GENERALIZED SUBSPACE BASED HIGH DIMENSIONAL DENSITY ESTIMATION

Karthikeyan Shanmuga Vadivel[⋆] *Mehmet Emre Sargin*[⋆] *Swapna Joshi*[⋆] *B.S. Manjunath*[⋆] *Scott Grafton*[†]

[⋆] Department of ECE, University of California Santa Barbara
[†] Department of Psychology, University of California Santa Barbara
[⋆]{karthikeyan,msargin,sjoshi,manj}@ece.ucsb.edu
[†]grafton@psych.ucsb.edu

ABSTRACT

Our paper presents a novel high dimensional probability density estimation technique using any dimensionality reduction method. Our method first performs subspace reduction using any matrix factorization algorithm and estimates the density in the low-dimensional space using sample-point variable bandwidth kernel density estimation. Subsequently, the high dimensional density is approximated from the low dimensional density parameters. The reconstruction error due to dimensionality reduction process is also modeled in a principled and efficient manner to obtain the high dimensional density estimate. We show the effectiveness of our technique by using two popular dimensionality reduction tools, principal component analysis and non-negative matrix factorization. This technique is applied to AT&T, Yale, Pointing'04 and CMU-PIE face recognition datasets and improved performance compared to other dimensionality reduction and density estimation algorithms is obtained.

Index Terms— Probability density function; Principal component analysis; Face recognition

1. INTRODUCTION

Probability density function estimation in the high dimensional input space is important for statistical analysis of images. However, to well represent the sample space, the number of required images should be exponential to the cardinality of the input space. Due to this curse of dimensionality, estimating density parameters accurately in the high dimensional space becomes difficult. To overcome this problem, we propose a method to effectively estimate the density parameters in any lower dimensional subspace and then efficiently map them back to its original input space.

Subspace analysis techniques aim to extract lower dimensional representations for the data which help in simplifying problems such as classification and regression. In this respect Principal Component Analysis(PCA) [1] uses eigenvalue decomposition based dimensionality reduction. Independent Component Analysis(ICA) [2] aims to extract statistically independent non-gaussian components of the training data. Non-negative Matrix Factorization(NMF) [3] extracts non-negative features which have intuitive parts based meaning associated with them. All these techniques aim to represent the high-dimensional data as a linear combination of basis images. However, each of these algorithms factorize the data subject to different constraints, yielding favorable feature extractions for different applications.

Thanks to the NSF award III-0808772.

Thanks to the MacArthur Foundation and Public Health Service grant NIMH 1 R01 MH070539-01

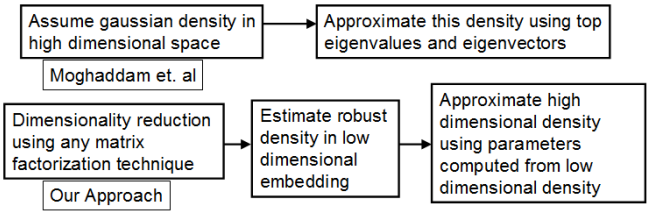

Fig. 1. Comparing flow diagrams of Moghaddam et. al. [4] and Our Approach

Probabilistic subspace models have been effective for visual recognition. In [8] and [4] Moghaddam et al. proposed a probabilistic high dimensional density estimation technique that assumes the data to have a gaussian mixture model in the input space and approximates the covariance matrix computations by the top eigenvalues and eigenvectors. Tipping and Bishop [5] proposed probabilistic principal component analysis (PPCA), which views PCA as a maximum likelihood procedure on a gaussian density model of the observed data. In [6], Wang et al. proposed a unified subspace analysis technique for face recognition. Lee et al. [7] proposed kernel extensions to the similarity measure used in [8]. In [9], Ramanathan et.al used the face similarity measure proposed in [8] for face verification across age.

All the above mentioned techniques assume a parametric density form in the high dimensional space and approximate these high dimensional parameters using top eigenvalues and eigenvectors similar in spirit to [4]. Our approach however, follows a three step process. First we compute the basis and the coefficients of the data matrix using any matrix factorization technique. This is followed by robust density estimation in the low dimensional space. Finally, the high dimensional density is efficiently approximated from the low dimensional density parameters. Figure 1 elucidates the difference in approach clearly. This approach has dual advantages. First we can leverage any matrix factorization technique for subspace reduction. Also, as we first estimate the density in the low dimensional subspace, complex modeling becomes feasible. The contributions of this paper are

- A novel methodology to estimate the density parameters in any subspace and effectively map them back to the original input space.
- Effective modeling of low dimensional density using the matrix factorization coefficients by sample-point variable bandwidth kernel density estimation.

- A novel approach of modeling the data matrix reconstruction error in a principled and efficient manner using eigenvalue perturbation.

2. REVIEW - DENSITY ESTIMATION USING EIGENSPACE DECOMPOSITION (DED)

Moghaddam et al. proposed an efficient density estimation technique for the high dimensional data $\mathbf{v} \in R^N$ in [4], which divides the vector space R^N into two complementary subspaces. This method estimates the complete probability distribution of an object's appearance using an eigenvector decomposition of the image space. The target density is decomposed into two parts: density in the principal subspace $\mathbf{F}$ and its orthogonal complement space $\bar{\mathbf{F}}$. The density in the principal subspace is obtained using the first M principal components $\mathbf{y} = \{y_i\}_{i=1...M}$. The complete optimal high-dimensional density estimate can be expressed as a product of two independent marginal gaussian densities

$$P(\mathbf{v}) = \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{{y_i}^2}{\lambda_i}\right)}{(2\pi)^{\frac{M}{2}}\prod_{i=1}^{M}{\lambda_i}^{1/2}}\right]\left[\frac{\exp\left(-\frac{\epsilon^2(\mathbf{v})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] = P_F(\mathbf{v})\hat{P}_{\bar{F}}(\mathbf{v}|\rho) \quad (1)$$

where $P_F(\mathbf{v})$ is the true marginal density in $\mathbf{F}$, and $\hat{P}_{\bar{F}}(\mathbf{v}|\rho)$ is the marginal density in the orthogonal complement space $\bar{\mathbf{F}}$. Here, $\epsilon^2(\mathbf{v})$ is the PCA residual and $\{\lambda_i\}$ are the top eigenvalues of the covariance matrix of $\mathbf{v}$. The optimal value of ρ is obtained by minimizing the divergence between the original probability density function and the approximation in (1). The optimal ρ is the average of the eigenvalues of $\bar{\mathbf{F}}$.

$$\rho = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i \quad (2)$$

There are several approximate ways to estimate ρ, as all the eigenvalues of $\bar{\mathbf{F}}$ are difficult to compute. This technique has been effective for face recognition and object detection problems. A detailed description of this method is presented in [4] and its application for face recognition is presented in [8]. A primary drawback with this technique is the high dimensional density is assumed to be gaussian and will not model datasets with high variability. To address this, our method uses a more general density model using variable bandwidth kernel density estimation. Also, DED is based on eigenspace decomposition and cannot utilize any general matrix factorization method for density estimation.

3. OUR APPROACH - GENERALIZED SUBSPACE BASED DENSITY ESTIMATION (GSD)

Our high dimensional density estimation technique GSD first performs matrix factorization on the data matrix. It decomposes a data matrix ($\mathbf{V}$) to a set of bases ($\mathbf{W}$) and corresponding coefficients ($\mathbf{H}$).

$$\mathbf{V}_{n\times n_t} \approx \mathbf{W}_{n\times m}\mathbf{H}_{m\times n_t} \quad (3)$$

where $\mathbf{V} = [v_{ij}] = [\mathbf{v}_1, ..., \mathbf{v}_{n_t}]$ is a $n \times n_t$ matrix , n is the total number of pixels in each image, $\mathbf{v}_j$ is the j^{th} input image represented as a column vector, and n_t is the number of training images. We denote the basis matrix $\mathbf{W} = [w_{ij}] = [\mathbf{w}_1, ..., \mathbf{w}_m]$ as an $n \times m$ matrix. The low dimensional embedding of every column of $\mathbf{V}$ is the corresponding column in $\mathbf{H} = [h_{ij}] = [\mathbf{h}_1, ..., \mathbf{h}_{n_t}]$.

In our appoach after dimensionality reduction, we initially estimate the probability density in low dimensional space ($\mathbf{H}$). In the current setting we use variable bandwidth kernel density estimation as proposed in [10] in the low dimensional space which estimates an effective and robust density. Hence, a generalized covariance bandwidth matrix $\mathbf{\Sigma}_{h_i}$ is estimated for the sample point $\mathbf{h}_i$, which is the covariance of the k nearest neighbors of $\mathbf{h}_i$ (using euclidean distance). Thus, using a Normal kernel, the density estimator in the coefficients subspace ($\mathbf{H}$) is

$$f_H(\mathbf{h}) = \frac{1}{n_t(2\pi)^{m/2}}\sum_{i=1}^{n_t}\frac{1}{|\mathbf{\Sigma}_{h_i}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{h}-\mathbf{h}_i)^T(\mathbf{\Sigma}_{h_i})^{-1}(\mathbf{h}-\mathbf{h}_i)\right) \quad (4)$$

Let $\mathbf{v}_i = \mathbf{W}\mathbf{h}_i + \delta_i$, and $\mathbf{V} = \mathbf{W}\mathbf{H} + \mathbf{\Delta}$. In this setting, $\mathbf{\Delta}$ is assumed to be uncorrelated with $\mathbf{H}$. In addition, δ_i's are assumed to be identically distributed as $\mathcal{N}(\delta_{mean}, \mathbf{\Sigma}_\Delta)$ and hence the high dimensional density $f_V(\mathbf{v})$ is expressed as follows

$$f_V(\mathbf{v}) = \frac{1}{n_t(2\pi)^{N/2}}\sum_{i=1}^{n_t}\frac{1}{|\mathbf{\Sigma}_{v_i}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{v}-\hat{\mathbf{v}}_i)^T(\mathbf{\Sigma}_{v_i})^{-1}(\mathbf{v}-\hat{\mathbf{v}}_i)\right) \quad (5)$$

where

$$\hat{\mathbf{v}}_i = \mathbf{W}\mathbf{h}_i + \delta_{mean}$$
$$\mathbf{\Sigma}_{v_i} = \mathbf{W}\mathbf{\Sigma}_{h_i}\mathbf{W}^T + \mathbf{\Sigma}_\Delta \quad (6)$$

In the above equation, the high-dimensional covariance ($\mathbf{\Sigma}_{v_i}$) consists of two parts, the covariance using the subspace $\mathbf{W}\mathbf{\Sigma}_{h_i}\mathbf{W}^T$, also denoted by $\mathbf{\Sigma}_{sub_i}$, and the covariance using the reconstruction error ($\mathbf{\Sigma}_\Delta$). Here, $|\mathbf{\Sigma}_{v_i}|$ and $\mathbf{\Sigma}_{v_i}^{-1}$ cannot be computed directly as $\mathbf{\Sigma}_{v_i}$ is not of full rank. Therefore, the top M eigenvalues and eigenvectors of $\mathbf{\Sigma}_{v_i}$ are used to approximate $|\mathbf{\Sigma}_{v_i}|$ as

$$|\mathbf{\Sigma}_{v_i}| \approx \lambda_{v_i}^1\lambda_{v_i}^2\lambda_{v_i}^3...\lambda_{v_i}^M \quad (7)$$

In the present analysis, M is assumed to be the same as the dimension of the coefficient subspace m in (3). In general, $M < m$. Also, the Mahalanobis distance $d(\mathbf{v}_i)$ is expressed as

$$d(\mathbf{v}_i) = (\mathbf{v}-\hat{\mathbf{v}}_i)^T(\mathbf{\Sigma}_{v_i})^{-1}(\mathbf{v}-\hat{\mathbf{v}}_i) = (\mathbf{v}-\hat{\mathbf{v}}_i)^T(\mathbf{\Phi}_{v_i}\mathbf{\Lambda}_{v_i}\mathbf{\Phi}_{v_i}^T)^{-1}(\mathbf{v}-\hat{\mathbf{v}}_i) = \mathbf{y}_i^T\mathbf{\Lambda}_{v_i}^{-1}\mathbf{y}_i \quad (8)$$

where $\mathbf{\Phi}_{v_i}$ and $\mathbf{\Lambda}_{v_i}$ are the eigenvector and eigenvalue matrices of $\mathbf{\Sigma}_{v_i}$. Let $\mathbf{y}_i = \mathbf{\Phi}_{v_i}^T(\mathbf{v}-\mathbf{v}_i)$ where $\mathbf{y}_i = \{y_i^k\}_{k=1...N}$. Here the eigenvalues and eigenvectors of $\mathbf{\Sigma}_{v_i}$ are denoted by $\{\lambda_{v_i}^k\}_{k=1...N}$ and $\{\phi_{v_i}^k\}_{k=1...N}$ respectively. Thus, the Mahalanobis distance $d(\mathbf{v}_i)$ can also be expressed as

$$d(\mathbf{v}_i) = \sum_{k=1}^{N}\frac{(y_i^k)^2}{\lambda_{v_i}^k} \quad (9)$$

Computation of (9) is not feasible as all the eigenvalues of the covariance matrix cannot be computed efficiently. Hence, the top M eigenvalues are used to approximate the Mahalanobis distance as

$$\hat{d}(\mathbf{v}_i) = \sum_{k=1}^{M} \frac{(y_i^k)^2}{\lambda_{v_i}^k} \tag{10}$$

We now propose an efficient way to compute the top M eigenvalues and eigenvectors of $\mathbf{\Sigma}_{v_i}$. In this regard, the eigenvalues and eigenvectors of $\mathbf{\Sigma}_{sub_i}(\mathbf{W}\mathbf{\Sigma}_{h_i}\mathbf{W}^T)$ are exactly and efficiently computed and then the effect of $\mathbf{\Sigma}_\Delta$ is considered from (6). The eigenvalues and eigenvectors of $\mathbf{\Sigma}_{sub_i}$ are exactly computed using the common PCA trick

$$\mathbf{W}\mathbf{\Sigma}_{h_i}\mathbf{W}^T = \frac{1}{n_t}(\mathbf{W}\tilde{\mathbf{H}})(\mathbf{W}\tilde{\mathbf{H}})^T \tag{11}$$

where $\tilde{\mathbf{H}}$ is the mean subtracted matrix obtained from $\mathbf{H}$. The eigenvalues and eigenvectors of $\frac{1}{n_t}(\mathbf{W}\tilde{\mathbf{H}})^T(\mathbf{W}\tilde{\mathbf{H}})$ are computed to find the eigenvalues and eigenvectors of $\mathbf{W}\mathbf{\Sigma}_{h_i}\mathbf{W}^T$ similar to PCA.

Let the eigenvalues and eigenvectors of $\mathbf{\Sigma}_{sub_i}$ be denoted by $\{\lambda_{sub_i}^k\}$ and $\{\phi_{sub_i}^k\}$. Now, $\mathbf{\Sigma}_\Delta$ is assumed to be a perturbation to $\mathbf{\Sigma}_{sub_i}$ ($\mathbf{\Sigma}_{v_i} = \mathbf{\Sigma}_{sub_i} + \mathbf{\Sigma}_\Delta$). Therefore, the eigenvalues and eigenvectors of $\mathbf{\Sigma}_{v_i}$ are obtained as follows

$$\lambda_{v_i}^k = \lambda_{sub_i}^k + (\phi_{sub_i}^k)^T \mathbf{\Sigma}_\Delta \phi_{sub_i}^k \tag{12}$$

$$\begin{aligned}\phi_{v_i}^k =& \phi_{sub_i}^k (1 - \frac{1}{2}(\phi_{sub_i}^k)^T \mathbf{\Sigma}_\Delta \phi_{sub_i}^k) \\ &+ \sum_{j=1, j\neq i}^{M} \frac{(\phi_{sub_i}^k)^T \mathbf{\Sigma}_\Delta \phi_{sub_i}^k}{\lambda_{sub_i}^k - \lambda_{sub_j}^k} \phi_{sub_j}^k\end{aligned} \tag{13}$$

The derivation of the above relation on eigenvalue perturbation is presented in [11].

Using $\{\lambda_{v_i}\}$ and $\{y_i\}$ (computed from $\{\phi_{v_i}\}$) from (12) and (13), the final density estimate GSD is approximated from (5) as

$$\begin{aligned}f_V(\mathbf{v}) =& \frac{1}{n_t(2\pi)^{N/2}} \sum_{i=1}^{n_t} \frac{1}{(\Pi_{k=1}^M \lambda_{v_i}^k)^{1/2}} \\ & \exp\left(-\frac{1}{2} \sum_{k=1}^{M} \frac{(y_i^k)^2}{\lambda_{v_i}^k} \right)\end{aligned} \tag{14}$$

4. EXPERIMENTS AND RESULTS

The performance of our algorithm on face recognition task is presented with four canonical face datasets, the AT&T (formerly ORL face database), Yale, Pointing'04 and CMU-PIE faces. The multi-class face recognition problem is modeled into multiple two class classification problems similar to [8]. Models are built to find if two faces belong to the same person or different people. Here, Ω_s is the class for modeling faces belonging to the same person and Ω_d is the class for modeling faces belonging to different people. In the current implementation, the absolute difference image between two people $|\mathbf{I}_1 - \mathbf{I}_2|$, same person or different people is considered to obtain Ω_s and Ω_d respectively. As our method can be used in conjunction with any matrix factorization technique, we choose two popular subspace reduction algorithms PCA and NMF. PCA captures global eigenfaces and NMF obtains parts-based features in their corresponding basis images respectively. The performance of our approach is compared with PCA, NMF, LNMF [12] and DED while changing the number of basis images.

4.1. Classification rule

In the following experiments, the training set for Ω_s and Ω_d are randomly chosen. Then, the models are trained to estimate the parameters $\{\lambda_{v_i}\}$ and $\{\phi_{v_i}\}$ for Ω_s and Ω_d from the training set. Given a test face, $\mathbf{v}_{test}$, the likelihoods of the test face and the training face $\mathbf{v}_{train}$ belonging to the same class (Ω_s) and the test face and the training face belonging to different classes (Ω_d) are computed. In this regard the difference image $|\mathbf{v}_{test} - \mathbf{v}_{train}|$ is used for $\mathbf{v}$ in (14). Subsequently, a subset of training faces are selected which have probability of Ω_s greater than probability of Ω_d after comparing the test data $\mathbf{v}_{test}$ with all the training images. Let the selected subset of training faces for $\mathbf{v}_{test}$ be denoted as $T_{v_{test}}$. Finally, the test class for $\mathbf{v}_{test}$ is inferred by a simple polling on the class labels of $T_{v_{test}}$.

4.2. AT&T and Yale datasets

The Cambridge AT&T database consists of 400 frontal face images of 40 people with varying facial expressions and details. A training set of 280 images, 7 from each class, are randomly selected and the remaining 120 are used for testing. In order to create the training set for Ω_s and Ω_d, 840 image pairs from the same person and 840 image pairs from different people are selected from the 280 images. The Yale face database contains 165 gray-scale frontal face images of 15 individuals. The images are taken under varying lighting conditions and facial expressions. To create the training set for Ω_s and Ω_d, 420 image pairs from the same person and 420 image pairs from different people are chosen from the training set of 120 images.

4.3. Pointing'04 face dataset

This head pose database [13] consists of 15 sets of images, 186 images per person. In our analysis, the dataset is preprocessed using Viola-Jones [14] frontal and profile face detectors to isolate the faces alone. In addition, false detections are manually discarded. The images are then resized to 40×40 pixels. The results after preprocessing are shown in Figure 2. We can notice that faces are unaligned due to which the recognition task is significantly difficult. We trained our algorithm on 675 training images (45 from every class) and tested on 724 images. To create the training set for Ω_s and Ω_d from the 675 training images, 1000 image pairs from the same person and 1000 image pairs from different people are randomly chosen.

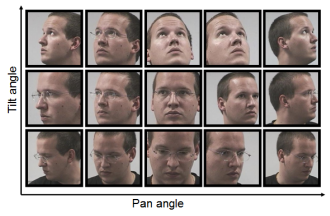

Fig. 2. Example of face from the Pointing'04 dataset.

4.4. CMU-PIE dataset

The CMU-PIE dataset [15] consists of face images of 68 subjects with 43 different illumination conditions across 13 poses. All extracted face images are aligned and normalized to size of 32×24 pixels. We select faces of 59 subjects, for every subject we randomly select 6 illuminations for each of the 9 pan angles (excluding c31 c25 c07 c09). An example subject is shown in Figure 3. Out of these 54 images per subject we randomly select 36 faces for training and the rest for testing. From the training images to create the training set for Ω_s and Ω_d, 1200 image pairs from the same person and 1200 image pairs from different people are chosen.

Fig. 3. Example face from the CMU-PIE dataset

Results A repeated random sub-sampling validation is performed using 5 repeats on all the methods. The performance of the different algorithms on AT&T and Yale faces is illustrated in Figure 4. In the AT&T database, among the dimensionality reduction algorithms, LNMF achieves the best accuracy of 94% for 35 basis images. In the Yale database, PCA achieves the best accuracy of around 77% for 25 basis images. DED gives an accuracy of around 95% and 94% for these datasets. GSD+PCA outperforms all the techniques on the Yale and AT&T datasets by achieving recognition accuracy of 98.8% and 99.1%(40 basis images) respectively. We note that the same feature $|\mathbf{I}_1 - \mathbf{I}_2|$ is used to compare the performance of DED and GSD. The key observation is that there is a significant performance difference in using the original NMF or PCA algorithm directly and using GSD+PCA or GSD+NMF.

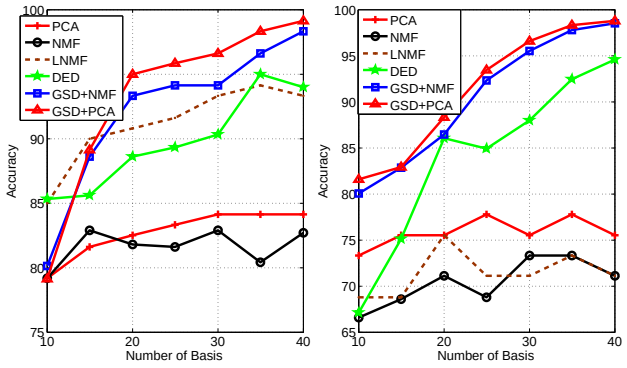

Fig. 4. Classification performance on AT&T (left) and Yale (right) datasets.(Best viewed in color)

In the more challenging Pointing'04 and CMU-PIE datasets with pose and illuminiation variations, GSD+NMF achieves the best performace of 84.7% and 92.1% for 60 and 70 basis images respectively. This is followed by GSD+PCA. Table 1 shows the performance of other methods on these datasets. Here we report the best results after changing the basis images images from 30 to 80 for both the datasets. In face recognition problems with extreme appearence, pose and illumination variations, parts based subspace reduction methods such as NMF are more effective than PCA. As our method can use any matrix factorization based subspace reduction technique to obtain the high dimensional density, using GSD+NMF we obtain the best results in Pointing'04 and CMU-PIE faces.

	PCA	NMF	LNMF [12]	DED [8]	GSD +PCA	GSD +NMF
Pointing	77.6	78.8	76.5	78.3	82.8	**84.7**
CMU-PIE	82.4	83.1	83.3	87.2	90.4	**92.1**

Table 1. Classification accuracies (in %) on Pointing'04 and CMU-PIE datasets.

5. CONCLUSION

A novel framework for high dimensional density estimation using variable bandwidth kernel density estimation in the lower dimensional space is proposed in this paper. This method can use any matrix factorization technique to obtain the high dimensional density. In conjunction with PCA and NMF our method outperforms other relevant subspace reduction and density estimation algorithms on popular face recognition datasets.

6. REFERENCES

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, 1991.

[2] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, 1994.

[3] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, 1999.

[4] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE TPAMI*, 1997.

[5] M.E. Tipping and C.M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society*, 1999.

[6] X. Wang and X. Tang, "A unified framework for subspace face recognition," *IEEE TPAMI*, 2004.

[7] J. Lee et al., "Visual object recognition using probabilistic kernel subspace similarity," *Pattern Recognition*, 2005.

[8] B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE TPAMI*, 2002.

[9] N. Ramanathan and R. Chellappa, "Face verification across age progression," *TIP*, 2006.

[10] H. Chen and P. Meer, "Robust Computer Vision through Kernel Density Estimation," in *ECCV*, 2002.

[11] L.N. Trefethen and D. Bau, *Numerical linear algebra*, 1997.

[12] S.Z. Li et al., "Learning spatially localized, parts-based representation," in *IEEE CVPR*, 2001.

[13] N. Gourier et al., "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures (POINTING)*, 2004.

[14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE CVPR*, 2001.

[15] T. Simon et al., "The CMU pose, illumination, and expression (PIE) database," in *IEEE Face and Gesture*, 2002.