FACE RECOGNITION BASED ON SEPARABLE LATTICE HMMS

Daisuke Kurata†, Yoshihiko Nankaku†, Keiichi Tokuda†,
Tadashi Kitamura†, and Zoubin Ghahramani‡

† Department of Computer Science and Engineering, Graduate School of Engineering
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
‡ Gatsby Computational Neuroscience Unit
University College London, Gower Street, London, WC1N 3AR, UK

ABSTRACT

In this paper, we propose separable lattice hidden Markov models, in which multiple hidden state sequences interact to model the observation on a lattice. The proposed model can be efficiently applied for modeling images, image sequences, 3-D object models and higher dimensional applications, due to the composite structure of Markov chains which reduces the complexity while retaining good properties for multi-dimensional data. In case of 2-D lattices, the proposed model performs an elastic matching in both horizontal and vertical directions; this makes it possible to model not only invariances to the size and location of an object but also nonlinear warping in each dimension. We present a training algorithm for separable lattice HMMs based on a variational approximation. Moreover, the deterministic annealing EM (DAEM) algorithm was applied to the variational algorithm for separable lattice HMMs. Face recognition experiments on the XM2VTS database show that the proposed model has good properties for face image modeling.

1. INTRODUCTION

Hidden Markov Models (HMMs) have been very successfully applied to numerous problems of one-dimensional data, particularly in speech recognition. Many attempts to use HMMs for modeling multi-dimensional data (e.g., images, image sequences, and 3-D object models) have also been presented. However, their extension to multi-dimensions leads to an exponential increase in the amount of computation required for the regular Baum-Welch and Viterbi algorithms. To reduce the computational complexity, we have to constrain the model structure by limiting the number of possible alignments and by assuming independence between hidden variables.

In previous work [1], planar hidden Markov models were developed, in which the probability of a particular state depends only on the state at adjacent observations in both horizontal and vertical directions. However, the computational complexity of planar HMMs is still exponential. On the other hand, a more restricted structure, pseudo 2-D HMMs (or embedded HMMs) have been proposed [2] and applied to many image applications. A pseudo 2-D HMM has the states of a superior HMM in the horizontal direction, called super-states, and each super-state has a one-dimensional HMM in the vertical direction instead of a probability density function. However, the state alignments of consecutive observation lines in the vertical direction are calculated independently of each other and this hypothesis does not always hold true in practice.

In this paper, we propose separable lattice hidden Markov models, in which multiple hidden state sequences interact to model the observation on a lattice. The proposed model can be efficiently applied for modeling images, image sequences, 3-D object models and higher dimensional applications. In the case of 2-D lattices, the proposed model performs an elastic matching in both horizontal and vertical directions; this makes it possible to model not only invariances to the size and location of an object but also nonlinear warping in each dimension. The parameters of a separable lattice HMM can be estimated via the expectation maximization (EM) algorithm for approximating the Maximum Likelihood (ML) estimate. Even though the constrained structure of separable lattice HMMs reduces the complexity of EM algorithm, the exact expectation step is still computationally intractable due to the dependency between Markov chains.

To derive a feasible problem, we applied the variational EM algorithm to separable lattice HMMs and present a structure approximation in which the hidden state sequences are decoupled. In order to represent complex data structures, a number of model structures on HMMs have been considered, e.g., factorial hidden Markov models [3]. These models have additional structural assumptions that are not available within the simple HMM framework, however the estimation of model parameters becomes computationally intractable. In recent years, variational methods have been used for approximating maximum likelihood estimation in such probabilistic graphical models [4], [5]. Variational methods approximate the posterior distribution over the hidden variables by a tractable distribution and provides a lower bound on the log-likelihood which is guaranteed to increase on each iteration.

The convergence point of the EM algorithm might depend on the initial model parameters. Moreover, in the variational EM algorithm for separable lattice HMMs, decoupled posterior distributions are updated individually based on the other distributions which are unreliable at an early stage of training. To overcome this problem, we applied the deterministic annealing EM (DAEM) algorithm [6] to the variational algorithm for separable lattice HMMs. We also show that the DAEM algorithm can improve the performance of separable lattice HMMs in face image recognition experiments using training data with variations in size and location.

In the following section, we define the structure of separable lattice HMMs. In Section 3, we present the exact EM algorithms for the proposed model and apply the variational approximation in Section 4. In Section 5, we present the application of the DAEM algorithm to the proposed model. In Section 6, we describe experiments of face recognition on the XM2VTS database and finally conclude in Section 7

2. SEPARABLE LATTICE HMMS

We define separable lattice hidden Markov models for modeling multi-dimensional data. The observations of M-dimensional data, e.g., pixel values of an image and image sequence, are assumed to be given on a M-dimensional lattice:

$$\boldsymbol{O} = \{\boldsymbol{O_t} | \boldsymbol{t} = (t^{(1)}, \ldots, t^{(m)}, \ldots, t^{(M)}) \in \boldsymbol{T}\} \qquad (1)$$

1-4244-0469-X/06/$20.00 ©2006 IEEE

where $\boldsymbol{t}$ denotes the coordinates of the lattice in M-dimensional space $\boldsymbol{T}$ and $t^{(m)} = 1, \ldots, T^{(m)}$ is the coordinate of the m-th dimension. The observation $\boldsymbol{O_t}$ is emitted from the state indicated by the hidden variable $\boldsymbol{S_t} \in \boldsymbol{K}$. The hidden variables $\boldsymbol{S_t} \in \boldsymbol{K}$ can take one of $K = \prod_m K^{(m)}$ states which assumed to be arranged on an M-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 the K states and each state corresponds to a segmented region in which the observation vectors are assumed to be generated from the same distribution. Since the observation $\boldsymbol{O_t}$ is dependent only on the state $\boldsymbol{S_t}$ as in ordinary HMMs, dependencies among hidden variables determine the properties and the modeling ability of multi-dimensional HMMs. To reduce the number of possible state sequences, we constrain the hidden variables to be composed of M Markov chains:

$$\boldsymbol{S} = \{\boldsymbol{S}^{(1)}, \ldots, \boldsymbol{S}^{(m)}, \ldots, \boldsymbol{S}^{(M)}\} \quad (2)$$

$$\boldsymbol{S}^{(m)} = \{S_1^{(m)}, \ldots, S_{t^{(m)}}^{(m)}, \ldots, S_{T^{(m)}}^{(m)}\} \quad (3)$$

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

$$\boldsymbol{S_t} = (S_{t^{(1)}}^{(1)}, S_{t^{(2)}}^{(2)}, \ldots, S_{t^{(M)}}^{(M)}) \quad (4)$$

This means that in the 2-D case, the segmented regions of observations are constrained to be rectangles and this allows an observation lattice to be elastic in both vertical and horizontal directions. Using this structure, the number of possible state sequences can be reduced from $\{\prod_m K^{(m)}\}^{\prod_m T^{(m)}}$ to $\prod_m \{K^{(m)}\}^{T^{(m)}}$.

The joint probability of observation vectors $\boldsymbol{O}$ and hidden variables $\boldsymbol{S}$ can be written as

$$\begin{aligned} & P(\boldsymbol{O}, \boldsymbol{S} | \Lambda) \\ &= P(\boldsymbol{O} | \boldsymbol{S}, \Lambda) \prod_{m=1}^{M} P(\boldsymbol{S}^{(m)} | \Lambda) \quad (5) \\ &= \prod_{\boldsymbol{t}} P(\boldsymbol{O_t} | \boldsymbol{S_t}, \Lambda) \times \\ & \quad \prod_{m=1}^{M} \left[P(S_1^{(m)} | \Lambda) \prod_{t^{(m)}=2}^{T^{(m)}} P(S_{t^{(m)}}^{(m)} | S_{t^{(m)}-1}^{(m)}, \Lambda) \right]. \quad (6) \end{aligned}$$

Model parameters of a separable lattice HMM are summarized as follows:

1) $\boldsymbol{\Pi}^{(m)} = \{\pi_i^{(m)} | 1 \le i \le K^{(m)}\}$: the initial state probability distribution, where $\pi_i^{(m)} = P(S_1^{(m)} = i | \Lambda)$ is the probability of state i at $t^{(m)} = 1$ in the m-th state sequence.

2) $\boldsymbol{A}^{(m)} = \{a_{ij}^{(m)} | 1 \le i \le K^{(m)}, 1 \le j \le K^{(m)}\}$: the transition probability matrix, where $a_{ij}^{(m)} = P(S_{t^{(m)}}^{(m)} = j | S_{t^{(m)}-1}^{(m)} = i | \Lambda)$ is the transition probability from state i to state j in the m-th state sequence.

3) $\boldsymbol{B} = \{b_{\boldsymbol{k}}(\boldsymbol{O_t}) | \boldsymbol{k} \in \boldsymbol{K}\}$: the output probability distributions, where $b_{\boldsymbol{k}}(\boldsymbol{O_t}) = P(\boldsymbol{O_t} | \boldsymbol{S_t} = \boldsymbol{k}, \Lambda)$ is the probability of observation vector $\boldsymbol{O_t}$ at the state $\boldsymbol{k}$ on the state lattice $\boldsymbol{K}$ and assumed to be a single Gaussian distribution: $P(\boldsymbol{O_t} | \boldsymbol{S_t} = \boldsymbol{k}, \Lambda) = \mathcal{N}(\boldsymbol{O_t}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$ where $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ are the mean vector and the covariance matrix, respectively. Note that separable lattice HMMs have $\prod_m K^{(m)}$ Gaussians directly as the parameters on a lattice unlike factorial HMMs[1].

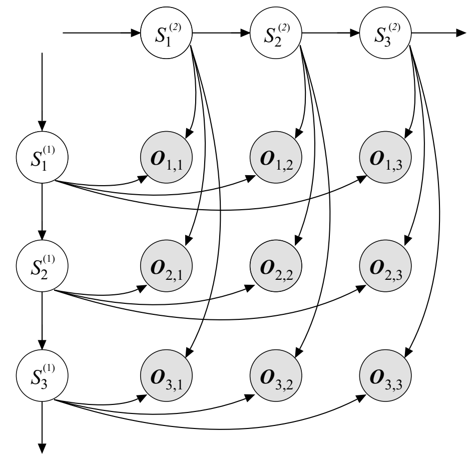


Fig. 1. Separable lattice HMM with two Markov chains

Using shorthand notation, a separable lattice HMM is defined as $\Lambda = \{\Lambda^{(1)}, \ldots, \Lambda^{(M)}, \boldsymbol{B}\}$, $\Lambda^{(m)} = \{\boldsymbol{\Pi}^{(m)}, \boldsymbol{A}^{(m)}\}$.

The separable 2-D lattice HMMs can be applied to image modeling and perform an elastic matching in both horizontal and vertical directions by assuming the transition probabilities with left-to-right and top-to-bottom topologies. Although the structure of the proposed model cannot represent rotations of images, it is still useful for image detection and the framework makes it possible to achieve size- and location-invariant image recognition. Furthermore, the proposed model can be used for 3-D and higher dimensional applications, e.g., image sequences, 3-D object models, etc., due to the composite structure which reduces the complexity of the algorithm while retaining the good properties for modeling multi-dimensional data.

Figure 1 shows the graphical representations of separable lattice HMMs for two-dimensional data. Although the embedded HMMs can be assumed to have different prior transition distributions for each super states, this figure shows the models which have the same transition distribution for all embedded state sequences, i.e., there is no arc from a super state to embedded states. The main difference between the proposed model and the embedded HMMs is that separable lattice HMMs have a symmetric structure in vertical and horizontal directions. Therefore, there is no need to determine which direction of two-dimensional data should be modeled as the super states or the embedded states. If the hidden variables of the embedded states are also shared for all observation sequences, an embedded HMM becomes equivalent to a separable 2-D lattice HMM. In the embedded HMMs, the exact EM algorithm can be performed in practice, because the state transitions of an embedded state sequence depend only on the corresponding super state. However, in the separable 2-D lattice HMMs, the state transitions of one direction depend on the all the hidden variables of the other direction; therefore the exact EM algorithm becomes infeasible.

[1]Factorial HMMs have $\sum_m K^{(m)}$ Gaussians along with Markov chains and they contribute linearly to the output probability distributions.

3. EM ALGORITHM

The parameters of separable lattice HMMs can be estimated via the expectation maximization (EM) algorithm which is an iterative procedure for approximating the Maximum Likelihood (ML) estimate. This procedure maximizes the expectation of the complete data log-likelihood so called $\mathcal{Q}$-function:

$$\mathcal{Q}(\Lambda, \Lambda') = \sum_{\boldsymbol{S}} P(\boldsymbol{S} \mid \boldsymbol{O}, \Lambda) \ln P(\boldsymbol{S}, \boldsymbol{O} \mid \Lambda') \tag{7}$$

The likelihood of the training data is guaranteed to increase by increasing the value of the $\mathcal{Q}$-function:

$$\mathcal{Q}(\Lambda, \Lambda') \geq \mathcal{Q}(\Lambda, \Lambda) \quad \Rightarrow \quad P(\boldsymbol{O} \mid \Lambda') \geq P(\boldsymbol{O} \mid \Lambda) \tag{8}$$

The EM algorithm starts with some initial model parameters and iterates between the following two steps:

$$\begin{aligned} &\text{(E step)} &:& \quad \text{compute } \mathcal{Q}(\Lambda^{(k)}, \Lambda) \\ &\text{(M step)} &:& \quad \Lambda^{(k+1)} = \arg\max_{\Lambda} \mathcal{Q}(\Lambda^{(k)}, \Lambda) \end{aligned}$$

where k denotes the iteration number. The E-step computes the posterior probabilities over the hidden states while keeping model parameters Λ fixed to current values. The M-step uses these probabilities to calculate the expected log-likelihood of the training data as function of the parameters and maximize the $\mathcal{Q}$-function with respect to model parameters Λ. In this procedure, each step increases the value of the $\mathcal{Q}$-function; hence the likelihood of the training data is also guaranteed to increase or remain unchanged on each iteration. The complexity of the exact E-step can be reduced to $O(\{K^{(n)}\}^2 T^{(n)} \prod_{m \neq n} \{K^{(m)}\}^{T^{(m)}})$, however, it is still infeasible.

4. VARIATIONAL EM ALGORITHM

Variational methods have been used for approximate maximum likelihood estimation in probabilistic graphical models with hidden variables. We present a structure approximation in which the hidden state sequences are decoupled. The variational methods approximate the posterior distribution over the hidden variables by a tractable distribution. Any distribution $Q(\boldsymbol{S})$[2] over the hidden variables defines a lower bound on the log-likelihood

$$\begin{aligned} \ln P(\boldsymbol{O}|\Lambda) &= \ln \sum_{\boldsymbol{S}} P(\boldsymbol{S}, \boldsymbol{O}|\Lambda) \\ &= \ln \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \frac{P(\boldsymbol{S}, \boldsymbol{O}|\Lambda)}{Q(\boldsymbol{S})} \\ &\geq \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \ln \frac{P(\boldsymbol{S}, \boldsymbol{O}|\Lambda)}{Q(\boldsymbol{S})} \\ &= \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \ln P(\boldsymbol{S}, \boldsymbol{O}|\Lambda) - \sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \ln Q(\boldsymbol{S}) \\ &= \mathcal{F}(Q, \Lambda) \end{aligned} \tag{9}$$

where we have applied Jensen's inequality. The difference between $\ln P(\boldsymbol{O}|\Lambda)$ and $\mathcal{F}$ is given by the Kullback-Leibler divergence between $Q(\boldsymbol{S})$ and the posterior distribution of the hidden variables $P(\boldsymbol{S}|\boldsymbol{O}, \Lambda)$. Since the true log-likelihood $\ln P(\boldsymbol{O}|\Lambda)$ is independent of $Q(\boldsymbol{S})$, maximizing the lower bound $\mathcal{F}$ is equivalent to minimizing the Kullback-Leibler divergence. If we allow $Q(\boldsymbol{S})$ to have complete flexibility then we see that the optimal $Q(\boldsymbol{S})$ distribution is given by the true posterior $P(\boldsymbol{S}|\boldsymbol{O}, \Lambda)$, in the case where the KL divergence is zero and the bound becomes exact. In order to yield a tractable algorithm, it is necessary to consider a more restricted structure of $Q(\boldsymbol{S})$ distributions. Given the structure, the parameters of $Q(\boldsymbol{S})$ are varied so as to obtain the tightest possible bound, which maximizes $\mathcal{F}$.

The variational EM algorithm iteratively maximizes $\mathcal{F}$ with respect to the Q and Λ holding the other parameters fixed:

$$\begin{aligned} &\text{(E step)} &:& \quad Q^{(k+1)} = \arg\max_{Q \in C} \mathcal{F}(Q, \Lambda^{(k)}) \\ &\text{(M step)} &:& \quad \Lambda^{(k+1)} = \arg\max_{\Lambda} \mathcal{F}(Q^{(k+1)}, \Lambda) \end{aligned}$$

where C is the set of constrained distributions. The maximum in the M-step is obtained by maximizing the term $\sum_{\boldsymbol{S}} Q(\boldsymbol{S}) \ln P(\boldsymbol{S}, \boldsymbol{O}|\Lambda)$ in (9), since the entropy of $Q(\boldsymbol{S})$ does not depend on model parameters Λ. In this procedure, the lower bound $\mathcal{F}$ is guaranteed to increase instead of the value of the $\mathcal{Q}$-function.

The complexity and the approximation property of the variational EM algorithm are dependent on a constraint to the posterior distribution $Q(\boldsymbol{S})$ and it should be determined for each structure of graphical models. Here we consider a constrained family of variational distributions for separable lattice HMMs by assuming that $Q(\boldsymbol{S})$ factorizes over subset $\boldsymbol{S}^{(m)}$ of the variables in $\boldsymbol{S}$, so that

$$Q(\boldsymbol{S}) = \prod_{m=1}^{M} Q(\boldsymbol{S}^{(m)}) \tag{10}$$

where $\sum_{\boldsymbol{S}^{(m)}} Q(\boldsymbol{S}^{(m)}) = 1$, $m = 1, \ldots, M$. The complexity of E-step with the variational approximation becomes $O(M \prod_m K^{(m)} T^{(m)})$. Note that the computational cost can be significantly reduced from the exact EM algorithm to polynomial time complexity.

5. VARIATIONAL DAEM ALGORITHM

The EM algorithm has the problem that the solution converges to a local optimum and the convergence point depends on the initial model parameters. In the variational EM algorithm for separable lattice HMMs, the decoupled posterior distributions are updated individually based no only on the initial model parameters but also on the other distributions, both of which are unreliable at an early stage of training. To avoid this problem, we apply the DAEM algorithm to the algorithm derived in the previous section and show that the expectations with respect to the decoupled posterior distributions for the DAEM can also be calculated by the forward-backward procedure.

In the DAEM algorithm, the problem of maximizing the log-likelihood is reformulated as minimizing the thermodynamic free energy defined as

$$\mathcal{F}_\beta = -\frac{1}{\beta} \ln \sum_{\boldsymbol{S}} P(\boldsymbol{S}, \boldsymbol{O}|\Lambda)^\beta \tag{11}$$

where $1/\beta$ called the "temperature" and maximizing $\mathcal{F}_\beta(Q_\beta, \Lambda)$ with a fixed temperature can be interpreted as the approach to thermodynamic equilibrium. In the algorithm, the temperature is gradually decreased and the function is deterministically optimized at each temperature. The procedure of the DAEM algorithm can be summarized as follows:

[2]The notation of distribution $Q(\boldsymbol{S})$ over states is distinct from the notation of $\mathcal{Q}$-function $\mathcal{Q}(\Lambda, \Lambda')$.

1 Give an initial model and set $\beta = \beta_{min}$.

2 Iterate EM-steps with β fixed until $\mathcal{F}_\beta$ converged:

(E step) : $\arg\max_{Q_\beta \in C} \mathcal{F}_\beta(Q_\beta, \Lambda^{(k)})$

(M step) : $\Lambda^{(k+1)} = \arg\max_{\Lambda} \mathcal{F}_\beta(Q_\beta^{(k+1)}, \Lambda)$

3 Increase β.

4 If $\beta > 1$, stop the procedure. Otherwise go to step 2.

where $1/\beta_{min}$ is an initial temperature and should be chosen as a high enough value that the EM-steps can achieve a single global maximum of $\mathcal{F}_\beta$. At the initial temperature, the entropy of $Q_\beta(\boldsymbol{S})$ is intended to be maximized rather than the $\mathcal{Q}$-function (the first term of equation (11)); therefore $Q_\beta(\boldsymbol{S})$ takes a form nearly uniform distribution. While the temperature is decreasing, the form of $Q_\beta(\boldsymbol{S})$ changes from uniform to the original posterior and at the final temperature $1/\beta = 1$, the negative free energy $\mathcal{F}_\beta$ becomes equal to the lower bound $\mathcal{F}$, accordingly the DAEM algorithm agrees with the original EM algorithm.

6. EXPERIMENTS

In order to demonstrate the modeling ability of separable lattice HMMs, face recognition experiments on the XM2VTS database [7] were conducted. We prepared 8 images of 100 subjects; 7 images are used for training and 1 image for testing. Face images of grayscale 64×64 pixels were extracted from the original images. In this process, two sets of data were prepared:

- "dataset1": the size- and location-normalized data.
- "dataset2": the data with size and location variations. The sizes and locations were randomly generated by Gaussian distributions almost within the location shift of 40×20 pixels from the center point and the range of size from 500×500 to 600×600 with fixed aspect.

The pixel intensity of images were used as the feature vectors, and modeled by a separable 2-D lattice HMM with 32×32 states, single-Gaussian distributions. The state sequences $\boldsymbol{S}^{(1)}$ and $\boldsymbol{S}^{(2)}$ correspond to horizontal and vertical directions, respectively. The transition probabilities for each state sequence are assumed to be a left-to-right and top-to-bottom no skip topology. The initial model was constructed from a linear segmentation for each observation lattice.

To confirm the effectiveness of separable lattice HMMs, we constructed other models summarized as follows:

- "Gauss": 64×64 dimensional Gaussian distributions without state transitions.
- "Emb-V," "Emb-H": the embedded HMMs which have superstates in the vertical and horizontal direction, respectively.
- "SL-EM, " "SL-DA": the separable 2-D lattice HMMs which are trained with variational EM algorithm and variational DAEM algorithm, respectively.

Figure 2 shows the visualized mean vectors of Gaussian distributions and the recognition rates for each model. Comparing "Gauss" between two datasets, the mean vector of "dataset2" becomes blurred than that of "dataset1", and the recognition performance was degraded by the variations. Since "Emb-V" and "Emb-H" ignore the correlation between embedded state sequences, the geometric continuity was not preserved in the mean vectors. On the other hand, "SL-EM" achieved better results than "Emb-V" and "Emb-H" in the both datasets. Although the mean vector of "SL-EM" keeps the continuity

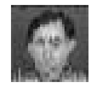
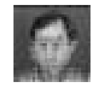
dataset	Gauss	Emb-V	Emb-H	SL-EM	SL-DA
dataset1					
	60%	69%	62%	80%	**84%**
dataset2					
	17%	56%	58%	70%	**76%**

Fig. 2. The mean vectors of the output probability distributions and the recognition rates for each model

of horizontal and vertical directions in "dataset1," that of "dataset2" is distorted around the center of the face after the re-estimation because of the inaccuracy of the initial estimate. However, "SL-DA" improves the performance of "SL-EM" in the both datasets due to reducing the dependency on initial models.

7. CONCLUSION

We propose separable lattice hidden Markov models for modeling the observations on a multi-dimensional lattice and presented a training algorithm based on a variational approximation. The face recognition experiments were performed on the XM2VTS database. In the experiments, the proposed model achieves better results than the embedded HMMs. This result suggests that the separable 2-D lattice HMMs are useful for applications of image detection and recognition. Furthermore, the DAEM algorithm improves the performance of the proposed model. Extensions to more flexible models will be future works.

ACKNOWLEDGMENTS
We would like to thank Tetsuya Nunome for support of this research.

8. REFERENCES

[1] E. Levin, and R. Pieraccini, "Dynamic Planar Warping for optical Character Recognition," Proc. ICASSP, vol.3, pp.149-152, 1992.

[2] S. Kuo, and O.E. Agazzi, "Keyword Spotting in Poorly Printed Documents Using Pseudo 2-D Hidden Markov Models," IEEE Trans. Pattern Analysis and Machine Intelligence, vol.16, no.8, pp.842-848, 1994.

[3] Z. Ghahramani, and M.I. Jordan, "Factorial Hidden Markov Models," Machine Learning, vol.29, pp.245-273, 1997.

[4] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, "An introduction to Variational Methods for Graphical Models," Machine Learning, vol.37, pp.183-233, 1999.

[5] Z. Ghahramani, "On Structured Variational Approximations," University of Toronto Technical Report, CRG-TR-97-1, 1997, revised 2002.

[6] N. Ueda, and R. Nakano, "Deterministic Annealing EM Algorithm," Neural Networks, vol.11, no.2, pp.271-282, 1998.

[7] K. Messer, J. Mates, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," Audio- and Video-Based Biometric Person Authentication, pp.72-77, 1999.