FACE RECOGNITION BASED ON SEPARABLE LATTICE 2-D HMMS USING VARIATIONAL BAYESIAN METHOD

Kei Sawada, Akira Tamamori, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan

ABSTRACT

This paper proposes an image recognition technique based on separable lattice 2-D HMMs (SL2D-HMMs) using the variational Bayesian method. SL2D-HMMs have been proposed to reduce the effect of geometric variations, e.g., size and location. The maximum likelihood criterion had previously been used in training SL2D-HMMs. However, in many image recognition tasks, it is difficult to use sufficient training data, and it suffers from the over-fitting problem. A higher generalization ability based on model marginalization is expected by applying the Bayesian criterion and useful prior information on model parameters can be utilized as prior distributions. Experiments on face recognition indicated that the proposed method improved image recognition.

Index Terms— face recognition, hidden Markov model, separable lattice 2-D HMMs, Bayesian criterion, variational Bayesian method

1. INTRODUCTION

Statistical approaches have been successfully applied in image recognition. Especially, principal component analysis (PCA) based approaches such as the eigenface method [1] and subspace method show good recognition performance in many applications. Although there are many significant classifiers and feature representations, some pre-processing is usually applied to input images prior to feature extraction and classifier training. The aim of pre-processing is to normalize image variations, e.g., geometric variations such as size, location, and rotation. These normalization processes are important because many classifiers cannot absorb such image variations and the accuracy of normalization significantly affects image recognition. However, task dependent heuristic techniques are applied independently of classifiers. Therefore, it is necessary to develop normalization techniques for each task. Furthermore, the final objective in image recognition is not to accurately normalize images for human perception but to achieve better recognition. Therefore, it is a good idea to integrate the normalization processes into classifiers and optimize them based on a consistent criterion to improve recognition.

Hidden Markov model (HMM) based techniques have been proposed to reduce the influence of geometric variations. Geometric matching between input images and model parameters is represented by discrete hidden variables and the normalization process is included in calculating probabilities. However, the extension of HMMs to multi-dimensions generally leads to an exponential increase in the amount of computation for the training algorithm. Separable lattice 2-D HMMs (SL2D-HMMs) have been proposed [2] to reduce computational complexity while retaining outstanding properties that model multi-dimensional data. SL2D-HMMs can perform elastic matching both horizontally and vertically, which makes it possible to model not only invariances to the size and location of an object but also nonlinear warping in all dimensions.

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 typically been used in image recognition using SL2D-HMMs. However, since the ML criterion produces a point estimate of model parameters, the estimation accuracy may be degraded due to the over-fitting problem when the amount of training data is insufficient. The Bayesian criterion, on the other hand, assumes that model parameters are random variables, and a high generalization ability can be obtained by marginalizing all model parameters in estimating predictive distributions. Moreover, the Bayesian criterion can utilize prior distributions representing useful prior information on model parameters. However, the Bayesian criterion requires complicated integral and expectation computations to obtain posterior distributions when models have hidden variables. To overcome this problem, maximum a posteriori (MAP) method [3] and variational Bayesian (VB) method [4] have been proposed as approximation methods. We applied the Bayesian criterion to image recognition based on SL2D-HMMs, and derived the training algorithm based on the VB method, both of which are explained in this paper.

The rest of the paper is organized as follows. Section 2 briefly explains the structure of SL2D-HMMs, and Section 3 describes the VB method for SL2D-HMMs. Section 4 presents face recognition experiments we did on the XM2VTS database and we finally conclude the paper in Section 5.

2. SEPARABLE LATTICE 2-D HMMS

Separable lattice 2-D hidden Markov models (SL2D-HMMs) are defined for modeling two-dimensional data. The observations of two-dimensional data, e.g., the pixel values of an image and image sequence, are assumed to be given on a two-dimensional lattice:

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

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 $m \in \{1, 2\}$. Observation $\boldsymbol{O_t}$ is emitted from the state indicated by hidden variable $\boldsymbol{S_t} \in \boldsymbol{K}$. Hidden variables $\boldsymbol{S_t} \in \boldsymbol{K}$ can take one of $K = K^{(1)}K^{(2)}$ states, which are assumed to be arranged on two-dimensional state lattice $\boldsymbol{K} = \{1, \ldots, K\}$. In other words, a set of hidden variables $\{\boldsymbol{S_t} \,|\, \boldsymbol{t} \in \boldsymbol{T}\}$ represents a segmentation of observations into K states, and each state corresponds to a segmented region in which the observation vectors are assumed to be generated from the same local deformation. Since observation $\boldsymbol{O_t}$ is only dependent on state $\boldsymbol{S_t}$ as in ordinary HMMs, dependencies among hidden variables determine the properties and modeling abilities of two-dimensional HMMs.

To reduce the number of possible state sequences, the hidden

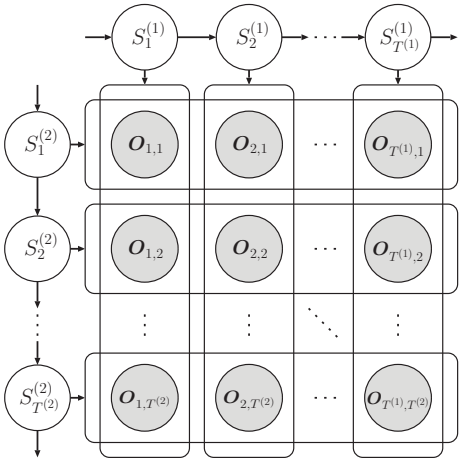

Fig. 1. Graphical representation of SL2D-HMMs

variables to be composed of two Markov chains by constraint are:

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

$$\boldsymbol{S}^{(m)} = \{S^{(m)}_{t^{(m)}} \,|\, 1 \leq t^{(m)} \leq T^{(m)}\}, \tag{3}$$

where $\boldsymbol{S}^{(m)}$ is the Markov chain along with the m-th coordinate and $S^{(m)}_{t^{(m)}} \in \{1, \ldots, K^{(m)}\}$. The composite structure of hidden variables in SL2D-HMMs is defined as the product of hidden state sequences:

$$\boldsymbol{S_t} = (S^{(1)}_{t^{(1)}}, S^{(2)}_{t^{(2)}}). \tag{4}$$

This means that the segmented regions of observations are constrained to rectangles, which allows an observation lattice to be elastic both horizontally and vertically. The number of possible state sequences can be reduced by using this structure from $\{\prod_m K^{(m)}\}^{\prod_m T^{(m)}}$ to $\prod_m \{K^{(m)}\}^{T^{(m)}}$.

Figure 1 shows the graphical representation of SL2D-HMMs. The joint probability of observation vectors $\boldsymbol{O}$ and hidden variables $\boldsymbol{S}$ can be written as:

$$P(\boldsymbol{O}, \boldsymbol{S} \,|\, \boldsymbol{\Lambda}) = P(\boldsymbol{O} \,|\, \boldsymbol{S}, \boldsymbol{\Lambda}) \prod_{m=1}^{2} P(\boldsymbol{S}^{(m)} \,|\, \boldsymbol{\Lambda}). \tag{5}$$

When it is assumed that the state output probability distributions are a single Gaussian distribution, a set of model parameters $\boldsymbol{\Lambda}$ is represented by $\{\boldsymbol{\pi}^{(m)}, \boldsymbol{a}^{(m)}, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\}$, where $\{\pi^{(m)}_i\}^{K^{(m)}}_{i=1}$ is the initial state probability, $\{a^{(m)}_{ij}\}^{K^{(m)}}_{i,j=1}$ is the state transition probability, and $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ are the mean vector and the covariance matrix of the Gaussian distribution at state $\boldsymbol{k}$ on 2-D state space $\boldsymbol{K}$, i.e., $P(S^{(m)}_1 = i \,|\, \boldsymbol{\Lambda}) = \pi^{(m)}_i$, $P(S^{(m)}_{t^{(m)}} = j \,|\, S^{(m)}_{t^{(m)}-1} = i, \boldsymbol{\Lambda}) = a^{(m)}_{ij}$, and $P(\boldsymbol{O_t} \,|\, \boldsymbol{S_t} = \boldsymbol{k}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{O_t} \,|\, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$.

3. SEPARABLE LATTICE 2-D HMMS USING VARIATIONAL BAYESIAN METHOD

3.1. Bayesian criterion

The maximum likelihood (ML) criterion has typically been used to train SL2D-HMMs in image recognition. The optimal model parameters are estimated in the ML criterion by maximizing the likelihood of training data as:

$$\boldsymbol{\Lambda}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{O} \,|\, \boldsymbol{\Lambda}). \tag{6}$$

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

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

$$P(\boldsymbol{X} \,|\, \boldsymbol{O}) = \int P(\boldsymbol{X} \,|\, \boldsymbol{\Lambda}) \, P(\boldsymbol{\Lambda} \,|\, \boldsymbol{O}) \, d\boldsymbol{\Lambda}. \tag{7}$$

Posterior distribution $P(\boldsymbol{\Lambda} \,|\, \boldsymbol{O})$ for a set of model parameters $\boldsymbol{\Lambda}$ can be written with the Bayes theorem.

$$P(\boldsymbol{\Lambda} \,|\, \boldsymbol{O}) = \frac{P(\boldsymbol{O} \,|\, \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda})}{P(\boldsymbol{O})}, \tag{8}$$

where $P(\boldsymbol{\Lambda})$ is a prior distribution for $\boldsymbol{\Lambda}$ and $P(\boldsymbol{O})$ is evidence. The model parameters are integrated out in Eq. (7) so that the effect of over-fitting is mitigated. That is, the Bayesian criterion has a higher generalization ability than the ML criterion when there are insufficient numbers of training data. However, the Bayesian criterion requires complicated integral and expectation computations to obtain posterior distributions when models have hidden variables. Maximum a posteriori (MAP) and variational Bayesian (VB) methods have been proposed as approaches to approximation to overcome this problem.

The optimal model parameters in the MAP method are estimated by maximizing the posterior probability for given training data as:

$$\boldsymbol{\Lambda}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{O} \,|\, \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}). \tag{9}$$

The MAP method can utilize prior distribution $P(\boldsymbol{\Lambda})$, and can be seen as an extension of the ML criterion. Testing in the MAP method is conducted using predictive distribution $P(\boldsymbol{X} \,|\, \boldsymbol{\Lambda}_{\mathrm{MAP}})$. However, it still suffers from the over-fitting problem because of point estimates, when there are insufficient numbers of training data.

3.2. Variational Bayesian method

The VB method is an approximate version of the Bayesian approach. Since the VB method does not use an asymptotic assumption for the amount of data, it is possible to overcome the problem in the MAP method. An approximate posterior distribution is estimated in the VB method by maximizing a lower bound for log marginal likelihood $\mathcal{F}$ instead of the true likelihood. A lower bound for log marginal likelihood is defined by using Jensen's inequality:

$$\begin{aligned}
\ln P(\boldsymbol{O}) &= \ln \sum_{\boldsymbol{S}} \int P(\boldsymbol{O}, \boldsymbol{S} \,|\, \boldsymbol{\Lambda}) \, P(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \\
&= \ln \sum_{\boldsymbol{S}} \int Q(\boldsymbol{S}, \boldsymbol{\Lambda}) \, \frac{P(\boldsymbol{O}, \boldsymbol{S} \,|\, \boldsymbol{\Lambda}) \, P(\boldsymbol{\Lambda})}{Q(\boldsymbol{S}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\
&\geq \sum_{\boldsymbol{S}} \int Q(\boldsymbol{S}, \boldsymbol{\Lambda}) \ln \frac{P(\boldsymbol{O}, \boldsymbol{S} \,|\, \boldsymbol{\Lambda}) \, P(\boldsymbol{\Lambda})}{Q(\boldsymbol{S}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\
&= \mathcal{F},
\end{aligned} \tag{10}$$

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

$$\mathcal{F} = \ln P(\boldsymbol{O}) - \mathrm{KL}(Q(\boldsymbol{S}, \boldsymbol{\Lambda}) \,||\, P(\boldsymbol{S}, \boldsymbol{\Lambda} \,|\, \boldsymbol{O})). \tag{11}$$

Therefore, maximizing $\mathcal{F}$ with respect to $Q(\boldsymbol{S}, \boldsymbol{\Lambda})$ provides a good approximation of posterior distribution $P(\boldsymbol{S}, \boldsymbol{\Lambda} \,|\, \boldsymbol{O})$ in terms of

minimizing the KL divergence. The solution can be obtained by functional approximation based on the variational method.

To obtain approximate posterior distribution (VB posterior distribution) $Q(\boldsymbol{S}, \boldsymbol{\Lambda})$, we assumed that random variables were conditionally independent of one another, i.e.,

$$Q(\boldsymbol{S}, \boldsymbol{\Lambda}) = Q(\boldsymbol{S}^{(1)})Q(\boldsymbol{S}^{(2)})Q(\boldsymbol{\Lambda}), \tag{12}$$

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

$$\begin{aligned} Q(\boldsymbol{S}^{(m)}) = C_{\boldsymbol{S}^{(m)}} \exp\Bigg[& \sum_{\boldsymbol{S}^{(m')}} \int Q(\boldsymbol{S}^{(m')})Q(\boldsymbol{\Lambda}) \\ & \times \ln P(\boldsymbol{O}, \boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)} \,|\, \boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \Bigg], \end{aligned} \tag{13}$$

$$\begin{aligned} Q(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}) \exp\Bigg[& \sum_{\boldsymbol{S}^{(1)}} \sum_{\boldsymbol{S}^{(2)}} Q(\boldsymbol{S}^{(1)})Q(\boldsymbol{S}^{(2)}) \\ & \times \ln P(\boldsymbol{O}, \boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)} \,|\, \boldsymbol{\Lambda}) \Bigg], \end{aligned} \tag{14}$$

where $C_{\boldsymbol{S}^{(m)}}$ is the normalization term for $Q(\boldsymbol{S}^{(m)})$ and $C_{\boldsymbol{\Lambda}}$ is that for $Q(\boldsymbol{\Lambda})$, and m' is the m'-th dimension different from the m-th dimension. Since the VB posterior distributions, $Q(\boldsymbol{S}^{(m)})$ and $Q(\boldsymbol{\Lambda})$, that are obtained are dependent on each other, these updates need to be iterated as the EM algorithm. The update equations increase the value of the objective function $\mathcal{F}$ at each iteration until convergence.

Predictive distribution $P(\boldsymbol{X} \,|\, \boldsymbol{O})$ is estimated using Eq. (7) in the testing stage of the VB method. Since $Q(\boldsymbol{\Lambda})$ is an approximation of posterior distribution $P(\boldsymbol{\Lambda} \,|\, \boldsymbol{O})$, $Q(\boldsymbol{\Lambda})$ can be substituted for $P(\boldsymbol{\Lambda} \,|\, \boldsymbol{O})$ in Eq. (7). Although Eq. (7) includes a complicated expectation calculation, the same approximation as that in training can be applied. In face recognition using SL2D-HMMs, SL2D-HMMs are trained for each class, i.e., subject, separately, and the likelihood of testing data, which is calculated by the predictive distribution of SL2D-HMMs, is compared among all subjects and the class which obtains the highest likelihood is chosen as the result.

3.3. Prior distribution

The Bayesian criterion has an advantage in 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 as prior distributions. A conjugate prior distribution is a distribution where the resulting posterior distribution belongs to the same distribution family as the prior distribution. The conjugate prior distribution of an SL2D-HMM is defined as:

$$\begin{aligned} P(\boldsymbol{\Lambda}) = & \prod_{m=1}^{2} \left[\mathcal{D}(\boldsymbol{\pi}^{(m)} \,|\, \boldsymbol{\phi}^{(m)}) \prod_{i=1}^{K^{(m)}} \mathcal{D}(\boldsymbol{a}_i^{(m)} \,|\, \boldsymbol{\alpha}_i^{(m)}) \right] \\ & \times \prod_{\boldsymbol{k}} \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{k}} \,|\, \boldsymbol{\nu}_{\boldsymbol{k}}, \xi_{\boldsymbol{k}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{k}}) \mathcal{W}(\boldsymbol{\Sigma}_{\boldsymbol{k}}^{-1} \,|\, \eta_{\boldsymbol{k}}, \boldsymbol{B}_{\boldsymbol{k}}), \end{aligned} \tag{15}$$

where $\mathcal{D}(\cdot)$ is a Dirichlet distribution, and $\mathcal{N}(\cdot)\mathcal{W}(\cdot)$ is a Gauss-Wishart distribution. These distributions can be represented by a set of hyper-parameters $\{\boldsymbol{\phi}^{(m)}, \boldsymbol{\alpha}_i^{(m)}, \boldsymbol{\nu}_{\boldsymbol{k}}, \xi_{\boldsymbol{k}}, \eta_{\boldsymbol{k}}, \boldsymbol{B}_{\boldsymbol{k}}\}$. The posterior distributions can also be represented by the same set of parameters $\{\bar{\boldsymbol{\phi}}^{(m)}, \bar{\boldsymbol{\alpha}}_i^{(m)}, \bar{\boldsymbol{\nu}}_{\boldsymbol{k}}, \bar{\xi}_{\boldsymbol{k}}, \bar{\eta}_{\boldsymbol{k}}, \bar{\boldsymbol{B}}_{\boldsymbol{k}}\}$ by using a conjugate prior distribution.

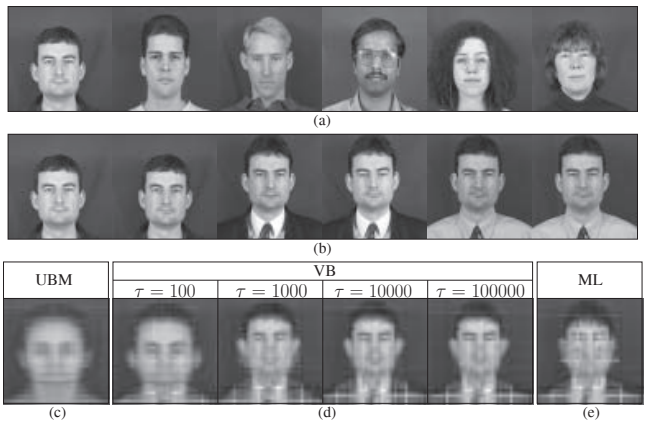

Fig. 2. (a) training images of different subjects, (b) all training images of one subject, (c) mean vector of UBM, (d) mean vector of posterior distribution, (e) mean vector of model obtained with ML method

Since the prior distributions of model parameters affect the estimation of posterior distributions in the Bayesian criterion, determining prior distributions is a serious problem in estimating appropriate models. We set the prior distribution to $P(\boldsymbol{\Lambda}) \propto P(\tilde{\boldsymbol{O}} \,|\, \boldsymbol{\Lambda})$ by using data $\tilde{\boldsymbol{O}}$ given in advance (we called this prior data). We used all training samples for all subjects as prior data in the research discussed in this paper. This is the same idea as that in the universal background model (UBM). The hyper-parameters based on a UBM are given as:

$$\begin{aligned} \phi_i^{(m)} = \frac{\tilde{T}_{0i}}{\tau} + 1, \quad \alpha_{ij}^{(m)} = \frac{\tilde{T}_{ij}}{\tau} + 1, \quad \boldsymbol{\nu}_{\boldsymbol{k}} = \tilde{\boldsymbol{O}}_{\boldsymbol{k}}, \\ \xi_{\boldsymbol{k}} = \frac{\tilde{T}_{\boldsymbol{k}}}{\tau}, \quad \eta_{\boldsymbol{k}} = \frac{\tilde{T}_{\boldsymbol{k}}}{\tau} + D, \quad \boldsymbol{B}_{\boldsymbol{k}} = \frac{\tilde{T}_{\boldsymbol{k}}}{\tau}\tilde{\boldsymbol{C}}_{\boldsymbol{k}}, \end{aligned} \tag{16}$$

where D is the dimension of a feature vector and τ is the tuning parameter. Statistics $\tilde{T}_{0i}, \tilde{T}_{ij}$, and $\tilde{T}_{\boldsymbol{k}}$ correspond to the occupancy probabilities of initial state i, state transition from i to j, and state $\boldsymbol{k}$ with respect to the prior data, respectively. Statistics $\tilde{\boldsymbol{O}}_{\boldsymbol{k}}$ and $\tilde{\boldsymbol{C}}_{\boldsymbol{k}}$ correspond to the mean vector and the covariance matrix of prior data in the $\boldsymbol{k}$-th state. We can control the degree of influence the prior distribution has on the posterior distribution by adjusting tuning parameter τ.

4. EXPERIMENTS

Face recognition experiments on the XM2VTS database [5] were conducted to evaluate the effectiveness of the proposed method. We prepared eight images of 100 subjects; six images were used for training and there were two images for testing. Face images of 64×64 grayscale pixels were extracted from the original images. SL2D-HMMs with 8×8, 16×16, 24×24, 32×32, 40×40, 48×48, 56×56, and 64×64 states were used in these experiments. The hyper-parameters of the prior distribution were determined by using statistics on UBM, which was trained from all training data. The ML (conventional), MAP, and VB methods (proposed) were compared to separately evaluate the two advantages of the Bayesian approaches, i.e., the use of the prior distribution and marginalization of model parameters.

Examples of training images and mean vectors have been given in Figure 2 to demonstrate what effect prior distributions had in the Bayesian approach. Figure 2(a) presents training images of different subjects, and Figure 2(b) presents all training images for one subject.

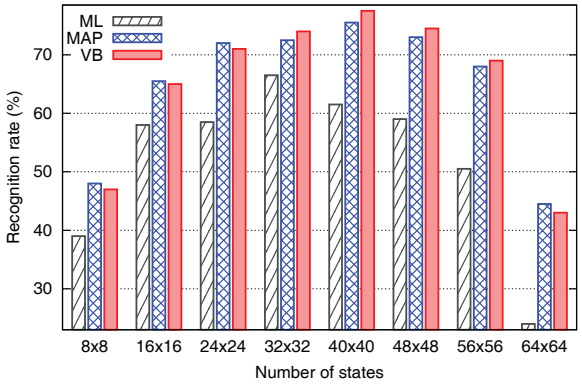

Fig. 3. Recognition rate (training data are six images)

Figures 2(c) and 2(e) present the mean vector of the model obtained with the UBM and ML method. Figure 2(d) has the mean vectors of the posterior distributions obtained with the VB methods by varying tuning parameter τ. The number of states is 40×40 for all models in Figure 2. Although we can see that UBM roughly represents a facial shape from Figure 2(c), it is difficult to identify the characteristics of a particular subject. As tuning parameter τ is increased in Figure 2(d), the mean vector gradually changes from UBM to the image of the subject in Figure 2(b). Using UBM as the prior distribution with appropriate tuning parameter τ, similar state alignments are expected to be obtained for all subject models and this therefore avoids the over-fitting problem. It can actually be seen in Figure 2(d) that the VB method preserved the shape of the face using UBM, even though the shape of the ML criterion in Figure 2(e) had collapsed due to over-fitting. However, tuning parameter τ needs to be carefully determined, because the optimal value depends on the number of training data and the number of states. Although a method of automatically determining the tuning parameter based on cross validation [6] and empirical Bayes have been proposed, we adjusted the value in this experiment to obtain the best recognition rate and the adjusted values were obtained in a range from $\tau = 100$ to $\tau = 6000$.

Figure 3 shows the recognition rates for ML, MAP and VB methods using six images as training data for each subject. We can see from the results that the Bayesian criterion achieved significantly better recognition rates than the ML criterion. Similar performance was obtained under all conditions by comparing the MAP and VB methods. However, the VB method was slightly better than MAP when the appropriate number of states was selected. The highest recognition rates for VB and MAP methods were obtained at 40×40 states (VB: 77.5%, MAP: 75.5%), and ML obtained the best results at 32×32 states (ML: 66.5%). Although the recognition rate for ML significantly decreased with the increasing number of states, the VB and MAP methods retained higher performance than the ML method. This implies that the prior distribution for the VB and MAP methods effectively avoided the over-fitting problem.

Figure 4 shows the recognition rates for ML, MAP, and VB methods while the numbers of training data were changed. The VB and MAP methods achieved higher recognition rates than ML for all numbers of training images. The difference between ML and VB/MAP especially became larger when small numbers of training images were used. Although the MAP and the VB methods had almost the same recognition rates, the VB method obtained slightly better recognition rates when two training images were used. These results suggest that the Bayesian approach mitigated the over-fitting problem and achieved higher generalization ability than the ML method. In addition, we confirmed that the use of a prior distribution was more effective than the marginalization of model parameters in this task.

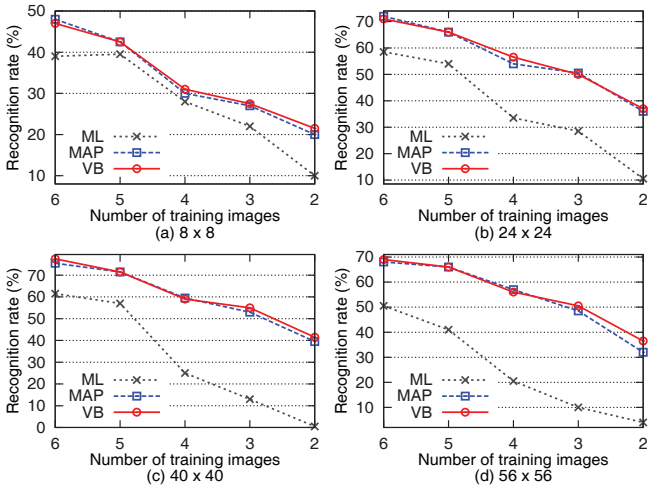

Fig. 4. Recognition rate (training data are six to two images) (a) 8×8 states, (b) 24×24 states, (c) 40×40 states, and (d) 56×56 states

5. CONCLUSION

This paper proposed image recognition based on SL2D-HMMs using the VB method. The face recognition experiments were performed on the XM2VTS database. The Bayesian criterion demonstrated better recognition than the ML criterion in the experiments. These results suggest that the Bayesian criterion is useful for applications of image recognition based on SL2D-HMMs. In addition, we confirmed from these results that the use of prior distributions was more effective than the marginalization of model parameters in this task. We intend to apply the Bayesian criterion to image recognition based on hidden Markov eigenface models, which integrate SL2D-HMMs and factor analyzers, in future work.

6. ACKNOWLEDGEMENTS

This work was partially supported by the Hori Sciences & Arts Foundation and the Artificial Intelligence Research Promotion Foundation.

7. REFERENCES

[1] M. Turk and A. Pentland, "Face recognition using eigenfaces," *IEEE Computer Society Conference*, pp. 586–591, 1991.

[2] 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.

[3] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. SAP*, vol. 2, pp. 291–298, 1994.

[4] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," *UAI*, pp. 21–30, 1999.

[5] K. Messer, J. Mates, J. Kitter, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," *AVBPA*, pp. 72–77, 1999.

[6] K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Bayesian context clustering using cross validation for speech recognition," *IEICE Trans. Inf. & Syst.*, vol. E94-D, pp. 668–678, 2011.