An Image-based Bayesian Framework for Face Detection

Lingmin Meng, Truong Q. Nguyen, David A. Castañon
Electrical and Computer Engineering Department, Boston University
8 St. Mary's Street, Boston, MA 02215
{lmeng, nguyent, dac}@bu.edu

Abstract

In this paper, we present a novel approach for frontal face detection in gray-scale images. We represent both faces and clutter by using two-dimensional wavelet decomposition. To characterize the statistical dependency between different levels of wavelet, we introduce a Hidden Markov Model (HMM), in which a number of discrete states at each level capture the diversity of faces as well as clutter. Our experiments indicate that the proposed algorithm outperforms conventional template-based methods such as matched filter and eigenface methods.

1 Introduction

Face detection and recognition from images is an active research area with numerous applications in user identification or verification, law enforcement and human-computer interaction. In many situations, face detection is required before face recognition. For real-time applications, it is important to locate faces in an image fast and accurately.

Many current face detection approaches are based on exploiting color and/or motion information in image sequence. In this work, we focus on detection in gray-scale still images, because color or motion information may not be available in many important applications.

The major difficulty in face detection is the large range of variations across possible faces. These variations arise because of illumination conditions, facial expressions, hair, hats and eyeglasses, making it difficult to develop robust face detection algorithms. In general, face detection in gray-scale images can be put into two categories: 1) Geometric feature-based algorithms [1, 2, 3, 4] rely on the accurate estimation of specific features; this is difficult to accomplish when there is no knowledge of the size or position of the face. 2) Image-based algorithms use the intensity values of all pixels, avoiding information loss which could arise through geometric feature extraction. In this paper, we focus on image-based algorithms.

The simplest image-based method is the matched filter or normalized correlation method. From a statistical point of view, the matched filter algorithm assumes that faces deviate from the average face by additive white Gaussian noise, which is is very inaccurate (see bottom left plot of Figure 2, white noise should have a diagonal covariance matrix). A more sophisticated image-based approach is the eigenface [5] method, which characterizes the deviations from the average face as independent Gaussian perturbations in a few image directions. These directions are the principal eigenvectors (popularly known as eigenfaces) of the autocovariance matrix of the training face samples.

The problem with the eigenface approach, as shown in [6], is that the eigenfaces reflect the variations in both the face and the background clutter. There are situations where a non-face block can match a linear combination of average face and its eigenfaces very well (in the sense of least squares error). To differentiate faces from background clutter, a single feature "linear discriminant" [7, 8] was developed by approximating the autocovariance matrices of face and nonfaces to be identical. Frey *et al.* [9] developed a mixture of local linear subspaces model to overcome the limitation of a fixed linear subspace for modeling different poses and expressions. However, all of these methods assume that the deviations from an average face or clutter can be represented by a Gaussian statistical model, which fundamentally limits the achievable recognition performance.

To overcome these limits, some methods decompose a face into its subbands using wavelet transforms. In [6], we decomposed the face spatially and spectrally and then applied normalized correlation in multiple subbands and segments. Recently, Rikert et al. [10] built a mixture of Gaussian model using the "parent vector" at multiple resolutions. This has the advantage of representing local characteristics of images in terms of wavelet coefficients. However, the statistical dependence between wavelet coefficients is difficult to model.

In this paper, we present a non-Gaussian statistical model for face detection. Our approach is based on wavelet representations of a sample image region, which captures the statistical relationship among subband images. This flexibility allows us to tune the model to compensate for different lighting and hair conditions of

a face, and to develop algorithms which recognize patterns such as edges which are naturally present at multiple scales.

Samaria and Harter [11] extracted top-bottom facial regions such as hair, forehead, eyes, nose and mouth, and the natural order in which the features appear is modeled using a top-bottom HMM. Later, Nefian and Hayes [12] used the Karhunen Loeve Transform (KLT) coefficients as observation vectors in the HMM to increase robustness and speed. Our approach is similar in spirit to these probabilistic approaches to recognition, except that these methods do not provide an explicit model of dependence between scales. Most of them rely on the local dependence between "spatially" neighboring components.

Other works ([13, 14, 15]) build the dependence across scale (also know as "Markov Tree"), but only among single wavelet coefficients. Colmenarez and Huang [13] used 4-grey-levels re-quantization of the wavelet coefficients. Choi *et al.* [15] used two states (large and small) for each wavelet coefficients. Some of them assume the "Markov Tree" is ergodic in space, which is true only for texture analysis.

In our framework, we model the statistical dependency between whole subband images at different levels. Our model captures efficiently, as a Gaussian mixture, the diversity of face objects and clutter. Our simulation results demonstrate the advantage of our proposed algorithm over conventional image-based face detection algorithms.

In the next section we describe the statistical modeling framework, and the resulting face detection algorithm. In section 3, we present simulation results and comparison with other face detection algorithms. In section 4, we discuss some extensions for this framework.

2 Proposed Framework

Our approach for face detection is based on testing each subblock image to determine the likelihood ratio that it is a face as opposed to clutter. Thus, locating a generic human face of certain size from a given image is essentially a binary decision problem of each candidate block. Let $\underline{y}$ be the observed sample, and hypothesis $\Omega = 1$ and $\Omega = 0$ denote face and non-face respectively. To minimize the binary decision error, one applies maximum a posterior (MAP) rule which leads to Likelihood Ratio Test (LRT).

The central idea of a statistical model is the definition of the likelihood function $P(\underline{y}|\Omega)$. In this section we describe in detail the model for both object and clutter. First we represent an image block by its multiscale decomposition using coarse approximation and details at all levels. Then we use a directed Bayesian net (Hidden Markov model with discrete states) to model the statistical dependency across scale in the wavelet representation.

2.1 Multiscale representation

Statistical signal modeling and processing methods based in the wavelet domain are, in many cases, much more effective than classical time-domain or frequency domain approaches. Our statistical model starts by iteratively decomposing an image block into its coarse approximation and details at various scales. Each level decomposition of an image gives its approximation and its detail. The detail is the sum of three wavelets of a 2-D wavelet decomposition, or equivalently, the difference between an original image and its coarse approximation. To remove the redundancy, the coarse approximation part is decimated by 2 in both directions. Then iteratively, the approximation part is decomposed into approximations and details. Specifically, an image block $\underline{y}$ is decomposed into:

$$\underline{y} = \uparrow 2(\ldots \uparrow 2(\uparrow 2(\underline{y}^{(0)}) + \underline{y}^{(1)}) + \ldots + \underline{y}^{(k-1)}) + \underline{y}^{(k)} \quad (1)$$

Figure 1 shows a five-level decomposition using 2-D Haar filters (one lowpass filter and three highpass filters) for four iterations. The 2-D Haar wavelet is chosen because it is orthogonal and has linear phase.

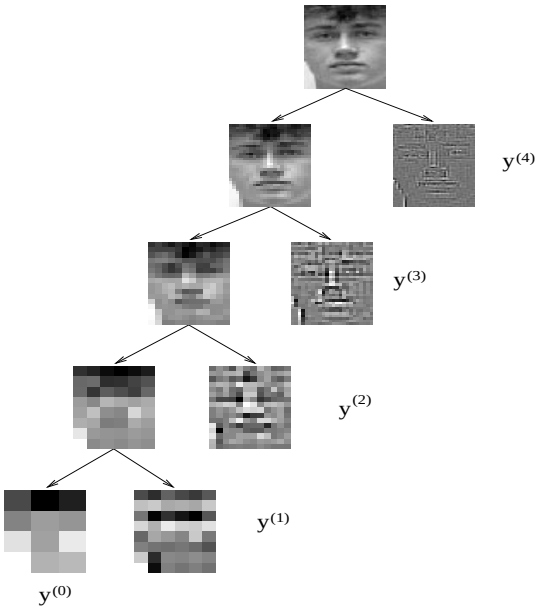


Figure 1: Multiscale decomposition of a face image

One of the primary properties of the wavelet transform is separation into orthogonal components. For Gaussian processes, this often results in independence of the wavelet coefficients across scale and time, whereby the representation of the statistical covariance in the wavelet basis is approximately block-diagonal. This property makes it simpler to represent the statistical structure of face variations.

To illustrate this, assume that the overall image statistics were modeled as Gaussian, which implies that each of the wavelet levels is also Gaussian, as

$$\underline{y}^{(i)} \sim N(\underline{\mu}^{(i)}, S^{(i)}) \tag{2}$$

Figure 2 shows the approximate statistical independence across wavelet levels. The figure shows a three-level decomposition of the sample covariance of a face image using 2-D Haar wavelets. Clearly, the covariance matrix of details tend to be sparse (concentrated on diagonal elements) and most of energy can be compressed into $S^{(0)}$.

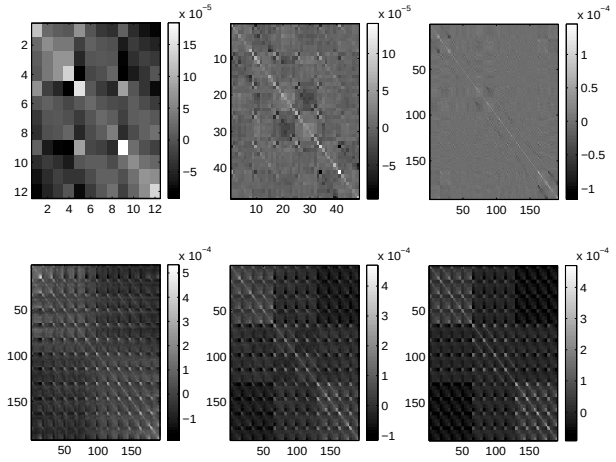


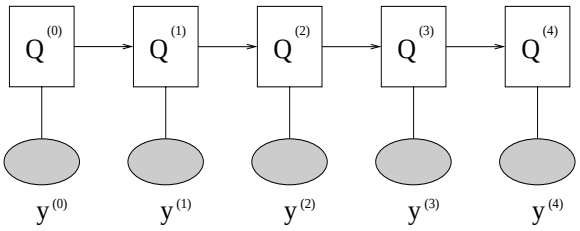
Figure 2: Top row: covariance matrices $S^{(0)}$, $S^{(1)}$, and $S^{(2)}$. Bottom row: true covariance matrix of the object before wavelet decomposition, covariance matrix synthesized from full covariance matrices $S^{(0)}$, $S^{(1)}$, $S^{(2)}$, and covariance matrix synthesized from full covariance matrix $S^{(0)}$ and the diagonal elements of $S^{(1)}$, $S^{(2)}$

2.2 Hidden Markov Model

The compression property of the covariance matrix is under the assumption that details at various levels are distributed independently and as Gaussian processes. As we discussed previously, this is not likely to be true for face images, where the distribution of each $\underline{y}^{(i)}$ is closely related with that of its parent or child level. For example, the hair condition may create a dark region in $\underline{y}^{(0)}$ and create edges on the forehead in the detail subband; the white teeth of a smiling face may create bright region in $\underline{y}^{(0)}$ and corresponding edges in detail subband; the direction of illumination (reflected in lowest resolution) may create certain patterns (reflected in detail subband) because the human face is a 3-D object. In this subsection, we describe a model to capture the statistical interdependency between the various levels of the wavelet decomposition.

Graphical models in statistics infer causal relationships between a set of random variables through the "conditional" independence structure in their joint probability density function. To capture the statistical dependency between levels, we use a graphical model in scale; specifically, we use a 1st-order Hidden Markov Model (Markov in scale) with conditionally Gaussian outputs to model the distribution of an image block. As shown in the graph below, each level of detail has a number of discrete states. The state of any level of detail depends only on the state of its parent (lower-resolution) level of detail, as

$$\begin{aligned} & Prob(Q^{(i+1)} = k | Q^{(i)} = j) \\ = \;& Prob(Q^{(i+1)} = k | Q^{(i)} = j, Q^{(i-1)} \ldots) \end{aligned} \tag{3}$$



At each scale i, the observation depends only on the discrete state $Q^{(i)}$. We assume the probability density function to be Gaussian:

$$f(\underline{y}^{(i)} | Q^{(i)} = k) = N(\underline{\mu}_k^{(i)}, S_k^{(i)}) \tag{4}$$

The complete model can be characterized by the parameter set: $\theta = \{\Pi, A^{(1)}, A^{(2)}, \ldots, A^{(m)}, \underline{\mu}_k^{(i)}, S_k^{(i)}\}$, where Π specifies the initial state probability of the lowest-scale approximation ($\Pi_k = Prob(Q^{(0)} = k)$), and matrix $A^{(i)}$ specifies the state transition probabilities ($A_{j,k}^{(i)} = Prob(Q^{(i)} = k | Q^{(i-1)} = j)$).

HMM's are very rich models, with many parameters; thus, we must ensure that we have enough training data. In particular, there are many parameters associated with the covariance matrices $S_k^{(i)}$. In order to obtain reliable parameter estimates, we assume that these covariance matrices are diagonal, thereby reducing the total number of model parameters. Another option would be using subspace methods to reduce the dimension and decorrelate the variation.

Using our model, the likelihood function can be written as:

$$\begin{aligned} & f(\underline{y}^{(0)}, \underline{y}^{(1)}, \ldots, \underline{y}^{(m)}) = \sum_{k_0=1}^{n_0} \sum_{k_1=1}^{n_1} \cdots \sum_{k_m=1}^{n_m} \\ & \Pi_{k_0} A_{k_0,k_1}^{(1)} \ldots A_{k_{m-1},k_m}^{(m)} \prod_{t=0}^{m} f(\underline{y}^{(t)} | Q^{(t)} = k_t) \end{aligned} \tag{5}$$

where n_i is the number of states at level i. The overall object image can be synthesized by summing up details at various levels (with proper upsampling operations). Given the fact that the sum of independent Gaussian variables is still Gaussian and the upsampled of a Gaussian sequence is still Gaussian, the overall object image

is distributed as Gaussian mixtures. The combination of all levels would give a total number $\prod_{i=1}^{m} n_i$ of mixtures. The weight on each mixture depends on the state transition matrices $A^{(i)}$ and Π. This Markov structure gives an approximate but efficient multivariate Gaussian mixture model. It is approximate because of the local dependency constraints. Figure 3 shows a few synthesized modes (centroids) of the Gaussian mixture trained from our face data set. This model structure also gives us flexibility of different treatment to different level of details, e.g., by choosing different number of states at different levels.

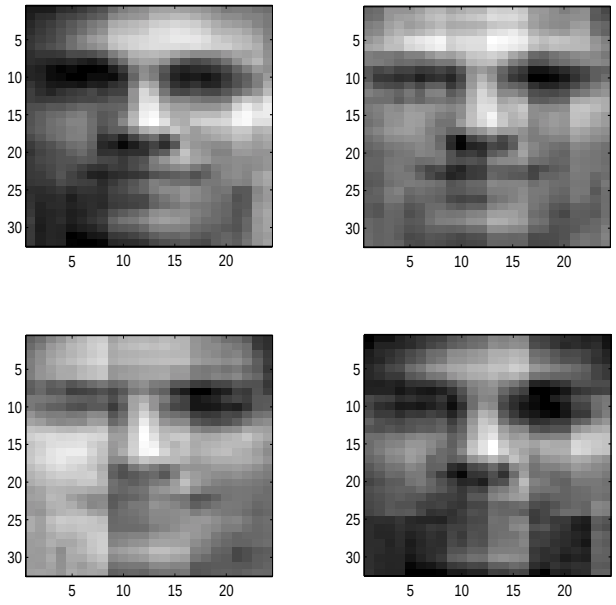


Figure 3: A few synthesized Gaussian modes with different probabilities (upper left one with largest probability)

2.3 Learning: parameter estimation

Maximum likelihood parameter estimation for HMMs requires the Expectation-Maximization (EM) algorithm because the state sequence $Q^{(i)}$ $(i = 0, 1, ...m)$ is hidden. Recall that the EM algorithm involves iterating **E-Step** (compute $Q(\theta|\theta^{(p)}) = E[\log p(\chi, Y|\theta)|Y, \theta^{(p)}]$) and **M-Step** (find new parameters $\theta^{(p+1)} = \arg\max_\theta Q(\theta|\theta^{(p)})$) where Y is the training observation set (wavelet decompositions of the data set) and χ is the associated unobserved state sequence. The E-step involves computing $\alpha_t(j) \equiv p(y^{(0)}, \ldots, y^{(t)}, x_t = j)$ by forward algorithm and computing $\beta_t(j) \equiv p(y^{(t+1)}, \ldots, y^{(m)}|x_t = j)$ by backward algorithm [16]. M-step involves updating all the parameters. For brevity, the derivations are omitted. The only difference from a standard HMM parameter estimation algorithm is that various resolutions do not share the same states.

In the training procedure, one can encounter very small values of the probability densities $f(y^{(i)}|Q^{(i)} = k)$ due to the large dimension of images. These very small values cause computer precision problems in the training algorithm. We use their relative values in the forward-backward algorithm, namely, use

$$\tilde{f}(y^{(i)}|Q^{(i)} = k) = \frac{f(y^{(i)}|Q^{(i)} = k)}{\min_s(f(y^{(i)}|Q^{(i)} = s))} \qquad (6)$$

as the observation density for each sample. One can show that this adjustment gives exactly the same result.

3 Simulation and results

We have compared the detection performance of different detectors on the "University of Michigan" face database. A group of 241 face images were used as the training set. All the faces were resized to 64×48 pixels (part of them are shown in Figure 4).

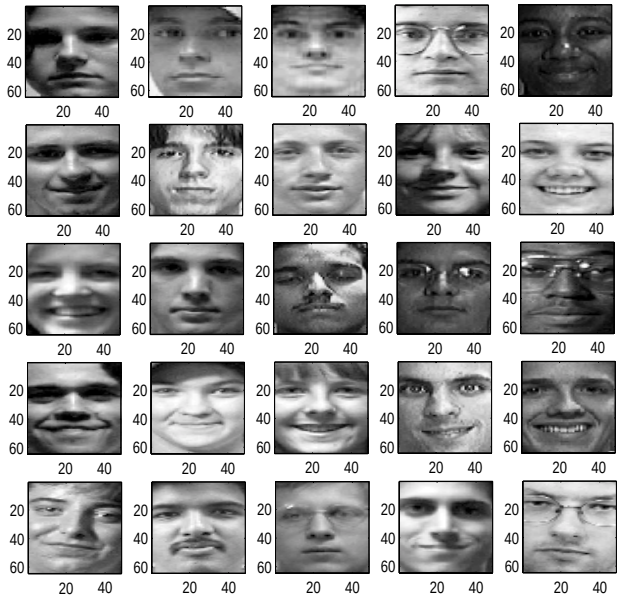


Figure 4: Some faces from training set

We used 5-level decomposition and used only 3 levels of them for both training and testing. Figure 5 shows the training result of face model using 5, 7, and 10 states for level 0, 1, and 2 respectively. The trained initial state probability Π and state transition from level 0 to level 1 $A^{(1)}$ are:

$$\Pi = \begin{bmatrix} 0.137 & 0.124 & 0.203 & 0.174 & 0.362 \end{bmatrix}$$

$$A^{(1)} = \begin{bmatrix} 0.85 & 0.00 & 0.00 & 0.00 & 0.00 & 0.15 & 0.00 \\ 0.07 & 0.00 & 0.24 & 0.53 & 0.00 & 0.10 & 0.07 \\ 0.00 & 0.33 & 0.00 & 0.00 & 0.22 & 0.43 & 0.02 \\ 0.00 & 0.00 & 0.16 & 0.39 & 0.30 & 0.02 & 0.12 \\ 0.00 & 0.19 & 0.14 & 0.07 & 0.15 & 0.01 & 0.43 \end{bmatrix}$$

The non-uniform distribution of state transition matrices verifies the local dependence across scale. A similar model for clutter was trained using the same procedure. We collected 480 randomly selected blocks of background as non-faces.

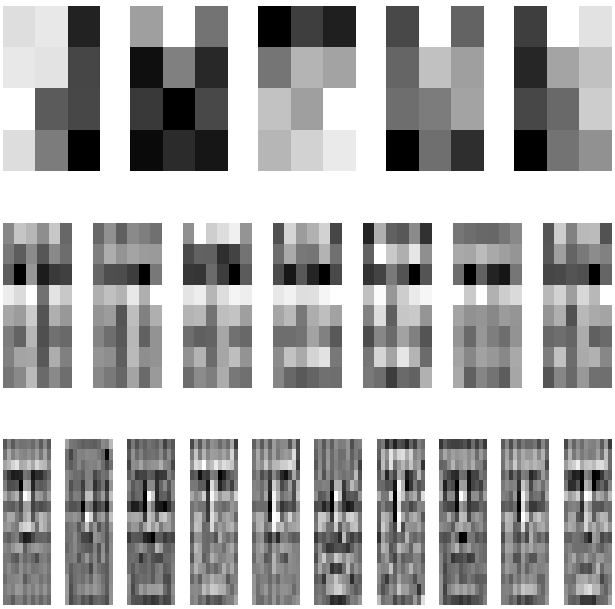


Figure 5: The training result for face model using $\underline{y}^{(0)}$, $\underline{y}^{(1)}$, $\underline{y}^{(2)}$: mean $\underline{\mu}_k^{(i)}$ at level 0, 1, and 2.

To enforce data independence, another group of 286 images (Figure 6) were used as test set. For simplicity, we test the algorithms for a single face size. All images in test set were resized so that all the faces inside were 64×48 pixels. In real applications where the actual face size is unknown, templates of all different sizes have to be applied. This is a disadvantage for all image-based approaches. In the horizontal direction of the face, we choose across the inner boundaries of two ears as the upper limit size and the lower limit size is across the outer boundaries of two eye-brows.

Usually, a detection algorithm is evaluated on an ROC curve by varying the threshold. For simplicity, and because there was only one face in each test image in our experiment, the Maximum Likelihood Ratio detection scheme was adopted. Basically, the rectangular window that gives the maximum likelihood ratio is detected as a face.

$$(i,j)^{MLT} = \arg\max_{(i,j)} \frac{f(\underline{y}_{(i,j)}|\Omega = 1)}{f(\underline{y}_{(i,j)}|\Omega = 0)} \qquad (7)$$

where $\underline{y}_{(i,j)}$ is the re-ordering of the pixel values in a local neighborhood of spatial location (i,j). It is shown [8] that if a method yields a better ROC curve, it is guaranteed to yield a lower error rate.

To illustrate the performance of our algorithm, we compared it with the following alternative image-based methods:

- Matched filter without normalization.
- Matched filter with normalization.
- Distance From Feature Space (DFFS) [17].
- Distance In Feature Space (DIFS) [17].
- Segmented distance computation [6] using 8 parts.
- Wavelet decomposition using 4 subbands [6].
- 8 part segmentation plus 4 subband wavelet decomposition [6].
- Linear discriminant [8].

The following table summarizes the error percentage of all the methods.

Approach	error
whole average face w/o normalization	46.3%
whole average face w/ normalization	29.1%
DFFS w/ 3 eigenfaces	23.1%
DFFS w/ 6 eigenfaces	20.3%
DFFS w/ 12 eigenfaces	35.3%
DIFS w/ 3 eigenfaces	23.1%
DIFS w/ 6 eigenfaces	20.3%
DIFS w/ 12 eigenfaces	35.0%
8 parts	12.9%
4 subbands	12.9%
8 parts and 4 subbands combined	7.0%
Linear Discriminant	7.3%
Proposed Hidden Markov Model	5.0%

The results show some interesting trends: the eigenface methods make more errors when large numbers of eigenfaces are used, because the extra eigenface directions can be matched to clutter. The results also illustrate that our proposed algorithm is significantly better than the alternatives in the experiments. Figure 6 shows some of the experiment faces, along with the detected faces indicated by our algorithm.

4 Discussion and Future Extensions

In summary, we built a statistical modeling framework to model the diversity of image objects by decomposing an image into orthogonal wavelet subbands and connecting them with HMM. This model, as an efficient Gaussian mixture, captures the statistical interdependency across scales. We used this framework to develop algorithms for face detection. Our experiments show that the proposed algorithm significantly outperforms comparable matched filter, eigenface and linear discriminant analysis methods in face detection tasks.

There are many directions in which our results can be extended. The most obvious limitation of this work is that we assumed all faces to have the same dimensions. In most applications, the size of the face is unknown. In practice, a limited number of sizes can be tested. Thus, the modeling framework has to be extended to incorporate moderate variations in size so that a number of detections can cover the continuous range of sizes.

A second direction is the extension of the discrete-state Hidden Markov Model used in our work to continuous states. Such models have been proposed in [18],

Figure 6: Some results using Hidden Markov Models and Maximum Likelihood Ratio detection. Note that the fourth face in the third row is not detected correctly.

and can model correlation across scale in terms of a dynamical system evolving in scale. One can extend this approach to perform statistical detection and estimation, as in [18]. A third direction is to incorporate spatial decompositions besides subband decompositions in the model. Such an approach was suggested in [6], and showed some advantages in face detection. A final direction for future research is extension of these ideas beyond face detection into face recognition and other recognition applications.

Acknowledgment

The authors acknowledge the work of the people who composed the face database website and each individual appearing in this face database.

References

[1] A.L. Yuille, P.W.Halilinman, and D.S. Cohen. Feature extraction from faces using deformable templates. *Intl. Journal of Computer Vision*, 8(2):99-111, 1992.

[2] I. Craw, D. Tock, and A. Bennet. Finding face features. *Proc. of the 2nd European Conf. on Computer Vision*, pp 92-96, Santa Margherita Ligure, Italy, May 1992.

[3] R. Brunelli and T. Poggio. Face recognition through geometrical features. *Proc. of the 2nd European Conf. on Computer Vision*, pp 972-800, Santa Margherita Ligure, Italy, May 1992.

[4] D. Reisfeld and Y. Yeshurun. Facial normalization using few anchor points. *Proc. of the 12th IAPR Intl. Conf. on Pattern Recognition*, Jerusalem, Israel, 1994.

[5] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, June 1994.

[6] L. Meng and T. Q. Nguyen. Frontal face detection using multi-segment and wavelet. *Conf. on Information Science and Systems*, 1999

[7] P. Belhumeur, J. Hespanha, and D. Kriegman, Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Analysis and Machine Intelligence* vol. 19, no. 7, pp. 711-720, July, 1997.

[8] L. Meng and T. Q. Nguyen. Frontal face localization using linear discriminant. *Asilomar Conf. on Signals, Systems, and Computers*, October, 1999.

[9] B.J. Frey, A. Colmenarez, T.S. Huang. Mixtures of local linear subspaces for face recognition. *Proc. of Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

[10] T.D. Rikert, M.J. Jones, P. Viola. A cluster-based statistical model for object detection. *Proc. of Intl Conf. on Computer Vision*, 1999.

[11] F.S. Samaria and A.C. Harter. Parameterization of a stochastic model for human face identification. *Proc. of the 2nd IEEE Workshop on Applications of Computer Vision.* December 1994.

[12] A. Nefian and M. Hayes. Face detection and recognition using hidden Markov models. *ICIP'98*

[13] A.J. Colmenarez and T.S. Huang. Face Detection with information-based maximum discrimination. *Proc. of Computer Vision and Pattern Recognition*, June 1997.

[14] J.S. De Bonet and Paul Viola. Texture recognition using a non-parametric multi-scale statistical model. *Proc. of Computer Vision and Pattern Recognition*, Santa Barbara, CA, June 1998.

[15] H Choi, B Hendricks, and R Baraniuk. Analysis of multiscale texture segmentation using wavelet-domain hidden Markov models. *Asilomar Conf. on Signals, Systems, and Computers*, October, 1999.

[16] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[17] B. Moghaddam. Probabilistic visual learning for object detection. *The 5th Intl. Conf. on Computer Vision*, Cambridge, MA, June 1995.

[18] M. Luettgen, W.C. Karl, A.S. Willsky, R.R. Tenney. Multiscale representations of Markov random fields. *IEEE Trans. on Signal Processing*, 41(12):3377-3396, 1993.