A Review of Maximum Confidence Hidden Markov Models in Face Recognition

Swati Raut and S. H. Patil

Abstract—**The work presented in this paper focuses on the use of Hidden markov models for face recognition. New discriminative training creation to assure model compactness and discriminability.hidden markov model(HMM) is statistical model in which the system being modeled is assumed to be markov processes with unoabserd state. Hmm can be considered as a simplest dynamics Bayesian network. In Hidden Marko model, the state is not directly visible but output dependent on state is visible. Accordingly w develop the maximum confidence hidden markov modeling (MC-HMM) for face recognition. In MC-HMM we merge transformation matrix to extract discriminative facial features. MC-HMM achieves higher recognition with lower feature dimensions.**

Index Terms—**hidden Markov model, confidence measure, discriminative feature extraction, discriminative training, classification, face recognition.**

I. Introduction

Most popular training method for hidden markov model (HMM) based speech recognition system. Principal component analysis (PCA) and linear recognition. Discriminate analysis (LDA) are popular for face recognition the recognition accuracies using these methods highly depend on training templates.HMM has been used for many pattern recognition applications including speech recognition, character recognition and face recognition. A HMM can be considered as the simplest dynamic Bayesian network. In a regular model, the state is directly visible to observer and therefore the state transition probabilities are the only parameters. In hidden markov model state is not directly visible but output depend on the state is visible. Each state has probability distribution over the possible output token, therefore the sequence of token generated by HMM gives some information about sequence of state, while face recognition we consider the pixel values for modeling HMM states. When image get capture some noise will get added in that image. In HMM raw pixel values used for modeling HMM states. Facial features get extracted while building HMMs. Discrete cosine transform acted as feature extraction for HMM modeling. In 1D HMM for face recognition the object image could be blocked in a sequence of sub images and characterized by one dimensional (1D) state sequence for recognition. In 2D HMM was motivated to model vertical horizontal features using matrix of HMM states, each state was allowed to flexibility skip neighboring states in vertical horizontal or diagonal directions the computation complexity grew rapidly when number of states increased. In 2D HMMs for object recognition two issue get consider, the first issue involves a hybrid process of feature extraction and model estimation. A new discriminative training creation derived from hypothesis test theory. The maximum confidence objective function is carried out for discriminative training algorithms e.g. minimum verification error(MVE). A new discriminative training creation derived by a statistical hypothesis test theory. Minimum classification error(MCE) and minimum mutual error(MCE) and minimum mutual information support to build large margin HMM. A new viterbi decoding is implemented to align features in to the corresponding HMM states[1].

The HMM can be applied to image processing. In consideration of the fact that the image can be seen as a two dimension matrix of data, according to Samaria, space sequences must be considered .The idea is again to exploit the vertical sequential structure of a human face. A sequence of overlapping horizontal stripes are built on the image and the sequence of these stripes is labeled by means of a 1DHMM. For a HMM-based face recognition system having a consistently sized face region is particularly important because the HMM requires regional analysis of the face with a scanning window of fixed size. A straightforward approach is to resize all determined face regions to a common size. To facilitate more efficient computation we seek the smallest sized face region possible without impacting the overall system recognition rate. This feature extraction technique is based on scanning the image with a fixed-size window from left-to-right and top-to-bottom.

The most popular training method for hidden Markov model (HMM)-based speech recognition systems is maximum likelihood (ML) estimation. Discriminative training methods such as maximum mutual information (MMI) corrective training and minimum classification error (MCE) attempt to minimize the error rate more effectively by utilizing both the correct and the other categories, and incorporating that into the training phase. Note that the MMI and MCE criteria were shown to be closely related . if the true distribution of the samples to be classified can be accurately described by the assumed statistical model, and the size of the training set tends to infinity, then ML estimation outperforms MMI estimation in the sense that it yields less variance in the parameter estimates. MMI estimation aims to maximize the posterior probability of the words in the training set given their corresponding utterances.

Unlike in the ML case, there is no simple optimization method to this problem. First experiments in MMI were reported by Bahl who used the gradient descent algorithm

Manuscript received September 13, 2011; revised February 3, 2012.

Swati Raut is with Bharati Vidyapeeth university, College of engineering, Maharashtra, India (e-mail:getdiya2008@gmail.com).

S. H. Patil is with Department of Computer Engineering, Bharathi Vidyapeeth University, College of engineering , Maharashtra, India (e-mail: shpatil@bvucoep.edu.in).

for the optimization of the objective function. Gradient descent algorithms are sensitive to the size of the update step. A large update step can cause an unstable behavior. However, a small update step might result in a prohibitively slow convergence rate. A proposed a method for maximizing the MMI objective function which is based on a generalization of the Baum–Eagon inequality. This method was proposed to discrete HMMs. Normandin proposed a useful approximated generalization of this method to HMMs with Gaussian output densities known as the extended Baum–Welch (EBW) algorithm [2].

Within the last several years, much progress has been made toward recognizing faces under small variations in lighting, facial expression and pose, reliable techniques for recognition under more extreme variations have proven elusive. a new approach for face recognition one that is insensitive to large variations in lighting and facial expressions. Note that lighting variability includes not only intensity, but also direction and number of light sources. As is evident from , the same person, with the same facial expression, and seen from the same viewpoint, can appear dramatically different when light sources illuminate the face from different directions.

II. Face recognition Exploits Two Observations

1) All of the images of a surface, taken from a fixed viewpoint, but under varying illumination, lie in a 3D linear subspace of the high-dimensional image space.

2) Because of regions of shadowing, specularities, and facial expressions, the above observation does not exactly hold. In practice, certain regions of the face may have variability from image to image that often deviates significantly from the linear subspace, and, consequently, are less reliable for recognition.

In Fisherfaces, a derivative of Fisher's Linear Discriminate (FLD) maximizes the ratio of between-class scatter to that of within-class scatter.

The Eigen face method is also based on linearly projecting the image space to a low dimensional feature space. However, the Eigen face method, which uses principal components analysis (PCA) for dimensionality reduction, yields projection directions that maximize the total scatter across all classes, i.e., across all images of all faces. In choosing the projection which maximizes total scatter, PCA retains unwanted variations due to lighting and facial expression. Fisher's Linear Discriminate is a "classical" technique in pattern recognition , first developed by Robert Fisher in 1936 for taxonomic classification [3]. Depending upon the features being used, it has been applied in different ways in computer vision and even in face recognition. Cheng et al. presented a method that used Fisher's discriminator for face recognition, where features were obtained by a polar quantization of the shape. Baker and Nayar have developed a theory of pattern rejection which is based on a two class linear discriminate. Eigenfaces computationally expensive and require great amounts of storage; it is natural to pursue dimensionality reduction schemes. the Fisher face method had lower error rates than any of the other three methods. Yet, no claim is made about the relative performance of these algorithms on much larger databases[4].

Another important approach is the *Elastic Matching* method, which proved to be invariant to expression changes. The idea is to build a lattice on image faces (rigid matching stage), and to apply, at each point of the lattice, a bank of Gabor filters. In case of variations of expression, this lattice can warp to adapt itself to the face (elastic matching stage).

The wavelet transform is a methodology emerged in the last years, useful in many applications, specially in the field of image compression. Wavelet-based coding provides substantial improvements in picture quality at higher compression ratios, with respect to standard DCT transform. Over the past few years, a variety of wavelet-based schemes for image compression have been developed and implemented. Because of the many advantages, the compression technologies used in the upcoming JPEG-2000 standard are all based on the wavelet technology. Wavelets could be defined as a mathematical tool for hierarchically decomposing functions. The wavelet transform is aimed at describing a function in terms of a coarse overall shape, plus details that range from broad to narrow. More formally, wavelets are functions defined over a finite interval and having an average zero value. The basic idea is to represent any arbitrary function f(y) as a superposition of a set of basis functions. These basis functions, or *baby wavelets*, are obtained from a single prototype wavelet called the *mother wavelet*, by dilations or contractions (scaling) and translations (shifts). The wavelet system calculates the coefficients representing the image with a normalized two-dimensional Haar basis, sorting these coefficients in order of decreasing magnitude. Subsequently, the first M coefficients are retained, performing a lossy image compression. By using wavelet and HMM give more accuracy and perfection[5].

III. Face Recognition System

In this section we outline the basic architecture of face recognition system based on Gonzalez's image analysis system the functional block diagram of face recognition is shown in Fig. 1.

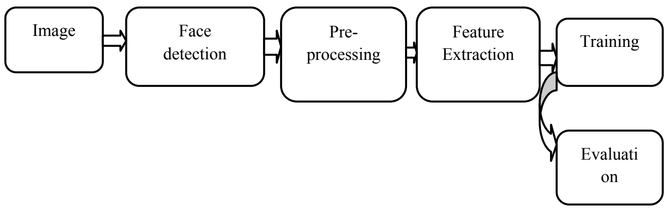

Fig. 1. Face Recognition System

In this paper feature extraction plays most important part. In Hmm we are using discriminative feature extraction algorithm[2].

A. Face Detection and Cropping Block

This is the first stage of any face recognition system and the key difference between a semi-automatic and a fully automatic face recognizer. In order to make the recognition system fully automatic, the detection and extraction of faces from an image should also be automatic. Face detection also represents a very important step before face recognition, because the accuracy of the recognition process is a direct function of the accuracy of the detection process.

B. Pre-Processing Block

The face image can be treated with a series of pre-processing techniques to minimize the effect of factors that can adversely influence the face recognition algorithm. The most critical of these are *facial pose* and *illumination*. A discussion on these factors and their significance w.r.t. HMM techniques

C. Feature Extraction Block

In this step the features used in the recognition phase are computed. These features vary depending on the automatic face recognition system used. For example, the first and most simplistic features used in face recognition were the geometrical relations and distances between important points in a face, and the recognition 'algorithm' matched these distances the most widely used features in face recognition are KL or eigenfaces, and the standard recognition 'algorithm' uses either the Euclidian or Mahalanobis distance to match features

D. Face recognition Block

This consists of 2 separate stages: a *training process*, where the algorithm is fed samples of the subjects to be learned and a distinct model for each subject is determined; and an *evaluation process* where a model of a newly acquired test subject is compared against all existing models in the database and the most closely corresponding model is determined. If these are sufficiently close a recognition event is triggered.

IV. Discriminative Training for 2D Pattern Classification

Discriminative training method such as minimum mutual information(MMI) and minimum classification error(MCE) mainly minimize the error rate more effectively by utilizing both correct and other categories and incorporating that into training phase. In discriminative training for 2D pattern classification MCE algorithm introduced.

A. Minimum Classification Error Training

Minimum classification error(MCE) training aims to estimate model distribution $\{p(X|\Lambda_c),\ c=1,\ldots,c\}$ in a fashion that classification errors of training data X=$\{X_1,\ldots X_T\}$ are minimized.

$$d(X,\Lambda_c) = -g(X,\Lambda_c) + G(X,\overline{\Lambda}_c)$$

$$= -\log p(X \mid \Lambda_c) + \left[\frac{1}{c-1}\sum_{\Lambda_j \neq \Lambda c} \rho \log p\left(X|\Lambda_j\right)\right]^{1/\rho} \quad (1)$$

where *p* was a tuning parameter in antidiscriminant function G(X,Λ_c) determined from c-1 no target models.

$$\Lambda_{MCE} = \arg_{\Lambda} \min E_X\left[l(X,\Lambda)\right] = \arg_{\Lambda} m\text{in}\, E_X\left[\sum_{c=1}^{C} l(X,\Lambda_c)1(X \in \Lambda_c\right] \quad (2)$$

B. Discriminative Feature Extraction

(1) was the indicator function. The solution to this optimization was obtained by a gradient descent algorithm. By using discriminative feature extraction where a transformation matrix W was estimated by maximizing Fisher's ratio criterion.

$$W_{LDA} = \arg_{W} \max \frac{\left|W^T S_b W\right|}{\left|W^T S_w W\right|} \quad (3)$$

where S_b and S_w were between-class and within-class scatter matrices, respectively. LDA extracted the most discriminative features by W^T_{LDA}. MCE criterion was feasible to extract discriminative features through a transformation matrix producing the minimum expected loss.

C. Hidden Markov Model

The Hidden Markov Models are stochastic models which provide a high level of flexibility for modeling the structure of an observation sequence. They allow for recovering the (hidden) structure of a sequence of observations by pairing each observation with a (hidden) state. Hidden Markov Models (HMMs) represent a most famous statistical pattern recognition technique and can be considered as the state-of-the-art in speech recognition. This is due to their excellent time warping capabilities, their effective self organizing learning capabilities and their ability to perform recognition and segmentation in one single step. They are used not only for speech and handwriting recognition but they are involved in modeling and processing images too. This is the case of their use in the face recognition field.

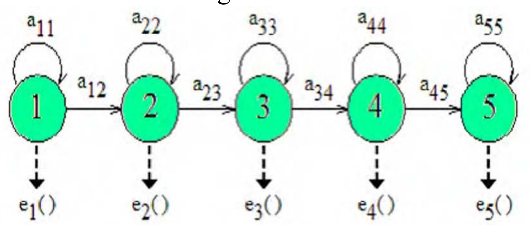

Fig. 2. "*Left to Right*" Hidden Markov Model – 5 state

Hidden Markov Models have been successfully used for speech recognition where data is essentially one dimensional. Extension to a fully connected two dimensional HMM has been shown to be computationally very complex. For frontal face images, the significant facial regions (hair, forehead, eyes, nose, mouth) come in a natural order from top to bottom, even if the images are taken under small rotations in the image plane and/or rotations in the plane perpendicular to the image plane. Each of these facial regions is assigned to a state in a left to right 1D continuous HMM.

D. *Training and Recognition algorithm*

Feature extraction: In the context of face identification problem, we need to make a few image pre-processing; we need transform the face image to a uniform by level and size normalized. Then we need do fractal coding, we change face image in to matrix of fractal code. We looked upon each row of the fractal matrix as a vector. Then take these vectors clustering analysis with K-means algorithm.

Training HMM: Train HMM by use of fact and record each HMM and cluster centers. This training process is conducted using the Baum-Welch algorithm As the detail image norms are real values, a continuous observation HMM is employed. Repeat it until all training-images are trained.

E. *Recognition algorithms*

Feature extraction: A transform the face image to a uniform by level and size normalized. Then transform the face image to fractal matrix, use the matrix of fractal coding into HMM to calculate the probability.

Choose the max matching probability recognition. If all of the matching probability is minor we consider as the face image can't find the matching face in this face database.

The HMM are characterized by two interrelated processes:

1) An unobservable Markov chain with a finite number of states, a state transition probability matrix and an initial state probability distribution.

2) A set of probability density functions for each state.

In this paper facial images X= $\{X_t\}=\{X_{tnm}\}$ of person b are block with spatial indices (m,n) and modeled by structural HMM containing of set of super states S_n along with set of embedded states $q_{mn.}$Here embedded and super states skipping is not allowed[1] .

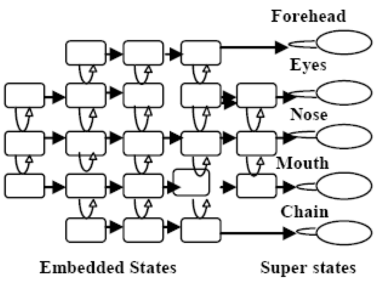

Fig. 3. Facial images representation using HMM

In above fig. five super states considered each super states characterizing forehead, eyes ,nose, mouth, and chain in the vertical direction with initial state probabilities and the state transition probabilities

$\{\Pi_{\upsilon,}\ A_{\upsilon}\}=\{\pi_{sn},a_{snSn+1}\}$. In 1D HMM total, 27 embedded states are considered. Parameters include initial state probabilities, $\Pi_h=\{\ \pi_{qnm}\}$state transition probabilities Ah = {aqnm,qn,m+1} and observation probabilities $B_h=\{b_{qnm}(x)\}$ with mixtures of Gaussian distributions.

$$b_{qnm}(x)=p(x\mid q_{qnm;}B_h)=\sum_{l_{nm}}^{N_l}\omega_{qnm}l_{nm}N(x\mid\mu_{qnm}l_{nm},\sum\nolimits_{q_{nm}l_{nm}}\ ) \tag{4}$$

where $w_{qnm}l_{nm}$, $\mu q_{nm}l_{nm}$ are the mixture weight, mean vector and covariance matrix of embedded state q_{nm} And mixture component l_{nm}[1].

F. *Feature selection methods*

Feature selection algorithm's aim is to select a subset of the extracted features that cause the smallest classification error. The importance of this error is what makes feature selection dependent to the classification method used. The most straightforward approach to this problem would be to examine every possible subset and choose the one that fulfills the criterion function

Feature selection is a NP-hard problem, so researchers make an afford towards a satisfactory algorithm, rather than an optimum one. The idea is to create an algorithm that selects the most satisfying feature subset, minimizing the dimensionality and complexity[7].

G. *Face classification*

Once the features are extracted and selected, the next step is to classify the image. Appearance-based face recognition algorithms use a wide variety of classification methods. Sometimes two or more classifiers are combined to achieve better results. On the other hand, most model-based algorithms match the samples with the model or template. Then, a learning method is can be used to improve the algorithm. One way or another, classifiers have a big impact in face recognition. Classification methods are used in many areas like data mining, finance, signal decoding, voice recognition, natural language processing or medicine. Therefore, there is many bibliography regarding this subject. Here classifiers will be addressed from a general pattern recognition point of view. Classification algorithms usually involve some learning - supervised, unsupervised or semi-supervised. Unsupervised learning is the most difficult approach, as there are no tagged examples. However, many face recognition applications include a tagged set of subjects. Consequently, most face recognition systems implement supervised learning methods. There are also cases where the labeled data set is small. Sometimes, the acquisition of new tagged samples can be infeasible. Therefore, semi-supervised learning is required[7].

H. *Face Recognition: Different Approaches*

Face recognition is an evolving area, changing and improving constantly. Many research areas affect face recognition - computer vision, optics, pattern recognition, neural networks, machine learning, psycology, etcetera. Previous sections explain the different steps of a face recognition process. However, these steps can overlap or change depending on the bibliography consulted. There is not a consensus on that regard. All these factors hinder the development of a unified face recognition algorithm classification scheme. This section explains the most cited criteria.

1) *Geometric/Template Based approaches*

The template based methods compare the input image with a set of templates. The geometry feature-based methods analyze local facial features and their geometric relationships. This approach is sometimes called feature based approach . Examples of this approach are some Elastic Bunch Graph Matching algorithms. This approach is less used nowadays . There are algorithms developed using both approaches. For instance, a 3D morph able model approach can use feature points or texture as well as PCA to build a recognition system .

2) Piecemeal/Wholistic approaches

Faces can often be identified from little information. Some algorithms follow this idea, processing facila features independently. In other words, the relation between the features or the relation of a feature with the whole face is not taken into account. Many early researchers followed this approach, trying to deduce the most relevant features. Some approaches tried to use the eyes , a combination of features , and so on. Some Hidden Markov Model (HMM) methods also fall in this category .Although feature processing is very important in face recognition, relation between features (configure processing) is also important. In fact, facial features are processed holistically

3) Appearance-based/Model-based approaches

Facial recognition methods can be divided into appearance-based or model based algorithms. The differential element of these methods is the presentation of the face. Appearance-based methods represent a face in terms of several raw intensity images. An image is considered as a high-dimensional vector. Then statistical techniques are usually used to derive a feature space from the image distribution. The sample image is compared to the training set. On the other hand, the model-based approach tries to model a human face. The new sample is fitted to the model, and the parameters of the fitted model used to recognize the image. Appearance methods can be classified as linear or non-linear, while model-based methods can be 2D or 3D. Linear appearance-based methods perform a linear dimension reduction.

The face vectors are projected to the basis vectors, the projection coefficients are used as the feature representation of each face image. Examples of thisapproach are PCA, LDA or ICA. Non-linear appearance methods are more complicate. In fact, linear subspace analysis is an approximation of a nonlinear manifold. KernelPCA (KPCA) is a method widely used . Model-based approaches can be 2-Dimensional or 3-Dimensional. These algorithms try to build a model of a human face. These models are often morphable. A morph able model allows to classify faces even when pose changes are present. 3D models are more complicate, as they try to capture the three dimensional nature of human faces. Examples of this approach are Elastic Bunch Graph Matching or 3D Morphable Models.

4) Template/statistical/neural network approaches

A similar separation of pattern recognition algorithms into four groups is proposed by Jain and colleges in We can group face recognition methods into three main groups. The following approaches are proposed[8].

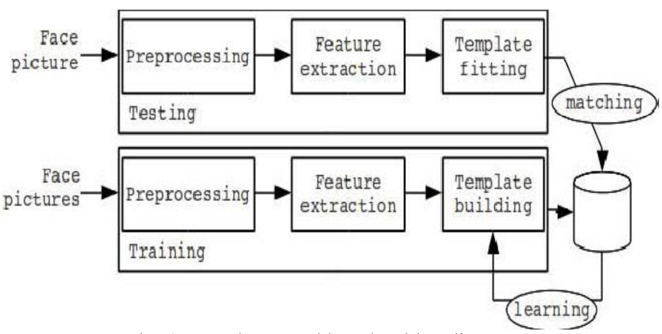

Fig. 4. Template-matching algorithm diagram

a) Template matching. Patterns are represented by samples, models, pix els, curves, textures. The recognition function is usually a correlation
or distance measure. .

b) Statistical approach. Patterns are represented as features. The recognition function is a discriminate function. .

c) Neural networks. The representation may vary. There is a network function in some point.

V. Maximum Confidence Hidden Markov Model

A. Hypothesis Test

In Hypothesis test theory mainly three steps considered [8].

1) Define null hypothesis H, and choose a significance level.

2) Calculate like hood function $p(X| H_o)$ and $p(X| H_1)$And determine creation region.

3) Make a decision of acceptance or rejection of H_0
we simultaneously estimate feature transformation matrix and HMM parameters for facial image recognition. When building new objective function, we setup null and alternative hypotheses as

H_0 : Observation X is from target HMM Gaussians.
H_1:Observation X is not from target HMM Gaussians.

$$LR = \frac{p(x \mid H_0)}{p(x \mid H_1)} \geq \kappa \tag{5}$$

K is decision threshold determined by a significance level for distribution of LR. In this paper unsupervised hierarchical adaption used for facial modeling through optimizing the accumulated models. Here we maximizing confidence of fitting X closer to target parameters Λ and farther from competing parameters $\bar{\Lambda}$ .Image is segmented in to forehead, eyes, nose, mouth, and chain. A reliable state alignment is desirable to estimate compact model parameters.

B. HMM modeling for feature extraction

We perform feature extraction $W^Td\ X_{tnm}$ prior to the estimation of d-dimensional HMM parameters Ω, $d \leq D$. Similar to PCA and LDA, transformation matrix Wd corresponds to the first d columns of matrix W = $[W_d\ W_{D-d}]$. We focus on estimating $\Lambda = \{W, \Pi_\upsilon, A_\upsilon, \Pi_h, A_h, B_h=\{\omega_{qnm}, l_{nm}, \mu_{qnm}l_{nm}, \sum q_{nm}l_{nm}\}\}$ from a sequence of frames $X=\{x_{tnm}\}$.The EM algorithm used for MC parameter estimation.

Parameter Estimation

1. $X = \{X_t\}=\{X_{tnm}\}$:D-dimensional observation vector of frames $t = 1,\cdots,T$ in spatial indices n=1,$\cdots$, N and m =1,$\cdots$, M

2. $\Lambda=\{W, \Omega=\{ \Pi_\upsilon, A_\upsilon\ \Pi_h, A_h, B_h\}\}$: model parameter of a class c,$1\leq c \leq C$

3. $W = [W_d\ W_{D-d}]$: feature transformation matrix with reduced dimension d

4. $\{\Pi_\upsilon, A_\upsilon\}=\{\pi_{sn}, a_{snSn+1}\}$: initial state probabilities and the

state transition probabilities of super (vertical)state (s_n, s_{n+1})=1,⋯⋯, N_s

5. {Π_h, A_h, B_h}={Π_{qnm}, $a_{\omega qnm}$, $\omega_{qnm,lnm}$, μ_{qnmlnm} , $\sum_{qnmlnm}$}}:initial state probabilities, state transition probabilities and observation probabilities of embedded (horizontal) states ($q_{nm}q_{n,m+1}$) =1,⋯⋯, N_q and mixture component l_{nm}=1,⋯, N_l observation probability $b_{qn}(x)$ consist of parameter of mixture weight ω_{qnm}, l_{nm}, Gaussian mean vector.

6. $\bar{\wedge}$, q_{mn}, l_{nm}: competing model, embedded state and mixture component

7. {s,q,l} ={s_n ,q_{nm}, l_{nm}} :sequence of super states, embedded states and mixture components

8. {s_n ,q_{nm}, l_{nm}} optimal sequence obtain by viterbi algorithm

9.Q(Λ'|Λ):auxiliary function of new estimate Λ'Given current estimate Λ

10. $y_t(q_{nm}, l_{nm})$: posterior probability of x_{nm} staying in state q_{nm} and mixture component l_{nm} given the current parameter Λ and X

11. $\delta_v(s_n)$, $\delta_h(q_{nm})$:The beast confidence scores of super and embedded states at vertical and horizontal indices(n, m).

12. $\Psi_v(s_n)$, $\Psi_v(q_{nm})$:backtracking variables recording the best emitting super and embedded states.

13. α, β: discriminative factor with $\alpha = \beta(N_g-1)$ where N_g total number of Gaussians

$$Q(\Lambda'|\Lambda)=E_{s,q,l}[LLR(X,s,q,l|\Lambda')|X,\Lambda]=\sum_{s,q,l}P(s,q,l\mid X,\Lambda)X\sum_{t=1}^{T}\left[\log\pi'_{s1}\sum_{n=1}^{N}\left[\log a'_{s_ns_{n+1}}+log\pi'_{qn1}+\sum_{m=1}^{M}\left[loga'_{qnmq_{n,m+1}}+\log N\left(W'^{T}_{d}x_{tnm}\middle|\mu'_{qnml_{nm}}\right)-\frac{\alpha}{N_{g-1}}\sum_{(\bar{q}_{nm,}\bar{l}_{nm})\neq(q_{nm},l_{nm})}\log N\left(W'^{T}_{d}x_{tnm}\middle|\mu'_{\bar{q}_{nm}\bar{l}_{nm}},\Sigma'_{\bar{q}_{nm}\bar{l}_{nm}}\right)\right]\right]\right] \quad (6)$$

Here Λ and Λ' are auxiliary function where Λ new estimate, and Λ' current estimate respectively . N_g total number of Gaussians. And Λ = {W, B_h={ω_{qnm}, l_{nm}, $\mu_{qnm}l_{nm}$, $\sum q_{nm}l_{nm}$}} considering the Gaussian parameters with reduced rank.

$$\sum_{t=1,n=1,m=1}^{T.N.M.}\sum_{q_{nm}=1,l_{nm}=1}^{N_qN_1}\gamma_t(q_{nm},l_{nm}).\left\{\begin{bmatrix}\log\omega'_{q_{nm}1_{nm}}+\log|W'|\\+\frac{1}{2}\log\left|\sum_{q_{nm}l_{nm}}^{l-1}\right|\\-\frac{1}{2}\left(W'^{T}_{d}X_{tnm}-\mu'_{q_{nm}l_{nm}}\right)^{T}\\\sum_{q_{nm}l_{nm}}^{l-1}(W'^{T}_{d}X_{tnm}-\mu'_{q_{nm}l_{nm}})\end{bmatrix}\right.$$

$$-\beta\sum_{(\bar{q}_{nm},\bar{l}_{nm})\neq(q_{nm},l_{nm})}\left.\begin{bmatrix}\log\omega'_{\bar{q}_{nm}\bar{l}_{nm}}+\log|W'|\\+\frac{1}{2}\log\left|\sum_{\bar{q}_{nm}\bar{l}_{nm}}^{l-1}\right|\\-\frac{1}{2}(W_d^TX_{tnm}-\mu'_{\bar{q}_{nm}\bar{l}_{nm}})^T\\\sum_{q_{nm}l_{nm}}^{l-1}(W'^{T}_{d}X_{tnm}-\mu'_{\bar{q}_{nm}\bar{l}_{nm}})\end{bmatrix}\right\} \quad (7)$$

Where $\gamma t(qnm, lnm) = p(qnm, lnm|Xt, \Lambda)$is the posterior probability of x_{tnm} staying in (q_{nm}, l_{nm}),given current parameters Λ and X.

$$\mu'_{q_{nm}l_{nm}}=W_d^T.\frac{\sum_{t=1,n=1,m=1}^{T.N.M.}\gamma_t(q_{nm,}l_{nm})X_{tnm}-\beta\sum_{t=1,n=1,m=1}^{T.N.M.}\sum_{(\bar{q}_{nm},\bar{l}_{nm})\neq(q_{nm}l_{nm})}\lambda_t(\bar{q}_{nm,}\bar{l}_{nm})X_{tnm}}{\sum_{t=1,n=1,m=1}^{T.N.M.}\gamma_t(q_{nm,}l_{nm})-\beta\sum_{t=1,n=1,m=1}^{T.N.M.}\sum_{(\bar{q}_{nm},\bar{l}_{nm})\neq(q_{nm}l_{nm})}\lambda_t(\bar{q}_{nm,}\bar{l}_{nm})}=W_d^T\tilde{\mu}_{q_{nm}l_{nm}} \quad (8)$$

The most widely used features for HMM in face recognition are 2D-DCT coefficients. These DCT coefficients combine excellent decorrelation properties with energy compaction. Indeed, the more correlated the image is, the more energy compaction increases. Thus a relatively small number of DCT coefficients contain the majority of information encapsulated in an image. A second advantage is the speed with which they can be computed since the basis vectors are independent of the database and are often precomputed and stored in an imaging device as part of the JPEG image compression standard[9].

Recognition rates obtained when using 2D DCT with HMM can achieve 100% success on smaller databases such as ORL. In our research we also introduce the use of Daubechies wavelets. Apart from the work of wavelets have not been previously used with HMMs for face recognition applications when test images were extracted from the same video sequence as the training images, proving that the proposed approach can cope with variations in facial features due to small orientation changes, provided the lighting and backgrounds are constant.

C. *Relation between MCHMM and MVE*

In general, MVE and MCHMM are developed fordiscriminative training for verification and classification problems, respectively. Specially, MCHMM is exploited for discriminative feature transformation and HMM training. In MVE framework, the gradient-based iterative procedure is used to minimize expected verification error and estimate utterance verification models. Here, MCHMM is derived using *EM algorithm* and come up with *closed-form solutions* to HMM parameters. Rapid parameter estimation is achievable. MCHMMis applied for the experiments on face recognition as described later. Using MCHMM, each frame window is aligned with the maximum confidence state so that image segmentation can be performed accordingly. Also, when we investigate the relation between training criteria using maximum confidence and MVE, it is interesting to see that log likelihood ratio criterion in (8) is similar to the accumulated mis-verification measure in MVE criterion using log likelihood function as discriminate function.

Correspondingly, maximizing confidence measure is comparable to minimizing verification error. Again, this is why the proposed MCHMM is able to achieve discriminative training performance.

D. MC-HMM viterbi algorithm and implementation procedure

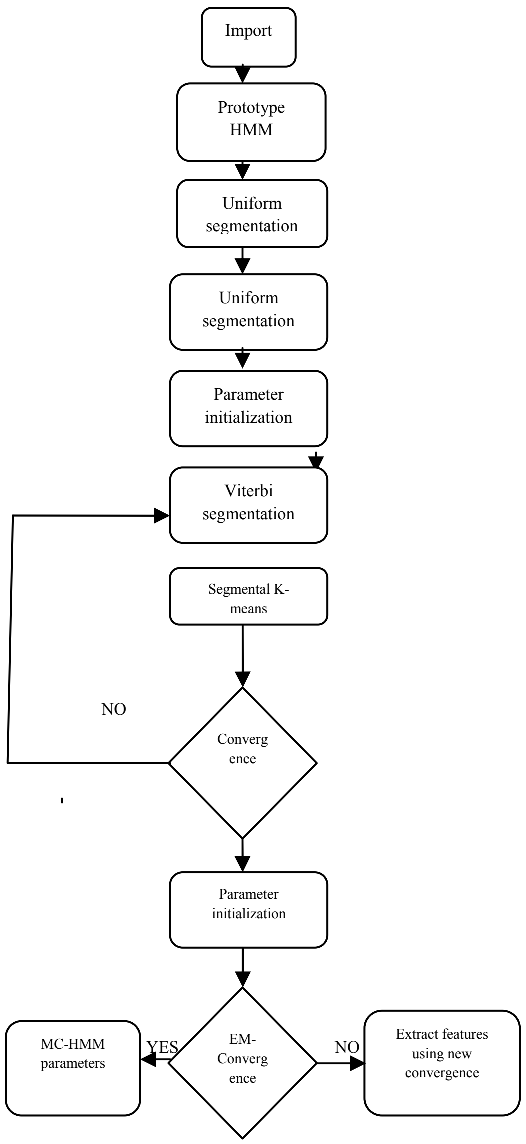

Fig. 5. The training process for an HMM

In MC-HMM viterbi algorithm applied in horizontal direction for feature vectors at each row $\{X_{tn1}, \cdots\cdots, X_{tnm}\}$ using HMM parameters $\{\Pi_h, A_h, B_h\}$
Of super states s_n.

$$\partial_h(q_{nm}) = \max_{q_{n1},\dots,q_{n,m-1}} LLR(X_{tn1},\dots,X_{tnm},q_{n1},\dots,q_{n,m-1},q_{nm} \mid \Lambda) \quad (9)$$

Which accounts for the first m feature vectors and ends with embedded state q_{nm}.

$$\partial_h(q_{nm}) = \max_{1\le q_{n,m-1}\le N_q} \partial_h(q_{n,m-1}).a_{q_{n,m-1}q_{nm}}.LLR(X_{tnm},q_{nm} \mid \Lambda) \quad (10)$$

$$\varphi_h(q_{nm}) = \arg\max_{1\le q_{n,m-1}\le N_q} \partial_h(q_{n,m-1}).a_{q_{n,m-1}q_{nm}} \quad (11)$$

We recursively find $\delta_{h(q_{nm})}$ and $\Psi_h(q_{nm})$ for $m=1,\cdots M$.

In second stage we apply viterbi algorithm in vertical direction similar to (10) and (11).

VI. THE PROBLEMS OF FACE RECOGNITION.

This work has presented the face recognition area, explaining different approaches, methods, tools and algorithms used since the 60's. Some algorithms are better, some are less accurate, some of the are more versatile and others are too computationally costly. Despite this variety, face recognition faces some issues inherent to the problem definition, environmental conditions and hardware constraints[8].

A. Illumination

Many algorithms rely on color information to recognize faces. Features are extracted from color images, although some of them may be gray-scale. The color that we perceive from a given surface depends not only on the surface's nature, but also on the light upon it. In fact, color derives from the perception of our light receptors of the spectrum of light -distribution of light energy versus wavelength. There can be relevant illumination variations on images taken under uncontrolled environment. That said, the chromacity is an essential factor in face recognition. The intensity of the color in a pixel can vary greatly depending on the lighting conditions.

Is not only the sole value of the pixels what varies with light changes. The relation or variations between pixels may also vary. As many feature extraction methods relay on color/intensity variability measures between pixels to obtain relevant data, they show an important dependency on lighting changes. Keep in mind that, not only light sources can vary, but also light intensities may increase or decrease, new light sources added. Entire face regions be obscured or in shadow, and also feature extraction can become impossible because of solarization. The big problem is that two faces of the same subject but with illumination variations may show more differences between them than compared to another subject. Summing up, illumination is one of the big challenges of automated face recognition systems.

B. Occlusion

We understand occlusion as the state of being obstructed. In the face recognition context, it involves that some parts of the face can't be obtained. For example, a face photograph taken from a surveillance camera could be partially hidden behind a column. The recognition process can rely heavily on the availability of a full input face. Therefore, the absence of some parts of the face may lead to a bad classification. This problem speaks in favor of a piecemeal approach to feature extraction, which doesn't depend on the whole face. There are also objects that can occlude facial features - glasses, hats, beards, certain hair cuts, etc.

C. Optical technology

A face recognition system should be aware of the format in which the input images are provided. There are different cameras, with different features, different weaknesses and problems. Usually, most of the recognition processes involve a preprocessing step that deals with this problem.

D. Expression

Facial expression is another variability provider. However, it isn't as strong as illumination or pose. Several algorithms don't deal with this problem in a explicit way, but they show a good performance when different facial expressions are present. On the other hand, the addition of expression variability to pose and illumination problems can become a real impediment for accurate face recognition.

VII. Review and Concluding Remarks

The focus of this paper is on the use of MC-HMM techniques for face recognition. For this we have presented Discriminative Feature Extraction where a transformation matrix W was estimated by maximizing Fisher's ratio criterion. Although additional papers treating specific aspects of this field can be found in the literature, these are invariably based on one or another of the key techniques presented and reviewed here. Our goal has been to quickly enable the interested reader to review and understand the state-of-art for MC-HMM models applied to face recognition problems. It is clear that different techniques balance certain trade-offs between computational complexity, speed and accuracy of recognition and overall practicality and ease-of-use. Our hope is that this paper will make it easier for new researchers to understand and adopt MC-HMM for face analysis and recognition applications and continue to improve and refine the underlying techniques.

References

[1] J.-T. Chien and C.-P. Liao "Maximum confidence hidden markov modeling for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 4, April 2008.

[2] A. Ben-Yishai and D. Burshtein, "A Discriminative Training Algorithm for Hidden Markov Models," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 3, pp. 204-217, 2004.

[3] D. Daleno, L. Cariello, M. Giannini, and G. Mastronardi, "Pseudo 2D Hidden Markov Model and Neural Network Coefficients in Face Recognition," *Via Orabona,* 4–70125 Bari – Italy.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Transaction on Pattern Analysis and Machine Intelligence* vol. 19, no. 7, July 1997.

[5] M. Bicego, U. Castellani, V. Murino, "Using Hidden Markov Models and Wavelets for face recognition," *in IEEE Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)* 0-7695-1948-2/03, 2003

[6] M. S. Kim, D. Kim, and S. Y. Lee, "Face Recognition Using the Embedded HMM with Second-Order Block-Specific Observations," *Pattern Recognition*, vol. 36, no. 11, pp. 2723-2735, 2003.

[7] P. F. de Carrera, *Face Recognition Algorithms*, June 16, 2010

[8] A. V. Nefian and M. H. Hayes, "Hidden markov models for face recognition," in *ICASSP98,* pp. 2721-2724, 1998.

[9] R. C. Gonzalez and R. E. Woods, "Digital image processing," *Addison-Wesley*, Reading, MA, 1992.

[10] C.-P. Liao and J.-T. Chien, "Maximum Confidence Hidden Markov Modeling,"Department of Computer Science and Information Engineering National Cheng Kung University, Tainan, Taiwan 70101, ROC

[11] B. Du, S. Shan, L. Qing, and W. Gao, "Empirical comparisons of several preprocessing methods for illumination insensitive face recognition," *Acoustics, Speech, and Signal Processing*, vol. 2, pp. 981–984, 2005.

[12] M. Gokmen, E. Vucini, and M.E. Groller, Face recognition under varying illumination, 15thWSCG, pages 57–64, 2007.

[13] T. Kanade, Computer recognition of human faces, *Interdisciplinary Systems Research*, 47, 1977.

[14] S. Kuo and O. Agazzi, "Keyword spotting in poorly printed documents using pseudo 2-d hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, pp. 842–848, August 1994.

[15] H. S. Le and H. Li, "Face identification system using single hidden markov model and single sample image per person," *IEEE International Joint Conference on Neural Networks*, vol. 1, 2004.

[16] S. H. Lin, S. Y. Kung, and L. J. Lin, "Face Recognition/Detection by Probabilistic Decision-Based Neural Network," *IEEE Transaction on Neural Networks*, vol. 8, pp. 114-132, 1997.

[17] C. Liu and H. Wechsler, "Evolutionary Pursuit and its Application to Face Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 570-582, 2000.

[18] Z. Sun and J. Sun, "Face Recognition Based on Fractal and Hidden Markov Model," *Third International Symposium on Electronic Commerce and Security*, 2010.

S. H. Patil received the PhD degree in computer engineering. He is currently working as Professor and head of department of computer engineering, Bharathi Vidyapeeth University, college of engineering Pune-21, Maharashtra, India

Swati Raut is a student in M.tech(Computer), Bharati vidyapeeth university, College of engineering pune-21, Maharashtra, India