SEPARABLE LATTICE 2-D HMMS INTRODUCING STATE DURATION CONTROL FOR RECOGNITION OF IMAGES WITH VARIOUS VARIATIONS

Takaya Makino, Shinji Takaki, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

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

ABSTRACT

In this paper, an extension of separable lattice HMMs (SL-HMM) is described that introduces state duration control for dealing with images with various variations. SL-HMM are generative models that have size and location invariances based on state transition of HMMs. An extended model that has the structure of hidden semi-Markov models (HSMMs) in which the state duration probability is explicitly modeled by parametric distributions is also proposed. However, in this model, each state duration in a Markov chain is independent. It is supposed that each state duration should have a correlation. Therefore, in this paper, we propose a novel model that solves this problem by introducing variables representing the correlation among the state durations. Face recognition experiments show that the proposed model improved the recognition performance for images with size, locational, and rotational variations.

Index Terms— face recognition, separable lattice HMMs, hidden semi Markov models, state duration control

1. INTRODUCTION

In image recognition, statistical approaches have been successfully applied, e.g., the eigenface [1] and subspace methods [2]. These methods can model images effectively. However, if images contain geometric variations such as size, location, and rotation, the recognition performance is degraded. Although it is possible to normalize images by using heuristic normalization techniques in the preprocess part of the classification, task dependent normalization techniques need to be developed for each application. Furthermore, the final objective of image recognition is not to accurately normalize images for human perception but to achieve a better recognition performance. Therefore, it is natural to use the same criterion for both training classifiers and normalization. This means that the normalization process should be integrated into classifiers.

Hidden Markov model (HMM) based techniques have been proposed as such kind of approaches for geometric variations. The geometric matching between input images and models is represented by discrete hidden variables, and the normalization process is included in the calculation of probabilities. However, the extension of HMMs to multi-dimensions generally leads to an exponential increase in the amount of computation for its training algorithm. To reduce the computational complexity while retaining the good properties for modeling multi-dimensional data, separable lattice Hidden Markov models (SL-HMM) were proposed [4]. Two dimensional structures are usually used for the image recognition, though this model can represent three or larger dimensional structures. SL-HMM can perform an elastic matching in both horizontal and vertical directions. This property makes it possible to model not only invariances to the size and location of an object but also nonlinear warping in each dimension. However, SL-HMM cannot deal with rotational variations and local deformation. To deal with such variations, extended separable lattice 2-D HMMs (ESL-HMM) that have state sequences corresponding to all rows and columns of an input image were proposed [5]. This structure enables ESL-HMM to align for images every row and column independently. Therefore, ESL-HMM can obtain more complicated state alignments such as rotational variations than can SL-HMM. However, geometric continuities in the images are not modeled in this structure. Geometric continuities are needed to be retained when the models of images are constructed.

Separable lattice hidden semi-Markov models (SL-HSMM) were proposed in order to model state duration [6]. The state duration probability of HMMs exponentially decreases as duration increases, therefore, it may not be appropriate for modeling image variations accurately. To overcome this problem, the structure of hidden semi-Markov models (HSMMs) in which the state duration probability is explicitly modeled by parametric distributions is used [7]. SL-HSMM are statistical models in which the state duration models are integrated into the structure of SL-HMM. However, state durations are independently generated from duration models. To model images with variations such as size and location, we should consider modeling images, including the correlation among the state durations.

In this paper, we propose novel models that solve the above problem by introducing hidden variables representing the correlation among the state durations into SL-HSMM. By adopting common linear regression to all state duration models, all state durations of whole lattice can have correlation. As a result, geometric continuities in the images are retained. Furthermore, we also propose models that introduce this state duration control into ESL-HMM, called ESL-HSMM-LR. Since ESL-HMM cannot retain the geometric continuities in the images due to aligning for images every row and column independently, it is expected that state duration control is effective. That is, ESL-HSMM-LR can align for images every row and column independently, retaining the geometric continuities of images. Therefore, we expect that ESL-HSMM-LR are suitable models for various variations.

The rest of the paper is organized as follows. In Section 2, SL-HMM and ESL-HMM are described briefly. In Section 3, the structure of the model with state duration control is defined. In Section 4, face recognition experiments on the XM2VTS database are described, and finally, the paper is concluded in Section 5.

2. SEPARABLE LATTICE 2-D HMMS

2.1. Separable lattice 2-D HMMs

Separable lattice 2-D HMMs are defined for modeling 2-D data [4]. Observations of 2-D data are assumed to be given on a 2-D lattice:

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

where $\boldsymbol{t}$ denotes the coordinates of the lattice in 2-D space $\boldsymbol{T}$ and $t^{(m)} = 1, \ldots, T^{(m)}$ is the coordinate of the m-th dimension. In SL-

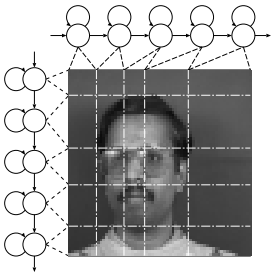

Fig. 1. Example of state alignments of SL-HMM

HMM an observation, $\boldsymbol{O_t}$, is emitted from a state indicated by a hidden variable, $\boldsymbol{S_t} \in \boldsymbol{K}$. This hidden variable $\boldsymbol{S_t} \in \boldsymbol{K}$ can take one of $K = K^{(1)}K^{(2)}$ states, which are assumed to be arranged on a 2-D state lattice, $\boldsymbol{K} = \{1, \ldots, K\}$. In other words, a set of hidden variables, $\{\boldsymbol{S_t} \mid \boldsymbol{t} \in \boldsymbol{T}\}$, represents a segmentation of observations into $\boldsymbol{K}$ states, and each state corresponds to a segmented region in which the observation is assumed to be generated from the same distribution. Since an observation, $\boldsymbol{O_t}$, is only dependent on a state, $\boldsymbol{S_t}$, as in ordinary HMMs, the dependencies of hidden variables determine the properties and modeling ability of multi-dimensional HMMs. In SL-HMM, to reduce the number of possible state sequences, hidden variables are constrained to be composed of two Markov chains:

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

$$\boldsymbol{S}^{(m)} = \{S^{(m)}(1), \ldots, S^{(m)}(t^{(m)}), \ldots, S^{(m)}(T^{(m)})\} \quad (3)$$

where $\boldsymbol{S}^{(m)}$ is the Markov chain along with the m-th dimension and $S^{(m)}(t^{(m)}) \in \{1, \ldots, K^{(m)}\}$ is the state of the $t^{(m)}$-th coordinate. The composite structure of hidden variables in SL-HMM is defined as the product of hidden state sequences: $\boldsymbol{S_t} = (S^{(1)}(t^{(1)}), S^{(2)}(t^{(2)}))$. This means that the segmented regions of observations are constrained to rectangles. Therefore, SL-HMM allow an observation lattice to be elastic both vertically and horizontally. The joint probability of observation $\boldsymbol{O}$ and hidden variables $\boldsymbol{S}$ can be written as:

$$\begin{aligned} P(\boldsymbol{O}, \boldsymbol{S} \mid \Lambda) &= P(\boldsymbol{S} \mid \Lambda) P(\boldsymbol{O} \mid \boldsymbol{S}, \Lambda) \\ &= P(\boldsymbol{S}^{(1)} \mid \Lambda) P(\boldsymbol{S}^{(2)} \mid \Lambda) \prod_{\boldsymbol{t}} P(\boldsymbol{O_t} \mid \boldsymbol{S_t}, \Lambda) \end{aligned} \quad (4)$$

where Λ is a set of model parameters. Figure 1 shows an example of the state alignments of SL-HMM.

2.2. Extended separable lattice 2-D HMMs

SL-HMM can model invariance to variations in the size and location of an object. However, since SL-HMM have only one state sequence in each dimension, segmented regions of observations are constrained to rectangles, i.e., SL-HMM cannot deal with rotational variations and local deformations. To deal with these, extended separable lattice 2-D HMMs were proposed [5]. ESL-HMM have state sequences corresponding to all rows and columns of an input image. Hidden variables representing state sequences are defined as:

$$\boldsymbol{S} = \{\boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)}\} \quad (5)$$

$$\boldsymbol{S}^{(m)} = \{\boldsymbol{S}^{(m)}_1, \ldots, \boldsymbol{S}^{(m)}_{t^{(n)}}, \ldots, \boldsymbol{S}^{(m)}_{T^{(n)}}\} \quad (6)$$

$$\boldsymbol{S}^{(m)}_{t^{(n)}} = \{S^{(m)}_{t^{(n)}}(1), \ldots, S^{(m)}_{t^{(n)}}(t^{(m)}), \ldots, S^{(m)}_{t^{(n)}}(T^{(m)})\} \quad (7)$$

$$S^{(m)}_{t^{(n)}}(t^{(m)}) \in \{1, 2, \ldots, K^{(m)}\} \quad (8)$$

where n is a variable representing a dimension that differs from m. The composite structure of hidden variables in ESL-HMM is defined

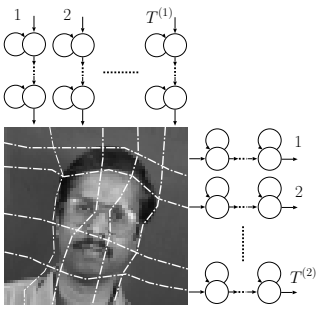

Fig. 2. Example of state alignments of ESL-HMM

as the product of hidden sequences: $\boldsymbol{S_t} = (S^{(1)}_{t^{(2)}}(t^{(1)}), S^{(2)}_{t^{(1)}}(t^{(2)}))$. The joint probability of observation $\boldsymbol{O}$ and hidden variables $\boldsymbol{S}$ can be written as:

$$\begin{aligned} &P(\boldsymbol{O}, \boldsymbol{S} \mid \Lambda) \\ &= \prod_{t^{(2)}} P(\boldsymbol{S}^{(1)}_{t^{(2)}} \mid \Lambda) \prod_{t^{(1)}} P(\boldsymbol{S}^{(2)}_{t^{(1)}} \mid \Lambda) \prod_{\boldsymbol{t}} P(\boldsymbol{O_t} \mid \boldsymbol{S_t}, \Lambda) \end{aligned} \quad (9)$$

Figure 2 shows the structure of ESL-HMM. ESL-HMM can obtain more complicated state alignments because this structure enables them to align for images every row and column independently.

3. SEPARABLE LATTICE 2-D HMMS INTRODUCING STATE DURATION CONTROL

3.1. Separable lattice 2-D HMMs with explicit state duration modeling

The state duration probability of HMMs exponentially decreases as duration increases because it depends only on the transition probability in the HMMs. The probability of d consecutive observations in the i-th state of HMMs is given by the probability of taking the self-loop a_{ii} at the i-th state for d times as:

$$P_i(d) = a_{ii}^{d-1}(1 - a_{ii}) \quad (10)$$

Therefore, HMMs cannot represent state duration distributions that peak at the middle accurately. To overcome this problem, hidden semi-Markov models (HSMMs) were proposed as models in which the state duration probability is explicitly modeled by parametric distributions [7]. This model can output the state duration probability in accordance with the state duration distributions. That is, this model has two parameters: the mean η_i and the variance σ_i of the state duration distributions. Now, the state duration distribution is modeled by Gaussian distributions as:

$$P_i(d) = \mathcal{N}(d \mid \eta_i, \sigma_i) \quad (11)$$

SL-HSMM were proposed to represent state durations that SL-HMM cannot represent appropriately due to the same reason as HMMs [6]. Figure 3 shows the structure of an SL-HSMM. By building the model structure of HSMMs into separable lattice 2-D HMMs, we can construct more suitable models for images with large variations. However, state durations are independently generated from duration models. To model images with variations such as size and location, we should consider modeling images, including the correlation among the state durations. That is to say, the ratio of each part in images is needed to be retained.

3.2. Separable lattice 2-D HMMs introducing state duration control

To solve the above problem, this paper proposes a novel model that introduces variables representing the correlation among the state durations into SL-HSMM. In this paper, the proposed model named SL-HSMM-LR (SL-HSMM with linear regression)), because the correlation of all state durations is represented by linear regression of the mean of duration distributions. The model can keep the ratio of state durations in each image: that is typical duration variations in size and location changes, and the geometric continuities of image mapping will be retained.

We represent the state sequence in the HMMs by using variables that show the state duration explicitly because HSMMs contain parameters with respect to the state duration distribution. In SL-HSMM-LR, by considering the independence of state duration and state sequence, each variable is defined as:

$$\boldsymbol{S}^{(m)} = \left\{ L^{(m)}, \boldsymbol{q}^{(m)}, \boldsymbol{d}^{(m)} \right\} \tag{12}$$

$$\boldsymbol{q}^{(m)} = (q^{(m)}(1), \ldots, q^{(m)}(L^{(m)})) \tag{13}$$

$$\boldsymbol{d}^{(m)} = (d^{(m)}(1), \ldots, d^{(m)}(L^{(m)})) \tag{14}$$

where $\boldsymbol{q}^{(m)}$ is a state number sequence that does not include continuation, $L^{(m)}$ is the length of the state number sequence, and $\boldsymbol{d}^{(m)}$ is the state of the duration sequence. Figure 4 shows the model structure of SL-HSMM-LR. In SL-HSMM-LR, the likelihood function is defined as:

$$\begin{aligned} & P(\boldsymbol{O}, \boldsymbol{\tau}, \boldsymbol{L}, \boldsymbol{q}, \boldsymbol{d} \mid \Lambda) \\ &= \int P(\boldsymbol{\tau}) P(\boldsymbol{L}, \boldsymbol{q} \mid \Lambda) P(\boldsymbol{d} \mid \boldsymbol{L}, \boldsymbol{q}, \boldsymbol{\tau}, \Lambda) P(\boldsymbol{O} \mid \boldsymbol{L}, \boldsymbol{q}, \boldsymbol{d}, \Lambda) d\boldsymbol{\tau} \\ &= \prod_{m=1,2} \left\{ \int P(\boldsymbol{\tau}^{(m)}) P(L^{(m)}, \boldsymbol{q}^{(m)} \mid \Lambda) \right. \\ &\quad \left. \times P(\boldsymbol{d}^{(m)} \mid L^{(m)}, \boldsymbol{q}^{(m)}, \boldsymbol{\tau}^{(m)}, \Lambda) d\boldsymbol{\tau}^{(m)} \right\} \\ &\quad \times \prod_{\boldsymbol{t}} P(\boldsymbol{O}_{\boldsymbol{t}} \mid \boldsymbol{L}, \boldsymbol{q}, \boldsymbol{d}, \Lambda) \end{aligned} \tag{15}$$

Each variable is constrained as:

$$\boldsymbol{L} = \left\{ L^{(1)}, L^{(2)} \right\}, \boldsymbol{q} = \left\{ \boldsymbol{q}^{(1)}, \boldsymbol{q}^{(2)} \right\}, \boldsymbol{d} = \left\{ \boldsymbol{d}^{(1)}, \boldsymbol{d}^{(2)} \right\} \tag{16}$$

SL-HSMM-LR have the linear regression coefficients $a^{(m)}$, $b^{(m)}$ generated from $\boldsymbol{\tau}^{(m)}$ for every dimension and every image, and the state duration probability is defined as:

$$P_i^{(m)}(d) \quad = \quad \mathcal{N}(d^{(m)} \mid a^{(m)} \eta_i^{(m)} + b^{(m)}, \ \sigma_i^{(m)}) \tag{17}$$

Equation (17) means that the mean of the state duration probability distribution is scaled and shifted by the parameters $a^{(m)}$ and $b^{(m)}$. By having such a probable structure, the state duration that is in accordance with the overall length of the output signal for every data is outputted. Moreover, the linear regression coefficients $a^{(m)}$ and $b^{(m)}$ are the random variables, and it is assumed that $P(\boldsymbol{\tau}^{(m)})$ is output from the prior distribution acquired from certain prior information.

The model parameters of SL-HSMM-LR can be re-estimated using the EM algorithm, which is an iterative procedure for approximating the maximum likelihood (ML) estimates. However, it is practically difficult to conduct, because the computational complexity of E-step becomes $O(\{K^{(1)}K^{(2)}\}^{T^{(1)}T^{(2)}})$ due to counting

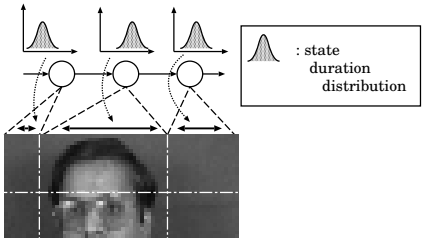

Fig. 3. Example of state alignments of SL-HSMM

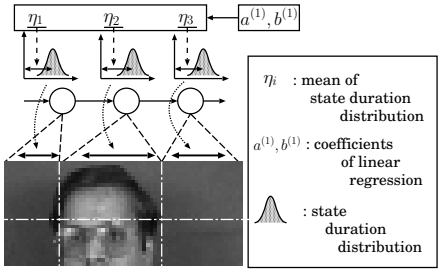

Fig. 4. Example of state alignments of SL-HSMM-LR

the all possible state sequences. Therefore, we apply the variational EM (VEM) algorithm [8] as in the training of SL-HMM and ESL-HMM. By applying the generalized forward-backward algorithm, the complexity of E-step becomes $O(\{K^{(1)}\}^2\{T^{(1)}\}^2)$ or $O(\{K^{(2)}\}^2\{T^{(2)}\}^2)$ while that of SL-HMM and ESL-HMM is $O(K^{(1)}K^{(2)}T^{(1)}T^{(2)})$.

Controlling the state duration by introducing the structure of HSMMs and linear regression can be applied not only to SL-HMM but also to ESL-HMM (ESL-HSMM-LR). Since the structure of ESL-HMM holds two or more Markov chains, it can flexibly align for images with variations. However, ESL-HMM cannot retain the geometric continuities in the images. Therefore, by introducing the control of the state duration to ESL-HMM, geometric continuities in the images are modeled, aligning flexibly. With this structure, it is expected that a suitable model for images with various variations can be constructed. The complexity of the ESL-HSMM-LR is $O(\{K^{(1)}\}^2\{T^{(1)}\}^2T^{(2)})$ or $O(\{K^{(2)}\}^2\{T^{(2)}\}^2T^{(1)})$.

4. EXPERIMENTS

To evaluate the effectiveness of the proposed models, we conducted image recognition experiments. Now, we chose the XM2VTS database [11] that is a kind of a complicated image. Table 1 shows the experimental conditions. In these experiments, five models were compared: SL-HMM, ESL-HMM, SL-HSMM, SL-HSMM with linear regression for state duration control (SL-HSMM-LR), and ESL-HMM with linear regression for state duration (ESL-HSMM-LR). To evaluate images that included some variations, two datasets were prepared. Table 2 shows the details of the datasets. We used two kinds of feature vectors as observations: the pixel values and 2-D discrete cosine transform (2-D DCT) coefficients of images. Two-dimensional DCT coefficients were calculated with the following method. Input images were scanned by using a 12×12 window that shifted by one pixel from left-to-right and top-to-bottom, and the 2-D DCT coefficients were calculated from each scanned block. Only 4×4 coefficients of the lowest frequencies in the 2-D DCT domain were used as feature vectors. Figures 5, 6, and 7 show the recognition rates on "dataset 1" (pixel values), "dataset 2" (pixel

Table 1. Experimental conditions

original image size	720×576
extracted image size	64×64, grayscale
training data	6 images per person $\times$ 100 subjects
test data	2 images per person $\times$ 100 subjects
method	SL-HMM , SL-HSMM, SL-HSMM-LR, ESL-HMM, ESL-HSMM-LR
state number	8×8, 16×16, 24×24, 32×32, 40×40

Table 2. Experimental datasets

dataset 1	images with variations normalized
dataset 2	locational variation: within 40×20 from center
	size variation: $550 \times 550 \sim 700 \times 700$
	rotational variations: within $-10 \sim 10$ degrees

values), and "dataset 2" (DCT coefficients), respectively.

Figure 5 shows that SL-HSMM-LR improved recognition rates from the SL-HMM and SL-HSMM. This result means that the state duration control is effective. However, the recognition rate of SL-HSMM-LR was inferior to that of the SL-HSMM when the number of states was 40×40. The reason is thought to be that if there is a high number of states, the ratio of each state duration becomes uniform to some extent. Therefore, linear regression does not have a sufficient effect. Next, the recognition rate of the ESL-HMM decreased relatively. The reason is thought to be that since the ESL-HMM could obtain more complicated state alignments than could the SL-HMM, geometric continuities were not modeled, and state alignments were over-fitted. The recognition rate was lower than that of the SL-HMM and SL-HSMM-LR because of over-fitting of the state alignments since ESL-HSMM-LR strongly reflects the structure of the ESL-HMM. However, ESL-HSMM-LR improved the recognition rate from the ESL-HMM because of the effect of state duration control. The recognition rates of ESL-HSMM-LR in Figure 6 were higher than those in Figure 5. The reason is that over-fitting of the state alignments was alleviated when images included the variations. Therefore, it is necessary to consider a method that solves the over-fitting of the state alignments in order to construct a more robust model.

In Figure 6, there were few differences between the SL-HMM, SL-HSMM, and SL-HSMM-LR. This is because the images of "dataset 2" included the rotational variations. That is, it was difficult for them to align for images successfully because their segmented regions of observations were constrained to rectangles. The recognition rate of ESL-HSMM-LR was the highest of all models when the number of states was 16×16. This means that ESL-HSMM-LR was robust for images with various variations. The reason is thought to be that ESL-HSMM-LR can align for images with variations flexibly, retaining the ratio of each part in images.

In Figure 7, the recognition rates of the ESL-HMM and ESL-HSMM-LR went up relatively compared with those in Figure 6. This is because 2-D DCT coefficients included shape information on all block regions. Because this prevented the state alignments from over-fitting, the recognition rates of the ESL-HMM and ESL-HSMM-LR in particular were improved. However, if 2-D DCT coefficients are used, the recognition rate is strongly influenced by the output probability. Therefore, since state duration control does not have a sufficient effect, there were no large differences between the ESL-HMM and ESL-HSMM-LR. However, the recognition rate of ESL-HSMM-LR was also the highest of all models. These results suggested that the ESL-HSMM-LR can recognize images with various variances more accurately. As a future work, it is necessary to examine results of other data sets and evaluation methods because we can not judge it whether the topology of the image is really kept by these results.

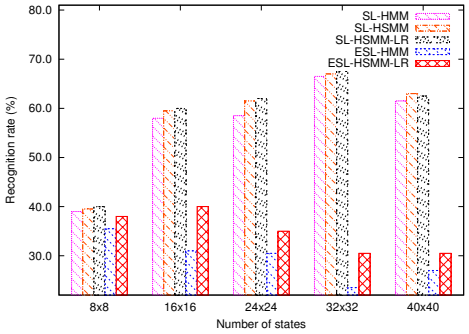

Fig. 5. Recognition rates on "dataset 1" (pixel values)

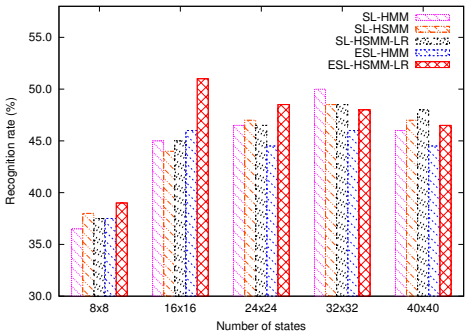

Fig. 6. Recognition rates on "dataset 2" (pixel values)

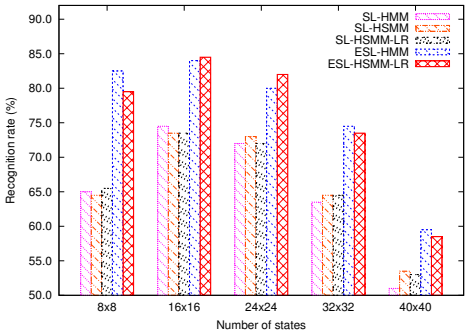

Fig. 7. Recognition rates on "dataset 2" (DCT coefficients)

5. CONCLUSION

We proposed separable lattice 2-D HMMs that introduce state duration control and a model that introduce state duration control into ESL-HMM. The state duration control can make the model maintain the ratio of each part of images. With this control, SL-HSMM-LR achieved the best results amongst all models in images with variations normalized. In images with variations, ESL-HSMM-LR was the most effective of all models. This result suggested that the proposed models can recognize face images with various variances more accurately.

6. ACKNOWLEDGEMENTS

This work was partially supported by the Artificial Intelligence Research Promotion Foundation.

7. REFERENCES

[1] M. Turk and A. Pentland, "Face Recognition Using Eigenfaces," IEEE Computer Vision and Pattern Recognition, pp.586–591, 1991.

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

[3] A. V. Nefian and M. H. Hayes, "A Hidden Markov Model for face recognition," ICASSP, vol.5, pp.2721–2724, 1998.

[4] D. Kurata, Y. Nankaku, K. Stoked, T. Kitamura, and Z. Ghahramani, "Face Recognition Based on Separable Lattice HMMs," ICASSP, pp.737–740, 2006.

[5] K. Kumaki, Y. Nankaku, K. Tokuda, "Face Recognition Based on Extended Separable Lattice 2-D HMMs," ICASSP, March 2012.

[6] Y. Takahashi, A. Tamamori, Y. Nankaku, K. Tokuda, "Face recognition based on Separable Lattice 2-D HMM with state duration modeling," ICASSP, pp.2162–2165, March 2010.

[7] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," Computer Speech and Language, vol.1, pp.29–45, 1986.

[8] 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.

[9] A. Tamamori, Y. Nankaku, and K. Tokuda, "An Extension of Separable Lattice 2-D HMMs for Rotational Data Variations," ICASSP, pp.2206–2209, 2010.

[10] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, No.9, pp.171–185, 1995.

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