EVENT RECOGNITION WITH TIME VARYING HIDDEN MARKOV MODEL

Zhaowen Wang[1], *Ercan E. Kuruoglu*[2], *Xiaokang Yang*[1], *Yi Xu*[1], *and Songyu Yu*[1]

[1]Institute of Image Communication and Information Processing, EE Dept
Shanghai Jiao Tong University, Shanghai, PRC 200240
[2]Istituto di Scienza e Tecnologie dell'Informazione (ISTI) -CNR
via G. Moruzzi 1, 56124, Pisa, Italy

ABSTRACT

Standard Hidden Markov Model (HMM) and the more general Dynamic Bayesian Network (DBN) models assume stationarity of state transition distribution. However, this assumption does not hold for many real life events of interest. In this paper, we propose a new time sequence model that extends HMM to time varying scenario. The time varying property is realized in our model by explicitly allowing the change of state transition density as the time spent in a particular state passes by. Rather than keeping transition densities at different time spots independent of each other, we exploit their temporal correlation by applying a hierarchical Dirichlet prior. This leads to a more robust time varying model, especially when training data are scarce. We also employ Markov Chain Monte Carlo (MCMC) sampling in learning the MAP estimate of time varying parameters, with a transition kernel incorporating linear optimization. The proposed model is applied to recognizing real video events, and is shown to outperform existing HMM-based methods.

Index Terms— event recognition, time varying, HMM, MCMC

1. INTRODUCTION

Recognizing events in video sequence has important applications in intelligent surveillance, traffic monitoring, human-computer interaction, robot learning, video summarization, and etc.

An event can be thought as a latent variable (e.g. an activity) that generates a time sequence of observations (e.g. 2D trajectory). The common approach to recognize an event comprises two steps: 1) extracting observation sequences from video; and 2) interpreting those sequences of a particular event with a model. Building such a model is usually difficult, since observation sequences are often featured with high dimension, multiple modality, variable duration and noise contamination.

Many efforts have been paid in previous works to bridge the gap between video observation and event model. Hidden Markov Model (HMM) and the more general Dynamic Bayesian Network (DBN) are popular tools in this field, because of their power of modeling statistical patterns evolving over time. Standard HMM and DBN models assume stationarity of state transition distribution, keeping the model structure and parameter constant all the time. However, the stationarity assumption does not hold for many real life events of interest. For example, two acquainted people will more likely split apart after they have talked for a while than after they have just met.

A few researchers have realized this limitation and tried to adapt HMM or DBN to the non-stationary scenario. In [1, 2], a non-stationary event model is decomposed into a cascade of stationary sub-DBNs, each of which has a distinct structure. The switching time between adjacent sub-DBNs is first determined by change detection, and then each sub-DBN can be constructed as in stationary case. The Non-stationary Hidden Semi-Markov Model (NHSMM) proposed in [3] augments HMM with time varying parameters, so that the transition probability to a new state varies in accordance with the time spent in current state. This amounts to build a HMM for each individual time epoch, and generally results in a more accurate description of non-stationary processes than the piecewise approximation in [1, 2]. However, NHSMM has an obvious drawback that the number of model parameters increases linearly with the length of maximum state duration. Such unrestricted growth of parameters will give rise to the problem of model over-fitting, especially when a large training set is unavailable.

In this paper we aim to solve the dilemma between model flexibility and robustness, and put forward a novel event model, called Time Varying HMM (TVHMM). Our model encodes non-stationarity into a finite sequence of time varying transition densities and can simulate state duration of infinite length. The temporal dependence between time varying parameters is taken into consideration by applying a

This work was supported in part by Shanghai Postdoctoral Foundation (06R214138), 863 (2006AA01Z124), NSFC (60502034, 60702044, 60828001) and the 111 Project (B07022). The second author gratefully acknowledges the financial support of the 111 Project jointly funded by Ministry of Education and State Administration of Foreign Experts Affairs, PRC.

hierarchical Dirichlet prior. In this way, transition densities of different time spots are optimized jointly and over-fitting can be avoided. We employ an efficient Markov Chain Monte Carlo (MCMC) method in learning the time varying parameters, which gives a better MAP estimate than simple EM algorithm. The proposed model is shown to perform better than other HMM-based methods in the recognition of real life activities.

2. TIME VARYING HMM

2.1. Time Varying Transition Matrix

Following the notations in HMM, we denote here the hidden state at time t as q_t, and its observation as O_t. The hidden state can have one of N possible values $\{S_1, ..., S_N\}$, and the observation value can be either discrete or continuous. The distribution of observation conditioned on hidden state, $P(O_t|q_t)$, is known as the measurement model. And the state transition distribution $P(q_{t+1}|q_t)$, also characterized by a $N \times N$ transition probability matrix $\mathbf{A}$, is the dynamic model. In TVHMM, the transition matrix $\mathbf{A}$ is defined to change as the time spent so far in current state elapses, and falls into one of the M predefined time-varying stages:

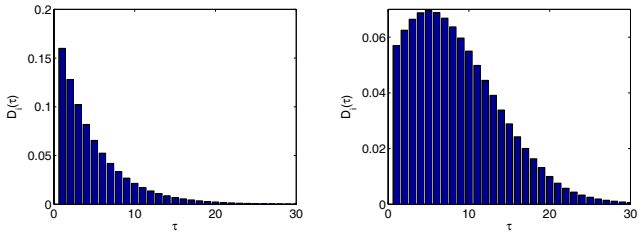
$$\begin{aligned} \mathbf{A}(\tau) &= \{a_{ij}(m)\} \quad i,j = 1...N \\ m &= \min(\tau, M) \end{aligned} \tag{1}$$

where τ is the length of time that current state has been kept, and $a_{ij}(m)$ is the probability of transiting from state S_i to state S_j in the m'th time varying stage.

Eq. (1) actually defines an individual transition matrix for each of the first M time epochs within a state duration, and uses $\{a_{ij}(M)\}$ as a constant transition matrix afterwards. This can be deemed as a generalization of NHSMM [3], whose non-stationary formulation can be recovered by setting the diagonal elements of $\{a_{ij}(M)\}$ to zero. With our scheme, the maximum state duration length is no longer constrained by the stage number M. This can be seen by explicitly evaluating the state duration distribution with the sequence of self-transition probabilities $\{a_{ii}(m)\}_{m=1...M}$ in Eq. (1):

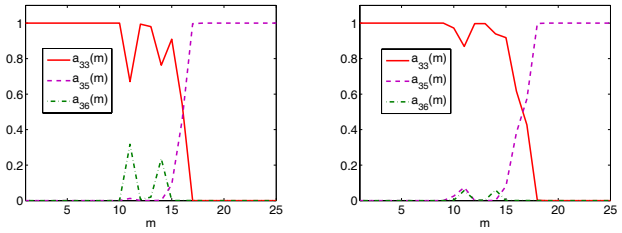
$$\mathcal{D}_i(\tau) = \begin{cases} \prod_{m=1}^{\tau-1} a_{ii}(m)(1-a_{ii}(\tau)), & \tau = 1...M \\ \prod_{m=1}^{M-1} a_{ii}(m)[a_{ii}(M)]^{\tau-M}(1-a_{ii}(M)), & \tau > M \end{cases} \tag{2}$$

where $\mathcal{D}_i(\tau)$ is the probability of staying in state i for exactly τ time epochs. With different choices of $a_{ii}(m)$'s, $\mathcal{D}_i(\tau)$ can take any probability for $\tau < M$, and diminishes exponentially for $\tau \geq M$. Therefore, TVHMM allows a state to survive for an arbitrarily long time, and the state duration distribution can be modeled accurately up to the first M time epochs. Two examples are illustrated in Fig. 1.



(a) (b)

Fig. 1. State duration distribution in TVHMM for (a) constant self-transition probability; (b) linearly decreasing self-transition probability.



(a) (b)

Fig. 2. Temporally independent transition probabilities versus time varying stage, learned with (a) 3 cases; (b) 10 cases.

2.2. Hierarchical Dirichlet Prior

To construct a time varying model as specified in Eq. (1), we are in fact required to build a series of M HMM's, each of which has a distinct $N \times N$ transition matrix. With such a great number of parameters, the model is in danger of over-fitting, especially when training data are scarce (which is often the case for video applications). For example, taking respectively 3 and 10 training cases from the data in Sec. 4, we construct two such time varying models and plot part of their parameters in Fig. 2. The parameter set estimated using only 3 cases is more jittered in time axis due to the statistical noise in small training data set. While the more accurate parameters estimated with 10 cases change smoothly through time. This should be a general property for most video-based applications, where the sampling interval is rather short compared with the length of any practical event.

We hereby propose to restrict the degree of freedom of $a_{ij}(m)$'s by imposing a hierarchical Dirichlet [4] prior:

$$\mathbf{a}_i(m) \sim Dir(\sigma \mathbf{a}_i(m-1)) \quad i = 1...N, m = 2...M \tag{3}$$

where $\mathbf{a}_i(m) = [a_{i1}(m), a_{i2}(m), ..., a_{iN}(m)]^T$, $Dir(\cdot)$ denotes Dirichlet distribution, and σ is a smoothing constant chosen in advance. Dirichlet distribution is well-suited to model probability density as it is supported over the probability simplex. And hierarchical Dirichlet distribution conveys the information that the current transition density of TVHMM is very likely to be close to the one at previous time epoch,

which conforms to the smooth parameter changing assumption mentioned above. With the dependence between temporally adjacent parameters taken into account, the statistical uncertainty introduced by small training set can be compensated, and the balance between model flexibility and robustness is achieved in TVHMM.

3. LEARNING MAP PARAMETERS

3.1. Dynamic Bayesian Network Representation

Since the transition probability in TVHMM depends on the duration of current state, we are required to keep the history of hidden state up to M epochs in analyzing our model. To simplify the computation, we propose to restore the first order Markovian property of TVHMM by designating an auxiliary hidden variable d_t to the time varying stage index. Then the model can be represented in the form of dynamic Bayesian network with a new dynamic model governing the joint transition distribution of q_t and d_t:

$$\begin{aligned} & P(q_t^j, d_t^n | q_{t-1}^i, d_{t-1}^m) \\ = \; & P(q_t^j | q_{t-1}^i, d_{t-1}^m) P(d_t^n | q_{t-1}^i, d_{t-1}^m, q_t^j) \\ = \; & \begin{cases} a_{ij}(m)\delta(n - \min(m+1, M)) & i = j \\ a_{ij}(m)\delta(n-1) & i \neq j \end{cases} \end{aligned} \tag{4}$$

where q_t^i stands for $q_t = S_i$, d_t^m stands for $d_t = m$, and $\delta(\cdot)$ is delta function. Note that d_t actually transits in a deterministic way, so its introduction will not expand the dimension of hidden state space too much.

With the model structure defined in Eq. (4), we can apply the general methods of learning and inferring DBN [5] on the proposed TVHMM. The MAP estimate of model parameter can be found using EM algorithm, which iteratively optimizes the following objective:

$$\begin{aligned} & \max_{\{\mathbf{a}_i(m)\}} P\left(\{\mathbf{a}_i(m)\} | \{q_t, d_t, O_t\}\right) \\ \rightarrow \; & \max_{\{\mathbf{a}_i(m)\}} P\left(\{q_t, d_t, O_t\} | \{\mathbf{a}_i(m)\}\right) P(\{\mathbf{a}_i(m)\}) \end{aligned} \tag{5}$$

where $\{\mathbf{a}_i(m)\}$ stands for the time varying parameter set $\{\mathbf{a}_i(1)...\mathbf{a}_i(M)\}$, $\{q_t, d_t, O_t\}$ is the complete data sequence expected with old parameters, and $P(\{\mathbf{a}_i(m)\})$ is the joint parameter distribution derived from Eq. (3).

3.2. MCMC Sampling

The joint optimization of Eq. (5) is intractable. The authors of [4] proposed to find an approximated solution with Linear Minimum Mean Square Error (LMMSE) estimator. However, the solution given by LMMSE is often far from true MAP parameter, due to the high dimensionality of the parameter set $\{\mathbf{a}_i(m)\}$. Therefore, we propose to simulate the full posterior of $P\left(\{\mathbf{a}_i(m)\} | \{q_t, d_t, O_t\}\right)$ with MCMC, and use the MCMC sample with the highest posterior probability as our MAP estimate.

Metropolis-Hastings algorithm [6] is employed in the MCMC sampling. Given the current parameter set $\{\mathbf{a}_i(m)\}$, a new sample $\{\mathbf{a}_i^*(m)\}$ is generated by sequentially drawing $\mathbf{a}_i^*(1)$, $\mathbf{a}_i^*(2)$, ... $\mathbf{a}_i^*(M)$ from Dirichlet distributions centered at the output of LMMSE:

$$\begin{aligned} \mathbf{a}_i^*(m) & \sim P(\mathbf{a}_i^*(m) | \mathbf{a}_i^*(m-1), \mathbf{a}_i(m+1)) \\ & \sim Dir[\sigma \mathbf{a}_i^*(m-1) + (\sigma+1)\mathbf{a}_i(m+1) + \mathbf{ss}_i(m)] \end{aligned} \tag{6}$$

where $\mathbf{ss}_i(m)$ is an N-dimensional vector whose jth element is the expected value of $\sum_t P(q_t^j, q_{t-1}^i, d_{t-1}^m)$. The terms of $\mathbf{a}_i^*(m-1)$ and $\mathbf{a}_i(m+1)$ are dropped from Eq. (6) when they are not defined for $m = 1$ and $m = M$. Thus the overall transition kernel is:

$$\begin{aligned} & P(\{\mathbf{a}_i^*(m)\} | \{\mathbf{a}_i(m)\}) \\ = \; & P(\mathbf{a}_i^*(1) | \mathbf{a}_i(2)) P(\mathbf{a}_i^*(M) | \mathbf{a}_i^*(M-1)) \\ & \times \prod_{m=2}^{M-1} P(\mathbf{a}_i^*(m) | \mathbf{a}_i^*(m-1), \mathbf{a}_i(m+1)) \end{aligned} \tag{7}$$

This transition kernel can efficiently explore parameter space with the knowledge provided by LMMSE; at the same time, the uncertainty in sampling grants it a potential to find better MAP estimate. This is why MCMC method is preferred here.

The new sample $\{\mathbf{a}_i^*(m)\}$ is accepted with probability

$$\begin{aligned} P_A = \min \Bigg(1, & \frac{P(\{q_t, d_t, O_t\} | \{\mathbf{a}_i^*(m)\})}{P(\{q_t, d_t, O_t\} | \{\mathbf{a}_i(m)\})} \times \\ & \frac{P(\{\mathbf{a}_i^*(m)\}) P(\{\mathbf{a}_i(m)\} | \{\mathbf{a}_i^*(m)\})}{P(\{\mathbf{a}_i(m)\}) P(\{\mathbf{a}_i^*(m)\} | \{\mathbf{a}_i(m)\})} \Bigg) \end{aligned} \tag{8}$$

If accepted, $\{\mathbf{a}_i^*(m)\}$ will take the place of $\{\mathbf{a}_i(m)\}$ in future sampling. The sampling process proceeds until the distribution of accepted samples converges to the parameter posterior of Eq. (5).

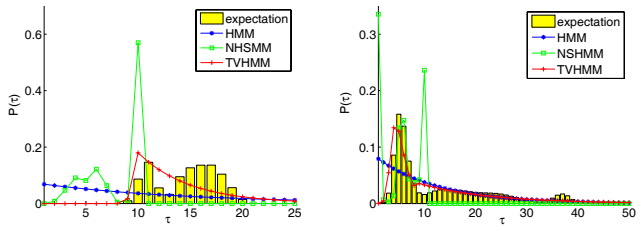
4. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of TVHMM through two real event recognition tasks.

We first try to recognize three basic human motions - running, walking and jumping - with articulation trajectory data provided by [7]. The dimension-reduced data serve as observation here and states are defined as cluster centers found by K-means. For each type of event, a model is learned with 4 training cases. In the testing phase, 6 unseen cases of each type are fed into all the three models, and the normalized likelihood is used as recognition probability. The recognition rates of the correct event averaged over all test cases are listed in Table 1, and the results of HMM and NHSMM are also shown for comparison. The proposed TVHMM achieves much higher recognition rate than the two competitors.

Table 1. Average recognition rate for human motion.

model	running	walking	jumping
HMM	0.994	0.551	0.414
NHSMM	0.996	0.791	0.858
TVHMM	0.999	0.856	0.984



(a) state 3, running event (b) state 6, walking event

Fig. 3. State duration distribution.

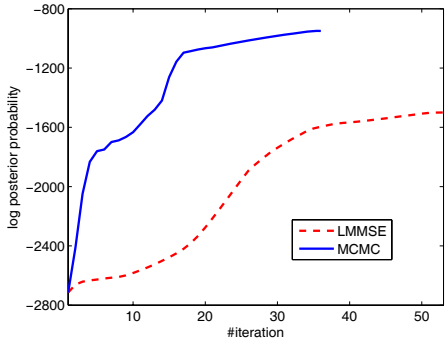


Fig. 4. MAP parameter probability versus EM iteration.

The superiority of TVHMM is attributed to its flexibility in modeling state duration distribution, as visualized in Fig. 3. The state duration distributions learnt with all the three models are compared with the 'truth' distribution estimated from the HMM-inferred most likely state sequence. It is shown that both NHSMM and TVHMM are better at modeling non-exponential duration distribution than HMM. TVHMM performs even better if the distribution spans beyond stage number M (set to 10 for both TVHMM and NHSMM here).

In our second experiment, moving trajectories of pedestrians [8] and vehicles [9] are subject to recognition. We build the models in a similar way as in the first experiment and get the recognition rates in Table 2. Although the two motion patterns are difficult to distinguish using HMM and NHSMM, our model still gives higher recognition rate. The underlying reason is revealed in Fig. 4, which plots the ascent of MAP parameter probability during EM learning. It is shown that, with our MCMC scheme employed in the maximization step of EM, the posterior probability converges to a higher value in fewer iterations than when LMMSE is used.

Table 2. Average recognition rate for moving object.

model	pedestrian	vehicle
HMM	0.582	0.542
NHSMM	0.588	0.568
TVHMM	0.610	0.639

5. CONCLUSION

We present a time varying version of HMM in which the transition matrix varies with the time spent in the current state. Transition density is defined for state duration of any length so that our model is capable of simulating duration distributions of diversified forms and infinite length. The temporal coherence of time varying parameters is considered by introducing a hierarchical Dirichlet prior and the MAP estimate is obtained with the aid of MCMC sampling, where the transition kernel is optimized by LMMSE. The proposed model is evaluated on two real event recognition problems, and shows substantially improved recognition rate over existing methods.

6. REFERENCES

[1] S. H. Nielsen and T. D. Nielsen, "Adapting bayes network structures to non-stationary domains," in *the 3rd European Workshop on Probabilistic Graphical Models*, 2006, pp. 223–230.

[2] A. Rao, A. O. Hero III, D. J. States, and J. D. Engel, "Inferring time-varying network topologies from gene expression data," *Journal on Bioinformatics and Systems Biology, EURASIP*, pp. 7–7, 2007.

[3] E. Marhasev, M. Hadad, and G. A. Kaminka, "Non-stationary hidden semi markov models in activity recognition," in *Workshop on Modeling Others from Observations*. AAAI, 2006.

[4] S. Veeramachaneni, D. Sona, and P. Avesani, "Hierarchical dirichlet model for document classification," in *International Conference on Machine Learning, Bonn Germany*, 2005, vol. 119, pp. 928–935.

[5] K. Murphy, *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. thesis, UC Berkeley, Computer Science Division, July 2002.

[6] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, pp. 97C109, 1970.

[7] MOCAP data set, http://mocap.cs.cmu.edu.

[8] CAVIAR, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[9] NGSIM program, http://ngsim.fhwa.dot.gov.