Probabilistic Visual Learning for Object Detection, a Summary

Daniel Ramyar, Tai Thorsen, Andrew Khoudi

March 6, 2019

Introduction

This paper compares the performance of two object detection techniques the first being based on "distance-from-feature-space" DFFS and later maximum likelihood. The training images are decomposed using principal component analysis and the eigenvalues and eigenvectors are then extracted.

This makes it possible to determine whether a new image with an object is a member of this training data class Ω by extracting the unique coefficients $\mathbf{y}$ representing the new object.

PCA & Maximum likelihood detection

In order to do object detection we first need to decompose our N images which are of size $m \times n$ pixels into $1 \times k$ arrays where $k = m \cdot n$. When all the images are in 1D array form we stack them to form a new data matrix of size $N \times k$.

From this data matrix we can calculate the covariance matrix Σ. Using Σ along with the eigenvector matrix of Σ, Φ, the corresponding diagonal matrix of eigenvalues, Λ is given by equation (1):

$$\Lambda = \Phi^T \Sigma \Phi \tag{1}$$

A PCA analysis is then peformed by using a Karhunen-Loeve tranform (KL) to extract the eigenvectors that corresponds to the largest eigenvalues in Λ.

From this we get a principal component feature vector $\mathbf{y} = \Phi_M^T \tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$ is the mean-normalized image vector, and Φ_M is a submatrix of Φ containing the M principal eigenvectors.

The vectorspace spanned by our principal eigenvectors, we call the feature space (or principal subspace), $F = \{\Phi_i\}_{i=1}^{M}$, and the remaining $N-M$ eigenvectors then span an orthogonal subspace $\bar{F} = \{\Phi_i\}_{i=M+1}^{N}$.

In a partial KL expansion, the residual reconstruction error is defined as

$$\epsilon^2(\mathbf{x}) = \sum_{i=M+1}^{N} y_i^2 = ||\tilde{\mathbf{x}}||^2 - \sum_{i=1}^{M} y_i^2 \tag{2}$$

and can be computed from the first M principal components along with the quadature sum of the mean-normalized image $\tilde{\mathbf{x}}$. The "distance-from-feature-space" (DFFS) is defined as the component in the orthogonal subspace $\bar{F}$, and is equivalent to the residual error $\epsilon^2(\mathbf{x})$ in Eq.(2). The component of $\mathbf{x}$ which lies in the feature space F is referred to as the "distance-in-feature-space" (DIFS).

Unimodal F-space Densities

By assuming that we have a good estimate of the mean $\bar{\mathbf{x}}$ and covariance Σ of the distribution from the given training set $\mathbf{x}^t$, the likelihood of an input pattern $\mathbf{x}$, for a high-dimensional unimodal Gaussian density, is given by

$$P(\mathbf{x}|\Omega) = \frac{e^{-\frac{1}{2}d(\mathbf{x})}}{(2\pi)^{N/2}|\Sigma|^{1/2}} \tag{3}$$

where $d(\mathbf{x})$ is the Mahalanobis distance, defined as:

$$\begin{aligned} d(\mathbf{x}) &\equiv \tilde{\mathbf{x}}^T \Sigma^{-1} \tilde{\mathbf{x}} = \tilde{\mathbf{x}}^T \left(\Phi \Lambda^{-1} \Phi^T\right) \tilde{\mathbf{x}} \\ &= \mathbf{y}^T \Lambda^{-1} \mathbf{y} = \sum_{i=1}^{N} \frac{y_i^2}{\lambda_i} \end{aligned} \tag{4}$$

From this sum an estimator of $d(\mathbf{x})$ is made, using the DFFS in equation (2), as follows

$$\hat{d}(\mathbf{x}) = \sum_{i=1}^{M} \frac{y_i^2}{\lambda_i} + \frac{1}{\rho}\epsilon^2(\mathbf{x}) \tag{5}$$

And the likelihood estimator from $\hat{d}(\mathbf{x})$ then becomes:

$$\begin{aligned} \hat{P}(x|\Omega) &= \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^M \prod_{i=1}^{M}\lambda^{1/2}}\right] \cdot \left[\frac{\exp\left(-\frac{\epsilon^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] \\ &= P_F(x|\Omega)\hat{P}_{\bar{F}}(x|\Omega) \end{aligned} \tag{6}$$

where $P_F(\mathbf{x}|\Omega)$ is the true marginal density in F-space and $P_{\bar{F}}(\mathbf{x}|\Omega)$ is the estimated marginal density in the orthogonal complement $\bar{F}$-space. The optimal value of ρ can now be determined by minimizing a suitable cost function $J(\rho)$, given by the Kullback-Leibler divergence:

$$\begin{aligned} J(\rho) &= E\left[\log\frac{P(\mathbf{x}|\Omega)}{\hat{P}(\mathbf{x}|\Omega)}\right] \\ &= \frac{1}{2}\sum_{i=M+1}^{N}\left[\frac{\lambda_i}{\rho} - 1 + \log\frac{\rho}{\lambda_i}\right] \end{aligned} \tag{7}$$

And this is minimized with respect to ρ, with the optimal weight ρ^*, that given by the arithmetic average of the eigenvalues in the orthogonal subspace $\bar{F}$:

$$\rho^* = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i \tag{8}$$

Multimodal F-space Densities

The data is seldom unimodal, and tends to lie on complex manifolds in image space. If we assume that the $\bar{F}$-space components are Gaussian and independent of the principal features in F, the DFFS remains the residual $\epsilon^2(\mathbf{x})$, while the DIFS$= -\log P(\mathbf{y})$ where $P(\mathbf{y})$ describes $P(\mathbf{x}|\Omega)$ as an arbitrary density in the principal component vector $\mathbf{y}$. The density $P(\mathbf{y})$ can be modelled using a Mixture-of-Gaussians:

$$P(\mathbf{y}|\Theta) = \sum_{i=1}^{N_c} \pi_i g(\mathbf{y}; \mu_i, \Sigma_i) \tag{9}$$

where $g(\mathbf{y}; \mu, \Sigma)$ is an M-dimensional Gaussian density, with mean vector μ and covariance matrix Σ, and π_i are the weights of the N_c gaussians, satisfying $\sum \pi_i = 1$. Θ is a set of parameters that completely specifies this mixture, $\Theta = \pi_i, \mu_i, \Sigma_i{}_{i=1}^{N_c}$. An estimate of these parameters can be obtained from a training set, $\{y^t\}_{t=1}^{N_T}$ using the ML principle:

$$\Theta^* = \text{argmax} \left[\prod_{t=1}^{N_T} P(y^t|\Theta) \right] \tag{10}$$

which is best solved using the Expectation-Maximization. And given our previous assumptions, the estimate of the complete likelihood becomes:

$$\hat{P}(\mathbf{x}|\Omega) = P(y|\Theta)\hat{P}_{\bar{F}}(x|\Omega) \tag{11}$$

where $\hat{P}_{\bar{F}}(x|\Omega)$, as before, is a Gaussian component density based on the DFFS.

Results

Training of the facial features is performed on 7,562 "mugshots". Performance analysis of three different detectors: sum-of-square-differences (SSD), distance-from-feature-space (DFFS) and Maximum Likelihood (ML) have been made based on 7,000 test images. The dimension for the principal subspace for DFFS and ML detectors was limited to five. By independently variating the detection threshold the receiver operating characteristics (ROC) curves was established for each detector, see figure 1. The performance of the ML detector (single-scale version) peaks with a detection-rate of 95% and in general exceeds the two other detection methods.

The multiscale ML detector version were also tested on an alternative database with 2,000 face images with a 10-dimensional principal subspace. This is resulted in a correction rate of 97%. As previously presented the detection is based on use of eigenvectors/eigenfaces. A generic set of eigenfaces can be calculated based on a set of training data, figure 2(a) presents an example of the eight highest prioritized eigenfaces. A unique set of coefficients multiplied with these eigenfaces results in a face image reconstruction that can be used to face detection. Figure 2(b) is an original face image normalized while 2(c) is the reconstruction with use of eigenfaces (in this case 100-dimensional eigenspace representation). The reconstructed eigenspace image requires approximately 1/5 bytes to encode compared to a standard image compression technique, figure 2(d) (JPEG).

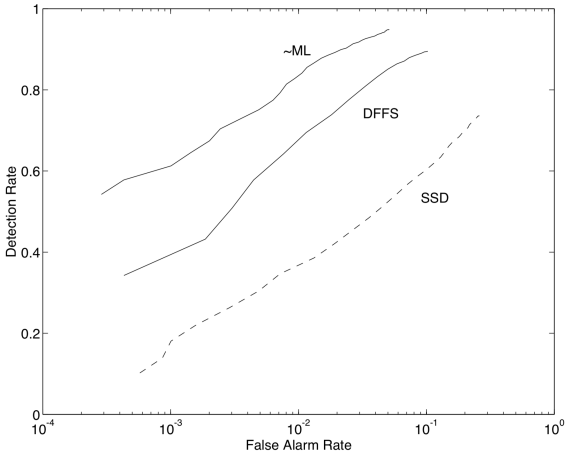

Figure 1: The ROC curves displaying the Performance of an SSD, DFFS and a ML detector.

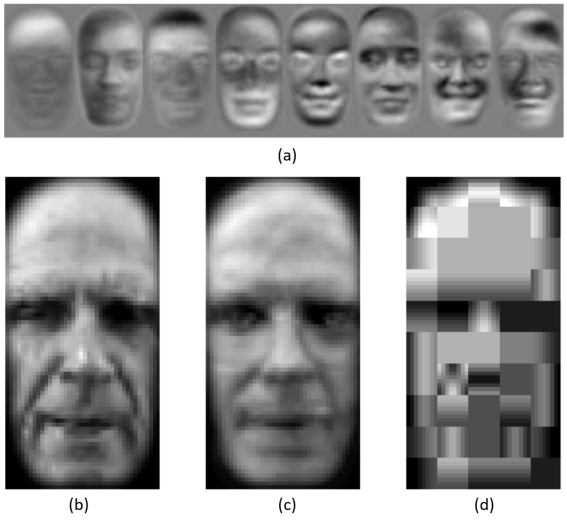

Figure 2: (a) Eight eigenfaces, (b) aligned face, (c) eigenspace reconstruction (85 bytes) (d) JPEG reconstruction (530 bytes).

Discussion

Using eigenspace decomposition and PCA for dimensionality reduction, density functions of high-dimensional images were estimated. Afterwards, a maximum likelihood approach was successfully experimented on the density estimates for object detection. To establish a more robust face detection algorithm a prior Bayesian probability can be developed to prevent use of negative image examples in face detection.