World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:12, 2016

Human Action Recognition Using Variational Bayesian HMM with Dirichlet Process Mixture of Gaussian Wishart Emission Model

Wanhyun Cho, Soonja Kang, Sangkyoon Kim, Soonyoung Park

***Abstract*—** In this paper, we present the human action recognition method using the variational Bayesian HMM with the Dirichlet process mixture (DPM) of the Gaussian-Wishart emission model (GWEM). First, we define the Bayesian HMM based on the Dirichlet process, which allows an infinite number of Gaussian-Wishart components to support continuous emission observations. Second, we have considered an efficient variational Bayesian inference method that can be applied to drive the posterior distribution of hidden variables and model parameters for the proposed model based on training data. And then we have derived the predictive distribution that may be used to classify new action. Third, the paper proposes a process of extracting appropriate spatial-temporal feature vectors that can be used to recognize a wide range of human behaviors from input video image. Finally, we have conducted experiments that can evaluate the performance of the proposed method. The experimental results show that the method presented is more efficient with human action recognition than existing methods.

***Keywords*—**Human action recognition, Bayesian HMM, Dirichlet process mixture model, Gaussian-Wishart emission model, Variational Bayesian inference, Prior distribution and approximate posterior distribution, KTH dataset

I. Introduction

HIDDEN Markov Models (HMMs) are widely used in a variety of fields for modeling time series data, with applications including speech recognition, natural language processing, protein sequence modeling and genetic alignment, general data compression, information retrieval, motion video analysis and object tracking, and financial time series prediction [1]. The core theory of HMMs was developed principally by Baum and Colleagues, with initial applications to elementary speech processing, integrating with linguistic models, and making use of insertion and deletion states for variable length sequences [2]. The popularity of HMMs soared in the following decade giving rise to a variety of elaborations, as reviewed in Juang and Rabiner [3]. Moreover, the realization that HMMs can be expressed as Bayesian networks [4] has given rise to more complex and interesting models, for example, factorial HMMs [5], tree-structured HMMs [6], and switching state-space models [7]. Beal [8] presents a unified variational Bayesian framework in his PhD dissertation which approximates true posterior distributions in models with latent variables using a lower bound on the marginal likelihood. On the other hand, several papers applying the HMM model with human action recognition have been published in recent years. Yin and Meng [9] present a novel hierarchical probability latent model to recognize human activities from a sequence of visual data. Their model consists of four layers from bottom-up: spatial-temporal visual features layer, atomic pattern layer, latent topic layer, and behavior pattern layer. Then, they applied the proposed model to represent the behavior patterns and latent topics as distributions over atomic patterns. Tian et al. [10] propose a Hierarchical Filtered Motion (HFM) method to recognize actions in crowded videos by using Motion History Image (MHI) as basic representation of motion due to its robustness and efficiency. Uddin et al. [11] present a novel approach for human activity recognition using the joint angles from a 3D model of the human body. They estimated body joint angles from time-series activity images acquired with a single stereo camera. The estimated joint-angle features are then mapped into code-words to generate discrete symbols for the HMM of each activity. Gaikward and Narawade [12] present novel HMM-based approach that uses threshold and voting to automatically and effectively segment and recognize complex activities. They also survey two hybrids of Neural Network and HMM, i.e. HMM-NN and NN-HMM, and compare their performance with that of the traditional HMM. Piyathilaka and Kodagoda [13] presented a human activity detection model that uses only 3-D skeleton features generated from an RGB-D sensor. To infer human activities, they implemented a Gaussian Mixture Model based HMM to capture the multimodal nature of the 3D positions of each skeleton joint. They tested their model in a publicly available dataset that consists of twelve different daily activities performed by four different people.

The main contribution of this study can be considered the following two facts. The first contribution will propose the DPM of GWEM that can be suitable to model a continuous feature vector. The second contribution will be the utilization of the variational Bayesian estimation method to derive the posterior distributions of the parameters vector and latent variables needed to define our model. In Section II, we have used the DPM theory to autonomously determine the number of components of the Gaussian mixture model. In Section III, we have considered an efficient variational Bayesian inference method to drive the posterior distributions of the parameters

Wanhyun Cho (professor) and Soonja Kang (professor) are with the Department of Statistics and Mathematical Education, Chonnam National University, Gwangju, 61186 South Korea (corresponding author, phone: +82-62-530-3443; fax +82-62-530-3449; e-mail {whcho, kangsj}@ chonnam.ac.kr,).

Sangkyoon Kim (researcher) and Soonyoung Park (professor) are with the Electrical Engineering Department, Mokpo National University, Chonnam, 58554 South Korea (e-mail: narciss76@mokpo.ac.kr, sypark@mokpo.ac.kr).

vector and latent variables in the proposed model, and then we have derived the predictive distribution that may be used to classify new observation. Section IV proposes the overall process of extracting feature vectors from given video images that is one of the most difficult problems in human behavior classification. As well, various experiments have been conducted to evaluate the performance of the proposed method. Section IV, outlines the conclusion of the paper.

II. Variational Bayesian HMM with DPM of GWEM

A. Bayesian HMM

An HMM models a sequence of p -valued discrete observations (symbols) $\mathbf{y}_{1:T} = \{\mathrm{y}_1, \cdots, \mathrm{y}_T\}$ by assuming that the observation y_t at time t was produced by a k -valued discrete hidden state s_t , and that the sequence of hidden states $\mathbf{s}_{1:T} = \{\mathrm{s}_1, \cdots, \mathrm{s}_T\}$ was generated by a first-order Markov process. That is to say the complete-data likelihood of a sequence of length T is given by:

$$p(\mathbf{s}_{1:T}, \mathbf{y}_{1:T}) = p(s_1)\mathrm{p}(\mathrm{y}_1 \mid \mathrm{s}_1)\prod_{t=2}^{T} p(s_t \mid s_{t-1})\, p(y_t \mid s_t) \tag{1}$$

where $p(s_1)$ is the prior probability of the first hidden state, $p(s_t \mid s_{t-1})$ denotes the probability of transition from state s_{t-1} to state s_t, and $p(y_t \mid s_t)$ are the emission probabilities for each of p symbols at each state. In this simple HMM, all the parameters are assumed stationary, and we assume a fixed finite number of hidden states and number of discrete symbol observations. Hence, the probability of the observations $\mathbf{y}_{1:T}$ results from summing over all possible hidden state sequences,

$$p(\mathbf{y}_{1:T}) = \sum_{\mathbf{s}_{1:T}} p(\mathbf{s}_{1:T}, \mathbf{y}_{1:T}) \tag{2}$$

Moreover, the set of parameters for the initial state prior $\boldsymbol{\pi}$, transition probabilities $\mathbf{A}$, and emission probabilities $\mathbf{B}$ are represented by the parameter Θ :

$$\Theta = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}\}$$

$\boldsymbol{\pi} = \{\pi_j\} : \pi_j = p(s_1 = j) : (k \times 1)$ initial hidden state prior; $\mathbf{A} = \{\mathrm{a}_{jj'}\} : \mathrm{a}_{jj'} = p(s_t = j' \mid s_{t-1} = j) : (k \times k)$ state transition matrix; $\mathbf{B} = \{b_{jm}\} : b_{jm} = p(\mathrm{y}_t = m \mid s_t = j) : (k \times p)$ symbol emission matrix. Here, the Bayesian approach to learning treats the model parameters as unknown quantities and, prior to observation of the data, assigns a set of beliefs over these quantities in the form of prior distributions. In the light of the data, Bayes' rule can be used to infer the posterior distribution over the parameters.

$$p(\Theta \mid \mathbf{y}_{1:T}) = \frac{p(\Theta)\, p(\mathbf{y}_{1:T} \mid \Theta)}{\int p(\Theta)\, p(\mathbf{y}_{1:T} \mid \Theta) d\Theta} \tag{3}$$

In this way, the parameters of the model are treated as hidden variables and are integrated to form the marginal likelihood:

$$p(\mathbf{y}_{1:T}) = \int p(\Theta)\, p(\mathbf{y}_{1:T} \mid \Theta) d\Theta \tag{4}$$

A natural choice for parameter priors over $\boldsymbol{\pi}$, the row of $\mathbf{A}$, and the row of $\mathbf{B}$ are Dirichlet distributions.

$$\begin{aligned} p(\Theta) &= p(\boldsymbol{\pi})\, \mathrm{p}(\mathbf{A})\, \mathrm{p}(\mathbf{B}) \\ p(\boldsymbol{\pi}) &= \mathrm{Dir}(\{\pi_1, \cdots, \pi_k\} \mid \mathbf{u}^{(\pi)})) \\ p(\mathbf{A}) &= \prod_{j=1}^{k} \mathrm{Dir}(\{\mathrm{a}_{j1}, \cdots, \mathrm{a}_{jk}\} \mid \mathbf{u}^{(\mathrm{A})})) \\ p(\mathbf{B}) &= \prod_{j=1}^{k} \mathrm{Dir}(\{b_{j1}, \cdots, b_{jp}\} \mid \mathbf{u}^{(\mathrm{B})})) \end{aligned} \tag{5}$$

Here, for each matrix **A** and **B**, the same single hyper-parameter vector $\mathbf{u}^{(\mathrm{A})}$ and $\mathbf{u}^{(\mathrm{B})}$ is used for every row. The use of these hyper parameters is motivated because the hidden states are identical to a prior. The form of the Dirichlet prior, using $p(\boldsymbol{\pi})$ as an example, is:

$$p(\boldsymbol{\pi}) = \frac{\Gamma(u_0^{(\pi)})}{\prod_{j=1}^{k} \Gamma(u_j^{(\pi)})} \prod_{j=1}^{k} \pi_j^{u_j^{(\pi)} - 1}, \quad u_j^{(\pi)} > 0, \quad \forall j \tag{6}$$

where $u_0^{(\pi)} = \sum_{j=1}^{k} u_j^{(\pi)}$ is the strength of the prior, and the positivity constraint on the hyper-parameters is required for the prior to be proper.

B. Bayesian HMM with DPM of GWEM

So far, it has considered that the observation vector is the discrete case. From now, we will consider the case of continuous observation vectors. As well, suppose that the probability distribution of the observed vectors can be expressed in a mixture of an infinite number of Gaussian distribution. Therefore, in order to implement a continuous observation and infinite number of Gaussian problem, we have to consider the use of Dirichlet process theory and the Gaussian mixture model. Here, we first review Drichlet process model. A Dirichlet process (DP) $DP(\alpha, H)$ with concentration parameter α and base distribution H , is a distribution over probability distributions. Formally, let Ω be the probability space underlying the distribution. Then, we say that $G \sim DP(\alpha, H)$ if, for any finite partition $A_1, \cdots, A_N$ of Ω , the distribution of $G^{'}s$ probability mass on this partition is given by

$$(G(A_1), \cdots, G(A_N)) \sim \mathrm{Dirichlet}(\alpha H(A_1), \cdots, \alpha H(A_N)) .$$

Here, we can see the role of the concentration parameter α. That is, larger values of α encourage G to more closely follow the base distribution H, whereas smaller values allow for more deviation.

A more informative definition of the Dirichlet process is the *stick-breaking construction*, which defined $G \sim DP(\alpha, H)$ in terms of *stick-breaking weights* $\{u_k\}_{k=1}^{\infty}$ and *stick-breaking pieces* $\{\mathrm{c}_k\}_{k=1}^{\infty}$. Specifically, the stick-breaking representation of $G \sim DP(\alpha, H)$ can be defined as follows:

$$u_k \sim \text{Beta}(1,\alpha),\ c_k(u_1,\cdots,u_k) = u_k\prod_{l=1}^{k-1}(1-u_l), k=1,\cdots,\infty, \quad (7)$$

$$\Theta_k \sim H,\ G(y) = \sum_{k=1}^{\infty} c_k \delta_{\Theta_k}(y) \quad (8)$$

The above equation shows the very important fact that the DP is discrete. The support of G consists of a countably infinite set of atoms, drawn independently from H. This makes the Dirichlet process a natural choice for the distribution over hidden states in many popular models.

Second, the DPM model is used as a nonparametric prior in a hierarchical Bayesian specification:

$$G \sim DP(\alpha, H),\ \Theta_k \sim G,\ y_n \sim p(y_n \mid \Theta_k) \quad (9)$$

Data generated from this model can be partitioned according to the distinct values of the parameter. Taking this view, the DP mixture model has a natural interpretation as a flexible mixture model in which the number of components is random and grows when new data is observed. In the DPM model, the vector $\mathbf{c} = (c_1, c_2, \cdots)$ comprises the infinite vector of mixing proportions and $(\Theta_1, \Theta_2, \cdots)$ are the atoms representing the mixture components. Let $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_T)$ be the set of observations modeled by a DPM model. Then, each one of the observations $\mathbf{y}_t$ is assumed to be drawn from its own probability density function $p(\mathbf{y}_t \mid \Theta_k)$ parameterized by the parameter set Θ_k. Let z_k be an assignment variable of the mixture component with which the data point $\mathbf{y}_t$ is associated. The data set can be described as arising from the following process:

Draw $u_k \sim \text{Beta}(1,\alpha)$; Compute $c_k = u_k\prod_{l=1}^{k-1}(1-u_l)$

Draw $\Theta_k \sim H$

For the t -th data point:

Draw $\mathbf{z}_t = (z_{t1}, z_{t2}, \cdots) \sim \text{Mult}(\mathrm{c}(\mathbf{u}))$,

Draw $\mathbf{y}_t \sim p(\mathbf{y}_t \mid \Theta_{z_t=k})$

Third, for many application domains, the data associated with each hidden state may have a complex, multimodal distribution. We want to model this data with such emission distributions in HMM non-parametrically, using an infinite DP mixture of Normal-Wishart distributions. The study augments the HMM state s_t with a term z_t indexing the mixture component of the s_t^{th} emission density. For each HMM state s_t, we assume that there is a unique stick-breaking construction $c_{jm} = u_{jm}\prod_{l=1}^{m-1}(1-u_{jl}),\ \ m = 1,\cdots,\infty$ defining the mixture weights of the m^{th} emission density so that $p(z_t = m \mid s_t = j) = \mathrm{c}_{jm}$. We also assume the m^{th} emission density as Gaussian distribution $N(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m^{-1})$ with mean vector $\boldsymbol{\mu}_m$ and precision matrix $\boldsymbol{\Lambda}_m$. In the Bayesian HMM with DPM, we have to consider a set of the appropriate prior distributions over model parameters. We first choose conjugate-exponential priors for the mean vector $\boldsymbol{\mu}_m$ and precision matrix $\boldsymbol{\Lambda}_m$. Hence, we impose a joint Gaussian-Wishart distribution over the means and precisions of the Gaussian emission likelihoods in the model as:

$$p(\Theta_m = (\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m)) \sim NW(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m \mid \boldsymbol{\mu}_0, \tau_0, \mathbf{W}_0, \upsilon_0) \quad (10)$$

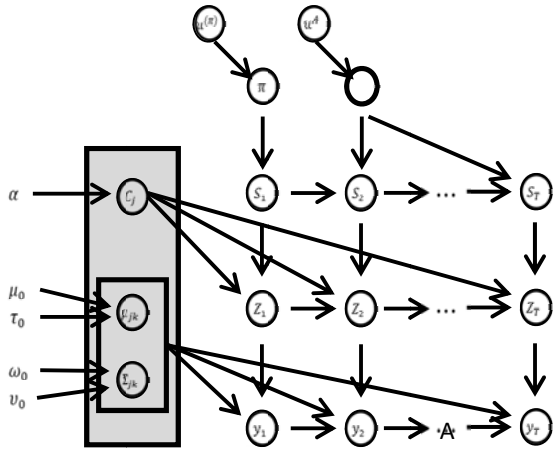

Fig. 1 Bayesian HMM with DPM of GW emission model

The Bayesian HMM with DPM of GWEM considered is formally described in Fig. 1. Therefore, the joint probability density function for observation vector $\mathbf{y}_{1:T}$, all hidden variables and parameters $\Omega = \{\mathbf{s}, \mathbf{z}, \boldsymbol{\pi}, \mathbf{A}, \mathbf{B}, \mathbf{C}(\mathbf{u}), \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ can be rewritten as:

$$\begin{aligned} &p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{s}_{1:T}, \boldsymbol{\pi}, \mathbf{A}, \mathrm{C}(\mathbf{u}), \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{s}_{1:T} \mid \boldsymbol{\pi}, \mathbf{A}, \mathrm{C}(\mathbf{u}), \boldsymbol{\mu}, \boldsymbol{\Lambda})\, p(\boldsymbol{\pi})\, p(\mathbf{A}) \\ &\qquad \times p(\mathrm{C}(\mathbf{u}))\, p(\boldsymbol{\mu} \mid \boldsymbol{\Lambda})\, p(\boldsymbol{\Lambda}), \end{aligned} \quad (11)$$

where the individual factors are:

$$\begin{aligned} &p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{s}_{1:T} \mid \boldsymbol{\pi}, \mathbf{A}, \mathrm{C}(\mathbf{u}), \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= p(s_1 \mid \boldsymbol{\pi}) \prod_{t=2}^{T} p(s_t \mid s_{t-1}, \mathbf{A}) \\ &\qquad \times \prod_{t=1}^{T} p(\mathrm{z}_t \mid s_t, \mathrm{C}(\mathbf{u})) p(y_t \mid \boldsymbol{\mu}_{s_t, z_t}, \boldsymbol{\Lambda}_{s_t, z_t}) \end{aligned} \quad (12)$$

$$p(\boldsymbol{\pi}) \sim \mathrm{Dir}(\{\pi_1,\cdots,\pi_K\} \mid \mathbf{u}^{(\pi)})),\ p(s_1) \sim \mathrm{Multi}(\boldsymbol{\pi}),$$

$$p(\mathbf{a}_j) \sim \mathrm{Dir}(\{\mathrm{a}_{j1},\cdots,\mathrm{a}_{jK}\} \mid \mathbf{u}_j^{(A)}),\ p(\mathbf{A}) = \prod_{j=1}^{K} p(\mathbf{a}_j),$$

$$p(s_t \mid s_{t-1}, \mathbf{A}) \sim \mathrm{Multi}(\mathbf{a}_{s_{t-1}}), s_{t-1} = 1,\cdots,K,$$

$$u_{jm} \sim Beta(1,\alpha),\ \mathrm{c}_{jm} = u_{jm}\prod_{l=1}^{m-1}(1-u_{jl}), m = 1,\cdots,\infty,$$

$$\mathbf{c}_j(\mathbf{u}) = (c_{j1}, c_{j2},\cdots),\ p(\mathrm{C}(\mathbf{u})) = \prod_{j=1}^{K} p(\mathbf{c}_j(\mathbf{u}))\ ,$$

$$p(\mathrm{z}_t \mid s_t) \sim \mathrm{Multi}(\mathrm{c}_{s_t}(\mathbf{u})),$$

$$p(\mathbf{y}_t \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{m=1}^{\infty} N(\boldsymbol{\mu}_{z_{tm}|s_t}, \boldsymbol{\Lambda}_{z_{tm}|s_t}^{-1})^{z_{tm}},$$

$$p(\boldsymbol{\mu} \mid \boldsymbol{\Lambda}) = \prod_{m=1}^{\infty} N(\boldsymbol{\mu}_m \mid \boldsymbol{\mu}_{0m}, (\tau_{0m}\boldsymbol{\Lambda}_m)^{-1})$$

$$p(\boldsymbol{\Lambda}) = \prod_{m=1}^{\infty} W(\boldsymbol{\Lambda}_m \mid \mathbf{W}_{0m}, \upsilon_{0m}).$$

III. Variational Bayesian inference

A. Variational Bayesian EM algorithm

The variational Bayesian inference problem of HMM with DPM of GWEM is to derive a family of variational posterior distributions over hidden variables and model parameters which can approximate the true posterior distributions with infinite number of mixture components. But, under this infinite dimensional setting, Bayesian inference is not apparently tractable. For this reason, we employ a common strategy in DPM literature, formulated on the basis of a truncated stick-breaking representation of the DP. That is, we fix a value M and we let the variational posterior over the u_{jm} have the property $q(u_{jM} = 1) = 1$. In other words, we set $c_{jm}(\mathbf{u})$ equal to zero for $m > M$. Note that, under this setting, the model is a full Dirichlet process and is not truncated, but only the variational distribution is truncated to allow for a tractable inference procedure. Hence, the truncation level M is a variational parameter which can be freely set, and not part of the prior model specification.

Let $\Omega = \{\mathbf{s}, \mathbf{z}, \boldsymbol{\pi}, A, C, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ be the set of all hidden variables and unknown parameters of the Bayesian HMM with DPM of the Gaussian-Wishart model over which a prior distribution has been imposed, and $\Psi = \{\boldsymbol{\mu}_{0m}, \tau_{0m}, \mathbf{W}_{0m}, \upsilon_{0m}\}_{m=1}^{M}$ be the set of the hyper-parameters of the imposed priors. Variational Bayesian inference consists in the introduction of an arbitrary (variational) distribution $q(\Omega)$ to approximate the actual posterior $p(\Omega \mid \mathbf{y}_{1:T}, \Psi)$, which is computationally intractable. Under this assumption, the log marginal likelihood $\log p(\mathbf{y}_{1:T})$ of the model can be written as:

$$\log p(\mathbf{y}_{1:T}) = F(q) + \mathrm{KL}(q \,\|\, p), \tag{13}$$

with

$$F(q,\Psi) = \int q(\Omega) \ln\left(\frac{p(\mathbf{y}_{1:T}, \Omega; \Psi)}{q(\Omega)}\right) d\Omega \tag{14}$$

and

$$\mathrm{KL}(q \,\|\, p) = -\int q(\Omega) \ln\left(\frac{p(\Omega \mid \mathbf{y}_{1:T}; \Psi)}{q(\Omega)}\right) d\Omega \tag{15}$$

Here, $\mathrm{KL}(q \,\|\, p)$ stands for the Kullback-Leibler (KL) divergence between the approximate variational posterior $q(\Omega)$ and the actual posterior $p(\Omega \mid \mathbf{y}_{1:T}, \Psi)$. Since KL divergence is nonnegative, $F(q,\Psi)$ forms a strict lower bound of the log marginal likelihood $\log p(\mathbf{y}_{1:T})$ defined as:

$$\log p(\mathbf{y}_{1:T}) \geq F(q) = \int q(\Omega) \ln\left(\frac{p(\mathbf{y}_{1:T}, \Omega; \Psi)}{q(\Omega)}\right) d\Omega \tag{16}$$

Hence, by maximizing this lower bound $F(q,\Psi)$ (variational free energy) so that it becomes as tight as possible, not only do we minimize the KL divergence between the true and variational posterior, but also implicitly integrate out the unknowns Ω.

For the approximate posterior distribution $q(\Omega)$, we consider two assumptions. First, we assume that we consider the conjugate prior distributions of all hidden variables and parameters in our model. Second, a set of parameters $\{\boldsymbol{\pi}, \mathbf{A}, \mathbf{C}(\mathbf{u}), \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ and hidden variables $\{\mathbf{s}, \mathbf{z}\}$ are mutually independent. Then, the approximate variational distribution of all hidden variables and parameters can be represented as;

$$\begin{aligned} q(\Omega) &= q(\mathbf{s},\mathbf{z})q(\boldsymbol{\pi})q(\mathbf{A})q(C(\mathbf{u}))q(\boldsymbol{\mu},\boldsymbol{\Lambda}) \\ &= q(s_1)\prod_{t=2}^{T} q(s_t \mid s_{t-1})\prod_{t=1}^{T} q(z_t \mid s_t)q(\boldsymbol{\pi})\prod_{j=1}^{K} q(\mathbf{a}_j) \\ &\quad \times \prod_{j=1}^{K}\prod_{m=1}^{M} q(\mathrm{c}_{jm})\prod_{j=1}^{K}\prod_{m=1}^{M} q(\boldsymbol{\mu}_{j,m}, \boldsymbol{\Lambda}_{j,m}). \end{aligned} \tag{17}$$

Here, using the calculus of variation principle, we can obtain the approximate posterior distributions of all hidden variables and parameters by minimizing the KL divergence or maximizing the free energy with the coordinate ascent algorithm. Then, the resulting variational posterior distributions can be given using the following two steps iteratively.

1) The Variational Bayesian M (VBM)-Step

The VBM step is obtained by taking functional derivatives of $F(q,\Psi)$ with respect to each of these distributions and equating them to zero, to yield the following approximate posterior distributions:

$$\begin{aligned} q(\boldsymbol{\pi}) &= \mathrm{Dir}(\{\pi_1,\cdots,\pi_k\} \mid \{w_1^{(\pi)},\cdots,w_k^{(\pi)}\})), \\ w_j^{(\pi)} &= u_j^{(\pi)} + q(s_1 = j), \end{aligned} \tag{18}$$

$$q(\mathbf{a}_j) = \mathrm{Dir}(\{\mathrm{a}_{j1},\cdots,\mathrm{a}_{jk}\} \mid \{w_{j1}^{(\mathrm{A})},\cdots,w_{jk}^{(\mathrm{A})}\})),$$
$$w_{jj'}^{(\mathrm{A})} = u_{j'}^{(\mathrm{A})} + \sum_{t=2}^{T} q(s_t = j' \mid s_{t-1} = j), \tag{19}$$

$$q(\mathrm{c}_{jm}) = Beta(w_{jm1}^{(\mathrm{c})}, w_{jm2}^{(\mathrm{c})}),$$
$$w_{jm1}^{(\mathrm{u})} = 1 + \sum_{t=1}^{T} q(z_t = m \mid s_t = j), \tag{20}$$
$$w_{jm2}^{(\mathrm{u})} = \alpha + \sum_{t=1}^{T} q(z_t > m \mid s_t = j).$$

Similar, regarding the posterior distributions over parameters of the Gaussian-Wishart, we have that

$$q(\boldsymbol{\mu}_{jm}, \Lambda_{jm}) = NW(\boldsymbol{\mu}_{jm}, \Lambda_{jm} \mid \mathbf{m}_{jm}, \tau_{jm}, \mathbf{W}_{jm}, \upsilon_{jm})$$
$$\xi_{jm} = \sum_{t=1}^{T} q(z_t = m \mid s_t = j),\ \tau_{jm} = \tau_{0m} + \xi_{jm},$$
$$\overline{\mathbf{y}}_{jm} = \frac{\sum_{t=1}^{T} q(z_t = m \mid s_t = j)\mathbf{y}_t}{\xi_{jm}},\ \mathbf{m}_{jm} = \frac{\tau_{0m}\boldsymbol{\mu}_{0m} + \xi_{jm}\overline{\mathbf{y}}_{jm}}{\tau_{jm}}, \tag{21}$$
$$\upsilon_{jm} = \upsilon_{0m} + \xi_{jm},$$
$$\Delta_{jm} = \sum_{t=1}^{T} q(z_t = m \mid s_t = j)(\mathbf{y}_t - \overline{\mathbf{y}}_{jm})(\mathbf{y}_t - \overline{\mathbf{y}}_{jm})^T,$$
$$\mathbf{W}_{jm} = \mathbf{W}_{0m} + \Delta_{jm} + \frac{\tau_{0m}\xi_{jm}}{\tau_{0m} + \xi_{jm}}(\mathbf{m}_{jm} - \overline{\mathbf{y}}_{jm})(\mathbf{m}_{jm} - \overline{\mathbf{y}}_{jm})^T.$$

2) The Variational Bayesian E (VBE)-Step

Taking derivatives of $F(q,\Psi)$ with respect to the variational posterior over the hidden variables yields:

$$q(s_1 = k) = \exp\left(\psi(w_k^{(\pi)}) - \psi(\sum_{j=1}^{K} w_j^{(\pi)})\right), \tag{22}$$

$$q(s_t = j' \mid s_{t-1} = j) = \exp\left(\psi(w_{jj'}^{(\mathrm{A})}) - \psi(\sum_{j'=1}^{K} w_{jj'}^{(\mathrm{A})})\right), \tag{23}$$

$$q(z_t = m \mid s_t = j) = c_{jm}^{*}(\mathbf{u})\, p^{*}(\mathbf{y}_t \mid \boldsymbol{\mu}_{jm}, \Lambda_{jm}), \tag{24}$$

where

$$c_{jm}^{*}(\mathbf{u}) = \exp(E(\ln c_{jm}(\mathbf{u})))$$
$$= \exp\left(\psi(w_{jm1}^{(\mathrm{u})}) - \psi(w_{jm1}^{(\mathrm{u})} + w_{jm2}^{(\mathrm{u})})\right) \tag{25}$$
$$\times \exp\left(\sum_{l=1}^{m-1}\left[\psi(w_{jl2}^{(\mathrm{u})}) - \psi(w_{jl1}^{(\mathrm{u})} + w_{jl2}^{(\mathrm{u})})\right]\right),$$

and

$$p^{*}(\mathbf{y}_t \mid \boldsymbol{\mu}_{jm}, \Lambda_{jm}) = \exp\left(E(\ln p(\mathbf{y}_t \mid \boldsymbol{\mu}_{jm}, \Lambda_{jm}))\right)$$
$$= \exp\left(-\frac{d}{2}\log 2\pi + \frac{1}{2}E(\ln|\Lambda_{jm}|)\right. \tag{26}$$
$$\left. -\frac{1}{2}E\left((\mathbf{y}_t - \boldsymbol{\mu}_{jm})^T \Lambda_{jm}(\mathbf{y}_t - \boldsymbol{\mu}_{jm})\right)\right),$$

$$E\left((\mathbf{y}_t - \boldsymbol{\mu}_{jm})^T \Lambda_{jm}(\mathbf{y}_t - \boldsymbol{\mu}_{jm})\right)$$
$$= \frac{d}{\tau_{jm}} + \upsilon_{jm}(\mathbf{y}_t - \mathbf{m}_{jm})^T \mathbf{W}_{jm}(\mathbf{y}_t - \mathbf{m}_{jm}) \tag{27}$$

$$E(\ln|\Lambda_{jm}|) = \sum_{i=1}^{d}\psi\left(\frac{\upsilon_{jm} + 1 - i}{2}\right) + d\ln 2 + \ln|\mathbf{W}_{jm}|. \tag{28}$$

where $\psi(\circ)$ denotes the digamma function.

B. Predictive Distribution

In the Bayesian scheme, the predictive probability of a test sequence $\mathbf{y}' = \mathbf{y}'_{1:T'}$, given a set of training cases denoted by $\mathbf{y} = \{\mathbf{y}_{i,1:T_i}\}_{i=1}^{n}$, is obtained by averaging the predictions of the HMM with DPM of GW model with respect to the posterior distributions over its parameters $\Theta = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}, \mathrm{C}(\mathbf{u}), \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$:

$$p(\mathbf{y}' \mid \mathbf{y}) = \int p(\Theta \mid \mathbf{y})\, p(\mathbf{y}' \mid \Theta) d\Theta \tag{29}$$

Unfortunately, for the very same reasons that the marginal likelihood of observations given by

$$p(\mathbf{y}_{1:T}) = \int \sum_{(\mathbf{s}_{1:T}, \mathbf{z}_{1:T})} p(\mathbf{s}_{1:T}, \mathbf{z}_{1:T}, \mathbf{y}_{1:T} \mid \Theta) p(\Theta) d\Theta \tag{30}$$

is in tractable, so is the predictive probability. Hence, we have to consider another method for approximating the predictive probability. One such method is to approximate the true posterior distribution with the variational posterior distribution resulting from the variational Bayesian optimization:

$$p(\mathbf{y}' \mid \mathbf{y}) \approx \int q(\Theta)\, p(\mathbf{y}' \mid \Theta) d\Theta \tag{31}$$

The variational posterior is a product of individual posterior of required parameters, which is in the same form as the prior, and so we are not able to doing anything anymore because we know that this integral is intractable. However, we can define the forward factor $\alpha_t(s'_t, z'_t)$ to be the posterior over the hidden variables (s'_t, z'_t) given the testing sequence up to and including time t and the trained parameters Θ:

$$\alpha_t(s'_t, z'_t) \equiv p(s'_t, z'_t \mid \mathbf{y}'_{1:t}; \Theta) \tag{32}$$

and form the forward recursion from $t = 1,\cdots,T'$:

$$\alpha_t(s'_t, z'_t) \equiv \frac{1}{p(y'_t \mid y'_{1:t-1};\Theta)} \times \sum_{(s'_{t-1}, z'_{t-1})} p(s'_{t-1}, z'_{t-1} \mid \mathbf{y}'_{1:t-1};\Theta)\, p(s'_t \mid s'_{t-1}) \times p(z'_t \mid s'_t)\, p(\mathbf{y}'_t \mid (s'_t, z'_t);\Theta) \quad (33)$$

$$= \frac{1}{\varsigma(y'_t;\Theta)} \left[\sum_{(s'_{t-1}, z'_{t-1})} \alpha_{t-1}(s'_{t-1}, z'_{t-1})\, p(s'_t \mid s'_{t-1}) \times p(z'_t \mid s'_t)\, p(\mathbf{y}'_t \mid (s'_t, z'_t);\Theta) \right)$$

where in the first time step $p(s'_t \mid s'_{t-1})$ is replaced with the prior $p(s'_1 \mid \boldsymbol{\pi})$, and for $t = 1$ we require the convention $\alpha_0(s'_0, z'_0) = 1$. Here, $\varsigma(y'_t;\Theta)$ is a normalization constant, a function of y'_t, given by,

$$\varsigma(y'_t;\Theta) \equiv p(y'_t \mid y'_{1:t-1};\Theta) \quad (34)$$

Note that as a by-product of computing these normalization constants, we can compute the probability of the sequence:

$$p(y'_{1:T};\Theta) = p(y'_1;\Theta)\, p(y'_2 \mid y'_1;\Theta) \cdots p(y'_T \mid y'_{1:T-1};\Theta) = \prod_{t=1}^{T} p(y'_t \mid y'_{1:t-1};\Theta) = \prod_{t=1}^{T} \varsigma(y'_t;\Theta) \quad (35)$$

Moreover, obtaining these normalization constants using a forward pass is simply equivalent to integrating out the hidden states one after the other in the forward ordering, as can be seen by writing the incomplete-data likelihood in the following way:

$$p(y'_{1:T};\Theta) = \sum_{s'_{1:T}, z'_{1:T}} p(s'_{1:T}, z'_{1:T}, \mathbf{y}'_{1:T};\Theta) = \sum_{s'_1, z'_1} \cdots \sum_{s'_T, z'_T} p(s'_1)\, p(z'_1 \mid s'_1)\, p(\mathbf{y}'_1 \mid s'_1, z'_1;\Theta) \times \prod_{t=2}^{T} p(s'_t \mid s'_{t-1})\, p(z'_t \mid s'_t)\, p(\mathbf{y}'_t \mid s'_t, z'_t;\Theta) \quad (36)$$

$$= \sum_{s'_1, z'_1} p(s'_1)\, p(z'_1 \mid s'_1)\, p(\mathbf{y}'_1 \mid s'_1, z'_1;\Theta) \cdots \sum_{s'_T, z'_T} p(s'_T)\, p(z'_T \mid s'_T)\, p(\mathbf{y}'_T \mid s'_T, z'_T;\Theta).$$

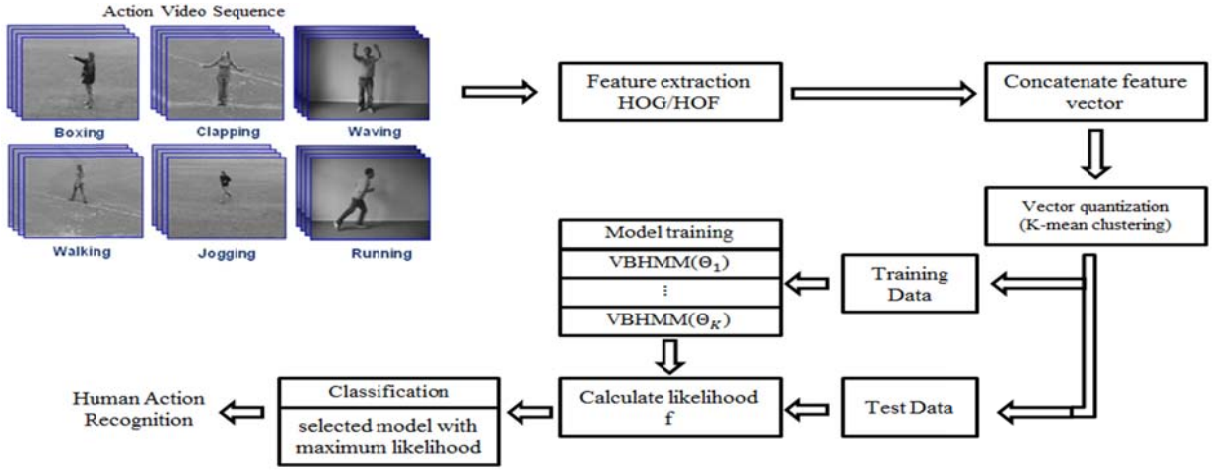

Fig. 2 Overall architecture of the proposed recognition system

IV. Human Action Recognition

A. KTH Dataset

In order to evaluate the performance of the proposed method, we use the KTH human action dataset. This dataset contains 25 people performing six action classes, namely: walking, running, jogging, hand waving, boxing, and hand clapping. Each video sequence contains one actor performing an action. In order to train the proposed model, we use the KTH human action dataset. This training dataset contained with a total of 384 video sets, in which 16 people repeatedly performed six action classes four times. And to test the performance of our model, the researchers have also used a test dataset consisting of 216 video sequences consisting of nine people repeatedly performing six different human behaviors four times.

B. Procedure for Human Action Recognition

Fig. 2 shows the overall architecture of the proposed recognition system for human actions. First, we extract the interest points in parallel, such as Hessian detectors from each frame image of the input video, and then we have configured the (3x3) square cells by using the extracted interest points to the rectangular. Second, we calculate two kinds of descriptors such as Histograms of Oriented Gradient (HOG) and Histogram of Optical Flow (HOF) in each of the points contained in this square, and then we construct feature histogram by concatenating 4 bins HOG and 5 bins HOF [14]. Third, we have clustered all of these feature histograms to 900 clusters using the K-means algorithm, and computed the mean vector and covariance matrix for each cluster. Fourth, we seek the cluster center with the minimum distance between the input feature vector and 900 clusters centers, and then we assigned a cluster mean vector number corresponding with the nearest cluster center into a feature vector. Fifth, by applying such a method with all frame images, it was possible to obtain a sequence of feature vectors corresponding to the input video.

Sixth, the classification probabilities were calculated corresponding to all human actions using a sequence of feature vector.

C. Recognition Results

From the results in Table I, it can be noted that six human behaviors can be mainly divided into two categories with similar behaviors. The first category of similar actions includes boxing, hand-clapping, and hand-waving, and the second category of similar behavior includes jogging, running, and walking. The results show that handclapping action is misclassified into boxing and hand-waving, and the jogging action misclassified into running and walking. However, it was noted that the correct classification rate of the proposed method appears to be 92.5 % on average. Finally, our model for obtaining this result is to be the number of state five, and the number of Gaussian mixture model components eight.

TABLE I
CLASSIFICATION RATE FOR PROPOSED METHOD

Classification rate	Boxing	Hand clapping	Hand waving	Jogging	Running	Walking
Boxing	1.0	0	0	0	0	0
Hand-clapping	0.31	0.58	0.11	0	0	0
Hand-waving	0	0	1.0	0	0	0
Jogging	0	0	0	0.69	0.25	0.06
Running	0	0	0	0.11	0.89	0
Walking	0	0	0	0	0	1.0

V. CONCLUSION

This paper shows that the VBHMM with DM of GWEM can be a useful tool for human action classification. First, the results have shown that a time series data of continuous feature vectors extracted from a human action video can be modeled by HMM with DPM of GWEM. Using the variational Bayesian inference approach, the researchers derived the approximate posterior distributions of all latent variables and parameters indicating a membership of class on the basis of the learning data. Second, we have derived the predictive distribution of the latent function corresponding to the new input vector by using both the existing training data and the new input vector. Next, we calculate the likelihood function for each class by using the predictive distribution corresponding to the new sample. Lastly, the study classifies the input video into the class which its likelihood function is maximized. The experimental results show that our method performs very well on public video datasets, such as the KTH dataset, more than others.

ACKNOWLEDGMENT

This study was supported by Korea Research Foundation (2014009398).

REFERENCES

[1] Z. Ghahramani, An Introduction to Hidden Markov Models and Bayesian Networks, International Journal of Pattern Recognition and Artificial Intelligence. Vol. 15(1), pp. 9-42, 2001.

[2] L. Bahl and F. Jelinek, Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition, IEEE Transaction on Information Theory, vol. 21, pp 404-411, 1975.

[3] B. H. Juang and L. R. Rabiner, Hidden Markov models for speech recognition, Technometrics, vol. 33, pp 251-272, 1991.

[4] P. Smyth, D. Heckerman, and M. I. Jordan, Probabilistic independent networks for hidden Markov probability models, Neural Computation, vol. 9, pp 227-269, 1997.

[5] Z. Ghahramani and M. I. Jordan, Factorial hidden Markov models, machine Learning, vol. 29, pp 245-273, 1997.

[6] M. J. Jordan, Z. Ghahramani, and L. K. saul, hidden Markov decision trees, In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, Advances in Neural Information Processing sytems, vol. 9, Cambridge, MA, 1997, MIT Press.

[7] Z. Ghahramani and G. E. Hinton, Variational learning for switching state-space models, Neural Computation, vol.24, no. 4, 2000.

[8] M. J. Beal, Variational Algorithm for Approximate Bayesian inference, A Thesis submitted for the degree of Doctor of Philosophy of the University of London, 2003.

[9] J. Yin and Y. Meng, Human Activity Recognition in Video using a Hierarchical Probabilistic Latent Model, Computer Vision and Pattern Recognition Workshop (CVPRW), San Francisco, CA, USA, 2010.

[10] Y. Tian, L. Cao, Z. Liu and Z. Zhang, "Hierarchical Filtered Motion for Action Recognition in Crowded Videos," IEEE Trans. On System, Man, and Cybernectics, Part C: Applications and Reviews, vol. 3, no. 2, pp. 1-11, 2011.

[11] M. Z. Uddin, N. D. Thang, J. T. Kim, and T.-S. Kim, "Human Activity Recognition Using Body Joint-Angle Features and Hidden Markov Model," ETRI Journal, vol. 33, no. 4, pp569-579, 2011.

[12] M. K. Gaikward, and Mr. V. Narawade, "HMM classifier for human activity recognition", Computer Science & Engineering: An International Journal, vol. 2, no. 4, pp 27-36, 2012

[13] L. Piyathilaka and S. Kodagoda, "Gaussian Mixture Based HMM for Human Daily Activity Recognition Using 3D Skeleton Features," 2013 IEEE 8th Conference on Industrial Electronics and Application, pp567-572,2013.

[14] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," pp1-11, 200.

Wan-Hyun Cho received both B.S. degree and M.S. degree from the Department of Mathematics, Chonnam National University, Korea in 1977 and 1981, respectively and Ph.D. degree from the Department of Statistics, Korea University, Korea in 1988. He is now teaching in Chonnam National University. His research interests are statistical modeling, pattern recognition, image processing, and medical image processing.

Soon-Ja Kang received both B.S. degree and M.S. degree from the Department of Mathematics, Chonnam National University, Korea in 1979 and 1981, respectively and Ph.D. degree from the Department of Mathematics, Seogang University, Korea in 1988. She is now teaching in Chonnam National University. Her research fields are mathematical education, advanced calculus, and education for the gifted children.

Sang-Kyoon Kim received the B.S., M.S. and Ph.D. degrees in Electronics Engineering, Mokpo National University, Korea in 1998, 2000 and 2015 respectively. From 2011 to 2015, he was a Visiting Professor in the Department of Information & Electronics Engineering, Mokpo National University, Korea. His research interests include image processing, pattern recognition and computer vision.

Soon-Young Park received B.S. degree in Electronics Engineering from Yonsei University, Korea in 1982 and M.S and Ph.D. degrees in Electrical and Computer Engineering from State University of New York at Buffalo, in 1986 and 1989, respectively. From 1989 to 1990 he was a Postdoctoral Research Fellow in the department of Electrical and Computer Engineering at the State University of New York at Buffalo. Since 1990, he has been a Professor with Department of Electronics Engineering, Mokpo National University, Korea. His research interests include image and video processing, image protection and authentication and image retrieval techniques.