One-sample Face Recognition Using HMM Model of Fiducial Areas

OJO, John Adedapo jaojo@lautech.edu.ng
Department of Electronic & Electrical Engineering,
Ladoke Akintola University of Technology (LAUTECH),
Ogbomoso, P.M.B. 4000, Nigeria.

Adeniran, Solomon A. sadenira@oauife.edu.ng
Department of Electronic & Electrical Engineering,
Obafemi Awolowo University (OAU),
Ile Ife, Nigeria.

Abstract

In most real world applications, multiple image samples of individuals are not easy to collate for recognition or verification. Therefore, there is a need to perform these tasks even if only one training sample per person is available. This paper describes an effective algorithm for recognition and verification with one sample image per class. It uses two dimensional discrete wavelet transform (2D DWT) to extract features from images; and hidden Markov model (HMM) was used for training, recognition and classification. It was tested with a subset of the AT&T database and up to 90% correct classification (Hit) and false acceptance rate (FAR) of 0.02% was achieved.

Keywords: Hidden Markov Model (HMM); Recognition Rate (RR); False Acceptance Rate (FAR); Face Recognition (FR)

1. INTRODUCTION

Face recognition has attracted attention from the research and industrial communities with a view to achieve a “hands-free” computer controlled systems for access control, security and surveillance. Many algorithms have been proposed to solve this problem starting from the geometric feature-based [1], holistic [2,3] to appearance-based approaches [4]. The performance of these algorithms however depends heavily on the largeness of the number of training set with the attendant problem of sample collection. Algorithms that have performed excellently well with multiple sample problem (MSP) may completely fail to work if only one training sample is used [5]. However, one sample problems are more real in everyday life than the MSP. National ID cards, smart cards, student ID cards and international passports should contain enough biometric information of individuals for recognition purposes. These cases fall under the one training sample per class problem or simply one sample problem (OSP). Many algorithms have been developed and comprehensive surveys are available [5,6].

In one sample problem, the idea is to get as much information as possible from the sample. One approach to this is to increase the size of the training set by projecting the image into more than one dimension space [7], using noise model to synthesise new faces [8] or generating virtual samples or geometric views of the sample image [9]. But the one sample problem has been changed to the multiple sample problem in these cases with increase in computational and storage costs. In addition, virtual samples generated may be highly correlated and can not be considered as independent training samples [10].

In appearance-based approaches, certain features of the image samples are extracted and presented to a classifier or classifying system, which uses a similarity measure (probabilistic measure, majority voting or linearly weighted summing) to ascertain the identity of the image [11,12]. The accuracy of the performance of these methods depends largely on the features extracted from the image [5]. Gray-value features are credited with the ability to retain texture information, while Gabor and other derived features are more robust against illumination and geometrical changes [13,14]. Since there are many combining classifiers with established high level of accuracy, good performance is expected with combination of appropriate feature selection technique.

In this paper, we present a one sample face recognition and verification system, which uses two dimensional discrete Wavelets transform (2D DWT) for feature extraction and one dimensional discrete hidden Markov models (1D DHMM) for classification.

2. PRELIMINARIES

- **Hidden Markov model (HMM)**

A signal that obeys the Markov process,

$$P(q_1, q_2, \dots q_n) = \prod_{i=1}^{T} P(q_i/q_{i-1}), \tag{1}$$

can be represented by a HMM, which consists of two interrelated processes; the observable symbol sequence and the underlying, unobservable Markov chain. A brief description of HMM is presented below, while the reader is referred to an extensive description in [15]. HMM is characterized by some elements; a specific number N of states {S}, while transition from one state S_i to another state S_j emits observation vectors O_t and the observation sequence is denoted as $O = O_1, O_2, \dots O_T$. Observable symbols in each state can take any value in the vector $V = \{v_1, v_2, \dots v_M\}$, where M is the number of the observable symbols. The probabilities of transition from a state i to j is expressed as,

$$a_{ij} = P\lfloor q_{t+1} = S_j/q_t = S_i\rfloor, \quad 1 \le i,\ j \le N, \tag{2}$$

$0 \le a_{ij} \le 1$ and $\sum_{j=1}^{N} a_{ij} = 1, \quad 1 \le i \le N,$

and $A = \{a_{ij}\}$

The likelihood of emitting a certain observation vector O at any state S_j is b_j, while the probability distribution $B = \{b_j(k)\}$ is expressed as,

$$b_j(k) = P\lfloor v_k at\, t \,/\, q_t = S_j\rfloor, \quad 1 \le j \le N,\ 1 \le\ K \le M \tag{3}$$

The initial state (prior) distribution $\pi = \pi_i$, where

$$\pi_i = P[q_i = S_i],\ 1 \le i \le N, \tag{4}$$

are the probabilities of S_i being the first state of sequence. Therefore a short notation for representing a model is,

$$\lambda = (A, B, \pi) \tag{5}$$

Given a model λ, and observation sequence O, the probability of the sequence given the model is $P(O \,/\, \lambda)$. This is calculated using the froward-backward algorithm,

$$P(O/\lambda) = \sum_{all\,Q} P(O/Q\,, \lambda)P(Q\,/\lambda), \tag{6}$$

$$P(O \,/\, \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i), \tag{7}$$

where $\alpha_t(i)$ is the forward variable and it is the probability of the partial observation sequence, $O = O_1, O_2, \ldots\ O_T$ and state S_j at time t, given the model λ;

$$\alpha_t(i) = P(O_1, O_2, \ldots O_t, q_t = S_i \,/\, \lambda), \tag{8}$$

$\beta_t(i)$ is the backward variable and it is is the probability of the partial observation sequence from $t+1$ to the end, given state S_i at time t and λ.

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \ldots O_T / q_t = S_i, \lambda). \tag{9}$$

The problem to solve for recognition purpose is to find the best state sequence Q that gives the maximum likelihood with respect to a given model. The viterbi algorithm is used for solving this problem. It finds the most probable path for each intermediate and finally for the terminating state. The algorithm uses two variables $\delta_t(i)$ and $\psi_t(i)$.

$$\delta_t(i) = max_{q_1, q_2, \ldots q_{t-1}}[P(q_1, q_2, \ldots\ q_t = i, O_1, O_2, \ldots\ O_t \,/\, \lambda)], \tag{10}$$

where $\delta_t(i)$ is the best score or highest probability along a single path, at time t, which accounts for the first t observations and ends in state S_j.

$$\psi_t(i) = argmax_{q_1, q_2, \ldots q_{t-1}} P(q_1, q_2, \ldots\ q_t = S_i, O_1, O_2, \ldots\ O_t \,/\, \lambda) \tag{11}$$

$\psi_t(i)$helps to keep tract of the “best path” ending in state S_i at time t.

- **Wavelets**

Wavelet transform uses multi resolution techniques to provide a time-frequency representation of the signal. It can be described as breaking up of a signal into shifted and scaled versions of the “mother” wavelet. Wavelet analysis is done by convolving the signal with wavelet kernels to obtain wavelet coefficients representing the contributions of wavelets in the signal at different scales and orientations [16,17].

Discrete wave transform (DWT) was developed to reduce the computation time and for easy implementation of the wavelet transform. It produces a time-scaled representation of the signal by using digital filtering techniques, the wavelet families. Unlike discrete Fourier transform that can be represented by a convolution equation, DWT comprises transformation kernels or equations that differ in its expansion functions, the nature of the functions (orthogonal or bi-orthogonal basis) and how many resolutions of the functions that are computed. A signal, which passes through the filter bank shown in Figure 2 is decomposed into four lower resolution components: the approximation $(cD_{j+1}^{(a)})$, horizontal $(cD_{j+1}^{(h)})$, vertical $(cD_{j+1}^{(v)})$ and diagonal coefficients $(cD_{j+1}^{(d)})$.

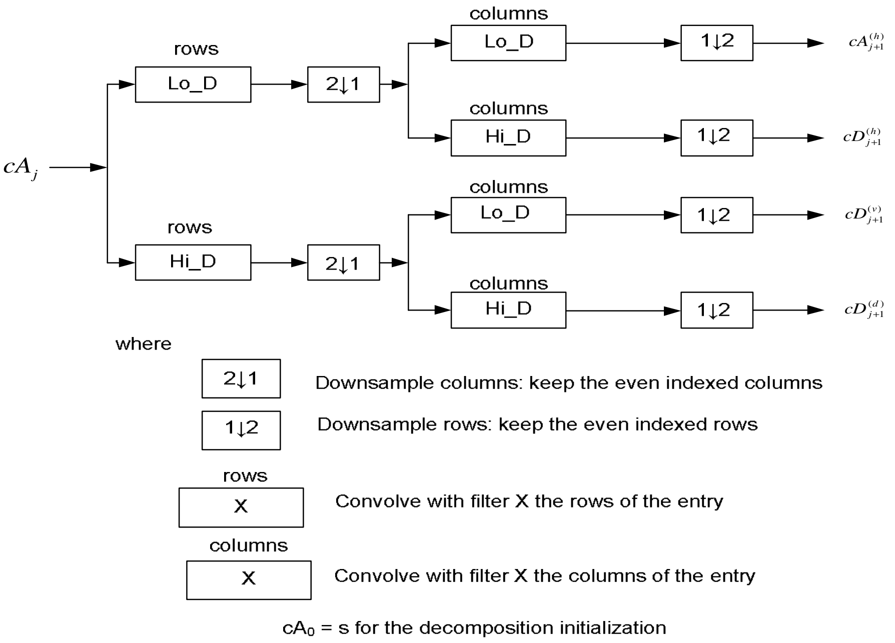
rows
Lo_D
2↓1
columns
Lo_D
1↓2
$cA_{j+1}^{(h)}$
columns
Hi_D
1↓2
$cD_{j+1}^{(h)}$
cA_j
columns
Lo_D
1↓2
$cD_{j+1}^{(v)}$
rows
Hi_D
2↓1
columns
Hi_D
1↓2
$cD_{j+1}^{(d)}$
where
2↓1
Downsample columns: keep the even indexed columns
1↓2
Downsample rows: keep the even indexed rows
rows
X
Convolve with filter X the rows of the entry
columns
X
Convolve with filter X the columns of the entry
cA_0 = s for the decomposition initialization

FIGURE 1: One-dimensional discrete top-to-bottom HMM (Source: MATLAB R2007a Help File)

3. METHOD

- **Modeling a Face and Feature Extraction**

A One dimensional (1D) discrete top-to-bottom (Bakis) HMM was used to segment each face into states as shown in Figure 2. The algorithm for feature extraction shown in Figure 3 was used to generate the observation vector. Two dimensional Discrete Wavelet Transform (2D DWT) was used to decompose the image into its approximation coefficients, horizontal details, vertical details and the diagonal details. ‘db1’, one of the Daubechies family of wavelets was used for decomposition. The approximation coefficient was coded using 256 gray levels thereby producing a coded (and reduced) form of the original or input image. The “coded” image was divided into sub-images and the overlap between successive sub-images was allowed to be up to 5 pixels less than the total height of the sub-image.

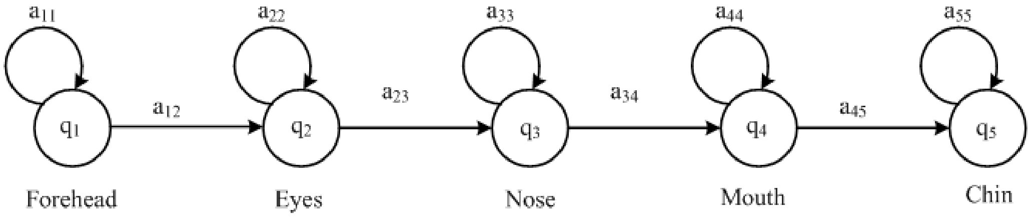

FIGURE 2: One-dimensional discrete top-to-bottom HMM

To generate the observation vector from each sub-image, the two dimensional sub-images were converted into a vector by extracting the coefficients column-wise. The number of features (NF) selected was varied to see its effect on the recognition ability of the system. The coefficients of the sub-images were stacked to form a vector, therefore a face image was represented by a vector ($Q \times NF$) in length, where Q is the number of states. Figure 4(a) shows the original image from the AT&T database while Figure 4(b) shows the gray-scale of the approximation coefficient of the same image with the sampling strip for segmenting the image into allowable states.

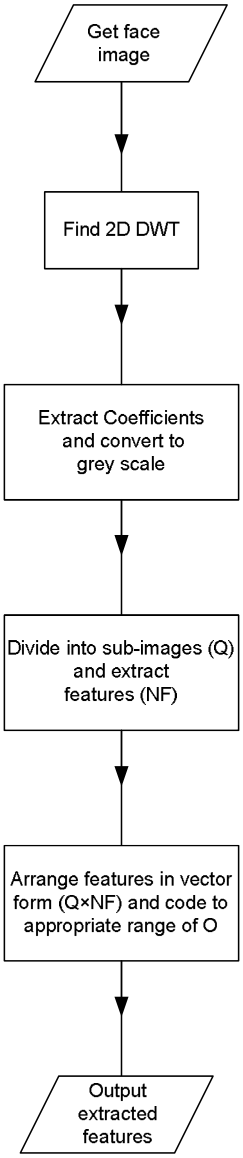
Get face image
Find 2D DWT
Extract Coefficients and convert to grey scale
Divide into sub-images (Q) and extract features (NF)
Arrange features in vector form (Q×NF) and code to appropriate range of O
Output extracted features

FIGURE 3: Algorithm for feature extraction

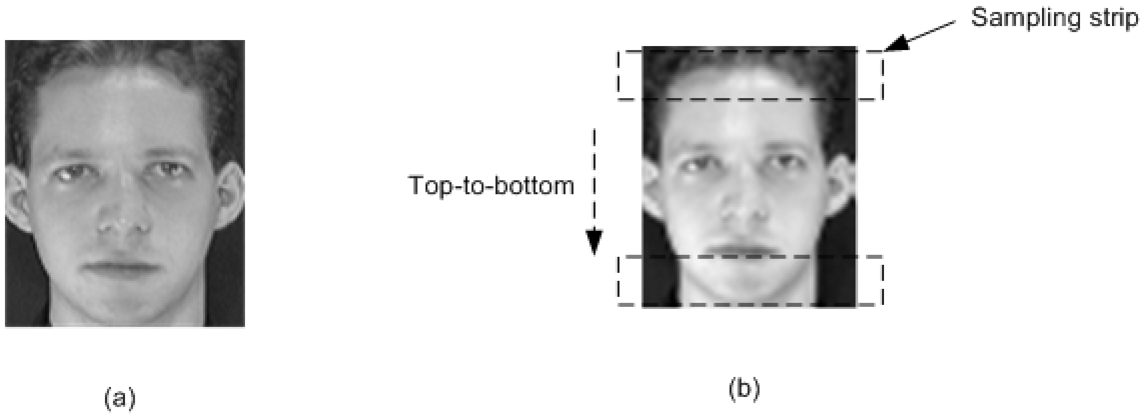

FIGURE 4: Grey-scale version of the approximation coefficients of an image and its segmentation into states

- **Training**

Features extracted from faces of individuals were used to train a model for each face using the algorithm shown in Figure 5. The initial parameter were generated randomly and improved using Baum-Welch re-estimation procedure [15] to get the parameters that optimised the likelihood of the training set observation vectors for the each face.
State transition probability (A) is defined as,

$$a_{ij} = 0, \qquad j < i \qquad (12)$$

$$a_{ij} = 0, \qquad j > i + \Delta \qquad (13)$$

where $\Delta = 1$ i.e. the model is not allowed to jump more than a state at a time. Since each face was divided into five sub-images, the resulting matrix is

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & a_{55} \end{bmatrix} \qquad (14)$$

$a_{NN} = p$, while $a_{Ni} = 0 \;\; for\ i < N\ and\ i > N + 1$ and the initial state probability (π) is defined as

$$\pi_i = \begin{cases} 0, & j \neq 1 \\ 1, & j = 1 \end{cases} \qquad (15)$$

$$\pi_i = [1,0,0,0,0,0]. \qquad (16)$$

The maximum number of iteration for the re-estimation is set to 5 or if the error between the initial and present value is less than 10^{-4}, then the model is taken to have converged and the model parameters are stored with appropriate class name or number (A_c, B_c, π_c).

- **Algorithm for Model Re-estimation**

```
(n is the maximum number of iteration allowed)
k = 1
initialise $\lambda = (A, B, \pi)$
compute $P(O/\lambda^k)$
while $k < n$ do
        estimate $P(O/\lambda^{k+1})$
        if $|P(O/\lambda^{k+1}) - P(O/\lambda^k)| < error$
            quit
        else
            $P(O/\lambda^k) \Leftarrow P(O/\lambda^{k+1})$
        end
end
```

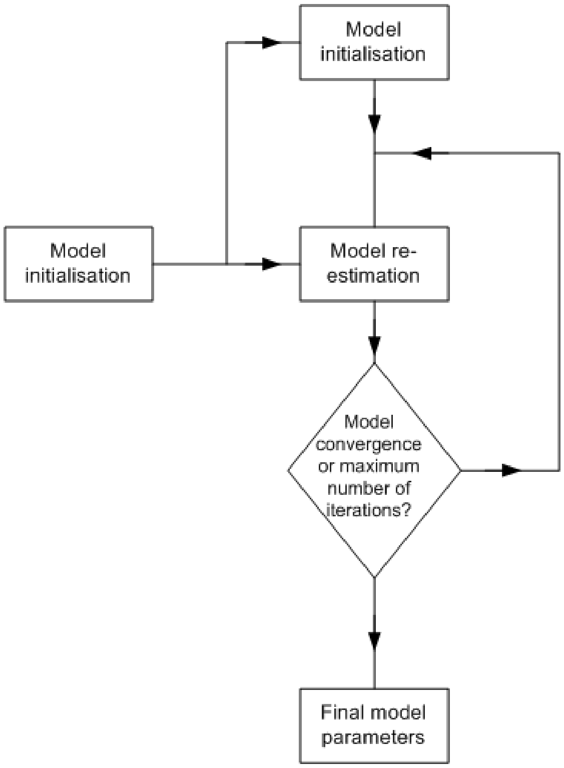

FIGURE 5: Algorithm for HMM training

- **Recognition and Verification**

Given a face to be tested or recognised, feature (observation vector) extraction is first performed as described in section 3.1. Model likelihoods (log-likelihood) for all the models in the training set (given the observation vectors) is calculated and the model with the highest log-likelihood is identified to be the model representing the face. Euclidean measure is used to test if a face is in the training set or database. If the log-likelihood is within the stated distance, the model (face) is recognised to be in the training set or in the database. However, in areas of applications such as access control, it is desired to know the exact identity of an individual, therefore the need to verify the correctness of the faces recognised.

For classification or verification, the Viterbi recogniser was used as shown in Figure 6. The test (face) image was converted to an observation sequence and then model likelihoods $P(O_{test} / \lambda_i)$

are computed for each $\lambda_i, i = 1,2, \dots c$. The model with highest likelihood reveals the identity of the unknown face.

$$v = argmax_{1 \le i \le c}[P(O_{test} \,/\, \lambda_i)] \quad (12)$$

4. RESULTS AND DISCUSSION

The algorithm was implemented in Matlab 7.4a on a HP AMD Turion 64 X2 TL-58, 1.90GHz, 2GB RAM on a Windows operating system. It was tested with a subset of the AT&T (formerly ORL) database [18]. A face image per person was used for training while five other images per person were used for testing, some of which are shown in Figure 7. The recognized images were verified for correctness, 80% correct classification (Hit) occurred while 20% were misclassified. The rest of the images that were not in the training set were used to test the false acceptance rate (FAR) i.e. the ratio of the numbers of images falsely accepted to the total number of images tested and 0.02 FAR occurred. The number of test images per class was reduced to two and 90% Hit, 0.025 FAR occurred as shown in Table 1. Furthermore, the algorithm was tested with ten subjects in the AT&T database. The general observation was that the percentage Hit and FAR were independent of number of subjects in class. For instance, 90% Hit, 0.05 FAR and 90% Hit, 0.05 FAR occurred when five and two test images per class were used respectively.

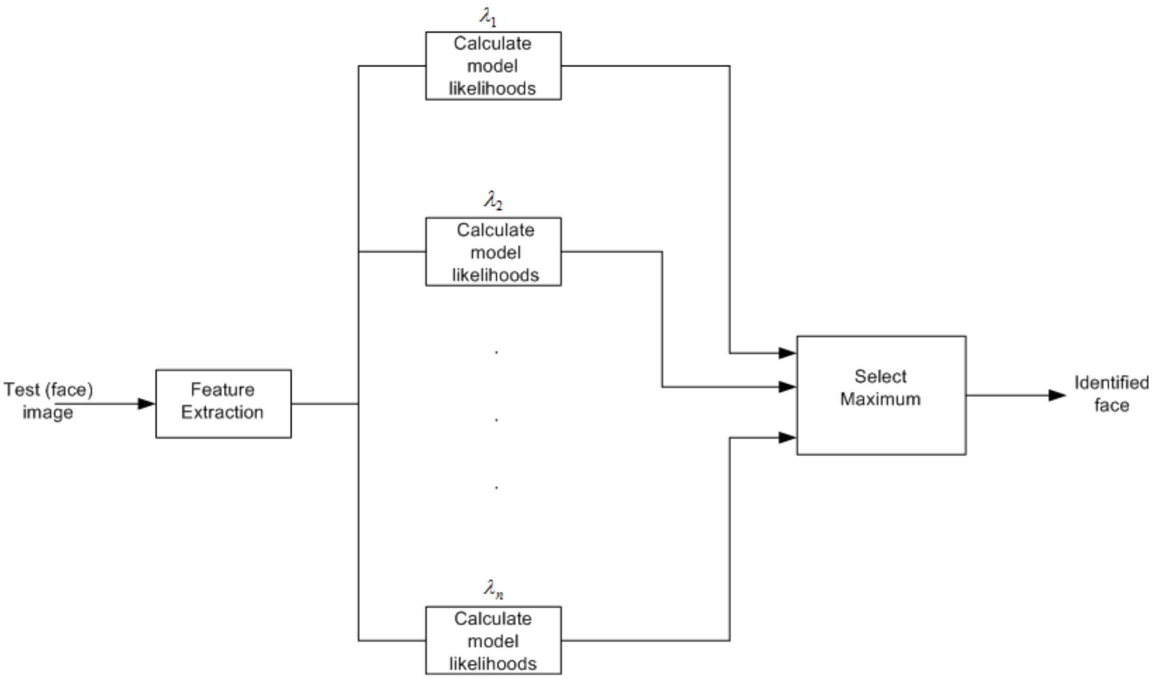

FIGURE 6: Viterbi recogniser

Figure 8 shows the effect that the number of features selected per each state (subimage) has on the number of correct classifications (Hit). The results show that 30 features per subimage were sufficient to give the best performance. In addition, the average time for testing a face was approximately 0.15s, which is near real-time. Going by these results, the algorithm is expected to be adequate for implementation in applications where small size database is required [19].

The performance of the algorithm when Two Dimensional Discrete Cosine Transform (2D-DCT) was used for feature extraction is compared with that of the Discrete Wavelet Transform (2D-DWT) and the results are shown in Table 2. The results show that there is a significant improvement in the recognition or classification accuracy when DWT was used for feature extraction.

Test Images	Number of class	Hit	Miss	FAR
5	20	80%	20%	0.02
2	20	90%	10%	0.025
5	10	90%	10%	0.04
2	10	90%	10%	0.05

TABLE 1: Classification accuracies achieved for a subset of AT&T database

Test Images	Number of class	Hit	
		DCT	DWT
5	20	39%	80%
2	20	50%	90%
5	10	46%	90%
2	10	45%	90%

TABLE 2: Classification accuracies for DCT and DWT

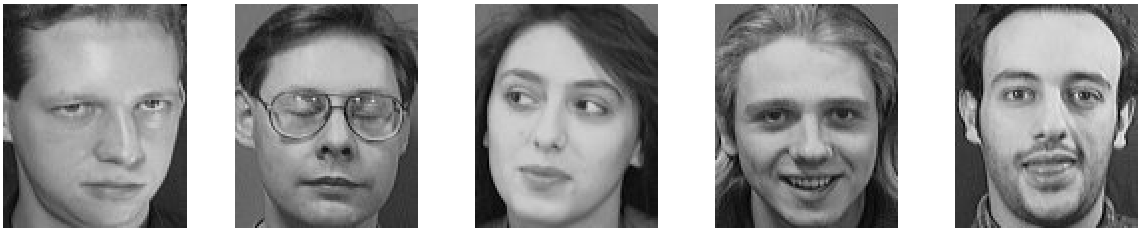

FIGURE 7: Some of the faces used for testing.

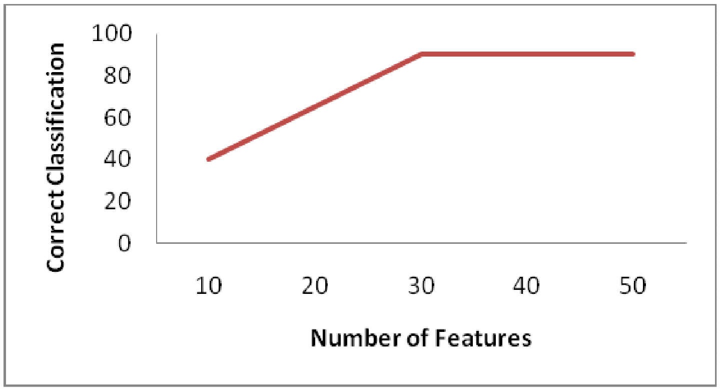

FIGURE 8: Graph of correct classifications against number of features per state

5. CONSLUSION

The paper presented a one sample face recognition system. Feature extraction was performed using 2D DWT and 1D top-to-bottom HMM was used for classification. When tested with a subset of the AT&T database, up to 90% correct classification (Hit) and as low as 0.02 FAR were achieved. The high recognition rate and the low FAR achieved shows that the new algorithm is suitable for face recognition problems with small-size database such as access control for personal computers (PCs) and personal digital assistants (PDAs).

6. REFERENCES

1. R. Brunelli and T. Poggio. *"Face recognition: Features versus templates".* IEEE Transaction on Pattern Analysis and Machine Intelligence, 15(10):1042-1062, 1993

2. L. Sirovich and M. Kirby. *"Low-Dimensional procedure for the characterization of human face".* Journal of the Optical Society of America A*,* 4(3):519–524, 1987

3. M. Turk and A. Pentland. *"Eigenfaces for Recognition".* Journal of Cognitive Neuroscience, 3(1):71-86, 1991

4. S. Lawrence, C.L. Giles, A. Tsoi and A. Back. *"Face recognition: A convolutional neural-network approach".* IEEE Transaction on Neural Networks, 8(1):98-113, 1997

5. W. Zhao, R. Chellappa, P.J. Philips and A. Rosenfeld. *"Face recognition: A literature survey".* ACM Computing Surveys 35(4):399-458, 2003

6. X. Tan, S. Chen, Z-H Zhou, and F. Zhang. *"Face recognition from a single image per person: a survey".* Pattern Recognition 39:1725-1745, 2006

7. J. Wu and Z.-H Zhou. *"Face recognition with one training image per person".* Pattern Recognition Letters, 23(2):1711-1719, 2001

8. H.C. Jung, B.W. Hwang and S.W. Lee. *"Authenticating corrupted face image based on noise model".* Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition, 272, 2004

9. F. Frade, De la Torre, R. Gross, S. Baker, and V. Kumar. *"Representational oriented component analysis (ROCA) for face recognition with one sample image per training class".* Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2:266-273, June 2005

10. A.M. Martinez. *“Recognizing imprecisely localised, partially occluded, and expression variant faces from a single sample per class”.* IEEE Transaction on Pattern Analysis and Machine Intelligence 25(6):748-763, 2002

11. B.S. Manjunath, R. Chellappa and C.V.D. Malsburg. *“A feature based approach to face recognition”.* In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 1:373-378, 1992

12. M. Lades, J.Vorbruggen, J. Buhmann, J. Lange, von der Malsburg and R. Wurtz. *“Distortion invariant object recognition in the dynamic link architecture”.* IEEE Transaction on Computers 42(3):300-311, 1993

13. X. Tan, S.C. Chen, Z-H Zhou, and F. Zhang. *“Recognising partially occluded, expression variant faces from single training image per person with SOM and soft kNN ensemble”.* IEEE Transactions on Neural Networks, 16(4):875-886, 2005

14. H.-S. Le and H. Li. *“Recognising frontal face images using Hidden Markov models with one training image per person”.* Proceedings of the 17th International Conference on Pattern Recognition (ICPR04), 1:318-321, 2004

15. L.R. Rabiner. *“A tutorial on Hidden Markov models and selected application in speech recognition”.* Proceedings of the IEEE, 77(2):257-286, 1989

16. I. Daubechies, *“Orthonormal bases of compactly supported wavelets”.* Communication on Pure & Applied Mathematics XLI, 41:909-996, 1988

17. I. Daubechies. *“Ten lectures on wavelets”.* CBMS-NSF Conference series in Applied Mathematics, No-61, SIAM, Philadelphia Pennsylvania, 1992

18. F. Samaria and A. Harter. *“Parameterization of a stochastic model for human face identification”.* 2nd IEEE Workshop on Applications of Computer Vision, Saratosa FL. pp.138-142, December, 1994

19. J. Roure, and M. Faundez-Zanuy. *“Face recognition with small and large size databases”.* Proceedings of 39th Annual International Carnahan Conference on Security Technology, pp 153-156, October 2005