Image recognition based on hidden Markov eigen-image models using variational Bayesian method

Kei Sawada*, Kei Hashimoto*, Yoshihiko Nankaku* and Keiichi Tokuda*
*Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan
E-mail: {swdkei, bonanza, nankaku, tokuda}@sp.nitech.ac.jp Tel: +81-52-735-7549

***Abstract*—An image recognition method based on hidden Markov eigen-image models (HMEMs) using the variational Bayesian method is proposed and experimentally evaluated. HMEMs have been proposed as a model with two advantageous properties: linear feature extraction based on statistical analysis and size-and-location-invariant image recognition. In many image recognition tasks, it is difficult to use sufficient training data, and complex models such as HMEMs suffer from the over-fitting problem. This study aims to accurately estimate HMEMs using the Bayesian criterion, which attains high generalization ability by using prior information and marginalization of model parameters. Face recognition experiments showed that the proposed method improves recognition performance.**

I. Introduction

Statistical approaches have been successfully applied in the field of image recognition. In particular, principal component analysis (PCA) based approaches, such as the eigenface (eigen-image) method [1] and subspace method [2], attain good recognition performance. There are many significant classifiers and feature representations. However, in the case of conventional methods, some pre-processing for normalizing image variations, e.g., geometric variations such as size, location, and rotation, is usually applied to input images because many classifiers cannot deal with such image variations. The accuracy of these normalization processes affects recognition performance. Task-dependent normalization techniques have thus been developed for each image recognition task. However, the final objective of image recognition is not to accurately normalize image variations for human perception but to achieve high recognition performance. It is therefore a good idea to integrate the normalization processes into classifiers and optimize them on the basis of a consistent criterion.

Statistical image recognition methods based on hidden Markov models (HMMs) have been proposed to reduce the effect of geometric variations [3], [4], [5]. Geometric matching between input images and models is represented by discrete hidden variables: i.e., the normalization process is included in the calculation of output probabilities. However, the extension of HMMs to multi-dimensions generally leads to an exponential increase in the amount of computation for model training. Separable lattice HMMs (SL-HMMs) have been proposed to reduce computational complexity while retaining good properties, i.e., in the case of two-dimensional data, elastic matching in the vertical and horizontal directions, for modeling multi-dimensional data [6]. This property enables modeling of not only invariances in the size and location of objects but also nonlinear warping in all dimensions. However, SL-HMMs still have a limitation in their application to image recognition: observations are assumed to be generated independently from corresponding states. It is insufficient to represent variations in images, e.g., lighting conditions and object deformation. To overcome the limitation, hidden Markov eigen-image models (HMEMs) have been proposed [7]. The basic idea of HMEMs is that eigen-images are generated from an SL-HMM. In the HMEMs, the eigen-images are represented by probabilistic hidden variable models, such as probabilistic PCA (PPCA) [8] and factor analysis (FA) [9], [10], [11], and geometrically transformed to match an input image by incorporating the state transition structure (into the loading matrix). HMEMs therefore have the advantageous properties of both eigen-images and SL-HMMs based methods: linear feature extraction based on statistical analysis and size-and-location-invariant image recognition.

In spite of the above-mentioned properties, HMEMs suffer from the over-fitting problem because they have a complex model structure compared to PPCA, FA, and SL-HMMs. Additionally, in many image recognition tasks, only a small amount of training data is available and the efforts to achieve high generalization ability are required. The maximum likelihood (ML) criterion has been used for training HMEMs. However, the ML criterion produces a point estimate of model parameters, so the estimation accuracy may be degraded due to the over-fitting problem when the amount of training data is insufficient. To overcome this problem, in the present study, an image recognition technique using HMEMs based on the Bayesian criterion and a training algorithm based on the variational Bayesian (VB) method [12] is proposed. The Bayesian criterion assumes that model parameters are random variables, and high generalization ability can be obtained by marginalizing all model parameters used in estimating predictive distributions. Moreover, the Bayesian criterion can utilize prior distributions representing useful prior information. SL-HMMs estimated by the Bayesian criterion demonstrated better recognition performance than those estimated by the ML criterion [13]. HMEMs based on the Bayesian criterion are therefore expected to achieve high generalization ability.

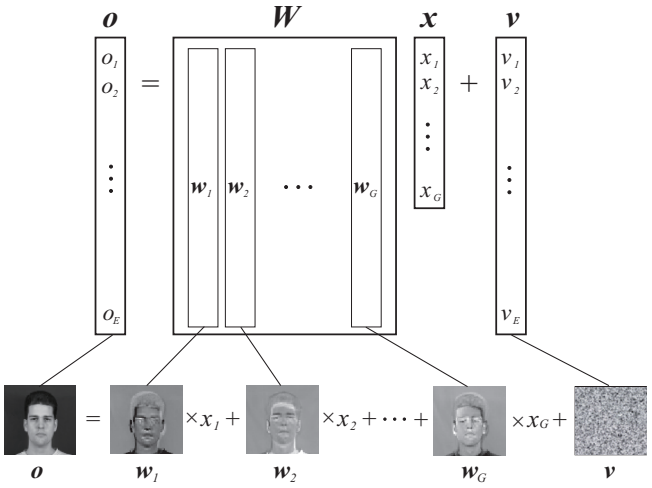

Fig. 1. Model structure of PEMs in face image modeling.

The rest of the paper is organized as follows. Section II explains the structure of HMEMs. Section III describes the VB method for HMEMs. Section IV describes face recognition experiments on the XM2VTS database [14], and Section V concluded the paper.

II. Hidden Markov eigen-image models

A. Probabilistic eigen-image models

Probabilistic principal component analysis (PPCA) [8] and factor analysis (FA) [9] are statistical methods for modeling the covariance structure with a small number of hidden variables. We called them probabilistic eigen-image models (PEMs). In the case of PEMs, an E-dimensional observation vector $\boldsymbol{o}$ is assumed to be generated from a G-dimensional factor vector $\boldsymbol{x}$ ($G < E$) and an E-dimensional noise vector $\boldsymbol{v}$ as follows:

$$\boldsymbol{o} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{v}, \tag{1}$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_G]$ is an $E \times G$ matrix known as a factor loading matrix. Factor vector $\boldsymbol{x}$ is a hidden variable assumed to be distributed in accordance with a standard Gaussian density $\mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{0}, \boldsymbol{I})$, and noise vector $\boldsymbol{v}$ is distributed in accordance with $\mathcal{N}(\boldsymbol{v} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma}$ is assumed to be a diagonal matrix, this model is called FA, and PPCA is a special case of FA in which the noise is isotropic, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. Figure 1 shows the model structure of PEMs in face image modeling. The likelihood of observation $\boldsymbol{o}$ given $\boldsymbol{x}$ can be written as

$$P(\boldsymbol{o} \,|\, \boldsymbol{x}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{o} \,|\, \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

because the product $\boldsymbol{W}\boldsymbol{x}$ becomes a constant vector added to noise vector $\boldsymbol{v}$. Figure 2 shows the graphical model of PEMs. The marginal distribution of observation $\boldsymbol{o}$ is obtained by integrating out the hidden variable $\boldsymbol{x}$ as follows:

$$\begin{aligned} P(\boldsymbol{o} \,|\, \boldsymbol{\Lambda}) &= \int P(\boldsymbol{o} \,|\, \boldsymbol{x}, \boldsymbol{\Lambda}) P(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \mathcal{N}(\boldsymbol{o} \,|\, \boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^\top + \boldsymbol{\Sigma}). \end{aligned} \tag{3}$$

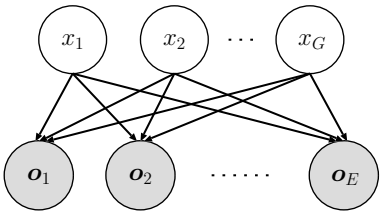

Fig. 2. Graphical model representation of PEMs. The circles represent random variables, clear ones means hidden variables, and shaded ones means observed variables.

From the above equation, it is obvious that PEMs are a Gaussian distribution whose covariance matrix is constrained by the loading matrix and the noise covariance matrix. That is, PEMs can capture the correlation structure among observations by a small number of parameters instead of using the full covariance matrix.

B. Separable lattice hidden Markov models

Separable lattice hidden Markov models (SL-HMMs) are used for modeling multi-dimensional data [6]. In the case that observations are two-dimensional data, e.g., pixel values of an image, observations are assumed to be given on a two-dimensional lattice as:

$$\boldsymbol{O} = \{\boldsymbol{O}_{\boldsymbol{t}} \,|\, \boldsymbol{t} = (t^{(1)}, t^{(2)}) \in \boldsymbol{T}\}, \tag{4}$$

where $\boldsymbol{t}$ denotes the coordinates of the lattice in two-dimensional space $\boldsymbol{T}$ and $t^{(m)} = 1, \ldots, T^{(m)}$ is the coordinate of the m-th dimension for $m \in \{1, 2\}$. In two-dimensional HMMs, observation $\boldsymbol{O}_{\boldsymbol{t}}$ is emitted from a state indicated by hidden variable $\boldsymbol{z}_{\boldsymbol{t}}$. The hidden variables $\boldsymbol{z}_{\boldsymbol{t}} \in \boldsymbol{K}$ can take one of $K = K^{(1)}K^{(2)}$ states, which are assumed to be arranged on a two-dimensional state lattice $\boldsymbol{K} = \{1, \ldots, K\}$. Since observation $\boldsymbol{O}_{\boldsymbol{t}}$ is only dependent on state $\boldsymbol{z}_{\boldsymbol{t}}$ as in ordinary HMMs, dependencies between hidden variables determine the properties and the modeling ability of two-dimensional HMMs.

In SL-HMMs, to reduce the number of possible state sequences, hidden variables are constrained to be composed of two Markov chains as follows:

$$\boldsymbol{z} = \{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}\}, \tag{5}$$

$$\boldsymbol{z}^{(m)} = \{z^{(m)}_{t^{(m)}} \,|\, 1 \leq t^{(m)} \leq T^{(m)}\}, \tag{6}$$

where $\boldsymbol{z}^{(m)}$ is the Markov chain along with the m-th coordinate, and $z^{(m)}_{t^{(m)}} \in \{1, \ldots, K^{(m)}\}$. The composite structure of hidden variables in SL-HMMs is defined as the product of hidden state sequences as:

$$\boldsymbol{z}_{\boldsymbol{t}} = (z^{(1)}_{t^{(1)}}, z^{(2)}_{t^{(2)}}). \tag{7}$$

This means that the segmented regions of observations are constrained to rectangles. That is, it allows an observation lattice to be elastic both horizontally and vertically. Figure 3 shows the model structure of SL-HMMs in face image modeling. The

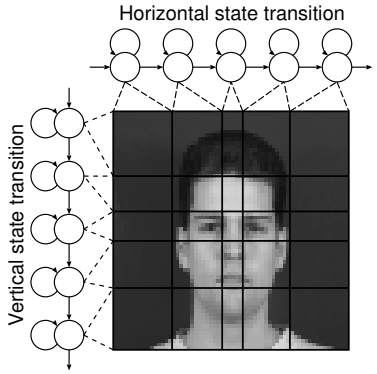

Fig. 3. Model structure of SL-HMMs in face image modeling.

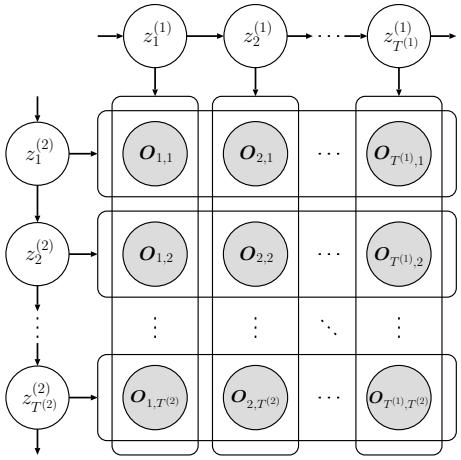

Fig. 4. Graphical model representation of SL-HMMs. The rounded boxes represent a group of variables, and the arrow to each box represents the dependency in regard to all variables in the box instead of drawing arrows to the all the variables.

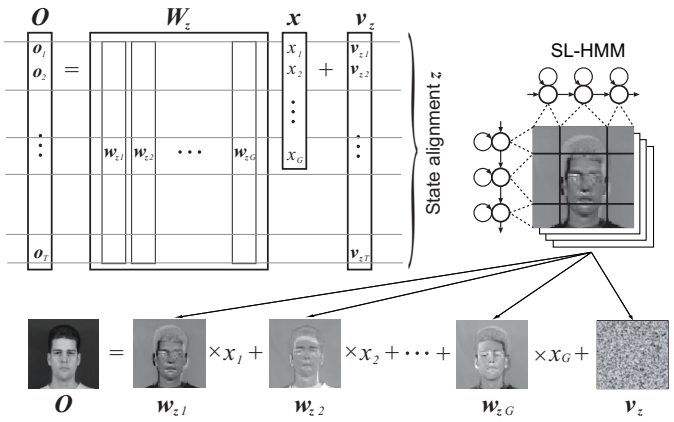

Fig. 5. Model structure of HMEMs in face image modeling.

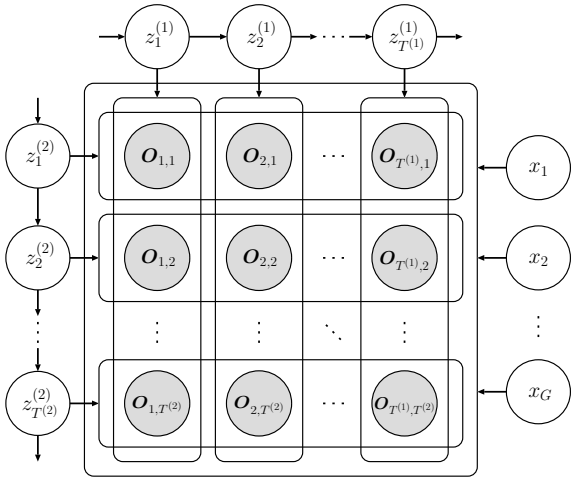

Fig. 6. Graphical model representation of HMEMs.

joint likelihood of observations $\boldsymbol{O}$ and hidden variables $\boldsymbol{z}$ can be written as:

$$P(\boldsymbol{O}, \boldsymbol{z} \,|\, \boldsymbol{\Lambda}) = P(\boldsymbol{O} \,|\, \boldsymbol{z}, \boldsymbol{\Lambda}) \prod_{m=1}^{2} P(\boldsymbol{z}^{(m)} \,|\, \boldsymbol{\Lambda})$$
$$= \prod_{\boldsymbol{t}} P(\boldsymbol{O_t} \,|\, \boldsymbol{z_t}, \boldsymbol{\Lambda}) \times \prod_{m=1}^{2} \left\{ P(z_1^{(m)} \,|\, \boldsymbol{\Lambda}) \prod_{t^{(m)}=2}^{T^{(m)}} P(z_{t^{(m)}}^{(m)} \,|\, z_{t^{(m)}-1}^{(m)}, \boldsymbol{\Lambda}) \right\}, \quad (8)$$

where $\boldsymbol{\Lambda}$ is a set of model parameters. Figure 4 shows graphical model representation of SL-HMMs. In the application of image modeling, SL-HMMs can perform an elastic matching in both horizontal and vertical directions by assuming the transition probabilities with left-to-right and top-to-bottom topologies. However, SL-HMMs have a limitation in their application to image recognition: observations are assumed to be generated independently of corresponding states. It is therefore insufficient to represent variations in images, e.g., lighting conditions and object deformation.

C. Hidden Markov eigen-image models

Hidden Markov eigen-image models (HMEMs) are defined as a model integrating a PEM and an SL-HMM [7]. The basic idea of HMEMs is that eigen-images are generated from an SL-HMM. Figures 5 and 6 show the model structure and graphical model representation of HMEMs, respectively. The likelihood function of HMEMs is defined as:

$$P(\boldsymbol{O} \,|\, \boldsymbol{\Lambda}) = \sum_{\boldsymbol{z}} \int P(\boldsymbol{O} \,|\, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\Lambda}) P(\boldsymbol{x}) P(\boldsymbol{z} \,|\, \boldsymbol{\Lambda}) \mathrm{d}\boldsymbol{x}. \quad (9)$$

where $\boldsymbol{x}$ is a factor vector distributed in accordance with $P(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{z}$ represents state variables as used in SL-HMMs. The transition probabilities are defined as:

$$P(\boldsymbol{z}^{(m)} \,|\, \boldsymbol{\Lambda}) = P(z_1^{(m)} \,|\, \boldsymbol{\Lambda}) \prod_{t^{(m)}=2}^{T^{(m)}} P(z_{t^{(m)}}^{(m)} \,|\, z_{t^{(m)}-1}^{(m)}, \boldsymbol{\Lambda}), \quad (10)$$

$$P(z_1^{(m)} = i \,|\, \boldsymbol{\Lambda}) = \pi_i^{(m)}, \quad (11)$$

$$P(z_{t^{(m)}}^{(m)} = j \,|\, z_{t^{(m)}-1}^{(m)} = i, \boldsymbol{\Lambda}) = a_{ij}^{(m)}. \quad (12)$$

The output probabilities, given as $\boldsymbol{x}$ and $\boldsymbol{z}$, are defined as:

$$P(\boldsymbol{O}\,|\,\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda}) = \prod_{\boldsymbol{t}} P(\boldsymbol{O_t}\,|\,\boldsymbol{x},\boldsymbol{z_t},\boldsymbol{\Lambda}), \quad (13)$$

$$P(\boldsymbol{O_t}\,|\,\boldsymbol{x},\boldsymbol{z_t}=\boldsymbol{k},\boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{O_t}\,|\,\boldsymbol{W_k}\boldsymbol{x}+\boldsymbol{\mu_k},\boldsymbol{\Sigma_k}), \quad (14)$$

where $\boldsymbol{\Lambda}$ is a set of model parameters $\{\boldsymbol{\pi}^{(m)},\boldsymbol{a}^{(m)},\boldsymbol{W_k},\boldsymbol{\mu_k},\boldsymbol{\Sigma_k}\}$: $\boldsymbol{\pi}^{(m)}$ is a set of initial state probabilities, $\boldsymbol{a}^{(m)}$ is a set of state transition probabilities, $\boldsymbol{W_k}$ is the loading matrix at state $\boldsymbol{k}$ in two-dimensional state space $\boldsymbol{K}$, and $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ denote the mean vector and covariance matrix of the noise vector at state $\boldsymbol{k}$. By incorporating the state transition structure into the loading matrix, eigen-images can be transformed to match an input image, and this state transition structure performs size and location normalization. Once the state sequences are given, HMEMs are regarded as PEMs which given normalized data. HMEMs therefore overcome the limitation of SL-HMMs (i.e., the correlation among all observations can be modeled through the factor variables) and thus share the advantageous properties of both PEMs and SL-HMMs: a linear feature extraction based on statistical analysis and invariances to size and location of images. Moreover, the structure of HMEMs includes conventional PEMs and SL-HMMs as special cases: HMEMs with the same number of states as the number of pixels of the input images become the conventional PEMs, and HMEMs with zero factor become the standard SL-HMMs.

III. Hidden Markov eigen-image models using variational Bayesian method

A. Bayesian criterion

The maximum likelihood (ML) criterion has been used to train HMEMs in image recognition [7]. The optimal model parameters $\boldsymbol{\Lambda}_{\mathrm{ML}}$ in the ML criterion are estimated by maximizing the likelihood of training data as follows:

$$\boldsymbol{\Lambda}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{O}\,|\,\boldsymbol{\Lambda}). \quad (15)$$

The predictive distribution for testing data $\boldsymbol{X}$ in the testing stage is given by $P(\boldsymbol{X}\,|\,\boldsymbol{\Lambda}_{\mathrm{ML}})$. However, the ML criterion produces a point estimate of model parameters, so the estimation accuracy may decreased by the over-fitting problem when there is an insufficient amount of training data.

On the other hand, the predictive distribution of the Bayesian criterion is given by:

$$P(\boldsymbol{X}\,|\,\boldsymbol{O}) = \int P(\boldsymbol{X}\,|\,\boldsymbol{\Lambda})P(\boldsymbol{\Lambda}\,|\,\boldsymbol{O})\mathrm{d}\boldsymbol{\Lambda}. \quad (16)$$

The posterior distribution $P(\boldsymbol{\Lambda}\,|\,\boldsymbol{O})$ for a set of model parameters $\boldsymbol{\Lambda}$ can be written with the Bayes' theorem as:

$$P(\boldsymbol{\Lambda}\,|\,\boldsymbol{O}) = \frac{P(\boldsymbol{O}\,|\,\boldsymbol{\Lambda})P(\boldsymbol{\Lambda})}{P(\boldsymbol{O})}, \quad (17)$$

where $P(\boldsymbol{\Lambda})$ is a prior distribution for $\boldsymbol{\Lambda}$, and $P(\boldsymbol{O})$ is an evidence. The model parameters are integrated out in Eq. (16) so that the effect of over-fitting is mitigated. That is, the Bayesian criterion has higher generalization ability than the ML criterion when there is an insufficient amount of training data. However, the Bayesian criterion requires complicated integral and expectation computations to obtain the posterior distributions when the models include hidden variables, such as HMEMs. The variational Bayesian (VB) method has been proposed as an approximation to overcome this problem [12]. In this study, the VB method was applied to HMEMs for image recognition.

B. Variational Bayesian method for HMEMs

1) Posterior distribution: The VB method is an approximate version of the Bayesian approach. That is, an approximate posterior distribution is estimated by maximizing a lower bound for log marginal likelihood $\mathcal{F}$ instead of the true likelihood. The lower bound of the log marginal likelihood is defined by using Jensen's inequality as:

$$\begin{aligned}\ln P(\boldsymbol{O}) &= \ln \sum_{\boldsymbol{z}} \iint P(\boldsymbol{O},\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\Lambda} \\ &= \ln \sum_{\boldsymbol{z}} \iint Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})\frac{P(\boldsymbol{O},\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})}{Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\Lambda} \\ &\geq \sum_{\boldsymbol{z}} \iint Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})\ln\frac{P(\boldsymbol{O},\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})}{Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\Lambda} \\ &= \mathcal{F}(Q), \end{aligned} \quad (18)$$

where $Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})$ is an arbitrary distribution. The relation between the log marginal likelihood and the lower bound $\mathcal{F}$ is represented by the Kullback-Leibler (KL) divergence between $Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})$ and the true posterior distribution $P(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda}\,|\,\boldsymbol{O})$ as:

$$\mathcal{F}(Q) = \ln P(\boldsymbol{O}) - \mathrm{KL}[Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})\,||\,P(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda}\,|\,\boldsymbol{O})]. \quad (19)$$

Maximizing $\mathcal{F}$ with respect to $Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})$ therefore provides a good approximation of posterior distribution $P(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda}\,|\,\boldsymbol{O})$ in terms of minimizing KL divergence. The solution can be obtained by using a functional approximation based on the variational method.

To obtain the approximate posterior distribution (VB posterior distribution) $Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda})$, hidden variables are assumed to be conditionally independent of one another, i.e.,

$$Q(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\Lambda}) = Q(\boldsymbol{x})Q(\boldsymbol{z}^{(1)})Q(\boldsymbol{z}^{(2)})Q(\boldsymbol{\Lambda}), \quad (20)$$

where $\int Q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1$, $\sum_{\boldsymbol{z}^{(m)}} Q(\boldsymbol{z}^{(m)}) = 1$, and $\int Q(\boldsymbol{\Lambda})\mathrm{d}\boldsymbol{\Lambda} = 1$. Under this assumption, the optimal VB posterior distributions that maximize objective function $\mathcal{F}$ are given by the variational method as:

$$\begin{aligned}Q(\boldsymbol{x}) \propto P(\boldsymbol{x})\exp\Bigg[&\sum_{\boldsymbol{z}^{(1)}}\sum_{\boldsymbol{z}^{(2)}}\int Q(\boldsymbol{z}^{(1)})Q(\boldsymbol{z}^{(2)})Q(\boldsymbol{\Lambda}) \\ &\times \ln P(\boldsymbol{O}\,|\,\boldsymbol{x},\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\boldsymbol{\Lambda})\mathrm{d}\boldsymbol{\Lambda}\Bigg], \end{aligned} \quad (21)$$

$$Q(\boldsymbol{z}^{(m)}) \propto \exp\Bigg[\sum_{\boldsymbol{z}^{(\tilde{m})}} \iint Q(\boldsymbol{x})Q(\boldsymbol{z}^{(\tilde{m})})Q(\boldsymbol{\Lambda}) \times \ln P(\boldsymbol{O}\,|\,\boldsymbol{x},\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\boldsymbol{\Lambda})P(\boldsymbol{z}^{(m)}\,|\,\boldsymbol{\Lambda})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\Lambda}\Bigg], \quad (22)$$

$$Q(\boldsymbol{\Lambda}) \propto P(\boldsymbol{\Lambda}) \exp\Bigg[\sum_{\boldsymbol{z}^{(1)}}\sum_{\boldsymbol{z}^{(2)}} \int Q(\boldsymbol{x})Q(\boldsymbol{z}^{(1)})Q(\boldsymbol{z}^{(2)}) \times \ln P(\boldsymbol{O}\,|\,\boldsymbol{x},\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\boldsymbol{\Lambda})P(\boldsymbol{z}^{(1)}\,|\,\boldsymbol{\Lambda})P(\boldsymbol{z}^{(2)}\,|\,\boldsymbol{\Lambda})\mathrm{d}\boldsymbol{x}\Bigg], \quad (23)$$

where $\tilde{m}$ represents the $\tilde{m}$-th dimension, which is an alternative to the m-th dimension. Since VB posterior distributions, $Q(\boldsymbol{x})$, $Q(\boldsymbol{z}^{(m)})$, and $Q(\boldsymbol{\Lambda})$ are dependent on each other, these updates need to be iterated as the expectation-maximization (EM) algorithm. The updates of the VB posterior distributions increase the value of objective function $\mathcal{F}$ at each iteration until convergence.

2) Prior distribution: The Bayesian criterion has the advantage that it can utilize prior distributions representing useful prior information on model parameters. Although arbitrary distributions can be used as prior distributions, conjugate prior distributions are widely used. A conjugate prior distribution is a distribution in which the resulting posterior distribution belongs to the same distribution family as the prior distribution. The conjugate prior distribution of an HMEM is defined as:

$$P(\boldsymbol{\Lambda}) = \prod_{m=1}^{2}\Bigg[\mathcal{D}(\boldsymbol{\pi}^{(m)}\,|\,\boldsymbol{\phi}^{(m)})\prod_{i=1}^{K^{(m)}}\mathcal{D}(\boldsymbol{a}_i^{(m)}\,|\,\boldsymbol{\alpha}_i^{(m)})\Bigg] \times \prod_{\boldsymbol{k}}\prod_{d=1}^{D}\mathcal{N}(\tilde{\boldsymbol{w}}_{\boldsymbol{k},d}\,|\,\boldsymbol{h}_{\boldsymbol{k},d},\boldsymbol{U}_{\boldsymbol{k}}^{-1}\sigma_{\boldsymbol{k},d}^2)\mathcal{G}((\sigma_{\boldsymbol{k},d}^2)^{-1}\,|\,\eta_{\boldsymbol{k}},\nu_{\boldsymbol{k},d}), \quad (24)$$

where D is the dimension of observation $\boldsymbol{O}_{\boldsymbol{t}}$, $\mathcal{D}(\cdot)$ and $\mathcal{N}(\cdot)\mathcal{G}(\cdot)$ are respectively a Dirichlet distribution and a Gauss-Gamma distribution, and $\tilde{\boldsymbol{w}}_{\boldsymbol{k},d}$ and $\sigma_{\boldsymbol{k},d}^2$ are defined as:

$$\boldsymbol{W}_{\boldsymbol{k}} = [\boldsymbol{w}_{\boldsymbol{k},1},\boldsymbol{w}_{\boldsymbol{k},2},\ldots,\boldsymbol{w}_{\boldsymbol{k},D}]^\top, \quad (25)$$

$$\boldsymbol{\mu}_{\boldsymbol{k}} = [\mu_{\boldsymbol{k},1},\mu_{\boldsymbol{k},2},\ldots,\mu_{\boldsymbol{k},D}]^\top, \quad (26)$$

$$\tilde{\boldsymbol{w}}_{\boldsymbol{k},d} = [\,\boldsymbol{w}_{\boldsymbol{k},d}^\top\ \mu_{\boldsymbol{k},d}\,]^\top, \quad (27)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{k}} = \mathrm{diag}(\sigma_{\boldsymbol{k},1}^2,\sigma_{\boldsymbol{k},2}^2,\ldots,\sigma_{\boldsymbol{k},D}^2), \quad (28)$$

where $\boldsymbol{W}_{\boldsymbol{k}}$ is assumed to be independent of each dimension. These distributions can be represented by a set of hyper-parameters $\{\boldsymbol{\phi}^{(m)},\boldsymbol{\alpha}_i^{(m)},\boldsymbol{h}_{\boldsymbol{k},d},\boldsymbol{U}_{\boldsymbol{k}},\eta_{\boldsymbol{k}},\nu_{\boldsymbol{k},d}\}$, where $\boldsymbol{h}_{\boldsymbol{k},d}$ and $\boldsymbol{U}_{\boldsymbol{k}}$ are defined as:

$$\boldsymbol{h}_{\boldsymbol{k},d} = \left[\,\boldsymbol{\omega}_{\boldsymbol{k},d}^\top\ \gamma_{\boldsymbol{k},d}\,\right]^\top, \quad (29)$$

$$\boldsymbol{U}_{\boldsymbol{k}}^{-1} = \begin{bmatrix}\boldsymbol{\Upsilon}_{\boldsymbol{k}} & \boldsymbol{u}_{\boldsymbol{k}}\\ \boldsymbol{u}_{\boldsymbol{k}}^\top & \upsilon_{\boldsymbol{k}}\end{bmatrix}. \quad (30)$$

The posterior distributions can also be represented by the same set of parameters $\{\bar{\boldsymbol{\phi}}^{(m)},\bar{\boldsymbol{\alpha}}_i^{(m)},\bar{\boldsymbol{h}}_{\boldsymbol{k},d},\bar{\boldsymbol{U}}_{\boldsymbol{k}},\bar{\eta}_{\boldsymbol{k}},\bar{\nu}_{\boldsymbol{k},d}\}$ because a conjugate prior distribution is used. Figures 7 and 8 show the graphical model representations of HMEMs for the ML and VB methods, respectively.

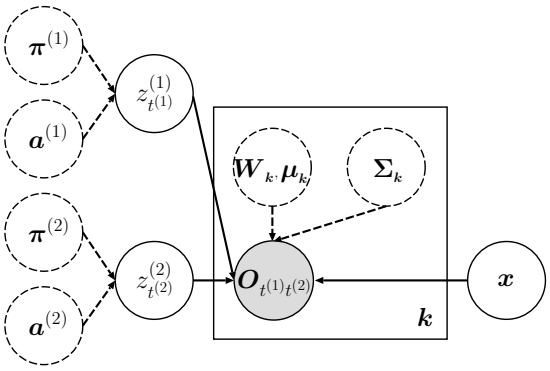

Fig. 7. Graphical model representation for HMEMs using the ML method. The dashed circles represent model parameters, and the rectangle represents the plate over the state $\boldsymbol{k}$.

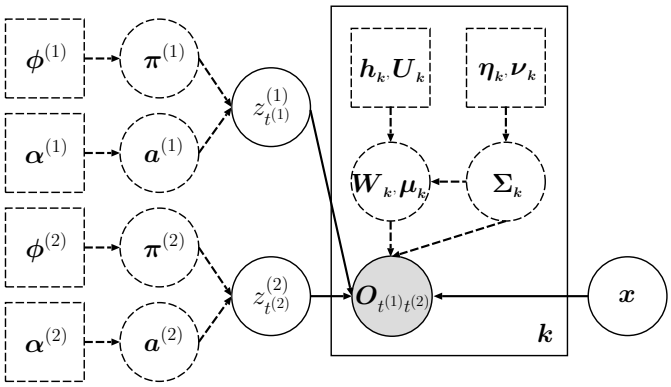

Fig. 8. Graphical model representation for HMEMs using the VB method. The dashed rectangles represent hyper-parameters.

3) Re-estimation formulae: In the HMEMs, the expectations with respect to $Q(\boldsymbol{z})$ is defined as:

$$\left\langle z_{i,t}^{(m)}\right\rangle = \sum_{\boldsymbol{z}^{(m)}} Q(\boldsymbol{z}^{(m)})z_{i,t^{(m)}}^{(m)}, \quad (31)$$

$$\left\langle z_{i,t^{(m)}-1}^{(m)}z_{j,t^{(m)}}^{(m)}\right\rangle = \sum_{\boldsymbol{z}^{(m)}} Q(\boldsymbol{z}^{(m)})z_{i,t^{(m)}-1}^{(m)}z_{j,t^{(m)}}^{(m)}, \quad (32)$$

$$\langle z_{\boldsymbol{k},\boldsymbol{t}}\rangle = \sum_{\boldsymbol{z}^{(1)}}\sum_{\boldsymbol{z}^{(2)}} Q(\boldsymbol{z}^{(1)})Q(\boldsymbol{z}^{(2)})z_{i,t^{(1)}}^{(1)}z_{j,t^{(2)}}^{(2)}, \quad (33)$$

$$z_{i,t^{(m)}}^{(m)} = \begin{cases}0 & (z_{t^{(m)}}^{(m)} \neq i)\\ 1 & (z_{t^{(m)}}^{(m)} = i)\end{cases}, \quad (34)$$

$$N_{\boldsymbol{k}} = \sum_{\boldsymbol{t}}\langle z_{\boldsymbol{k},\boldsymbol{t}}\rangle. \quad (35)$$

The VB posterior distribution in Eq. (21), i.e., $Q(\boldsymbol{x})$, can be written as a Gaussian distribution:

$$Q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}\,|\,\bar{\boldsymbol{\mu}}_{\boldsymbol{x}},\bar{\boldsymbol{\Sigma}}_{\boldsymbol{x}}), \quad (36)$$

where $\bar{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ are the mean and covariance matrix, respectively. The re-estimation formulae of the VB posterior

distribution $Q(\boldsymbol{x})$ are derived as:

$$\bar{\boldsymbol{\mu}}_{\boldsymbol{x}} = \bar{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\left\{\sum_{\boldsymbol{k}}\left[\sum_{d=1}^{D}\left\{\bar{\boldsymbol{\omega}}_{\boldsymbol{k},d}\bar{\eta}_{\boldsymbol{k}}\bar{\nu}_{\boldsymbol{k},d}^{-1}\sum_{\boldsymbol{t}}\langle z_{\boldsymbol{k},\boldsymbol{t}}\rangle O_{\boldsymbol{t},d}\right\}\right.\right. \\ \left.\left. -N_{\boldsymbol{k}}\sum_{d=1}^{D}(\bar{\boldsymbol{\omega}}_{\boldsymbol{k},d}\bar{\eta}_{\boldsymbol{k}}\bar{\nu}_{\boldsymbol{k},d}^{-1}\bar{\gamma}_{\boldsymbol{k},d}+\bar{\boldsymbol{u}}_{\boldsymbol{k}})\right]\right\}, \quad (37)$$

$$\bar{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \left[\boldsymbol{I}+\sum_{\boldsymbol{k}}\left\{N_{\boldsymbol{k}}\sum_{d=1}^{D}(\bar{\boldsymbol{\omega}}_{\boldsymbol{k},d}\bar{\eta}_{\boldsymbol{k}}\bar{\nu}_{\boldsymbol{k},d}^{-1}\bar{\boldsymbol{\omega}}_{\boldsymbol{k},d}^{\top}+\bar{\boldsymbol{\Upsilon}}_{\boldsymbol{k}})\right\}\right]^{-1}, \quad (38)$$

where $O_{\boldsymbol{t},d}$ is $\boldsymbol{O}_{\boldsymbol{t}} = \{O_{\boldsymbol{t},d}\,|\,d = 1,2,\ldots,D\}$. The VB posterior distribution $Q(\boldsymbol{z}^{(m)})$ in Eq. (22) has a Markovian structure as the likelihood function of an HMM. Therefore, Eqs. (31) and (32) can be computed efficiently by the Forward-Backward algorithm. The updates of the expectations with respect to VB posterior distribution $Q(\boldsymbol{z}^{(m)})$ are derived as:

$$\left\langle \log \pi_i^{(m)} \right\rangle = \Psi(\bar{\phi}_i^{(m)}) - \Psi\left(\sum_{l=1}^{K^{(m)}}\bar{\phi}_l^{(m)}\right), \quad (39)$$

$$\left\langle \log a_{ij}^{(m)} \right\rangle = \Psi(\bar{\alpha}_{ij}^{(m)}) - \Psi\left(\sum_{l=1}^{K^{(m)}}\bar{\alpha}_{il}^{(m)}\right), \quad (40)$$

$$\langle \ln \mathcal{N}(\boldsymbol{O}_{\boldsymbol{t}}\,|\,\boldsymbol{W}_{\boldsymbol{k}}\boldsymbol{x}+\boldsymbol{\mu}_{\boldsymbol{k}},\boldsymbol{\Sigma}_{\boldsymbol{k}})\rangle \\ = \sum_{d=1}^{D}\Bigg[\ln \mathcal{N}(O_{\boldsymbol{t},d}\,|\,\bar{\boldsymbol{h}}_{\boldsymbol{k},d}^{\top}\langle\tilde{\boldsymbol{x}}\rangle, \bar{\eta}_{\boldsymbol{k},d}^{-1}\bar{\nu}_{\boldsymbol{k},d}) \\ -\frac{1}{2}\ln\bar{\eta}_{\boldsymbol{k},d}+\frac{1}{2}\Psi(\bar{\eta}_{\boldsymbol{k},d})-\frac{1}{2}\mathrm{Tr}\Big\{\bar{\boldsymbol{h}}_{\boldsymbol{k},d}^{\top}\left\langle\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^{\top}\right\rangle\bar{\boldsymbol{h}}_{\boldsymbol{k},d} \\ -\bar{\boldsymbol{h}}_{\boldsymbol{k},d}^{\top}\langle\tilde{\boldsymbol{x}}\rangle\langle\tilde{\boldsymbol{x}}\rangle^{\top}\bar{\boldsymbol{h}}_{\boldsymbol{k},d}+\bar{\boldsymbol{U}}_{\boldsymbol{k}}^{-1}\left\langle\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^{\top}\right\rangle\Big\}\Bigg], \quad (41)$$

where $\Psi(\cdot)$ is a digamma function, and $\langle\tilde{\boldsymbol{x}}\rangle$ and $\left\langle\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^{\top}\right\rangle$ are expectations with respect to $Q(\boldsymbol{x})$, calculated from Eqs. (37) and (38) as follows:

$$\langle\tilde{\boldsymbol{x}}\rangle = \left[\,\bar{\boldsymbol{\mu}}_{\boldsymbol{x}}^{\top}\;\;1\,\right]^{\top}, \quad (42)$$

$$\left\langle\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^{\top}\right\rangle = \begin{bmatrix}\bar{\boldsymbol{\Sigma}}_{\boldsymbol{x}}+\bar{\boldsymbol{\mu}}_{\boldsymbol{x}}\bar{\boldsymbol{\mu}}_{\boldsymbol{x}}^{\top} & \bar{\boldsymbol{\mu}}_{\boldsymbol{x}} \\ \bar{\boldsymbol{\mu}}_{\boldsymbol{x}}^{\top} & 1\end{bmatrix}. \quad (43)$$

The VB posterior distribution in Eq. (23), i.e., $Q(\boldsymbol{\Lambda})$, can be written by a Dirichlet distribution and a Gauss-Gamma distribution. The re-estimation formulae of the VB posterior distribution $Q(\boldsymbol{\Lambda})$ are derived as:

$$\bar{\phi}_i^{(m)} = \phi_i^{(m)} + \left\langle z_{i,1}^{(m)}\right\rangle, \quad (44)$$

$$\bar{\alpha}_{ij}^{(m)} = \alpha_{ij}^{(m)} + \sum_{t^{(m)}=2}^{T^{(m)}}\left\langle z_{i,t^{(m)}-1}^{(m)} z_{j,t^{(m)}}^{(m)}\right\rangle, \quad (45)$$

Fig. 9. Examples of faces in the XM2VTS database.

$$\bar{\boldsymbol{h}}_{\boldsymbol{k},d} = \bar{\boldsymbol{U}}_{\boldsymbol{k}}^{-1}\left\{\boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{h}_{\boldsymbol{k},d}+\sum_{\boldsymbol{t}}\langle z_{\boldsymbol{k},\boldsymbol{t}}\rangle O_{\boldsymbol{t},d}\langle\tilde{\boldsymbol{x}}\rangle\right\}, \quad (46)$$

$$\bar{\boldsymbol{U}}_{\boldsymbol{k}} = \boldsymbol{U}_{\boldsymbol{k}}+N_{\boldsymbol{k}}\left\langle\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^{\top}\right\rangle, \quad (47)$$

$$\bar{\eta}_{\boldsymbol{k}} = \eta_{\boldsymbol{k}}+\frac{1}{2}N_{\boldsymbol{k}}, \quad (48)$$

$$\bar{\nu}_{\boldsymbol{k},d} = \nu_{\boldsymbol{k},d}+\frac{1}{2}\sum_{\boldsymbol{t}}\langle z_{\boldsymbol{k},\boldsymbol{t}}\rangle O_{\boldsymbol{t},d}O_{\boldsymbol{t},d} \\ +\frac{1}{2}\boldsymbol{h}_{\boldsymbol{k},d}^{\top}\boldsymbol{U}_{\boldsymbol{k}}\boldsymbol{h}_{\boldsymbol{k},d}-\frac{1}{2}\bar{\boldsymbol{h}}_{\boldsymbol{k},d}^{\top}\bar{\boldsymbol{U}}_{\boldsymbol{k}}\bar{\boldsymbol{h}}_{\boldsymbol{k},d}. \quad (49)$$

As for the proposed method, HMEMs with the same number of states as the number of pixels of the input images are equivalent to the conventional FA using the VB method [11], and HMEMs with zero factor become the standard SL-HMMs using the VB method [13].

C. Identification using variational Bayesian method

Predictive distribution $P(\boldsymbol{X}\,|\,\boldsymbol{O})$ is computed by using Eq. (16) in the testing stage of the VB method. Since $Q(\boldsymbol{\Lambda})$ is an approximation of true posterior distribution $P(\boldsymbol{\Lambda}\,|\,\boldsymbol{O})$, $Q(\boldsymbol{\Lambda})$ can be substituted for $P(\boldsymbol{\Lambda}\,|\,\boldsymbol{O})$ in Eq. (16). Although Eq. (16) includes a complicated expectation calculation, the same approximation as in training can be used. In image recognition using HMEMs, the HMEMs are separately trained for each class, i.e., subject, and the likelihood of testing data, which is calculated by the predictive distribution of HMEMs, is compared. The class that obtains the highest likelihood is then chosen as the identification result.

IV. Experiments

Face recognition experiments on the XM2VTS database [14] were conducted to evaluate the effectiveness of the proposed method. We prepared eight images of 100 subjects; six images were used for training, and two images were used for testing. Face images composed of 64×64 grayscale pixels were extracted from the original images. Figure 9 shows some examples of face image in the XM2VTS database. HMEMs with 32×32 states were used in the experiments. Two noise variance structures of the HMEMs were compared: PPCA structure (HMEM-PPCA) and FA structure (HMEM-FA).

Since prior distributions of model parameters affect the estimation of posterior distributions in the Bayesian criterion, determining prior distributions is a serious problem from the

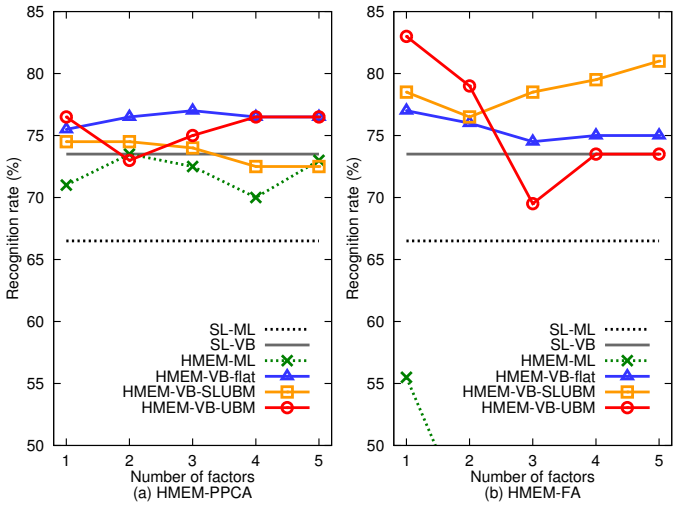

Fig. 10. Recognition rates obtained in face recognition experiments.

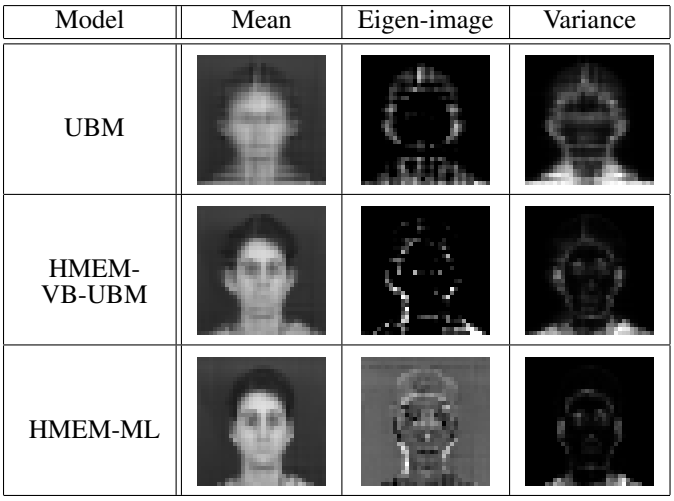

Fig. 11. The values of the mean, eigen-image, and variance were represented by gray-levels (the number of factors is one in the case of HMEM-FA).

viewpoint of estimating appropriate models. In the experiments, a uniform distribution and a universal background model (UBM) [15] that was constructed using all training samples of all subjects were used. In addition, a tuning parameter was used to control the degree of influence of the prior distribution.

In the experiments, the recognition rates were compared among the following four methods: SL-HMMs using ML ("SL-ML") and VB ("SL-VB"), and HMEMs using ML ("HMEM-ML") and VB (HMEM-VB, the proposed method). As the prior distribution for the HMEM-VB, three kinds of prior distributions were prepared: a uniform distribution ("HMEM-VB-flat"), a UBM representing statistics of SL-HMMs ("HMEM-VB-SLUBM"), and a UBM representing statistics of HMEMs ("HMEM-VB-UBM"). The tuning parameter was chosen to obtain the best recognition rate under the condition that the number of factors was one.

Figure 10(a) and 10(b) show the recognition rates of HMEM-PPCA and HMEM-FA, respectively. The VB method achieved higher recognition performance than the ML method in the case of both variance structures. The highest recognition rates of the VB method was 83.0% when using "HMEM-VB-UBM" with the FA structure (one factor) and while 73.5% was obtained by the ML method when using "HMEM-ML" with the PPCA structure (two factors). Additionally, in the case of the VB method, the HMEMs outperformed the SL-HMMs method. These results suggest that the proposed method can be effectively applied to image recognition. Comparing PPCA and FA structures, in the case of the ML method, HMEM-PPCA showed better recognition performance than HMEM-FA. Contrary when using the VB method, HMEM-FA was better than HMEM-PPCA. The highest recognition rates of HMEM-FA and HMEM-PPCA with VB method were 83.0% ("HMEM-VB-UBM" with one factor) and 77.0% ("HMEM-VB-flat" with three factors), respectively. This is because in the estimation of HMEM-FA, the ML criterion suffered from the over-fitting problem due to insufficient training data due to the more complex structure of HMEM-FA than HMEM-PPCA. In contrast, the VB method mitigated the over-fitting problem because it uses of prior distributions and marginalization of model parameters. These results show that the estimation accuracy of the noise variance affected the recognition performance and the proposed method can estimate the noise variance reliably.

Three prior distributions (flat, SLUBM, and UBM) for HMEMs using the VB method were compared in the experiments. Although in the case of PPCA, "HMEM-VB-flat" obtained high recognition rates. In the case of FA, "HMEM-VB-SLUBM" obtained high recognition rates. This result indicates that the UBM obtained from SL-HMMs is effective for determining the prior distribution for HMEM-FA. This is because the variance structures of HMEMs were assumed to be diagonal and have the same structure as SL-HMMs. In addition, there was no significant difference between "HMEM-VB-UBM" and "HMEM-VB-flat" when the number of factors was more than one. A possible reason for this result is that the tuning parameter was chosen under the condition that the number of factors is one. This result suggests that it is difficult to set the prior distribution of the loading matrix reasonably. However, high recognition performance can be expected if the loading matrix of the prior distribution can be set adequately.

Figure 11 shows the visualized mean, eigen-image and variance when the number of factors in HMEM-FA is one. It can be seen from Figure 11 that the UBM roughly represents a facial shape. Moreover, "HMEM-VB-UBM" and "HMEM-ML" represent the facial shape of the subject. Since "HMEM-VB-UBM" shows a clearer outline of the face than "HMEM-ML", "HMEM-VB-UBM" can estimate the accurate model.

V. Conclusion

An image recognition method based on hidden Markov eigen-image models (HMEMs) using the variational Bayesian

method was proposed. In face recognition experiments, HMEMs based on the Bayesian criterion demonstrated higher recognition performance than the ML criterion. These results suggest that the Bayesian criterion is useful for applications of image recognition based on HMEMs. Investigation of appropriate parameter sharing structures of HMEMs and experiments on various image recognition tasks will be future work.

REFERENCES

[1] M. Turk, and A. Pentland, "Face recognition using eigenfaces," *IEEE Computer Society Conference*, pp. 586-591, 1955.

[2] S. Watanabe, and N. Pakvasa, "Subspace Method of Pattern Recognition," *1st International Joint Conference on Pattern Recognition*, pp. 25-32, 1973.

[3] E. Levin, and R. Pieraccini, "DYNAMIC PLANAR WARPING FOR OPTICAL CHARACTER RECOGNITION," *ICASSP*, vol.3, pp. 149-152, 1992.

[4] S. Kuo, and O. E. Agazzi, "Keyword Spotting in Poorly Printed Documents Using Pseudo 2-D Hidden Markov Models," *TPAMI*, vol. 16, pp. 842-848, 1994.

[5] A. V. Nefian, and M. H. Hayes III, "Maximum Likelihood Training of Embedded HMM for Face Detection and Recognition," *ICIP*, vol. 1, pp. 33-36, 2000.

[6] D. Kurata, Y. Nankaku, K. Tokuda, T. Kitamura, and Z. Ghahramani, "Face Recognition based on Separable Lattice HMMs," *ICASSP*, vol. 5, pp. 737-740, 2006.

[7] Y. Nankaku, and K. Tokuda, "FACE RECOGNITION USING HIDDEN MARKOV EIGENFACE MODELS," *ICASSP*, vol. 2, pp. 469-472, 2007.

[8] M. E. Tipping, and C. M. Bishop, "Mixtures of Probabilistic Principal Component Analyzers," *Neural Computation*, vol. 11(2), pp. 443-482, 1999.

[9] D. Rubin, and D. Thayer, "EM Algorithms for ML Factor Analysis," *Psychometrika*, vol. 47, pp. 69-76, 1982.

[10] A. I. Rosti, and M. J. F. Gales, "Generalised linear Gaussian models," *Technical Report CUED/F-INFENG/TR.420*, 2001.

[11] Z. Ghahramani, and M. J. Beal, "Variational Inference for Bayesian Mixtures of Factor Analysers," *NIPS*, pp. 449-455, 1999.

[12] H. Attias, "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes," *UAI*, pp. 21-30, 1999.

[13] K. Sawada, A. Tamamori, K. Hashimoto, Y. Nankaku, and K. Tokuda, "FACE RECOGNITION BASED ON SEPARABLE LATTICE 2-D HMMS USING VARIATIONAL BAYESIAN METHOD," *ICASSP*, pp. 2205-2208, 2012.

[14] K. Messer, J. Mates, J. Kitter, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," *AVBPA*, pp. 72-77, 1999.

[15] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," *Eurospeech*, pp. 963-966, 1997.