Maximum Confidence Hidden Markov Modeling

Chih-Pin Liao and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan 70101, ROC
{cpliao, chien}@chien.csie.ncku.edu.tw

ABSTRACT

This paper presents a compact and discriminative hidden Markov model (HMM) approach for general pattern classification. To achieve model compactness and discriminability, we simultaneously perform feature dimension reduction and HMM parameter estimation via maximizing the confidence of accepting the hypothesis that observations are from target HMM states rather than competing HMM states. A new discriminative training criterion is derived using hypothesis test theory. Particularly, we develop the *maximum confidence hidden Markov modeling* (MCHMM) framework for face recognition. Using this framework, we incorporate a transformation matrix to extract discriminative facial features. The continuous-density HMM parameters are estimated using the extracted features. Importantly, we adopt a consistent criterion to build whole framework including feature extraction and model estimation. From the experiments on ORL facial databases, we find that the proposed method obtains robust image segmentation performance in presence of different variations of facial expressions, orientations, etc. In comparison of previous HMM approaches, the proposed MCHMM achieves better recognition accuracies and image segmentation.

1. INTRODUCTION

Face recognition has been known as the key technology for many information security systems. Although template based methods such as principal component analysis (PCA) and linear discriminant analysis (LDA) are popular for face recognition, the recognition accuracies using these methods highly depend on the training templates [6]. If the variations of face images are not contained in template set, system performance is deteriorated seriously. In contrast to template based approaches, hidden Markov model (HMM) first developed for speech recognition [11] has been powerful for many pattern recognition applications, e.g. optical character recognition [9], face recognition, etc. In general, the continuous-density HMM parameters are constructed by a set of state transition probabilities, Gaussian mean vectors and covariance matrices. The state structure is effective to represent time series signals such as speech data where time alignment is a series and critical problem. Due to the great success in speech recognition, HMM has been extensively employed in modeling data variations for face recognition [13], e.g. the variations from face expressions, orientations, with/without eyeglasses and beard, etc. Similar to speech signal, facial image can be blocked in a sequence of sub-images in a top-down left-right manner. This sequence can be characterized by one dimensional (1-D) state sequence in 1-D HMM. Using two-dimensional (2-D) HMM, the vertical and horizontal information of a face image is represented by a 2-D $m \times n$ state matrix. Namely, each face image is divided into m horizontal states and n vertical states. Each state is related to its neighboring state. Since 2-D HMM uses two-dimensional structure information of face image, the computation complexity grows rapidly when the number of states increases in a full connected 2-D HMM. To deal with the issue of extensive computation, several approximated 2-D HMM architectures including pseudo 2D-HMM [4], embedded HMM [7] and low complexity HMM [8] have been proposed.

Conventionally, PCA, LDA, discrete cosine transform or wavelet transform [1] performs linear transformation to extract low dimensional features from high dimensional pixel data. High frequency signal bearing noise information was discarded. Previously, there was no evidence of showing discriminability of using these features. The extracted features directly served as observation data for maximum likelihood (ML) estimation of HMM parameters. ML HMM parameters were assured to optimally match the observed features in terms of likelihood measure. However, model discriminability was not guaranteed. In this paper, we concern on two issues in establishing HMMs for high dimensional observation data. The first issue involves the *hybrid process of feature extraction and model estimation*. We aim to merge two operations under the common objective function. Secondly, we exploit a *new discriminative training* criterion derived from hypothesis test theory. The maximum confidence objective function is carried out for discriminative HMM estimation, which is different from other discriminative training algorithms e.g. minimum verification error (MVE) [12]0 training. We simultaneously estimate transformation matrix and HMM parameters through maximizing confidence criterion. Performing discriminative transformation is crucial to extract salient features for segmenting observation data into corresponding HMM states. Having desirable data alignment, we are able to estimate robust models covering different variations and improve face recognition accuracies in the experiments.

2. RELATED WORKS

Before describing new HMM framework, we first survey two related studies; hypothesis test theory and MVE discriminative training. Because maximum confidence criterion is originated from hypothesis test theory, we introduce background knowledge as follows.

2.1. Hypothesis Test

Hypothesis test principle is used to decide whether we should accept a claimed hypothesis or not. There are three steps in hypothesis test problem; 1. Define null hypothesis H_0 and alternative hypothesis H_1 and choose a significance level. 2. Calculate likelihood functions $P(X|H_0)$ and $P(X|H_1)$ and determine critical region. 3. Make a decision of acceptance or rejection of H_0. According to Neyman-Pearson Lemma, when likelihood functions $P(X \mid H_0)$ and $P(X \mid H_1)$ are given, the optimal solution is obtained by a likelihood ratio test

$$\mathrm{LR} = \frac{P(X|H_0)}{P(X|H_1)} \geq \alpha \,, \tag{1}$$

where α is a decision threshold determined by a significance level for distribution of LR . Generally, likelihood ratio LR has the physical meaning of measuring confidence of choosing null hypothesis. Null hypothesis is accepted if likelihood ratio is greater than α . The higher the likelihood ratio is measured, the stronger the evidence showing observations X belong to null hypothesis H_0 . Here, we concern on the hypothesis whether observation data $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ are from target model or not. We are motivated to construct a new discriminative training criterion for estimating feature transformation matrix and HMM parameters. The estimated parameters are desirable because the maximum confidence towards choosing target models rather than competing models is achieved. For comparison, we also introduce a related discriminative training algorithm called minimum verification error [12]0 We will illustrate the relation between MVE and MCM algorithms later.

2.2. Discriminative Training

Minimum classification error (MCE) [2] is a popular discriminative training algorithm in which the classification error is minimized to estimate discriminative models. When switching classification (multi-class) problem to verification (binary class) problem, we deal with two classes/outcomes which are acceptance and rejection. In [12]0, MCE was extended to minimum verification error (MVE) algorithm that total detection errors were minimized to achieve discriminative verification. MVE aims to estimate discriminative models for verifying test data corresponds to a target model λ_j against non-target model $\overline{\lambda}_j$. This verification problem is solvable using hypothesis test principle. For the purpose of utterance verification, target and competing models represent keyword and non-keyword (cohort) events, respectively. In MVE procedure of training λ_j and $\overline{\lambda}_j$, a discriminant function is defined using log likelihood function

$$\mathrm{g}(X, \lambda_j) = \log P(X|\lambda_j) \,, \tag{2}$$

and the anti-discriminant function is formed by averaging the other discriminant functions

$$G(X, \overline{\lambda}_j) = \log P(X|\overline{\lambda}_j) = \left[\frac{1}{C-1} \sum_{\lambda_c \neq \lambda_j} \eta \log P(X|\lambda_c) \right]^{1/\eta} , \tag{3}$$

where $P(X|\lambda_j)$ is the likelihood function of X given model $\lambda_j \in \Lambda = \{\lambda_1, \cdots, \lambda_C\}$. In case of $\eta = 1$ we have

$$G(X, \overline{\lambda}_j) = \frac{1}{C-1} \sum_{\lambda_c \neq \lambda_j} \log P(X|\lambda_c) = \log \left[\prod_{\lambda_c \neq \lambda_j} P(X|\lambda_c) \right]^{1/C-1} . \tag{4}$$

Mis-verification measure is expressed by

$$d(X, \lambda_j) = -\mathrm{g}(X, \lambda_j) + G(X, \overline{\lambda}_j) \,. \tag{5}$$

Then, the detection error based loss function is calculated by mapping $d(X, \lambda_j)$ into a range of zero to one through a sigmoid function

$$l(X, \lambda_j) = \frac{1}{1 + \exp(-d(X, \lambda_j))} = \frac{1}{1 + \dfrac{\exp(g(X, \lambda_j))}{\exp(G(X, \overline{\lambda}_j))}} \,. \tag{6}$$

Consequently, we minimize the expected detection error to find discriminative model

$$\hat{\Lambda} = \arg\min_{\Lambda} E_X[l(X, \Lambda)] = \arg\min_{\Lambda} E_X \left[\sum_{j=1}^{C} l(X, \lambda_j) \, 1(X \in \lambda_j) \right], \tag{7}$$

where $1(\cdot)$ is the indicator function.

3. MAXIMUM CONFIDENCE HIDDEN MARKOV MODEL

Hereafter, we are introducing a new discriminative training criterion for transforming features and estimating compact hidden Markov models.

3.1. Maximum Confidence Criterion

In [12]0, hypothesis test theory was applied to build verification system. Discriminative training algorithm was derived through minimizing expected verification error. In this study, we directly apply hypothesis test principle for deriving discriminative training criterion. Importantly, 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 state j*

H_1 : *Observation* X *is not from target HMM state j*

We aim to estimate model parameters by optimizing the likelihood ratio of target state to competing state, or equivalently, maximizing the confidence of deciding X is from target state λ_j rather than competing state $\overline{\lambda}_j$. Practically, our goal is to enlarge the separability between HMM states. The estimated parameters are able to improve the correctness of data alignment in HMM state level. Namely, this problem turns out to verify the result of data alignment to the corresponding HMM states. For face recognition, we are aligning facial image into different segments representing forehead, eyes, nose, mouth and chin. It is desirable to use correct state alignment to estimate good model parameters.

Considering high-dimensional facial image data, we perform a class-dependent linear transformation prior to HMM estimation. This transformation is used for dimension reduction and similar to many face recognition approaches using PCA and LDA. Thus, parameter set Λ include transformation matrices $\{W\}$, HMM mixture weights $\{\omega_{jk}\}$, mean vectors $\{\mu_{jk}\}$ and covariance matrices $\{\Sigma_{jk}\}$. For simplicity, we ignore HMM state initial probabilities and transition probabilities. We neglect class label in all expressions. MCHMM is estimated by maximizing the following log likelihood ratio

$$\hat{\Lambda} = \arg\max_{\Lambda} \mathrm{LLR}(X|\Lambda) = \arg\max_{\Lambda} \log P(X|\Lambda) - \log P(X|\overline{\Lambda}) \,, \tag{8}$$

where $\overline{\Lambda}$ is anti- or competing model set. Similar to (3), $\log P(X|\overline{\Lambda})$ can be calculated accordingly.

3.2. Parameter Estimation

When estimating MCHMM, we have hidden variable, i.e. HMM state label j. We should apply expectation-maximization (EM) algorithm to tackle missing data problem for maximum confidence estimation. In E-step, we calculate an expectation function of log likelihood ratio over state variable $s_t = j$

$$Q(\Lambda'|\Lambda) = E_S[\mathrm{LLR}(X,S|\Lambda')|X,\Lambda] = \sum_S P(S|X,\Lambda)\cdot \mathrm{LLR}(X,S|\Lambda')$$
$$= \sum_{t=1}^{T}\sum_{j=1}^{C} P(s_t = j|X,\Lambda)\cdot\left[\log P(\mathbf{x}_t|\lambda'_j) - \frac{1}{C-1}\sum_{\lambda_c\neq\lambda_j} P(\mathbf{x}_t|\lambda'_c)\right] \quad (9)$$

In M-step, we maximize this auxiliary function with respect to new estimate Λ'. Current estimate Λ is used to determine posterior probability $P(s_t = j|X,\Lambda) = \gamma_t(j)$. Before performing maximization step, we expand the expectation function $Q(\Lambda'|\Lambda)$ using K mixtures of Gaussian distributions as

$$\sum_{t=1}^{T}\sum_{j=1}^{C}\sum_{k=1}^{K}\gamma_t(j,k)\left\{\begin{bmatrix}\log\omega'_{ik} + \log|W'| - \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma'^{-1}_{ik}| \\ -\frac{1}{2}(W'\mathbf{x}_t - \mu'_{ik})^T\Sigma'^{-1}_{ik}(W'\mathbf{x}_t - \mu'_{ik})\end{bmatrix}\right.$$
$$\left. -\frac{1}{C-1}\sum_{\lambda_c\neq\lambda_j}\begin{bmatrix}\log\omega'_{ck} + \log|W'| - \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma'^{-1}_{ck}| \\ -\frac{1}{2}(W'\mathbf{x}_t - \mu'_{ck})^T\Sigma'^{-1}_{ck}(W'\mathbf{x}_t - \mu'_{ck})\end{bmatrix}\right\} \quad (10)$$
$$= Q_\omega(\{\omega'_{jk}\}|\{\omega_{jk}\}) + Q_g(\{\mu'_{jk},\Sigma'_{jk},W'\}|\{\mu_{jk},\Sigma_{jk},W\})$$

which can be separated into two sub-functions $Q_\omega(\{\omega'_{jk}\}|\{\omega_{jk}\})$ and $Q_g(\{\mu'_{jk},\Sigma'_{jk},W'\}|\{\mu_{jk},\Sigma_{jk},W\})$. Finally, we find new estimates $\Lambda' = \{\omega'_{jk},\mu'_{jk},\Sigma'_{jk},W'\}$ by taking differentials with respect to the corresponding parameters and setting them to zero. We obtain closed-form solutions to ω'_{jk}, μ'_{jk} and Σ'_{jk} as follows

$$\omega'_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k) - \frac{1}{C-1}\sum_{t=1}^{T}\sum_{\lambda_c\neq\lambda_j}\gamma_t(c,k)}{\sum_{t=1}^{T}\sum_{k=1}^{K}\gamma_t(j,k) - \frac{1}{C-1}\sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{\lambda_c\neq\lambda_j}\gamma_t(c,k)}. \quad (11)$$

$$\mu'_{jk} = W'\left[\frac{\sum_{t=1}^{T}\gamma_t(j,k)\cdot\mathbf{x}_t - \frac{1}{C-1}\sum_{t=1}^{T}\sum_{\lambda_c\neq\lambda_j}\gamma_t(c,k)\cdot\mathbf{x}_t}{\sum_{t=1}^{T}\gamma_t(j,k) - \frac{1}{C-1}\sum_{t=1}^{T}\sum_{\lambda_c\neq\lambda_j}\gamma_t(c,k)}\right]. \quad (12)$$

$$\Sigma'_{jk} = W'^T\cdot\left[\frac{\sum_{t=1}^{T}\gamma_t(j,k)\cdot(\mathbf{x}_t - \mu'_{jk})(\mathbf{x}_t - \mu'_{jk})^T - \frac{1}{C-1}\sum_{t=1}^{T}\sum_{\lambda_c\neq\lambda_j}\gamma_t(c,k)\cdot(\mathbf{x}_t - \mu'_{ck})(\mathbf{x}_t - \mu'_{ck})^T}{\sum_{t=1}^{T}\gamma_t(j,k) - \frac{1}{C-1}\sum_{t=1}^{T}\sum_{\lambda_c\neq\lambda_j}\gamma_t(c,k)}\right]\cdot W \quad (13)$$

However, there is no closed-form solution to feature transformation matrix W'. We adopt gradient decent algorithm to iteratively find optimal new estimate by

$$W'^{(i+1)} = W'^{(i)} + \varepsilon\cdot\frac{\partial Q_g(W'^{(i)}|W)}{\partial W'^{(i)}}, \quad (14)$$

where ε is the learning rate and

$$\frac{\partial Q_g(W'^{(i)}|W)}{\partial W'^{(i)}} = \sum_{j=1}^{C}\sum_{k=1}^{K} T\cdot\Sigma'_{jk}W'^{(i)}\cdot(W'^{(i)^T}\Sigma'_{jk}W'^{(i)})^{-1}. \quad (15)$$

3.3. Relation between MCHMM and MVE

In general, MVE and MCHMM are developed for discriminative 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. MCHMM is 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 discriminant 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.

4. EXPERIMENTS

4.1. Experimental Setup and Implementation Issues

In the experiments, we carried out baseline HMM and MCHMM for face recognition evaluation using two public-domain facial databases; ORL and FERET [10]. Baseline HMM system was also referred as embedded HMM [3] where each face image was characterized by five vertical super-states. Considering three states for forehead, seven states for eyes, seven states for nose, seven states for mouth and three states for chin, there were totally 27 HMM states used to model a human class. A 6x6 sample window with four overlapped pixels was specified to extract feature vectors from left top to right down for a face image. For simplicity, each state was modeled by a single Gaussian distribution. Using ORL database, we randomly selected five images as training images and the other five images as test images as referred to [5]. There were forty classes in ORL database. To verify the robustness of using HMM approaches, FERET database containing sufficient facial variations was adopted for evaluation. We selected 153 human classes with at most eight images provided for each class. All images were resized to 92x104 to match image size in ORL database. For each class, three frontal pose images were selected for training and the remaining images of each class containing different facial variations were included in test data set. Totally, there were 569 test images. To conduct fair comparison, we performed five-fold cross validation for all experiments. When realizing HMM approaches to face recognition, we developed Viterbi search algorithm to perform state alignment for a test image using model parameters from different classes. Using MCHMM, we calculated confidence scores of each frame matching with HMM states.

4.2. Experimental Results

As shown in Table 1, we find that MCHMM (97%) significantly outperforms baseline HMM (95.5%) when using ORL face database. To investigate the effect of feature dimension on computation time and recognition accuracy, we report recognition time per image and averaged recognition rate using FERET database at Table 2 and Figure 1, respectively. Here, baseline HMM is performed. Feature dimension is determined through the estimated transformation matrix W. The cases of 12, 14, 16, 18, 20 and 36-dimensional features are implemented. 36-dim case is marked as baseline curve because the sample window size is 6x6. We can see that computation time is increased by merging more features in face recognition. However, when examining the recognition accuracies, it is attractive to see that the best recognition rates are achieved using 16-dimensional features. In most cases, recognition rates are decreased when considering more classes. In general, the classification robustness is guaranteed using HMM approaches. Finally, we compare the HMM state alignments using baseline HMM and MCHMM in Figure 2. It is obvious that MCHMM discriminative models obtain better segmentation compared to baseline HMM models.

Table 1 Recognition rate using different models on ORL database.

	Baseline HMM	MCHMM
Recognition Rate (%)	95.5	97

Table 2 Recognition time using different feature dimensions. (Platform: CPU: P4 2.4GHz, RAM: 1 GB)

Dimension	10	12	14	16	18	36
Time (sec.)	0.25	0.30	0.38	0.43	0.55	1.67

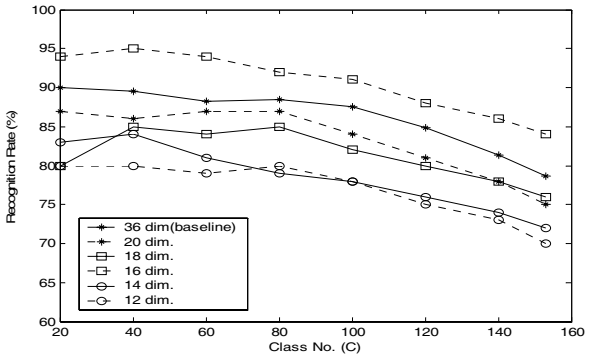

Figure 1: Recognition accuracy using different feature dimensions and class numbers on FERET database.

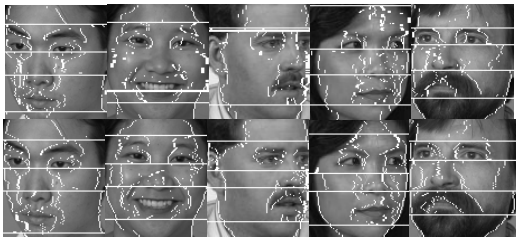

Figure 2: HMM state alignment for FERET face images using baseline HMM (Top Row) and MCHMM (Bottom Row).

5. CONCLUSION

We have presented a new maximum confidence HMM framework for face recognition. The novelty of this paper aimed to develop a discriminative and rapid estimation of feature transformation matrix and HMM parameters. This new objective function was derived from hypothesis test principle. Through verifying HMM states aligned for observation data, we obtained better state alignment for estimating desirable HMM parameters. Different PCA or LDA facial feature extraction, we performed discriminative transformation. EM algorithm was fulfilled to find all model parameters. From experimental results, we confirmed the goodness and robustness of using proposed MCHMM for face recognition.

6. REFERENCES

[1] M. Bicego, U. Castellani and V. Murino, "Using hidden Markov models and wavelets for face recognition", in *Proc. of IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 257-262, 2000.

[2] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error training", *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 40, no. 12, pp. 3043-3054, 1992.

[3] 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.

[4] S. S. Kuo and O. E. Agazzi, "Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov model", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 8, pp. 842-848, 1994.

[5] C.-P. Liao and J.-T. Chien, "Nonsingular discriminant feature extraction for face recognition", in *Proc. of ICASSP*, vol. 2, pp.957-960, 2005.

[6] A. M. Martinez and A. C. Kak, "PCA versus LDA", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 228-233, 2001.

[7] A. V. Nefian and M. H. Hayes III, "An embedded HMM-based approach for face detection and recognition", in *Proc. of ICASSP*, vol. 6, pp. 3553-3556, 1999.

[8] H. Othman and T. Aboulnasr, "A separable low complexity 2D HMM with application to face recognition", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1229-1238, 2003.

[9] H. S. Park and S. W. Lee, "A truly 2D hidden Markov model for off-line handwritten character recognition", *Pattern Recognition*, vol. 31, no. 12, pp. 1849-1864, 1998.

[10] P. J. Phillips, H. Moon, P. J. Rauss and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, 2000.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[12] M. G. Rahim, C.-H. Lee and B.-H. Juang, "Discriminative utterance verification for connected digits recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 266-277, 1997.

[13] F. S. Samaria and S. Young, "HMM-based architecture for face identification", *Image and Vision Computing*, vol. 12, no. 8, pp. 537-543, 1994.

[14] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 6, pp. 420-429, 1996.