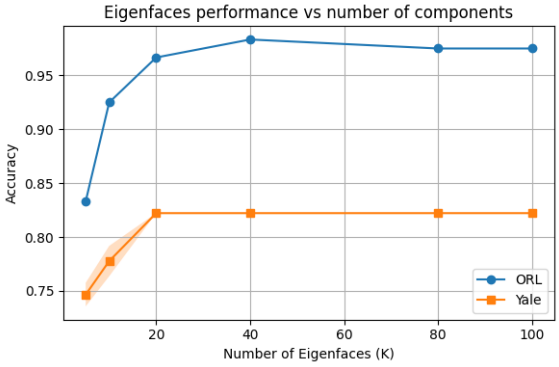


Figure 3: Impact of K on ORL and Yale dataset

Global robustness evaluation protocol. To assess the robustness of Eigenfaces, we evaluate its performance under a range of controlled image transformations designed to violate the method's core assumptions. For each transformation, we apply the same perturbation to both training and test images and report mean accuracy over multiple random train–test splits with a fixed test set.

Figure 4 presents a global comparison of several transformation families, including Gaussian blur, photometric jitter with impulsive noise, random rotations, and random cropping. These transformations probe complementary failure modes such as loss of high-frequency information, photometric instability, and geometric misalignment.

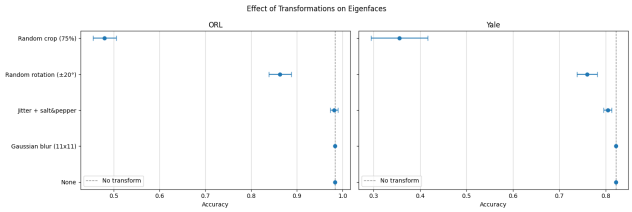


Figure 4: Impact of transformations on the accuray of Eigenfaces

Breakdown by transformation type. The global comparison reveals that Eigenfaces are particularly sensitive to geometric perturbations. Random rotations and random crops cause the most severe drops in accuracy on both datasets, even at moderate transformation strengths. This confirms that Eigenfaces rely heavily on strict pixel-level alignment and lack any form of geometric invariance. In contrast, Gaussian blur induces almost n not degradation, suggesting that while high-frequency details contribute to identity discrimination, they are not the sole determining factor.

Photometric jitter combined with salt-and-pepper noise also significantly impacts performance and leads to increased variance across runs. This reflects the global nature of PCA-based representations: localized corruption or intensity shifts affect the entire projected feature vector.

5.2.2 Fisherfaces

Experimental protocol (what our curves mean). Unless stated otherwise, Fisherfaces is evaluated using: (i) a **fixed test set** (fixed test indices per dataset), (ii) varying the number of training images per subject (train size), (iii) repeated experiments over multiple random training draws from the remaining pool (`n_exp`) to estimate a mean accuracy and a 95% confidence interval, and (iv) a **1-NN Euclidean classifier** in Fisherface space. For robustness studies, degradations (rotation, Gaussian noise, blur, flips, brightness scaling) are applied **only to test images**, keeping training images unchanged.

Core empirical phenomenon: dimensionality-driven "peaking". A consistent and dominant effect across our Fisherfaces experiments is the impact of how the PCA dimension k_{pca} is chosen. In the paper-aligned Fisherfaces construction, $k_{\text{pca}} = N - c$ grows with the number of training samples N. In practice, this means that as we increase

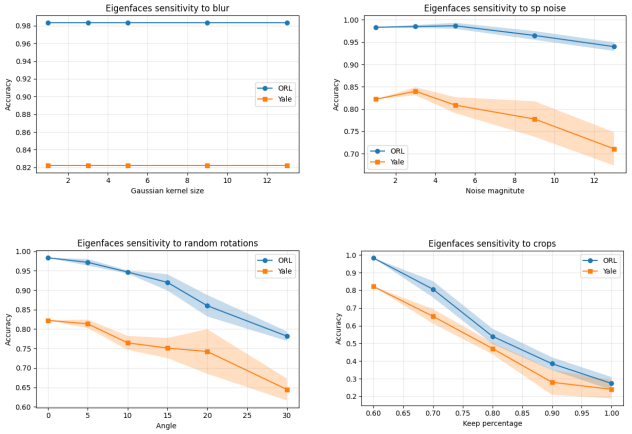


Figure 5: Severity analysis of Eigenfaces under different transformation families. Top-left: Gaussian blur. Top-right: photometric jitter with salt-and-pepper noise. Bottom-left: random rotations. Bottom-right: random cropping. Accuracy is reported as a function of transformation severity (mean $\pm$ standard deviation).

the number of training images per subject, the PCA subspace dimension increases automatically.

Empirically, this can induce a **peaking effect** (also known as a Hughes-type phenomenon in small-sample discriminant analysis): beyond some point, adding degrees of freedom increases estimation variance and/or numerical instability in the subsequent LDA step and can *reduce* generalization accuracy, even though more training data is available. Importantly, we verified this is **not** merely an artifact of random train/test splits by using **nested training sets** (train sets that grow by inclusion). Under nested splits, the non-monotonicity remains for the baseline choice $k_{\text{pca}} = N - c$, confirming that the behavior is primarily methodological.

ORL (AT&T) dataset: baseline Fisherfaces is unstable, fixed PCA resolves it. On ORL (40 subjects, 10 images/subject), the baseline Fisherfaces curve (with the default $k_{\text{pca}} = N - c$ growth) is **low and non-monotone**. Even when averaged across repetitions, the mean accuracy remains modest and the confidence intervals overlap heavily across train sizes, indicating that "more training images" does not reliably improve performance under this configuration.

In contrast, when we **fix (or cap)** k_{pca} **across train sizes** (i.e., keep the PCA subspace dimension constant while increasing the number of training images), Fisherfaces becomes **highly stable and almost monotone**, reaching substantially higher accuracy (in our plots, approximately in the 0.9 range for most train sizes). This provides strong evidence that the main driver of poor/non-monotone ORL performance is the **dimension growth** of the PCA stage, not the dataset being "in-the-wild" or the split variance.

Strict vs regularized vs normalized variants on ORL. We additionally compare controlled Fisherfaces variants:

- **Strict (paper-aligned) Fisherfaces** can perform reasonably at smaller train sizes but exhibits sharp drops at larger sizes in our ORL experiments. This indicates that, despite the theoretical nonsingularity argument in the PCA space, the estimated within-class scatter $\widetilde{S}_W$ can still be **numerically ill-conditioned**, making the generalized eigenproblem unstable without regularization.
- **Regularization** of $\widetilde{S}_W$ improves stability compared to strict mode but does not, by itself, eliminate the non-monotonicity caused by increasing k_{pca}.
- **Per-image normalization** provides only modest and inconsistent gains on ORL; it does not fix the dominant dimension-driven instability.

Yale dataset: Fisherfaces benefits from more data, but fixed PCA still helps. On Yale (15 subjects, systematic illumination changes and expression/accessory variations), Fisherfaces shows a clearer **increase in accuracy with train size** under the baseline configuration, and the confidence intervals typically tighten as train size grows. This behavior is consistent with Fisherfaces' objective: LDA explicitly aims to reduce within-class scatter caused by nuisance variations (notably illumination) while maintaining between-class separation.

However, even on Yale, fixing/capping the PCA dimension yields the **best overall performance and the most monotone behavior**. In our experiments, the gap between baseline and fixed-PCA variants is smaller than on ORL, indicating that Yale is *less sensitive* to the peaking effect under $k_{\text{pca}} = N - c$ growth, but the effect is still measurable.

Why can Yale look "easier" than ORL for Fisherfaces despite being described as challenging? A key objective explanation is the interaction between (i) the **number of classes** and (ii) the induced PCA dimension growth:

- ORL has $c = 40$ subjects; for a fixed number of training images per subject, N and thus $N - c$ become large quickly, increasing k_{pca} and amplifying the peaking effect.
- Yale has $c = 15$ subjects, so $N - c$ grows more slowly and the discriminant estimation is less fragile in our regime.

Thus, Yale can be "challenging" for naive global appearance methods (e.g., raw PCA/Eigenfaces) because illumination dominates, while being comparatively favorable to Fisherfaces which is designed to reduce within-class illumination scatter.

Robustness to degradations (test-only transforms). When applying degradations only to the test set, Fisherfaces exhibits characteristic sensitivities:

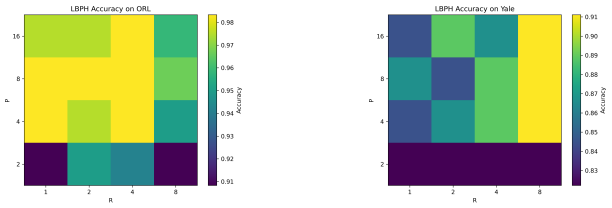
- **Rotation:** accuracy drops quickly with moderate rotations, indicating sensitivity to geometric misalignment (expected for global linear subspace models).
- **Blur:** increasing blur degrades performance, consistent with the loss of high-frequency discriminative information in global appearance representations.
- **Flips:** horizontal flips degrade accuracy; vertical flips severely break recognition (strong distribution shift).
- **Brightness scaling:** strong deterioration for extreme brightness factors, especially without per-image normalization.
- **Gaussian noise:** within the tested noise range, the degradation can be less pronounced than rotation/blur/flip, suggesting that geometric and photometric shifts dominate the failure modes.

Summary of Fisherfaces conclusions. Across both ORL and Yale, our most robust conclusion is that **controlling the PCA dimension is the primary lever** for Fisherfaces stability and accuracy. The baseline, paper-aligned choice $k_{\text{pca}} = N - c$ is theoretically motivated (to avoid singular S_W), but in practice it can produce non-monotone performance via dimension-driven peaking and numerical conditioning effects. A fixed/capped PCA variant yields more stable, interpretable trends with training size, while strict (fully paper-aligned) Fisherfaces is best treated as a reference configuration rather than the most reliable practical setting in our experimental regime.

5.2.3 Local Binary Pattern Histogram

Influence of the parameters P and R. We first analyze the sensitivity of LBPH to its key hyperparameters: the number of sampling points P and the radius R of the circular neighborhood. These parameters control the granularity and scale of local texture encoding. The result for the ORL and Yale datasets can respectively be seen in 6(a) and 6(b).

On both datasets, we observe that LBPH performance is relatively robust to moderate variations in P and R. Recognition accuracy remains high across a wide range of parameter combinations, but the best performance is typically



(a) ORL dataset (b) Yale dataset

Figure 6: Impact of LBPH parameters P (number of sampling points) and R (radius) on recognition accuracy.

achieved for P in the range of 4 to 8 and R in the range of 2 to 4 pixels. This suggests that capturing local texture information at a moderate scale is sufficient, and even optimal, for discriminating between subjects in these datasets. Extreme values of P and R tend to slightly degrade performance, likely due to either insufficient local detail or excessive smoothing of texture patterns. To mitigate between performace and computational cost, we select $P = 8$ and $R = 2$ for all subsequent LBPH experiments, as these values consistently yield near-optimal accuracy across both datasets while keeping the feature dimensionality manageable.

Global robustness evaluation protocol and Results by transformation type. We evaluate the robustness of LBPH using the same global protocol as for Eigenfaces, applying controlled transformations to the test set and measuring accuracy degradation. We consider 4 transformation families: Gaussian blur, photometric jitter with salt-and-pepper noise, random rotations, and random cropping. The results are presented in Figure 7.

Our results indicate that LBPH behaves differently from Eigenfaces and Fisherfaces under these transformations. LBPH is seems sensitive to Gaussian blur, especially on the Yale dataset where the performance decreases to around 60% at the highest blur level.

However, LBPH is relatively robust to photometric jitter and salt-and-pepper noise, with only drops on the Yale dataset at the highest noise levels. This suggests that the local texture encoding is less affected by global intensity shifts and localized pixel corruption compared to global appearance-based methods.

For geometric transformations, LBPH shows robustness to rotation until the angle reaches around 15 degrees, beyond which accuracy slightly degrades but remains above 75% even at 30 degrees. This indicates that while LBPH does not have explicit geometric invariance, its local encoding provides some tolerance to moderate misalignment.

Random cropping has the most significant impact on LBPH, on both datasets, as it can remove critical facial re-

gions and disrupt the spatial structure of the histograms. This transformation achieves the single most severe degradation for LBPH, making it a key failure mode to consider in practical applications.

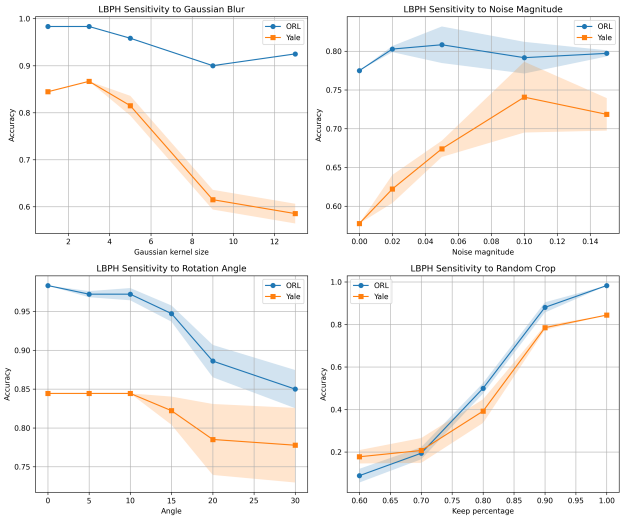


Figure 7: Impact of transformations on the accuracy of LBPH.

5.3. Deep Learning

We additionally evaluate a simple convolutional neural network as a deep learning baseline. The model (`TinyCNN`) consists of two convolutional layers followed by max-pooling, adaptive average pooling, and two fully connected layers. It operates on grayscale images and is trained from scratch using cross-entropy loss and the Adam optimizer, without data augmentation or pretraining.

Figure 8 and Figure 9 report recognition accuracy as a function of the number of training images per subject on the ORL and Yale datasets.

Results. On ORL, the deep learning model improves as the number of training images increases and reaches reasonable performance for larger training sets, but remains inferior to LBPH and Eigenfaces in the low-data regime. We hypothesize that this behavior is due to overfitting, as ORL exhibits limited variability and therefore does not reflect realistic generalization performance.

On Yale, performance remains significantly lower across all training regimes and exhibits high variance. This confirms our hypothesis regarding the limitations observed on ORL: deep learning models require substantially larger amounts of data to generalize effectively and are not well suited to the present task.

Figure 8: Recognition accuracy on the ORL dataset as a function of the number of training images per subject.

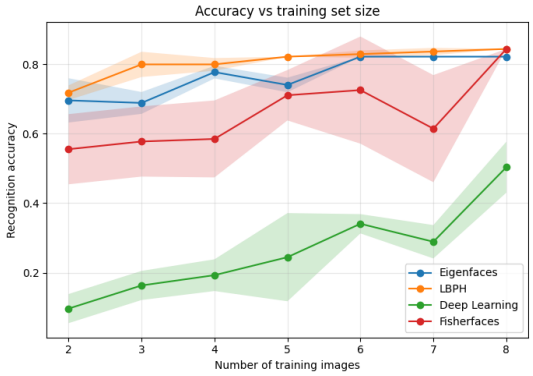


Figure 9: Recognition accuracy on the Yale dataset as a function of the number of training images per subject.

Discussion. Overall, these results highlight the strong data dependency of deep learning approaches. While convolutional models have high representational capacity, they require larger datasets or additional regularization strategies to outperform classical methods in small-sample face recognition settings.

6. Conclusions

In this study, we presented a comprehensive comparative evaluation of three classical face recognition methods, namely Eigenfaces, Fisherfaces, and Local Binary Pattern Histograms (LBPH), alongside a simple deep learning baseline. By testing these algorithms on the ORL and Yale datasets, we assessed their robustness to real-world degradations and their dependency on training data volume.

Our findings demonstrate a clear divide between global and local appearance models. Global methods, such as Eigenfaces, are highly sensitive to geometric perturbations (like rotations and cropping) and struggle with illumination changes. While Fisherfaces mitigate some of these photo-

metric issues by maximizing class separability, our experiments revealed that its stability is heavily dependent on controlling the PCA dimension. Conversely, the local texture-based LBPH method consistently outperformed the global approaches. It proved highly resilient in low-data scenarios and under significant illumination variations, though it remains vulnerable to severe spatial cropping and heavy blurring.

Finally, the evaluation of our CNN baseline underscored the limitations of deep learning in small-sample face recognition tasks. Trained from scratch on limited data, the network suffered from severe overfitting and failed to generalize effectively. Ultimately, this project highlights that while deep learning dominates large-scale vision tasks, classical methods still hold significant value in constrained settings, and that understanding their failure modes is crucial for designing robust face recognition systems.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision (ECCV)*, pages 469–481, 2004.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[4] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[5] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[7] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of the Second IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.

[8] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[10] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991.

Proposal (to be removed later)

1. Motivation and Problem Definition

Face recognition is a fundamental problem in computer vision with applications in security, identity verification, and human - computer interaction.

In this project, we focus on appearance-based face recognition techniques developed prior to deep learning. These methods rely on explicit feature extraction and linear subspace modeling.

The objective of this project is to compare three classical face recognition methods:

- Eigenfaces (PCA-based)[10]
- Fisherfaces (LDA-based)[2]
- Local Binary Pattern Histograms (LBPH) [1]

and to analyze their robustness under realistic image degradations such as illumination changes, noise, blur, and partial occlusions.

Problem statement: Given a labeled face image dataset, how do different classical face recognition algorithms perform under varying acquisition conditions, and what do their successes and failures reveal about global versus local visual representations?

2. Related Work

The Eigenfaces method introduced by Turk and Pentland applies Principal Component Analysis (PCA) to project face images into a low-dimensional subspace that captures the main modes of variation. While effective in controlled settings, Eigenfaces are highly sensitive to illumination changes.

Fisherfaces extend this approach by using Linear Discriminant Analysis (LDA) to maximize class separability, leading to improved robustness under varying lighting conditions, provided that sufficient training samples are available.

Local Binary Pattern Histograms (LBPH) represent faces using local texture descriptors computed over small neighborhoods. This local representation has been shown to be more robust to lighting variations and partial occlusions, at the expense of increased sensitivity to noise.

3. Methodology

3.1. Algorithms

We will implement and evaluate the following methods:

- **Eigenfaces**: PCA-based dimensionality reduction followed by nearest-neighbor classification.
- **Fisherfaces**: LDA applied after PCA to improve class discrimination.
- **LBPH**: Local Binary Pattern feature extraction with histogram-based comparison.

3.2. Datasets

We conduct our experiments on two standard face recognition datasets that are widely used in the evaluation of classical appearance-based methods.

ORL (AT&T) Face Dataset. The ORL face dataset [7] contains images of 40 individuals, with 10 grayscale images per subject. The images exhibit moderate variations in facial expression, pose, and the presence of accessories such as glasses. Due to its controlled nature and limited variability, this dataset is well suited for analyzing baseline performance and the impact of training set size on recognition accuracy.

Yale Face Dataset. The Yale face dataset [4] includes frontal face images captured under systematically varying illumination conditions. This dataset is particularly challenging for global appearance-based methods and is therefore well suited for studying robustness to lighting changes. It provides a classical benchmark for evaluating the effectiveness of discriminative subspace methods and local texture-based representations.

4. Evaluation

The proposed methods are evaluated using a combination of quantitative and qualitative analyses, with an emphasis on robustness and interpretability rather than raw recognition performance.

Experimental Protocol. For each dataset, we vary the number of training images per subject in order to study sample efficiency and sensitivity to limited supervision. Test images are kept fixed across experiments to ensure fair comparison between methods. When applicable, results are averaged over multiple random training splits.

To assess robustness, controlled degradations are applied to the test images only. These include illumination changes, additive Gaussian noise, image blur, and partial occlusions. Each degradation is applied at increasing levels of severity, yielding robustness curves that characterize performance decay under progressively more challenging conditions.

Evaluation Metrics. Performance is primarily measured using recognition accuracy, complemented by confusion

matrices to analyze class-specific failure patterns. Robustness curves plot recognition accuracy as a function of degradation strength, providing a compact visualization of stability under adverse conditions.

Qualitative Analysis. Beyond quantitative metrics, we perform qualitative analysis by visualizing learned Eigenfaces and Fisherfaces, as well as representative success and failure cases. These visualizations help interpret what variations are captured by global subspace methods (illumination, identity, or noise) and contrast them with the locality-driven behavior of LBP-based descriptors, thereby explaining observed performance trends.

Reference Comparison with Deep Learning Methods. For reference, we also include a comparison with a lightweight deep learning-based face recognition model. This comparison is not intended to achieve state-of-the-art performance, but to provide contextual insight into the gap between classical appearance-based methods and modern learned representations, particularly under unconstrained acquisition conditions. The deep learning model is evaluated using the same experimental protocol and test sets when possible, and its results are reported solely as a point of reference.

5. Expected Contributions

The project aims to deliver:

- A reproducible experimental comparison of classical face recognition methods
- An analysis of global versus local visual representations
- Clean, well-documented code and experimental protocols