Image Recognition Based on Convolutional Neural Networks Using Features Generated from Separable Lattice Hidden Markov Models

Takayuki Kasugai*, Yoshinari Tsuzuki*, Kei Sawada*, Kei Hashimoto*, Keiichiro Oura*, Yoshihiko Nankaku*, Keiichi Tokuda*
* Department of Computer Science, Nagoya Institute of Technology, Japan
E-mail:{kasugai,tsuzuki,swdkei,bonanza,uratec,nankaku,tokuda}@sp.nitech.ac.jp Tel/Fax: +81-52-735-5479/5840

***Abstract*—An image recognition method based on convolutional neural networks (CNNs) using features generated from separable lattice hidden Markov models (SLHMMs) is proposed. A major problem in image recognition is that the recognition performance is degraded by geometric variations in the size and position of the object to be recognized. To solve this problem, SLHMMs have been proposed as an extension of HMMs with size and locational invariances based on state transitions. Although SLHMMs are generative models that can represent the generation processes of observations well, there is a possibility that they are not specialized for discrimination compared to discriminative models. Our method integrates SLHMMs that extract features invariant to geometric variations with CNNs that build an accurate classifier based on discriminative models with the extracted features. Face recognition experiments showed that the proposed method improves recognition performance.**

I. INTRODUCTION

Recently, statistical methods using large-scale datasets have been increasing. In the field of image recognition, statistical methods such as eigenface methods [1] and subspace methods [2] have achieved good recognition performance. However, such statistical methods have a problem in terms of geometric variations, i.e., the size, position, and rotation of the target objects. One solution to this problem is to normalize geometric variations using heuristic normalization techniques in the preprocessing part of the classification process. However, this is very expensive and is task-dependent. Therefore, it is necessary to use the same criterion for both normalization and training classifiers.

Hidden Markov model (HMM)-based methods have been proposed as such an approach to geometric variations [3], [4]. In these methods, the geometric normalization is represented by discrete hidden variables, and the normalization process is performed through the calculation of probabilities. However, the extension of HMMs to multiple dimensions generally leads to an exponential increase in the computational complexity, and some efficient approximations of likelihood calculation and model structures have therefore been proposed [5]–[10]. To reduce the computational complexity while retaining the good properties for modeling multiple dimensional data, separable lattice HMMs (SLHMMs) have been proposed [11]. An SLHMM is a feasible model that can perform elastic matching in both the horizontal and vertical directions, making it possible to model invariances to the size and location of target objects. Although SLHMMs are generative models and can represent the generation processes of observations well, there is possibility that they are not specialized for discrimination compared to discriminative models. In other words, the recognition performance of generative models is likely to be inferior to that of discriminative models because discriminative models are specialized to discrimination.

Recently, discriminative models have been intensively studied, and neural networks in particular have shown great success in many applications. In image recognition, convolutional neural networks (CNNs) have successfully been used as discriminative models because of their robustness against geometric variations based on multiple convolutional and pooling layers [13]. Further, the network structure of CNNs has a feature-extraction process that is simultaneously optimized by training the classifier using the discriminative criterion. However, in terms of invariances of geometric variations, CNNs still have a weakness in that it is difficult to represent global geometric variations over an entire image because pooling is independently performed in each local window. Therefore, a structure that represents explicit image variations as in generative models should be useful to construct discriminative models with higher invariance to geometric variations.

In this work, we propose image recognition based on CNNs using features generated from SLHMMs. Combining an SLHMM that can consider the geometric variations of target objects and a CNN that is specialized for discrimination should be able to compensate for the disadvantages of both models. We recently proposed a method using log linear models (LLMs) as a discriminator to input features based on derivatives [14] with respect to the parameters of SLHMMs [15]. We expect that using CNNs as discriminative models will result in higher recognition rates than using LLMs. Furthermore, we investigate the optimal structure of CNNs under the assumption of combining them with SLHMMs. In the proposed model, invariances to geometric variations are dealt with by SLHMMs, so CNNs would no longer need to have the ability to perform invariances, which are mainly performed by pooling layers.

In sections 2 and 3 of this paper, SLHMMs and CNNs are briefly explained. Section 4 describes the flow of the proposed

method and the features generated from SLHMMs. Section 5 presents the face recognition experiments using the XM2VTS database [16]. We conclude with a brief summary in section 6.

II. SEPARABLE LATTICE HIDDEN MARKOV MODELS

Separable lattice hidden Markov models (SLHMMs) are defined for modeling multi-dimensional data. In the case that observations are 2D data, e.g., pixel values of an image, observations are assumed to be given on a 2D lattice as

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

where $\boldsymbol{t}$ denotes the coordinates of the lattice in 2D space $\boldsymbol{T}$ and $t^{(m)} = 1, \ldots, T^{(m)}$ is the coordinate of the m-th dimension for $m \in \{1, 2\}$. In 2D HMMs, observation $\boldsymbol{O_t}$ is emitted from the state indicated by hidden variable $\boldsymbol{S_t} \in \boldsymbol{K}$. The hidden variables $\boldsymbol{S_t} \in \boldsymbol{K}$ can take one of $K^{(1)}K^{(2)}$ states, which are assumed to be arranged on a 2D state lattice $\boldsymbol{K} = \{(1,1), (1,2), \ldots, (K^{(1)}, K^{(2)})\}$. $K^{(1)}$ and $K^{(2)}$ are the states of the HMM in the vertical and horizontal directions.

In SLHMMs, the hidden variables are constrained to be composed of two Markov chains in order to reduce the number of possible state sequences, as

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

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

where $\boldsymbol{S}^{(m)}$ is the Markov chain along with the m-th coordinate and $S^{(m)}_{t^{(m)}} \in \{1, 2, \ldots, K^{(m)}\}$. The composite structure of hidden variables in SLHMMs is defined as the product of hidden state sequences: $\boldsymbol{S_t} = (S^{(1)}_{t^{(1)}}, S^{(2)}_{t^{(2)}}) \in \boldsymbol{K}$. This means that the segmented regions of observations are constrained to be rectangles, which allows an observation lattice to be elastic in both the vertical and horizontal directions. Figures 1 and 2 show a graphical model of an SLHMM and the model structure of an SLHMM in face-image modeling. The joint probability of observation vectors $\boldsymbol{O}$ and hidden variables $\boldsymbol{S}$ can be written as

$$\begin{aligned} P(\boldsymbol{O}, \boldsymbol{S} | \boldsymbol{\Lambda}) &= P(\boldsymbol{O}, \boldsymbol{S}^{(1)}, \boldsymbol{S}^{(2)} | \boldsymbol{\Lambda}) \\ &= \prod_{\boldsymbol{t}} P(\boldsymbol{O_t} | \boldsymbol{S_t}, \boldsymbol{\Lambda}) \times \prod_{m=1}^{2} \left\{ P(S^{(m)}_1 | \boldsymbol{\Lambda}) \right. \\ &\quad \left. \prod_{t^{(m)}=2}^{T^{(m)}} P(S^{(m)}_{t^{(m)}} | S^{(m)}_{t^{(m)}-1}, \boldsymbol{\Lambda}) \right\}, \end{aligned} \tag{4}$$

where $\boldsymbol{\Lambda} = \{\boldsymbol{\pi}^{(m)}, \boldsymbol{a}^{(m)}, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\}$ is a set of model parameters, $\boldsymbol{k}$ is the 2D state index in the 2D state lattice $\boldsymbol{K}$, $\boldsymbol{\pi}^{(m)}$ is the initial state probability, $\boldsymbol{a}^{(m)}$ 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 that is an output probability distribution on a 2D state lattice.

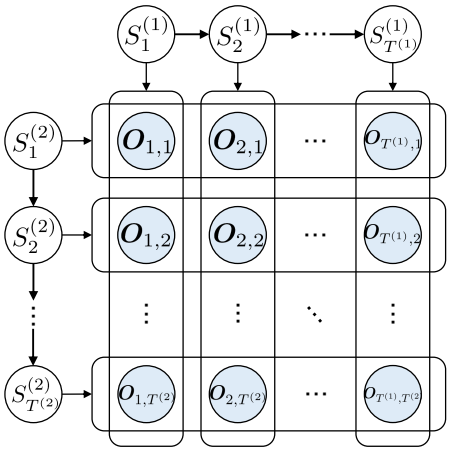


Fig. 1. Graphical model of an SLHMM. The rounded boxes represent a group of variables, and the arrow to each box represents the dependency in regard to all variables in the box instead of drawing arrows to all variables.

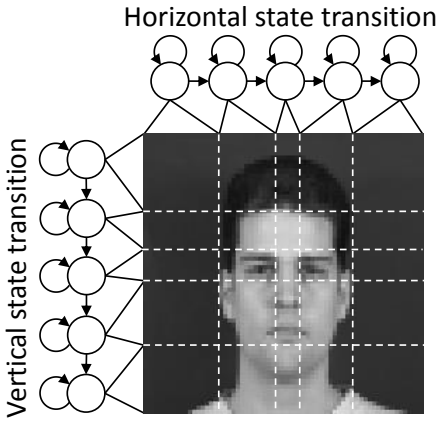


Fig. 2. Model structure of SLHMMs in face-image modeling

III. CONVOLUTIONAL NEURAL NETWORK

A CNN has a structure in which a convolutional layer and a pooling layer are repeated. In a convolutional layer, the coordinates $u_{i,j}$ in the feature map can be written as

$$u^{[k]}_{i,j} = f\left(\sum_{c} \sum_{s=0}^{m-1} \sum_{t=0}^{n-1} w^{[k,c]}_{s,t} x^{[c]}_{(i+s),(j+t)} + b^{[k]} \right), \tag{5}$$

where $\boldsymbol{u}^{[k]}$ is the feature map, k is the index of the filter, $\boldsymbol{x}^{[c]}$ is the previous feature map, c is a channel, $\boldsymbol{w}^{[k,c]}$ is the weight, and $b^{[k]}$ is the bias. The size of the filter is $m \times n$ and f is an activation function. A convolutional layer is the core building block of CNNs and does most of the computational heavy lifting. In a pooling layer, the spatial size of the representation is reduced progressively, so the amount of parameters and computation in the network are reduced.

IV. Image recognition using features generated from SLHMMs

SLHMMs make it possible to model invariances to the size and location of a target object. However, there is possibility that SLHMMs are not specialized for discrimination compared to discriminative models because SLHMMs are generative models representing the generation processes of observations.

In generative models, the posterior probability in the multinomial classification is transformed by Bayes' theorem as

$$P(C|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|C)P(C)}{P(\boldsymbol{O})}, \tag{6}$$

where C is a class, $\boldsymbol{O}$ is an observation data, and $P(\boldsymbol{O}) \neq 0$. Therefore, the posterior probability is indirectly estimated by estimating $P(\boldsymbol{O}|C)P(C)$. Then, $P(\boldsymbol{O}|C)P(C)$ indicates the generation processes of observations. By estimation these processes, it is possible to model distribution shapes. However, it is difficult to obtain true generative processes and distribution shapes of observations because training data is finite. Hence, there is a problem in recognition accuracy in discrimination using generative models based on (6). To solve this problem, methods using discriminative models that learn with criteria specialized for discrimination by directly estimating the posterior probability have been proposed, and the effectiveness of these methods has been confirmed.

CNNs have successfully been used as discriminative models due to their robustness against geometric variations based on multiple convolutional and pooling layers. However, it is difficult to represent global geometric variations over an entire image because pooling is independently performed in each local window. Therefore, we propose image recognition based on CNNs using features generated from SLHMMs. The proposed method can extract features invariant to geometric variations by using SLHMMs and build an accurate classifier based on CNNs with the extracted features.

A. Features based on SLHMMs

Some of the features based on generative models are derivative feature, which are defined by derivatives of the log-likelihood function with respect to the model parameters. Log-likelihood features for the image data are defined as

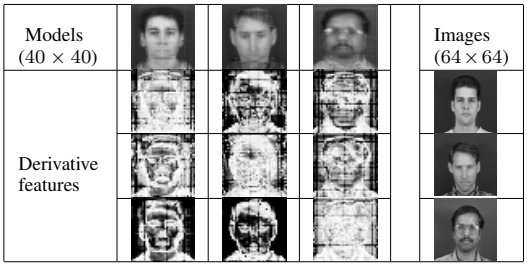
$$\boldsymbol{f}^{[i]} = \begin{cases} \ln P(\boldsymbol{O}^{[i]}|\boldsymbol{\Lambda}^{[1]}) \\ \vdots \\ \ln P(\boldsymbol{O}^{[i]}|\boldsymbol{\Lambda}^{[I]}), \end{cases} \tag{7}$$

where $\boldsymbol{O}^{[i]}$ is the i-th image data and $\boldsymbol{\Lambda}^{[i]}$ is the i-th model parameter. The derivative features of means that are a model parameter are thus represented as

$$\nabla_{\boldsymbol{\mu_k}} \boldsymbol{f}^{[i]} = \begin{cases} \nabla_{\boldsymbol{\mu_k}} \ln P(\boldsymbol{O}^{[i]}|\boldsymbol{\Lambda}^{[1]}) \\ \vdots \\ \nabla_{\boldsymbol{\mu_k}} \ln P(\boldsymbol{O}^{[i]}|\boldsymbol{\Lambda}^{[I]}), \end{cases} \tag{8}$$

$$\nabla_{\boldsymbol{\mu_k}} \ln P(\boldsymbol{O}^{[i]}|\boldsymbol{\Lambda}^{[i]}) = \sum_{\boldsymbol{t}} \langle S_{\boldsymbol{k},\boldsymbol{t}} \rangle_{Q(\boldsymbol{S})} \boldsymbol{\Sigma}_{\boldsymbol{k}}^{-1} (\boldsymbol{O_t} - \boldsymbol{\mu_k}), \tag{9}$$

TABLE I
Examples of derivative features



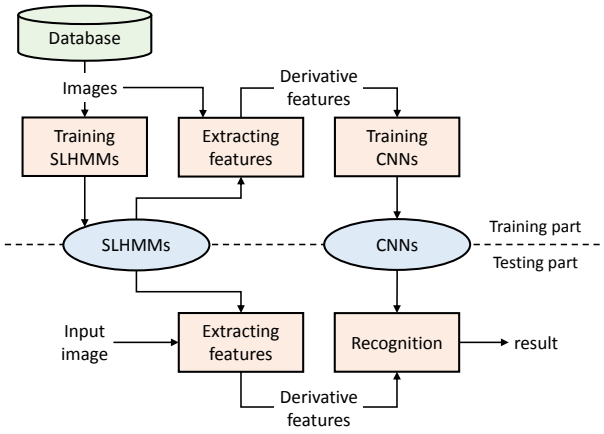


Fig. 3. Overview of proposed method

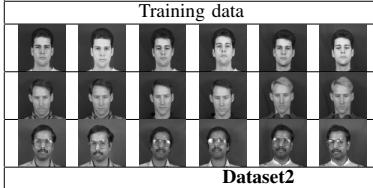
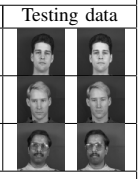
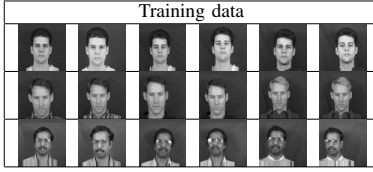
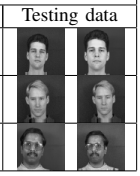
where $\langle S_{\boldsymbol{k},\boldsymbol{t}} \rangle_{Q(\boldsymbol{S})}$ is the posterior distribution of state $\boldsymbol{k}$ at coordinate $\boldsymbol{t}$ and $Q(\boldsymbol{S})$ is the approximate posterior probability of $P(\boldsymbol{S}|\boldsymbol{O}, \boldsymbol{\Lambda})$. The derivative features of state $\boldsymbol{k}$ are derived using the statistics related to the model parameters of state $\boldsymbol{k}$.

Table I lists examples of the visualization of derivative features when inputting an image to SLHMMs. The number of pixels of the visualization matches that of the states of the SLHMMs. The white area indicates that the gradient is small and close to the maximum likelihood point. It can be confirmed that the visualization between the image and each model is the whitest.

B. Image recognition based on CNNs with features generated from SLHMMs

Figure 3 shows an overview of the proposed method. First, SLHMMs are trained from training data in the training part, and second, derivative features described in Section IV-A are extracted by using the trained SLHMMs. Third, CNNs are trained from the extracted derivative features. Then, derivative features of all classes are superimposed as channels and input to CNNs. In the testing part, features corresponding to the testing data are extracted by the same procedures as the feature extracting in the training part. Recognition is executed by calculating the posterior probabilities of all classes from the trained CNNs and the extracted features.

TABLE II
EXAMPLES OF IMAGES IN EXPERIMENTS

Dataset1	
Training data	Testing data
	
Dataset2	
Training data	Testing data
	

V. EXPERIMENTS

A. Experimental conditions

To evaluate the effectiveness of the proposed method, face-recognition experiments on the XM2VTS database [16] were conducted. We prepared 8 images of 100 subjects for experiments; 6 images were used for training and 2 images were used for testing. Face images composed of 64×64 grayscale pixels were extracted from the original images. The example images are shown in Table II. We prepared two datasets for the experiments: **Dataset1**, which was size and location normalized, and **Dataset2**, which contained randomly generated size and location variations.

Three methods were compared: **SLHMM**, **CNN**, **SLHMM-LLM**, and **SLHMM-CNN** (proposed method). The details of **SLHMM**, **CNN**, and **SLHMM-CNN** are shown below. For the parameters of SLHMM and CNN, the optimum one with the highest recognition rate in the preliminary experiment was selected.

SLHMM:

Experimental conditions for **SLHMM** are shown in Table III below.

TABLE III
EXPERIMENTAL CONDITIONS FOR **SLHMM**.

The number of states	40×40
Estimation method	The maximum posteriori (MAP) estimation [17]
Training algorithm	Deterministic annealing EM (DAEM) algorithm

CNN:

The architecture for **CNN** was as follows:

- $I(64,1) - F(250) - O(100)$.
- $I(64,1) - C(5,8,1) - F(250) - O(100)$.
- $I(64,1) - C(5,8,1) - P(4,2) - F(250) - O(100)$.
- $I(64,1) - C(5,8,1) - C(2,16,1) - F(250) - O(100)$.
- $I(64,1) - C(5,8,1) - P(4,2) - C(2,16,1) - P(4,2) - F(250) - O(100)$.

The presentation of the CNN parameters is shown in Table IV

TABLE IV
PRESENTATION OF CNN PARAMETERS

$I(i,d)$	An input layer with d dimensional $i \times i$ sized image
$C(f,w,s)$	A convolutional layer with f filters of a $w \times w$ sized window with a stride of s
$P(w,s)$	A pooling layer
$F(n)$	A fully connected layer with n units
$O(c)$	An output layer with c classes

When using **Dataset2**, the first convolutional layer was $C(5,16,1)$ and the second convolutional layer was $C(5,32,1)$. The ReLU function and dropout with probability 0.5 were used in the convolutional and fully-connected layers.

SLHMM-CNN:

The architecture for **SLHMM-CNN** was as follows:

- $I(40,100) - F(1000) - O(100)$.
- $I(40,100) - C(2,8,1) - F(1000) - O(100)$.
- $I(40,100) - C(2,8,1) - P(4,2) - F(1000) - O(100)$.
- $I(40,100) - C(2,8,1) - C(2,16,1) - F(1000) - O(100)$.
- $I(40,100) - C(2,8,1) - P(4,2) - C(2,16,1) - P(4,2) - F(1000) - O(100)$.

When using **Dataset2**, the first convolutional layer was $C(5,32,1)$ and the second convolutional layer was $C(5,64,1)$. The sigmoid function and dropout with probability 0.5 were used in the convolutional and fully-connected layers.

B. Results

Figure 4 shows the results of the face recognition experiments on **Dataset1** with size and location normalized. It is confirmed from Fig. 4 that the recognition rates of **CNN** were lower than those of **SLHMM**. It seems that learning was not performed sufficiently in **CNN** due to the small amount of training data. **SLHMM-CNN** achieved a higher performance than **SLHMM** because an SLHMM-part in **SLHMM-CNN** dealt with a small amount of training data. This suggests that the effective feature extraction based on SLHMMs would help the training of CNNs. Comparing **SLHMM-CNN**, **SLHMM-LLM** and **SLHMM**, structures with a fully connected layer and convolutional layers obtained higher recognition rates than **SLHMM-LLM** and **SLHMM**. **SLHMM-CNN** is effective for discrimination because neural networks are specialized for discrimination. Hence, it is confirmed that recognition rates improve by adding a convolutional layer. In contrast, recognition rates decrease by adding the pooling layer. It seems that the effect of geometric invariances made no additional improvement due to the invariances of the feature extraction based on SLHMMs. Moreover, the negative aspect of max-pooling, which simply discarded information, appeared. We conclude that variations by the pooling layer are not effective for variations, and recognition rates decreased due to the loss of data information.

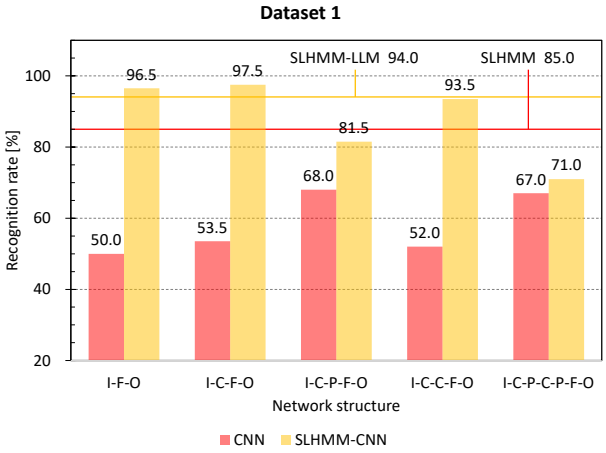


Fig. 4. Recognition rates on **Dataset1** with size and location normalized

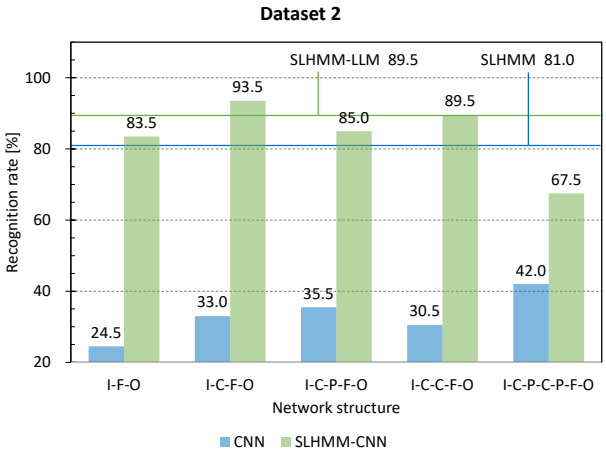


Fig. 5. Recognition rates on **Dataset2** with size and location randomly generated.

Figure 4 shows the results of the face recognition experiments on **Dataset2** with size and location randomly generated. It is confirmed from Figs. 4 and 5 that **SLHMM-CNN** and **SLHMM** obtained high recognition rates even on **Dataset2**. This is because SLHMMs represent global geometric variations over an entire image. However, in **CNN**, recognition rates in **Dataset2** were lower than those in **Dataset1**. This is likely because it is difficult for pooling layers to deal with global geometric variations.

VI. CONCLUSION

We proposed image recognition based on convolutional neural networks (CNNs) using features generated from separable lattice hidden Markov models (SLHMMs). The proposed method can obtain features by account geometric variations using SLHMMs as a feature generator. The results obtained in this study indicate that features extracted from SLHMMs are effective for classification and robust against geometric variations. Furthermore, the proposed method enables highly accurate recognition even when the amount of training data is small. For future work, we will construct SLHMMs with CNNs to generate features more discriminatively.

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] F. S. Samaria, "Face recognition using hidden Markov models," Ph. D. dissertation, University of Cambridge, 1994.
[4] A. V. Nefian and M. H. Hayes, "A Hidden Markov Model for face recognition," *International Conference on Acoustics, Speech and Signal Processing,* vol. 5, pp. 2721–2724, 1998.
[5] S. S. Kuo and O. E. Agazzi, "Keyword spotting in poorly printed documentions using pseudo 2-D hidden Markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 16, no. 8, pp. 842–848, 1994.
[6] A. V. Nefian and M. H. Hayes III, "Maximum likelihood training of the embedded HMM for face detection and recognition," *International Conference on Image Processing,* vol. 1, pp. 33–36, 2000.
[7] X. Ma, D. Schonfeld, and A. Khokhar, "Image segmentation and classfication based on a 2D distributed hidden Markov model," *Society of Photo-optical Instrumentation Engineers,* vol. 6822, 2008.
[8] J. Li, A. Najmi, and R. M. Gra, "Image classification by a two dimensional hidden Markov model," *IEEE Transactions on Signal Processing,* vol. 48, no. 2, pp. 517–533, 2000.
[9] H. Othman and T. Aboiilnasr, "A simplified second-order HMM with application to face recognition," *International Symposium on Circuits and Systems,* vol. 2, pp. 161–164, 2001.
[10] J. T. Chien and C. P. Liao, "Maximum confidence hidden Markov modeling for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, no. 4, pp. 606–616, 2008.
[11] D. Kurata, Y. Nankaku, K. Tokuda, T. Kitamura, and Z. Gharamani, "Face Recognition based on Separable Lattice HMMs," *International Conference on Acoustics, Speech and Signal Processing,* vol. 5, pp. 737–740, 2006.
[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradientbased learning applied to document recognition," *Processings of the IEEE,* vol. 86, pp. 2278–2324, 1998.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Conference on Neural Infomation Processing Systems,* pp. 1097–1105, 2012.
[14] C. Longworth and M. J. F. Gales, "Combining Derivative and Parametric Kernels for Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing,* pp. 748–757, 2009.
[15] Y. Tsuzuki, K. Sawada, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Image recognition based on discriminative models using features generated separable lattice HMMs." *International Conference on Acoustics, Speech and Signal Processing,* pp. 2607–2611, 2017.
[16] 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.
[17] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. SAP,* 2:291–298, 1994.