

DIC Structural HMM based IWAK-Means to Enclosed Face Data

Mohammed Alhanjouri
Asst. Prof. at Islamic university of Gaza
Gaza, Palestine

Hana Hejazi
Islamic university of Gaza
Gaza, Palestine

ABSTRACT

This paper identifies two novel techniques for face features extraction based on two different multi-resolution analysis tools; the first called curvelet transform while the second is waveatom transform. The resultant features are trained and tested via three improved hidden Markov Model (HMM) classifiers, such as: Structural HMM (SHMM), Deviance Information Criterion-Inverse Weighted Average K-mean-SHMM (DIC-IWAK-SHMM), and Enclosed Model Selection Criterion (EMC) coupled with DIC-IWAK-SHMM as the proposed methods for face recognition.

A comparative studies for DIC-IWAK-SHMM approach to recognize the face were achieved by using two type of features; one method using Waveatom features and the other method uses 2-level Curvelet features, these two methods compared with a six methods that used in previous researches.

The goal of the paper is twofold; using Deviance information criterion and IWAK-means clustering algorithm based on SHMM.

Keywords

HMM, Curvelet, Waveatom, Face Recognition, Structural HMM

1. INTRODUCTION

Face recognition has been studied extensively for more than 20 years now. Since the beginning of 90's the subject has become a major issue; mainly due to its important real-world applications in areas like video surveillance, smart cards, database security, internet and intranet access. Multiresolution analysis tools, notably wavelets, have been found quite useful for analyzing the information content of images; hence they enjoyed wide-spread popularity in areas like image processing, pattern recognition and computer vision.

After wavelets, many multiresolution tools were developed like contourlets, ridgelets, Curvelet etc. [1]. 'Waveatom Transform' is a recent addition to this list of multiscale transforms. It has already been used to resolve image processing problems but not much work has been done to explore the potential of Waveatom transform to solve pattern recognition problems. In some recent works, Waveatom transform used in image processing in the field of image denoising, and the results obtained are the best one when compared to the state of art [2].

In the stage of classification, the HMM has a good capability, the first usage of Hidden Markov models has been in speech recognition for few decades [3]. Later HMM are being applied to face recognition area. In 2000, the maximum likelihood training for the continuous mixture embedded HMM was presented and used for face detection and recognition [4]. On the

parallel line, the wavelet multiresolution analysis and HMM were combined in 2003 for face recognition. In this approach a face image is divided into a number of overlapping subimages and wavelet decomposition is performed on each of the subimages, and the performance was better than the original DCT based HMM [5].

Since HMMs are one-dimensional in nature, many researchers have tried to represent the two dimensional structural. In (2002), a generalization of the embedded hidden Markov models was used for face recognition. An application of the embedded Bayesian networks (EBNs) is presented for face recognition and introduced the improvement of this approach versus the "eigenface" and the embedded HMM approaches [6]. Later in (2003), low-complexity 2D-HMM (LC 2D-HMM) was proposed, which consists of a rectangular constellation of states, where both vertical and horizontal transitions are supported. In (2004), another approach is the 1D discrete HMM (1D-DHMM), which models a face image using two standard HMMs, one for observations in the vertical direction and one for the horizontal direction [7].

One recently developed model for pattern recognition is the structural hidden Markov models (SHMMs) [8]. This approach allows the user to weight substantially the local structures within a pattern that are difficult to disguise. This provides a SHMM recognizer with a higher degree of robustness. The concept of SHMMs has been shown to outperform HMMs in a number of applications including handwriting recognition.

Curvelet transform becomes a very popular multi-resolution transform after implementing its second generation. In face recognition, Curvelet transform seems to be promising [9-15]. The beginning was in [9], the face images were quantized from 256 to 16 and 4 gray scale resolutions, the quantized images were decomposed using Curvelet transform. Three support vector machines SVM were trained using Curvelet coefficients and the decision was made by simple majority voting. In [10] the face image undergoes Curvelet transform. PCA was performed on the approximated coefficients. K-Nearest Neighbor classifier was employed to perform the classification task. In [11] as preprocessing step researchers converted face images from 8 bit into 4 bit and 2 bit representations. Curvelet transform was performed to extract feature vectors from these representations, and then the approximated components were used to train different SVMs. Researchers in [12] addressed the problem of identifying faces when the training face database contains one face image of each person. The Curvelet approximated coefficients was framed as a minimization problem. The original image and the reconstructed images of the non-linear approximations were used to generate the training set. A comparative study amongst Wavelet and Curvelet was found

in [13]. In [14] the Curvelet sub-bands were divided into small sub-blocks. Means, variance and entropy were calculated from these sub-blocks as statistical measures. Feature vector was constructed by concatenating each block measure. Local discriminant analyses (LDA) was carried out on feature vectors and the city-block distance was used for classification. Researcher in [15] decomposed a face image using Curvelet transform at scale 4. Next Least Square Support Vector Machine (LS-SVM) was trained using Curvelet features. The results in [11, 12, 13, 14, 15] have showed Curvelet based schemes were better than wavelet based recognition schemes[9].

Wavelet Packet, Cosine Packet and Wave Atom Transforms based electrocardiogram (ECG) compression is presented in 2009 [16].

2. FEATURE EXTRACTION IN TRANSFORM DOMAIN

Feature extraction is the most important step for any face recognition system. In reality, using local features is a mature approach to face recognition problem. In this paper we will explain the extracted feature based Curvelet and Waveatom transform.

2.1 Curvelet transform

Curvelets was proposed by E. Candes and D. Donoho (2000) [17]. The idea of Curvelets is to represent a curve as a superposition of functions of various lengths and widths obeying the scaling parabolic law: $width \cong (length)^2$.

There is two generations of Curvelet transform. The first generation defines Curvelet between Wavelet and multiscale Ridgelet.

In the second generation, two different implementations of Curvelet were founded: The first digital transformation is based on Unequally Spaced Fast Fourier Transform (USFFT), while the second is based on the wrapping of specially selected Fourier samples. The two implementations essentially differ by the choice of spatial grid used to translate Curvelets at each scale and angle. Where, a tilted grid mostly aligned with the axes of the window which leads to the USFFT. On the other hand, a grid aligned with the input Cartesian grid which leads to the wrapping-based. Both digital transformations having the same output, but the Wrapping Algorithm gives a more intuitive algorithm and faster computation time [18]. Therefore, Curvelet via wrapping will be used for this work.

If we have the object $g[t_1, t_2]$, $t_1 \geq 0$, $t_2 < n$ as Cartesian array and $\hat{g}[n_1, n_2]$ to denote its 2D Discrete Fourier Transform, then the architecture of Curvelets via wrapping is as follows:

1. 2D Fast Fourier Transform (FFT) is applied to $g[t_1, t_2]$ to obtain Fourier samples $\hat{g}[n_1, n_2]$.
2. For each scale j and angle l , the product $\tilde{U}_{j,l}[n_1, n_2] \hat{g}[n_1, n_2]$ is formed, where $\tilde{U}_{j,l}[n_1, n_2]$ is the discrete localizing window.
3. This product is wrapped around the origin to obtain $\tilde{g}_{j,l}[n_1, n_2] = W(\tilde{U}_{j,l} \hat{g})[n_1, n_2]$; where the range for n_1, n_2 is now $0 \leq n_1 < L_{1,j}$ and $0 \leq n_2 < L_{2,j}$; $L_{1,j} \approx 2^j$ and $L_{2,j} \approx 2^{j/2}$ are constants.
4. Inverse 2D FFT is applied to each $\tilde{g}_{j,l}$, hence creating the discrete Curvelet coefficients.

2.2 Waveatom transform

L. Demanety and L. Ying presented a new member in the family of oriented, multiscale transforms for image processing and numerical analysis. This is called Waveatom transform [19].

Suppose j, m, n are integer valued where j is the cutoff in scale, n is the cutoff in space and m labels the different wedges within each scale. Consider a one-dimensional family of wave packets $\psi_{m,n}^j(x)$, $j \geq 0, m \geq 0, n \in \mathbb{N}$, centered in frequency around $\pm\omega_{j,m} = \pm\pi 2^j m$ with $c_1 2^j \leq m \leq c_2 2^j$ where $c_1 < c_2$ are positive constants, and centered in space around $x_{j,n} = 2^{-j}n$. One-dimensional version of the parabolic scaling states that the support of each bump of $\psi_{m,n}^j(\omega)$ is of length $O(2^j)$ while $\omega_{j,m} = O(2^{2j})$. Dyadic dilates and translates of $\hat{\psi}_m^0$ on the frequency axes are combined and basis functions, written as:

$$\psi_{m,n}^j(x) = \psi_m^j(x - 2^{-j}n) = 2^{j/2} \psi_m^0(2^j x - n) \quad (1)$$

The transform $WA : L^2(\mathbb{R}) \rightarrow l^2(\mathbb{Z})$ maps a function u onto a sequence of waveatom coefficients

$$c_{j,m,n} = \int_{-\infty}^{\infty} u(x) \psi_{m,n}^j(x) dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iz^{-j}n\omega} \overline{\hat{\psi}_m^j(\omega)} \hat{u}(\omega) d\omega \quad (2)$$

If the function u is discretized at $xk = kh$, $h=1/N$, $k=1 \dots N$, then with a small truncation error (3) is modified as:

$$C_{j,m,n}^D = \sum_{k=2\pi(-N+2+1:1:N=2)} e^{-iz^{-j}nk} \overline{\hat{\psi}_m^j(k)} \hat{u}(k) \quad (3)$$

A simple wrapping trick is used for the implementation of discrete wavelet packets and the steps involved are:

1. Perform an FFT of size N on the samples of (k) .
2. For each pair (j, m) wrap the product $\hat{\psi}_m^j \hat{u}$ by periodically inside the interval $[-2^j \pi, 2^j \pi]$ then perform inverse FFT of size 2^j of the result to obtain $C_{j,m,n}^D$.
3. Repeat step 2 for all pairs (j, m) .

The positive and negative frequency components represented by:

$$\hat{\psi}_{m,n}^j(\omega) = \hat{\psi}_{m,n}^+(\omega) + \hat{\psi}_{m,n}^-(\omega) \quad (4)$$

Hilbert transform $H\hat{\psi}_{m,n}^j(\omega)$ of eq. (4) represents an orthonormal basis $L^2(\mathbb{R})$ and is obtained through a linear combination of positive and negative frequency bumps weighted by i and $-i$ respectively.

$$H\hat{\psi}_{m,n}^j(\omega) = -i\hat{\psi}_{m,n}^+(\omega) + i\hat{\psi}_{m,n}^-(\omega) \quad (5)$$

To extend Waveatom to be 2D, let $\mu = (j, \mathbf{m}, \mathbf{k})$, where $\mathbf{m} = (m_1, m_2)$ and $\mathbf{n} = (n_1, n_2)$, so from equation 1.

$$\psi_{\mu}^+(x_1, x_2) = \psi_{m_1}^j(x_1 - 2^{-j}n_1) \psi_{m_2}^j(x_2 - 2^{-j}n_2) \quad (6)$$

A dual orthonormal basis, which is defined from the Hilbert-transformed,

$$\psi_{\mu}^-(x_1, x_2) = H\psi_{m_1}^j(x_1 - 2^{-j}n_1) H\psi_{m_2}^j(x_2 - 2^{-j}n_2) \quad (7)$$

By now, combine the primal and dual (Hilbert-transformed) basis. More precisely, the recombination

$$\psi_{\mu}^{(1)} = \frac{\psi_{\mu}^{+} + \psi_{\mu}^{-}}{2}, \psi_{\mu}^{(2)} = \frac{\psi_{\mu}^{+} - \psi_{\mu}^{-}}{2}$$

This combination provides basis functions with two bumps in the frequency plane, symmetric with respect to the origin, hence purely directional wave atoms. Together, $\psi_{\mu}^{(1)}$ and $\psi_{\mu}^{(2)}$ form the wave atom frame and may be denoted jointly as ψ_{μ} [19].

3. HMM BASED FACE RECOGNITION

Each feature vector is modeled by continuous left-to-right HMMs. Each HMM state generates a mixture of Gaussian densities. The number of states and the number of densities per state that are appropriate to model each class depend on the amount of training data available for that class. Due to this, these numbers need some empirical tuning.

In the training phase of face recognition system, each individual class in the database is represented by a HMM face model. First, the HMM is initialized. The image coefficients are segmented from top to bottom where each segment corresponds to a state, and the image data within a region is modeled by a multivariate Gaussian distribution. An observation sequence consists of all intensity values from each block. Next, model parameters are re-estimated by a process called E-M procedure to maximize the model probability until convergence. In other words, one state is responsible for characterizing the observation vectors of human foreheads, and another state is responsible for characterizing the observation vectors of human eyes. For face localization, an HMM is trained for a generic model of human faces from a large collection of face images (as shown in Figure 1). [20]

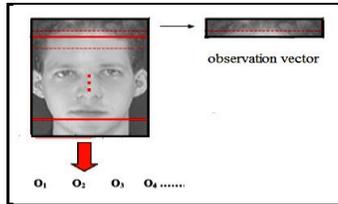


Fig 1: HMM observation vector

To classify a tested face image, we have to find the HMM with the highest probability in the Viterbi decoding. So each face image would classify with independence of each other. A varied window feature vector is used by using five states and optimizing both of number of feature vector in the window, and the number of Gaussians densities per state.

1. THE PROPOSED METHOD

In this section, the approaches that used in our proposal methods for classifying the features of human face are explained.

4.1 The Deviance Information Criterion

One of the problems standard HMM suffers from is the need to know the number of parameter in advance. To overcome this shortcoming, the Deviance information criterion (DIC) was proposed. The DIC was defined as:

$$DIC = \overline{D(\tilde{M})} + P_d \quad (8)$$

Where $\overline{D(\tilde{M})}$ is a measure of how well the model M fits the data which correspond to the expectation with respect to $P(M|X)$, P_d

is the effective number of parameters of the model, and $D(M)$ was defined as:

$$D(M) = -2 \log P(X|M) \quad (9)$$

Where $X = \{x_1, x_2, x_3 \dots x_c\}$. is the pattern contains sequence of local structure $S = \{s_1, s_2, s_3 \dots s_c\}$. Let \tilde{M} is the posterior mean of effective parameter,

$$P_d = \overline{D(\tilde{M})} - D(\tilde{M}) = E(-2 \log P(X|\tilde{M})) + 2 \log P(X|\tilde{M}) \quad (10)$$

The best model fits such data will have larger likelihood and smaller deviance. DIC was reformulated to,

$$DIC = -2 \log P(X|\tilde{M}) + 2P_d \quad (11)$$

$P(X|\tilde{M})$ can be obtained using forward algorithm for HMM. In order to calculate P_d , the expected value $E(-2 \log P(X|M))$ has to be approximated ,

$$P(X|M) = \sum_S P(X, S|M) \approx P(X, S) \quad (12)$$

$$P(X, S) = P(S, X) = P(S|X) \times P(X) = P(s_c | s_{c-1} \dots s_2 s_1 x_c \dots x_1) \times P(s_c s_{c-1} \dots s_2 s_1 | x_c \dots x_1) \times P(X) \quad (13)$$

When it is assumed that s_c depends only on x_c and s_{c-1} , and the structure probability distribution is a Markov chain of order 1 [21].

4.2 DIC-SHMM

A structural hidden Markov model is $M = [\pi, A, B, S, D]$, where: π is the initial state probability vector, A is the state transition probability matrix, B is the state conditional probability matrix of the visible observations, S is the posterior probability matrix of a structure given a sequence of observations, and D is the structure transition probability matrix.

The DIC-SHMM is characterized by the following elements:

1. H is the number of hidden states in the model. The individual states are labeled as $1, 2, \dots, H$, and denote the state at time t as q_t .
2. M , is the number of distinct observations o_i .
3. $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = i)$, $1 \leq i \leq H$, $\sum_i \pi_i = 1$.
4. $A = \{a_{ij}\}$ where $a_{ij} = P\{q_{t+1} = j | q_t = i\}$, $1 < i, j < H$, $0 < a_{ij} < 1$, $\sum_{j=1}^H a_{ij} = 1$.
5. $B = \{b_j(l)\}$ where $b_j(l) = P(o_l | q_j)$, $1 \leq l \leq M$, $1 \leq j, H$, $\sum_l b_j(l) = 1$
6. F is the number of distinct local structure.
7. $S = s_i(j)$ where $s_i(j) = P(S_j | O_t)$, $\sum_j s_i(j) = 1$ for each particular string O_i .
8. $D = \{d_{ij}\}$ where $d_{ij} = P\{S_{t+1} = j | S_t = i\}$, $1 < i, j < F$, $\sum_j d_{ij} = 1$

The DIC-HMM has four problems: probability evaluation, statistical decoding, structural decoding, and parameter estimation.

The first problem is to determine the probability for such a model to produce such a sequence X . It has been proven in [22] that the likelihood function of the observation sequence in equation (14) can be expressed as:

$$P(X|M) = \sum_S \left[\prod_{i=1}^c \frac{s_i(i) \times d_{i-1,i}}{P(s_i)} \right] \times \sum_q \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \dots a_{q_{(T-1)q_T}} b_{q_T}(o_T) \quad (16)$$

Then calculate P_d by substituting Equation 16 in Equation 10. The DIC value would be calculated by Equation 11. An $P(X|\tilde{M})$ can be obtained by summing over forward probability in the forward algorithm as in traditional HMM.

The second problem is to determine the optimal state sequence. It is computed using Viterbi algorithm as in HMM.

The third problem is to determine the optimal structure sequence. It can be computed using Viterbi algorithm.

The last one is parameter estimation:

1. D was estimated by Baum-Welch optimization technique.
2. S was estimated by constrained Gaussian mixture posterior probability estimation technique.
3. B was estimated by DIC instead of using ML in HMM.
4. A and π were estimated as in traditional HMM.

4.3 Inverse Weighted Average K-means Clustering Algorithm

Another weakness of HMM is that it depends on k-means clustering algorithm to defined Gaussian mixture for each model. In few words, k-means algorithm tries to partitioning a data set into K prototypes where these prototypes in some way best represent the data. K-means suffers of many problems; sensitivity to prototypes initialization, dead prototypes, converges to local optimum and needs specifying the number of clusters in advance [23]. EM algorithm solves the last problem, but HMM model still suffer of the other ones. Actually, k-means uses Euclidian distance to centralize the prototypes, first we need to normalize this distance to work with classes have different number of data points. The average distance seems to be compromised choice.

Suppose x is the data point, m is the prototype, N is the number of data points, and k is the number of prototypes,

$$perf_{k_{means}} = \sum_{i=1}^N \min_{j=1}^k \|x_i - m_j\|^2 \quad (17)$$

$$perf_{avg} = \frac{1}{N} \sum_{i=1}^N \min_{j=1}^k \|x_i - m_j\|^2 \quad (18)$$

The performance function was needed should have the following properties: first it would be minimum performance function to give good clustering. Second, it would contain relationships between all data points and all prototypes. Third, the distance between points and nearest prototype and maximize the distance between the prototypes.

$$perf_{new} = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{k-1} \sum_{\substack{j=1 \\ j \neq i}}^k \|m_i - m_j\|^2 \right] \min_{j=1}^k \|x_i - m_j\|^2 \quad (19)$$

Now, the problem with this formula is that the function maximizes the distance between prototypes diverge, since the limit when goes to infinity of its integral does not exist. Let $\|m_i - m_j\| = n$,

$$\lim_{n \rightarrow \infty} \int_{n=1}^k n^2 dn = \lim_{n \rightarrow \infty} \left(\frac{n^2}{2} - \frac{1}{2} \right) = \infty \quad (20)$$

So the maximization function would be the distance between the data points in such cluster and the other prototypes. This distance would be used as a weight in the minimization function, so the absolute value to the maximization function could be used. Both of the two distances (in-class distance and out-class distance) are convergence functions; since the limit when goes to infinity (zero) of out-class (in-class) distance integral exists and equals some constant.

Let $n = (x_i - m_j)$,

$$\lim_{n \rightarrow \infty} \int_{n=1}^k \frac{1}{n^2} dn = \ln(n(k-1)) - \ln 1 = constant \quad (21)$$

$$\lim_{n \rightarrow 0} \int_{n=1}^k n^2 dn = \lim_{n \rightarrow 0} \left(k^3 - \frac{1}{3} \right) = constant \quad (22)$$

This will yield to the following performance function:

$$perf_{new} = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{N(k-1)} \sum_{\substack{j=1 \\ j \neq i}}^{Nk} \frac{1}{|x_i - m_j|} \right] \min_{j=1}^k \|x_i - m_j\|^2 \quad (23)$$

First, the prototypes will distributed amongst the data points, then the distance between data points and un-nearest prototypes will calculated. Finally, this distance will be used as a weighted factor to update the prototypes position, this is will continue till convergence.

In order to optimize the performance, the partial derivative of this performance function with respect to such prototype would be calculated first, assign to zero, and solved for that prototype. Let m_c is the closest prototype to x_i and m_r is the set of other prototypes:

$$perf = \left[\frac{1}{N(k-1)} \sum_{\substack{r=1 \\ r \neq c}}^{Nk} \frac{1}{|x_i - m_r|} \right] \times |x_i - m_c|^2 \quad (24)$$

$$\begin{aligned} \frac{\partial perf(x_i)}{\partial m_c} &= -(x_i - m_c) \times \left[\frac{(2)}{N(k-1)} \sum_{\substack{r=1 \\ r \neq c}}^{Nk} \frac{1}{|x_i - m_r|} \right] \\ &= -(x_i - m_c) \times a_{ic} \end{aligned} \quad (25)$$

$$\begin{aligned} \frac{\partial perf(x_i)}{\partial m_r} &= -(x_i - m_c) \times \left[\frac{(x_i - m_c)}{N(k-1)} \sum_{\substack{r=1 \\ r \neq c}}^{Nk} \frac{1}{|x_i - m_r|^2} \right] \\ &= -(x_i - m_c) \times b_{ir} \end{aligned} \quad (26)$$

$$\frac{\partial perf}{\partial m_j} = \sum_{i=1}^N \frac{\partial perf(x_i)}{\partial m_j} = 0 \quad (27)$$

$$m_j = \frac{\sum_{i \in V_j} x_i a_{ij} + \sum_{i \in V_k, k \neq j} x_i b_{ij}}{\sum_{i \in V_j} a_{ij} + \sum_{i \in V_k, k \neq j} b_{ij}} \quad (28)$$

Where V_j is a set contains all of data points belong to cluster j , and V_k is a set contains all data points.

4.4 Enclosed Model Criteria

HMM chooses the Maximum Likelihood model (ML) using the Maximum Likelihood criterion as follows:

$$ML = \frac{1}{N} \log p(X|\theta_{ml}) \quad (29)$$

Where X is the dataset, and θ_{ml} is the maximum likelihood estimates parameters. HMM likelihood increases with the number of parameters in the model. The shortcoming of ML criterion is that there is no penalty to prevent the number of model parameters from increasing. On the other side, maximum likelihood estimation methods two attractive points; the first point is that they have good convergence properties as the number of training samples increases. The second point is that maximum likelihood estimation often can be found simply.

HMM is good to deal with noisy or distorted data but it tends towered outputting high probability not for the proper class data, but for out-class data too. So HMM discriminant is reduced.

To overcome this problem, it's needed to build enclosed models (out-class) for each class using included and occluded data. The procedure as follows: first of all, HMM standard model was trained using in-class data. Second, standard models recognition results are used to defined the confusion sets data. Again, HMM standard model was trained using out-class data. Finally, for each class, the final model (ML_{fin}) was calculated using the following equation:

$$\begin{aligned} ML_{fin} &= \frac{1}{N} \log p(X|\theta_{ml}) - \frac{1}{N'} \log p(X'|\theta'_{ml}) \\ &= ML_{in} - ML_{out} \end{aligned} \quad (30)$$

When θ_{ml} and θ'_{ml} , X and X' and N and N' , are the maximum likelihood estimate of parameters for in-class and out-class, in-class and out-class training data sets, and number for in-class and out-class training data sets respectively. ML_{in} and ML_{out} in-class model and out-class model respectively [24].

For DIC-SHMM, the final model (DIC_{fin}) is the difference between in-class model (DIC_{in}) and out-class model (DIC_{out}).

$$DIC_{fin} = DIC_{in} - DIC_{out} \quad (31)$$

5. EXPERIMENTAL RESULTS

This section shows the Experimental results obtained and how it obtained, discusses it in deep, and compares it to show the benefit to use the proposed method.

5.1 Data Collection

Experiments were carried out using three datasets from different sources: ORL (AT&T) database, Essex Grimace database and Yale database, all sets are used to implement different Algorithms to recognize the human face.

ORL (AT&T) database [25] contains distinct face images sets for 40 persons with dimension of 92×112 , and each set consists of 10 different images for the same person. For some persons, images were taken at different times varying the lighting, facial expression (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). All the images were taken against a dark homogeneous background with the faces in an upright,

frontal position (with tolerance for some side movement). Sample images of this dataset are shown in Figure 2.



Fig 2: Sample Images from ORL database.

Essex Grimace database [26] contains sequence face images for 18 persons each one has 20 images (180×200), all images taken with a fixed camera for male and female. During the sequence, the subject moves his/her head and makes grimaces which get more extreme towards the end of the sequence. Images are taken against a plain background, with very little variation in illumination. Sample images of this database are shown in Figure 3. For the purposes of the experiments carried out, the Essex faces were converted to grayscale before the training step.

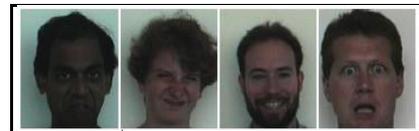


Fig 3: Sample Images from Essex Grimace Database.

Yale dataset [27], contains 165 single light source images of 15 persons under different viewing conditions (poses and illuminations). The image dimension is 510×364 . Images are taken against a white homogenous background. Sample images of this database are shown in Figure 4.

As preprocessing, first Color images of Essex Grimace database are converted to gray scale images using equation 32. Let R , G , and B be red, green, and blue value of colored image,

$$gray_{value} = 0.2989 \times R + 0.587 \times G + 0.114 \times B \quad (32)$$

For each dataset, all face images are quantized into 8 gray levels. The intensity image was scaled and rounds produce an equivalent indexed image. Then cropped face region and resized it into 120×120 .



Fig 4: Sample Images from Yale database.

In order to build a enclosed datasets; included data sets were built by add gray level to face image in different parts of faces while occluded data sets were built by cropping different parts of faces. Bellow, Figure 5 shows samples of out-class image from ORL and Yale datasets. Indeed, for each dataset, each person has two enclosed data classes; the first class is included data class which contains four included face images and six normal face images, the second class is occluded data class which contains four occluded face images and six normal face images.



Fig 5: Out-class Image Samples.

5.2 Simulation Results

In order to assess the efficiency of the proposed technique described in the previous section, a series of experiments were carried out using all databases separately. The experiments were carried out using five-fold cross validation. This involved dividing the training set images for each individual into five equally sized sets and using four of the sets for system training and the remainder being used for testing. The experiments were repeated five times with a different set being used for testing each time, to provide a more accurate recognition figure. Therefore, with both of ORL and Yale databases, eight images were used for training and two for testing during each run (one person). When using the Essex 95 database, 16 images were used for training and four for testing during each run. One SHMM model was trained for each individual in the database. During testing, an image was assigned an identity according to the SHMM model that produced the highest likelihood value. It was assumed that all testing individuals were known individuals. Accuracy of an individual run is thus defined as the ratio of correct matches to the total number of tested face images, with final accuracy equaling to the average accuracy from all of the five cross validation runs.

The initial set of experiments was designed to establish if DIC-IWAK-SHMM provided a benefit over SHMM for face recognition. DIC-IWAK-SHMM does not need to initialize the number of parameter like traditional SHMM. The experiments were carried out in both the Curvelet domain and Waveatom domain. The recognition accuracy for DIC-IWAK-SHMM face recognition is presented in Table 1. As can be seen from the table, the use of DIC-IWAK-SHMM instead of SHMM increases the recognition accuracy in all cases tested.

Table 1 (a) reported the accuracy rate when tested using ORL database. For Curvelet features, the accuracy rate increased from 96.75% for SHMM to 98% for DIC-IWAK-SHMM. The incorrect match rate for DIC-IWAK-SHMM is near 39% lower than SHMM model. On the Waveatom side, the performance increased to from 97.5% for SHMM versus 99% for DIC-IWAK-SHMM. There is an evident decreasing of about 40% in the rate of false classification.

Table 1 (b) summarized the correct classification rate when tested using Yale database. On the Curvelet side, the performance increased from 91.33% when SHMM was used to 94% when DIC-IWAK-SHMM was used. On Waveatom side, the performance increased from 93.33% when using SHMM against 95.33% when using DIC-IWAK-SHMM.

Table 1 (c) illustrated Essex Grimace database accuracy rate when used for testing. On the Curvelet side, the accuracy rate increased from 89.72% when SHMM was used up to 93.06% when DIC-IWAK-SHMM used. On Waveatom side, the

accuracy rate increased from 91.94% when using SHMM up to 94.44% when using DIC-IWAK-SHMM model.

Table (1): Comparison of face identification accuracy (%) using SHMM and DIC-IWAK-SHMM

(a): On ORL database		
Feature Extraction method	SHMM	DIC-IWAK-SHMM
Two levels Curvelet	96.75	98
Waveatom Transform	97.5	99
(b): On Yale database		
Feature Extraction method	SHMM	DIC-IWAK-SHMM
Two levels Curvelet	91.33	94
Waveatom Transform	93.33	95.33
(c): On Essex Grimace database		
Feature Extraction method	SHMM	DIC-IWAK-SHMM
Two levels Curvelet	89.72	93.06
Waveatom Transform	91.94	94.44

5.3 Enclosed Model Selection criteria

The last set of experiments was performed to study the effect of using DIC-IWAK-SHMM on enclosed data sets, for face recognition. The experiments were carried out in both the Curvelet domain and Waveatom domain. Table 2 summarizes the obtained results from included and occluded data sets.

Tables 2 (a) listed the results obtained for Curvelet features based DIC-IWAK-SHMM. It can be seen the big range in performance dropping due to included and occluded data sets. Actually, with ORL dataset the accuracy decreased of 10.75% and 18.75% for included and occluded data sets respectively. Also, the accuracy of Yale database decreased of 14.67% for included set and of 20.67% for occluded set. Finally for Essex Grimace, the accuracy decreased of 13.89% and 21.67% for included and occluded data sets respectively.

Tables 2 (b) viewed the results obtained for Waveatom features based DIC-IWAK-SHMM. It can notice that exist of huge decreasing in accurate rate due to inclusion and occlusion effects. Actually, with ORL dataset the success rate decreased from 99% to 88.5% and 85.5% for included and occluded data sets respectively. Also, the accuracy rate of Yale database decreased from 95% to 84% for included set and to 78.67% for occluded set. Finally for Essex Grimace, the accuracy rate decreased from 94.44% to 82.5% and 77.22% for included and occluded data sets respectively.

Table (2): Comparison of DIC-IWAK-SHMM identification accuracy (%) using both normal and enclosed datasets

(a): Curvelet features.			
Dataset	ORL	Yale	Essex Grimace
Normal dataset	98	94	93.06
Inclusion dataset	87.25	79.33	79.17
Occlusion dataset	79.25	73.33	71.39
(b): Waveatom features.			
Dataset	ORL	Yale	Essex Grimace
Normal dataset	99	95.33	94.44
Inclusion dataset	88.5	84	82.5
Occlusion dataset	85.5	78.67	77.22

The remaining set of experiments was performed to show the benefit of using enclosed model selection criterion (EMC) coupled with DIC-IWAK-SHMM for face recognition, as the proposed method. The experiments were carried out in both the

Curvelet domain and Waveatom domain. The recognition accuracy is presented in Table 3 for included data sets. On the side of occluded data sets, the recognition accuracy is presented in Table 4. As can be noticed from the tables, the use of enclosed model selection criterion DIC-IWAK-SHMM increases recognition accuracy in all cases tested.

Table 3 (a) reported the accuracy rate when tested using ORL database. For Curvelet features, the accuracy rate increased from 87.25% for DIC-IWAK-SHMM to 88.75% for proposed method. The incorrect match rate for proposed method is near 11.8% lower than DIC-IWAK-SHMM model. On the Waveatom side, the performance increased to from 88.5% for DIC-IWAK-SHMM to 89.75% for the proposed method. There is an evident decreasing of about 10.8% in the rate of false classification.

Table (3): Comparison of DIC-IWAK-SHMM and proposed method (EMC coupled with DIC-IWAK-SHMM) with

Inclusion dataset		
(a): On ORL database		
Extraction method	DIC-IWAK-SHMM	Proposed method
2 levels Curvelet	87.25	88.75
Waveatom	88.5	89.75
(b): On Yale database		
Extraction method	DIC-IWAK-SHMM	Proposed method
2 levels Curvelet	79.33	82
Waveatom	84	86
(c): On Essex Grimace database		
Extraction method	DIC-IWAK-SHMM	Proposed method
2 levels Curvelet	79.17	81.67
Waveatom	82.5	84.44

Table 3 (b) summarized the correct classification rate when tested using Yale database. On the Curvelet side, the performance increased from 79.33% when DIC-IWAK-SHMM was used to 82% when the proposed method was used. On Waveatom side, the performance increased from 84% when using DIC-IWAK-SHMM to 86% when using the proposed method.

Table 3 (c) illustrated Essex Grimace database accuracy rate when used for testing. On the Curvelet side, the accuracy rate increased from 79.17% when DIC-IWAK-SHMM used up to 81.67% for the proposed method. On Waveatom side, the accuracy rate increased from 82.5% when using DIC-IWAK-SHMM model up to 84.44% when using the proposed method.

Table 4 (a) reported the accuracy rate when tested using ORL database. For Curvelet features, the accuracy rate increased from 89.25% for DIC-IWAK-SHMM to 81.5% for proposed method. The incorrect match rate for proposed method is near 10.8% lower than DIC-IWAK-SHMM model. On the Waveatom side, the performance increased to from 85.5% for EMC-SHMM to 87% for the proposed method. There is decreasing of about 10.3% in the rate of false classification.

Table 4 (b) summarized the correct classification rate when tested using Yale database. On the Curvelet side, the performance increased from 73.33% when DIC-IWAK-SHMM was used to 76% when the proposed method was used. On Waveatom side, the performance increased from 78.67% when using DIC-IWAK-SHMM to 80.67% when using the proposed method.

Table 4 (c) illustrated Essex Grimace database accuracy rate when used for testing. On the Curvelet side, the accuracy rate increased from 71.39% when DIC-IWAK-SHMM used up to 74.17% for the proposed method. On Waveatom side, the accuracy rate increased from 77.22% when using DIC-IWAK-SHMM model up to 79.44% when using the proposed method.

Table (4): Comparison of DIC-IWAK-SHMM and proposed method (EMC coupled with DIC-IWAK-SHMM) with Occlusion dataset

(a): On ORL database		
Extraction method	DIC-IWAK-SHMM	Proposed method
2 levels Curvelet	79.25	81.5
Waveatom	85.5	87
(b): On Yale database		
Extraction method	DIC-IWAK-SHMM	Proposed method
2 levels Curvelet	73.33	76
Waveatom	78.67	80.67
(c): On Essex Grimace database		
Extraction method	DIC-IWAK-SHMM	Proposed method
2 levels Curvelet	71.39	74.17
Waveatom	77.22	79.44

5.4 Comparative Study

For comparison purposes, an experiment was performed to find the accuracy for DWT/SHMM when using five images from the ORL database and Yale for training and five images for testing. As Table 5 shows, the DIC-IWAK-SHMM approach to face recognition compares well with other techniques from the literature that have used these training sets.

The results demonstrate that high recognition accuracy can be achieved using DIC-IWAK-SHMM approach for face recognition. The proposed technique has been found to be robust against significant variation in illumination and facial details (present in ORL and Yale) as well. When compared to the best performing method in the table (Curveletface + PCA + LDA [13]), DIC-IWAK-SHMM features work the best and show well gain in accuracy for ORL and YALE respectively.

Table 5: Comparative results

Method	ORL	YALE
Curveletface + PCA [27]	96.6%	83.9%
Curveletface + PCA + LDA [27]	97.7%	92%
Curvelet-LDA [14]	98%	93.3%
DCT/HMM [4]	84%	NA
DWT/SHMM[22]	97%	NA
Ridgelet/SHMM[22]	94.7%	NA
Curvelet/ DIC-IWAK-SHMM (Proposed)	98%	94%
Waveatom/ DIC-IWAK-SHMM (Proposed)	99%	95.3%

6. CONCLUSION

This paper describes an DIC-IWAK-SHMM based approach for face recognition that uses Curvelet and Waveatom coefficients as features, this method is compared to the earlier HMM-based face recognition systems in [4, 14, 22, 28], where the DCT/HMM, Curvelet-LDA, Curvelet+PCA+LDA, DWT/SHMM, and Ridgelet/SHMM approaches are used.

The two proposed methods were carried out using three common databases: ORL (A&AT), Essex Grimace, and Yale databases, that are used to implement most of previous researches.

The proposed methods introduce the best recognition rate for all three databases, and the best result was obtained by using waveatom features for DIC-IWAK-SHMM and the second highest rate was achieved using curvelet feature for the same classifier. We can conclude that the proposed improved SHMM is more efficient and robust for face recognition, this improvement accomplished by using IWAK Means Clustering Algorithm, and Deviance Information Criterion.

7. REFERENCES

- [1] Mandal T. and Wu Q., 2008. Face Recognition using Curvelet Based PCA, 19th International Conference on Pattern Recognition (ICPR).
- [2] Waveatom: <http://www.waveatom.org>
- [3] Rabiner L., 1989. A tutorial on hidden markov models and selected applications in speech recognition”, IEEE Proc., vol. 77, no. 2, pp: 257–286.
- [4] Nefian A., Hayes M., 2000. Maximum likelihood training of the embedded HMM for face detection and recognition, Proc. Of the IEEE International Conference on Image Processing, ICIP, Vol. 1, 10-13 September 2000, Vancouver, BC, Canada, pp. 33-36.
- [5] Bai L., Shen L., 2003. Combining Wavelets with HMM for Face Recognition, 23rd International Conference on Innovative Techniques and Applications of Artificial Intelligence (SGAI '03), Cambridge, UK, 13-15 December, pp. 227-234.
- [6] Nefian A., 2002. Embedded Bayesian networks for face recognition, Proc. of the IEEE International Conference on Multimedia and Expo, Vol. 2, , Lusanne, Switzerland, 26-29 August 2002, pp. 133-136.
- [7] Le H., Li H., 2004. Recognizing frontal face images using hidden Markov models with one training image per person, ICPR, vol. 1, p: 318–321.
- [8] Bouchaffra D., Tan J., 2006. Introduction to structural hidden markov models: application to handwritten numeral recognition, Intelligent Data Analysis Journal, Vol. 10, No.1.
- [9] Mandal T., Majumdar A., Wu Q., 2007. Face Recognition by Curvelet Based Feature Extraction, Proc of ICIAR, Montreal, Canada, vol. 4633, 22-24 August 2007, pp 806-817.
- [10] Mandal T., Wu Q., 2008. Face Recognition using Curvelet Based PCA, Pattern Recognition, ICPR, 19th international conference, Tampa, Florida, USA.
- [11] Majumdar A., Bhattacharya A., 2007. Face Recognition by Multiresolution Curvelet Transform on Bit Quantized Facial Images, International Conference on Computational Intelligence and Multimedia Applications, vol. 2, 13.
- [12] Majumdar A., Ward R., 2008. Single image per person face recognition with images synthesized by non-linear approximation, 15th IEEE International Conference on Image Processing, ICIP 2008, P: 2740–2743.
- [13] Majumdar A., Ward R., 2008. Multiresolution Methods in Face Recognition. Delac K., Grgic M., Bartlett M.: Recent Advances in Face Recognition, InTech, Publisher, p: 79-96
- [14] Rziza M., El Aroussi M., et al, 2009. Local Curvelet Based Classification Using Linear Discriminant Analysis for Face Recognition, International Journal of Computer Science, Vol. 4, No. 1, pp. 72
- [15] Xie J., 2009. Face Recognition Based on Curvelet Transform and LS-SVM, Proceedings of the International Symposium on Information Processing (ISIP'09), Huangshan, China, 21-23 August 2009, p: 140-143.
- [16] Aggarwal V., Patterh M., 2009. ECG Compression using Wavelet Packet, Cosine Packet and Wave Atom Transforms, International Journal of Electronic Engineering Research, Vol. 1, No. 3, p: 259–268.
- [17] Candes E., Demanet L., et al., 2006. Fast Discrete Curvelet Transforms, Technical Report, Cal Tech, March.
- [18] Curvelets: A surprisingly effective non-adaptive representation for objects with edges: <http://www.Curvelet.org/papers/Curve99.pdf>
- [19] Demaney L., Ying L., 2007. Wave atoms and sparsely of oscillatory patterns, Appl. Comput. Harm. Anal. , February, 2007.
- [20] Young S., et al., The HTK Tools and Reference Manuals, version 3.4, Cambridge University Engineering Department, 2009.
- [21] Bouchaffra D., Tan J., 2006. Introduction to structural hidden markov models: application to handwritten numeral recognition, Intell. Data Anal. J., Vol. 10, Number 1, 2006.
- [22] Bouchaffra D., Amira A., 2008. Structural hidden Markov models for biometrics: fusion of face and fingerprint, Pattern Recognition, Vol. 41, p: 852– 867.
- [23] Han J., Kamber M., Data mining: concepts and techniques, 2nd edition, Elsevier, 2006.
- [24] Rama A., Tarrés F., Rurainsky J., Eisert P., 2008. 2D-3D Mixed Face Recognition Schemes, Delac K., Grgic M., Bartlett M.: Recent Advances in Face Recognition, In_Tech publisher, p: 125-148.
- [25] Face images database: <http://www.cl.cam.ac.uk/Research/DTG/attarchive>
- [26] Description of the Collection of Facial Images: <http://cswww.essex.ac.uk/mv/allfaces/grimace.zip>
- [27] Yale face database: <http://cvc.yale.edu>
- [28] Mandal T., Wu Q., Yuan Y., 2008. Curvelet based face recognition via dimension reduction, Signal Processing, Vol. 89, No. 12, pp. 2345-2353.