Evaluating the Informativity of Features in Dimensionality Reduction Methods

Mohammad Bagher Akbari Haghighat and Ehsan Namjoo
Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran
haghighat@ieee.org , e.namjoo@ieee.org

Abstract—**The ultimate goal of pattern recognition is to discriminate different classes with minimum misclassification rate. The feature vector used in classification should be as short as possible to reduce the algorithm complexity and informative enough to be able to discriminate complicated patterns. In this regard, dimensionality reduction methods are utilized to reduce the raw feature vector length and also to make the features more discriminative. In this paper, a face detection scheme is proposed by using discrete cosine transform (DCT) features in Bayesian discriminating features (BDF) classifier. Low redundancy of DCT features, optimal reconstruction property of Hotelling transform as the dimensionality reduction method, and the minimum error rate of Bayesian classifier, all in all, bring about a high detection rate in the proposed scheme. Various experiments, performed on different databases, certify that using more informative feature vectors results in a higher dimensionality reduction and improves the classifier's detection rate.**

Keywords—Bayes decision theory; BDF classifier; DCT features; dimensionality reduction; feature extraction.

I. Introduction

There are two major stages in any pattern recognition system: feature extraction and classification. Role of the feature extraction is to provide feature vectors with proper information content from the corresponding patterns. And, classification is a decision process. These two stages are highly interconnected in a successful pattern recognition system. If the feature vectors provided by the feature extraction step are not informative enough, it will be impossible to design an efficient and reliable pattern recognition system, even using the strongest classifiers, because some vital information describing a specific pattern might be ignored in the inappropriate feature extraction step. Therefore, feature extraction should be in a way so that the selected features contain the most important information for pattern description.

Many intelligent decision systems utilize dimensionality reduction methods in order to extract feature vectors with high information content [1-4]. In these methods, a transform is applied to raw feature vectors in order to construct a feature vector with highly discriminative components, usually in a decreasing order. That is, the first element is the most discriminative one, and the discriminating abilities of the following elements are decreasing along the transformed vector. After this transformation, M largest elements of the transformed feature vector are selected as the principal components to be used in the decision process, and the other elements are discarded. However, in some methods, the average of the remaining elements is also employed [5, 6].

Dimensionality reduction in feature extraction has two objectives. Firstly, reducing the feature vector length leads to a significant reduction in computational load of classifier. And secondly, using the important features with high discriminating abilities enhances the accuracy and robustness of the classifier which makes it more reliable [7]. In a raw feature vector, the content of the vector might be composed of a combination of different elements with different levels of importance that the classifier treats them equally. The unimportant elements often decrease the classifier's detection rate by shading the effect of the important ones on classification. Furthermore, sometimes they cause a situation in which the classifier gets over-trained by bulk unrelated training data. However, methods based on dimensionality reduction pick out the most important, discriminative, and relating features. So, the unimportant ones do not get involved in final decision process and never find a chance to devastate the classifier's detection rate. On the other hand, using dimensionality reduced feature vectors which are much shorter and more discriminative than the classic ones, the classifier would have enough flexibility in easily getting trained and also it would not get clamped by over-training.

In previous works, in order to attain all discriminating features, the raw feature vectors usually consist of several feature vectors extracted by different methods, and subsequently, the dimensionality reduction methods are utilized to reduce the vector length [5]. In this paper, it will be shown that the method used for the extraction of the raw vector has a significant effect on classification performance. In order to prove the claim, two feature extraction methods, Liu's features introduced in [5] and DCT features presented in Section II, are considered in the classification of the face as a complicated pattern. Regarding the property of DCT in reducing the spatial redundancy of the image pixels, there is no need to use several feature extraction methods as done in [5]. Moreover, zigzag ordering the DCT block and eliminating the insignificant coefficients in the raw vector present a more concentrated and meaningful representation of the face pattern, so that, after utilizing dimensionality reduction methods, the generated discriminative features result in a noteworthy improvement in classifier detection rate, in comparison with Liu's features [5].

Principal component analysis (PCA) is used as an effective tool for dimensionality reduction and the BDF classifier is utilized in decision process. The experiments performed on

978-1-61284-832-7/11/$26.00 ©2011 IEEE

different face databases prove the efficiency of the proposed approach. This paper is organized as follows: in Section II, details of the feature extraction methods used in this paper are discussed. Section III studies the dimensionality reduction in BDF classifier. In section IV, the proposed method and the experimental results are presented. Finally, the conclusion is drawn in Section V.

II. Feature Extraction

Feature extractors take proper information from rare patterns. The vectors achieved after feature extraction step, are fed to the classifier to be used in making decisions about a specific input pattern. There are two major points in feature vector selection: firstly, the selected feature vector must contain the most beneficial information about the patterns (informativity), and secondly, the feature vector length should be as short as possible in order to reduce the algorithm complexity. In this work, patterns are considered as rectangular 16×16 face sub-images. Two kinds of features are extracted and compared in our experiments. First one is the feature vector proposed by Liu in [5]; and the second one is DCT features [8].

A. Liu's Features

Let $I(i,j)\epsilon R^{m\times n}$ represent an $M\times N$ sub-image matrix and $X\epsilon R^{mn}$ be a vector that is formed by concatenating the rows or columns of $I(i,j)$. As defined below, the one dimensional Haar representation of $I(i,j)$ yields two images, $I_h(i,j)$ *and* $I_v(i,j)$, corresponding to the horizontal and vertical difference images, respectively.

$$I_h(i,j) = I(i+1,j) - I(i,j) \quad 1 \le i \le m-1, 1 \le j \le n \qquad (1)$$

$$I_v(i,j) = I(i,j+1) - I(i,j) \quad 1 \le i \le m, 1 \le j \le n-1 \qquad (2)$$

As defined, similar vectors $X^h\epsilon R^{(m-1)n}$ and $X^v\epsilon R^{m(n-1)}$ are constructed by concatenating the rows or columns of $I_h(i,j)$ *and* $I_v(i,j)$, respectively. The horizontal (row) and vertical (column) projection vectors of $I(i,j)$ are calculated as:

$$X_r(i) = \sum_{j=1}^{n} I(i,j), \quad 1 \le i \le m \qquad (3)$$

$$X_c(j) = \sum_{i=1}^{m} I(i,j), \quad 1 \le j \le n \qquad (4)$$

In order to normalize these vectors, each vector is subtracted from the mean of its components and then divided by their standard deviation. Let $\hat{X},\hat{X}_h,\hat{X}_v,\hat{X}_r,\hat{X}_c$ be the normalized vectors. A new feature vector is constructed by concatenating these normalized vectors as in (5).

$$\tilde{Y} = \left(\hat{X},\hat{X}_h,\hat{X}_v,\hat{X}_r,\hat{X}_c\right)^T \qquad (5)$$

where $(\ )^T$ represents the transpose operation. Finally the Liu's feature is generated by normalizing the above vector as below:

$$Y = \frac{\tilde{Y} - \mu}{\sigma} \qquad (6)$$

where μ and σ are the mean and the standard deviation of $\tilde{Y}$, respectively [5]. Note that, in this method, the feature vector length of a 16×16 sub-image is 768.

B. DCT Features

Here, two dimensional DCT coefficients of the face pattern are going to be used as features. There is a high correlation between pixels of an image, which results in a high spatial redundancy. Using pixel-based features, the size of feature vectors increases while a huge amount of redundancy does exist that leads to an uninformativity in the feature vector. DCT is the best approximation of Karhunen-Loeve transform (KLT) as the most efficient tool to decorrelate the image pixels and pack the block energy into a minimal number of coefficients.

In two dimensional DCT, coefficients are the representations of intensity changes in an image on different frequencies. Since the intensity variations often occur in approximately fixed places in different face patterns, these changes could provide an identical rhythm in the alternations of DCT coefficients which makes it a suitable option to create a face feature [8]. Low frequency coefficients which are more important are gathered in the top-left corner of the DCT block. Therefore, these coefficients are rearranged in zigzag form to fit in the importance order (Fig. 1).

In this feature extraction method, two dimensional DCT coefficients of the 16×16 sub-images are calculated and then zigzag reordered to provide a 256 length feature vector.

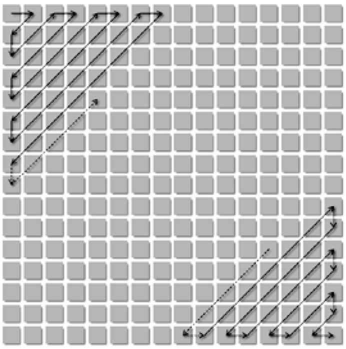

Fig. 1. Zigzag reordering of DCT coefficients.

III. Dimensionality Reduction in BDF Classifier

Principal component analysis (PCA), also called Hotelling transform, is an effective tool for dimensionality reduction in classification approaches. The most important property of PCA is its optimal reconstruction property. That is, PCA reduces the feature vector length in a way that the new vector contains the maximum amount of information from the previous one [1, 9].

The classifiers based on probability theory are more flexible among other appearance-based classifiers like neural networks. These classifiers are usually based on Bayes decision theory. If a proper probability density function is assigned to the training samples, Bayes decision theory will guarantee a minimum error rate classifier [10].

BDF classifier effectively utilizes PCA for dimensionality reduction. Employing Hotelling transform helps BDF to reduce the computational complexity by carrying data in a low-dimensional subspace. On the other hand, BDF classifier is based on Bayes decision theory, so it yields the minimum error rate if the proper probability density function is fitted to the train data set. In BDF classifier, the *M* largest eigenvalues of the covariance matrix, which contain large amount of information about the training patterns, are individually involved in classification while the other eigenvalues take part in classification by their arithmetic average just as a single parameter [5].

Gaussian probability density function is a powerful choice in modeling many natural events [11]. The conditional probability density function of the face feature vector can be modelled by this distribution as:

$$p(Y \mid \omega_f) = \frac{1}{(2\pi)^{\frac{N}{2}} \left|\Sigma_f\right|^{\frac{1}{2}}} \times exp\left\{-\frac{1}{2}(Y - M_f)^T \Sigma_f^{-1} (Y - M_f)\right\} \qquad (7)$$

where M_f, Σ_f and N are the mean vector, covariance matrix, and the length of the feature vector of the face class ω_f, respectively. Taking the natural logarithm on both sides of the equation (7), we will have:

$$\ln\left[p(Y \mid \omega_f)\right] = -\frac{1}{2}\left\{(Y - M_f)^T \Sigma_f^{-1} (Y - M_f) + N \ln(2\pi) + \ln(\mid \Sigma_f \mid)\right\} \qquad (8)$$

Using PCA, the covariance matrix, Σ_f, can be factorized as:

$$\Sigma_f = \varphi_f \Lambda_f \varphi_f^T \qquad (9)$$

where:

$$\begin{cases} \varphi_f^T \varphi_f = \varphi_f \varphi_f^T = I_N \\ \Lambda_f = diag\left\{\lambda_1, \lambda_2, ..., \lambda_N\right\} \end{cases}$$

where φ_f is an $N \times N$ eigenvector matrix of the covariance matrix of face feature vectors. Λ_f is a diagonal matrix with the same size containing the corresponding eigenvalues (λ_i) of Σ_f in decreasing order, and I_N is an $N \times N$ identity matrix.

The Hotelling transform is defined in (10):

$$Z = \varphi^T (Y - m_y) \qquad (10)$$

where m_y is the mean vector of Y feature vectors. Now, using the Hotelling transform and regarding the fact that φ is an orthogonal matrix, ($\varphi^T = \varphi^{-1}$), we will have:

$$(Y - M_f)^T \Sigma_f^{-1} (Y - M_f) = (Y - M_f)^T (\varphi_f \Lambda_f \varphi_f^T)^{-1} (Y - M_f)$$
$$= (Y - M_f)^T (\varphi_f^T)^{-1} \Lambda_f^{-1} \varphi_f^{-1} (Y - M_f) = Z^T \Lambda_f^{-1} Z$$

So, the natural logarithm of the probability density function of the face samples can be rewritten as:

$$\ln\left[p(Y \mid \omega_f)\right] = -\frac{1}{2}\left\{Z^T \Lambda_f^{-1} Z + N \ln(2\pi) + \ln\left|\Lambda_f\right|\right\} \qquad (11)$$

Z components are the principal components. Relying on the optimal reconstruction property of PCA, just M major eigenvalues are enough to estimate the probability density function for face class, where M is much smaller than N. In BDF, the remaining N-M values are estimated by their arithmetic average as:

$$\rho = \frac{1}{N - M} \sum_{k=M+1}^{N} \lambda_k \qquad (12)$$

where λ_i's are the eigenvalues of the covariance matrix for face class. Finally, the natural logarithm of the probability density function for the face class will be rewritten as below:

$$\ln\left[p(Y \mid \omega_f)\right] = -\frac{1}{2}\left\{ \sum_{i=1}^{M} \frac{z_i^2}{\lambda_i} + \frac{\left\|Y - M_f\right\|^2 - \sum_{i=1}^{M} z_i^2}{\rho} + \ln\left(\prod_{i=1}^{M} \lambda_i\right) + (N - M)\ln\rho + N\ln(2\pi) \right\} \qquad (13)$$

Similarly, the natural logarithm of the probability density function for the non-face class is calculated as:

$$\ln\left[p(Y \mid \omega_f)\right] = -\frac{1}{2}\left\{ \sum_{i=1}^{M} \frac{u_i^2}{\lambda_i^{(n)}} + \frac{\left\|Y - M_n\right\|^2 - \sum_{i=1}^{M} u_i^2}{\varepsilon} + \ln\left(\prod_{i=1}^{M} \lambda_i^{(n)}\right) + (N - M)\ln\varepsilon + N\ln(2\pi) \right\} \qquad (14)$$

where u_i's are the principal components, $\lambda_i^{(n)}$'s are the eigenvalues of the covariance matrix, and M_n is the mean feature vector for non-face class. ε is the estimation of the remaining N-M values, and ω_n represents the non-face class.

The Hotelling transform for the non-face class is:

$$U = \varphi_n^T (Y - M_n) \qquad (15)$$

where φ_n is an eigenvectors matrix that their columns are the eigenvectors of the covariance matrix of the non-face class.

Y represents a face pattern if $P(\omega_f \mid Y) > P(\omega_n \mid Y)$. In this case, Bayes decision theory is defined as:

$$P(\omega_f \mid Y) = \frac{P(\omega_f) p(Y \mid \omega_f)}{p(Y)}$$
$$P(\omega_n \mid Y) = \frac{P(\omega_n) p(Y \mid \omega_n)}{p(Y)} \qquad (16)$$

where $P(\omega_f)$ and $P(\omega_n)$ are the priori probabilities of the face and the non-faces classes, respectively, and $p(Y)$ is the mixture probability density function of Y. Finally, BDF classifier can be written as in (17):

$$Y \in \begin{cases} \omega_f & if \quad \delta_f + \tau < \delta_n, \ \delta_f < \theta \\ \omega_n & \qquad otherwise \end{cases} \qquad (17)$$

where τ and θ are two control parameters which are determined empirically based on training set, and:

$$\delta_f = \sum_{i=1}^{M}\frac{z_i^2}{\lambda_i} + \frac{\left\|Y - M_f\right\|^2 - \sum_{i=1}^{M} z_i^2}{\rho} + \ln\left(\prod_{i=1}^{M}\lambda_i\right) + (N-M)\ln\rho \quad (18)$$

$$\delta_n = \sum_{i=1}^{M}\frac{u_i^2}{\lambda_i^{(n)}} + \frac{\left\|Y - M_n\right\|^2 - \sum_{i=1}^{M} u_i^2}{\varepsilon} + \ln\left(\prod_{i=1}^{M}\lambda_i^{(n)}\right) + (N-M)\ln\varepsilon \quad (19)$$

IV. Proposed Method, Experimental Results and Analysis

In this section, performance of the proposed pattern recognition scheme with DCT features and BDF classifier is evaluated. Proposed method is compared with the well-known Liu's method in face detection [5]. Face images used in our experiments are from BioID database [12] containing 1500 face images and CMU database [13] containing 483 images. In order to have a class of non-face images, 2500 other nature images are selected. A few samples of face and non-face images are shown in Fig. 2. Training procedure of the BDF classifier is as follows.

Firstly, the covariance matrix and the mean vector of the training samples are calculated. Then, eigenvalues and eigenvectors of the covariance matrix are calculated. Afterwards, the appropriate number of eigenvalues (*M*) is selected; and finally, classifier parameters are determined as discussed in Section III. Calculating the covariance matrix and selecting the proper number of eigenvalues are common training procedures in both classifiers. Assuming that the eigenvalues of the covariance matrix are sorted decreasingly, the mean square error for the reconstructed vectors will be computed by equation (20) [6,14].

$$e_{ms} = \sum_{i=1}^{N}\lambda_i - \sum_{i=1}^{k}\lambda_i = \sum_{i=k+1}^{N}\lambda_i \quad (20)$$

where *N* is the feature vector length. Using equation (20) and according to the allowed range of error, we can determine the appropriate number of eigenvalues. For the first experiments, we have randomly chosen 1300 training samples from BioID database, and the number of test samples is 200. From non-face set, we have chosen 2000 training samples and 500 test samples.

γ is defined as *reconstruction ratio* in (21) [7]. That is, γ is the ratio of the sum of the *M* selected eigenvalues to the sum of the all eigenvalues. So, 1-γ will be defined as *reconstruction error*. Fig. 3 reveals the diagram of reconstruction ratio (γ) for different number of selected eigenvalues. As it can be seen, selecting more eigenvalues decreases the reconstruction error. Using Liu's features, this error is about 0.1 when 50 eigenvalues are selected. However, using DCT features, error reaches less than 0.1 when only 4 eigenvalues are selected. As it is shown in Fig. 3, these numbers of eigenvalues contain about 90% of the training patterns information. Considering the length of the feature vector in Liu (768) and in DCT (256) for a 16×16 image, and so the number of selected eigenvalues, dimensionality reduction of PCA on Liu's features is about 93.5%; however, using DCT features, dimensionality reduction reaches 98.4%.

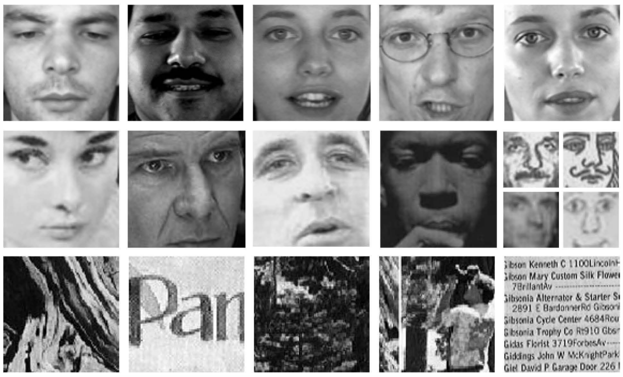

Fig. 2. Face and non-face sample images. First row: face sample images from BioID database. Second row: face sample images from CMU database. Third row: some non-face image samples.

$$\gamma = \sum_{i=1}^{M}\lambda_i \Big/ \sum_{i=1}^{N}\lambda_i \quad (21)$$

Selecting 50 major eigenvalues of the covariance matrix of Liu's features, and 4 eigenvalues using DCT features, TABLE I demonstrates the dimensionality reduction and detection rates of the BDF classifier.

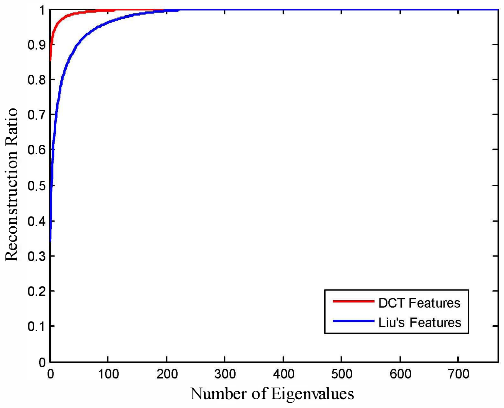

Fig 3. Reconstruction ratio (γ) in terms of the number of selected eigenvalues (*M*) (BioID Database).

TABLE I

Performance Evaluation of Liu and DCT Features in BDF Classifier (BioID Database)

Features	Liu	DCT
Feature Vector Length	768	256
Number of Selected Eigenvalues	50	4
Dimensionality Reduction (%)	93.49	98.44
Face Detection Rate (%)	99	100
False Positive Rate (%)	0	0.6

Another experiment is performed on CMU database containing 483 face images. 400 samples are randomly chosen for training, and 83 remaining face images are used in test process. Non-face dataset is also divided into 2000 training and 500 test samples. The diagram of the reconstruction ratio (γ) for different number of selected eigenvalues is demonstrated in Fig. 4. In this database, using Liu's features, the reconstruction error gets under 0.1 when 77 eigenvalues are selected. However, using DCT features, error reaches less than 0.1 when only 5 eigenvalues are selected.

Using 77 major eigenvalues of the covariance matrix of Liu's features, and 5 eigenvalues using DCT features, TABLE II reveals the dimensionality reduction and detection rate of the BDF classifier. As it can be seen, using DCT features lead to a higher detection rate as well as a much shorter feature vector which results in a considerable complexity reduction.

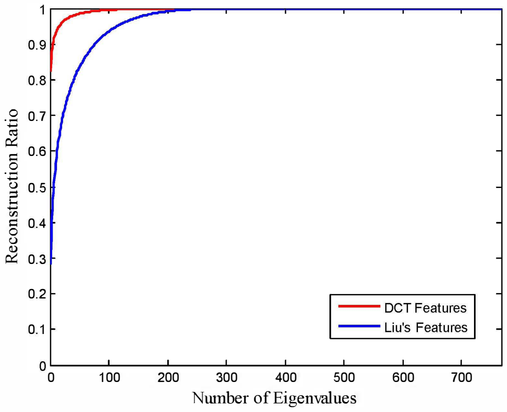

Fig 4. Reconstruction ratio (γ) in terms of the number of selected eigenvalues (M) (CMU Database).

TABLE II

PERFORMANCE EVALUATION OF LIU AND DCT FEATURES IN BDF CLASSIFIER (CMU DATABASE)

Features	Liu	DCT
Feature Vector Length	768	256
Number of Selected Eigenvalues	77	5
Dimensionality Reduction (%)	89.97	98.05
Face Detection Rate (%)	97.59	98.79
False Positive Rate (%)	3.2	2

V. CONCLUSION

In this paper, the effect of the raw feature vector selection in dimensionality reduction methods is evaluated. In this regard, a new face detection scheme is proposed which uses DCT features in BDF classifier. The experiments are performed on two well-known face databases: BioID and CMU. Comparing the results of the proposed approach with the method presented in [5], proves that DCT features which are less redundant and more informative have a better performance than the pixel-based features both in dimensionality reduction and classifier's detection rate.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, "*Pattern Classification*," John Wiley and Sons, 2nd Edition, 2000.

[2] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative Common Vectors for Face Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 4-13, 2005.

[3] S. Yan, D. Xu, B. Zhang, Q. Yang, H. Zhang, and S. Lin, "Graph embedding and extensions: A General Framework for Dimensionality Reduction," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[4] M. Belkin and P. Niyogi, "Towards a Theoretical Foundation for Laplacian-based Manifold Methods," *Journal of Computer and System Sciences,* vol. 74, no. 8, pp. 1289-1308, 2008.

[5] C. Liu, "A Bayesian Discriminating Features Method for Face Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 25, no. 6, pp. 725-740, 2003.

[6] B. Moghaddam, "Principal Manifolds and Probabilistic Subspaces for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 780-788, 2002.

[7] X. Jiang, "Linear Subspace Learning-Based Dimensionality Reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16-26, 2011.

[8] M.B.A. Haghighat, H. Seyedarabi, A. Aghagolzadeh, "Face Classification Using DCT Coefficients," *The 6th Iranian Conference on Machine Vision and Image Processing (MVIP),* pp. 379-383, Oct. 2010, Isfahan, Iran.

[9] S. Theodoridis and K. Koutroumbas, "*Pattern Recognition,*" Academic Press, 3rd Edition, 2006.

[10] K. Fukunaga, "Introduction to Statistical Pattern Recognition," Academic Press, 2nd Edition, 1990.

[11] K. Kim and G. Shevlyakov, "Why Gaussianity?," *IEEE Signal Processing Magazine,* vol. 25, no. 2, pp. 102-113, 2008.

[12] Available from: http://www.bioid.com/download-center/software/bioid-face-database.html .

[13] Available from: http://vasc.ri.cmu.edu/idb/html/face .

[14] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation, "*IEEE Transactions on Pattern Analysis and Machine Intelligence*," vol. 19, no. 7, pp. 696-710, 1997.