Dynamic Gesture Recognition Based on Dynamic Bayesian Networks

WEI-HUA ANDREW WANG*, CHUN-LIANG TUNG
Department of Industrial Engineering and Enterprise Information
Tunghai University
No. 181 Section 3, Taichung Harbor Road, Taichung, 407-04, Taiwan
wangwh@thu.edu.tw*, cltung@ncut.edu.tw

Abstract: Techniques for recognizing and matching dynamic human gestures are becoming increasingly important with the CCTV surveillance system. To provide consistent dynamic gesture recognition system, Hierarchical Dynamic Vision System (HDVS) which based on dynamic Bayesian networks (DBNs) is proposed for automatically identifying human gestures in this paper. DBNs, directed graphical models of stochastic process, generalize HMM by representing the hidden and observed state in terms of state variables in which can have more complex interdependencies than HMMs systems do. In this paper, hierarchical hidden Markov model (HHMM) is used as the underlying topology in the proposed dynamic system to recognize human gestures with motion trajectories in an indoor scene. A hierarchical HMM, represented by DBN, is structured multi-level stochastic processes. In the low-level processing, both motion trajectories and motion directions generated from hand part is used as features after watershed segmentation. In the high-level processing, human gestures are automatically recognized form the inference of HHMM-DBNs. In this paper, we focus on the following aspects of both system modeling and high-level processing: (1) Completed DBNs structure with HHMM, (2) approaches to human gesture recognition.

Keywords: Hierarchical Dynamic Vision System, dynamic Bayesian network

1 Introduction

CCTV, closed-circuit television, is widely used for surveillance in areas which need security, such as workshop blocks, public transportations and military installations. In these security areas, one needs advanced technology for recognizing particular events or human activities. Due to the requirements of many automatic surveillance applications, e.g. intrusion detection and hand gestures recognition, automatic event detection in video streams has been studied extensively in recent years. A central issue in automatic event detection is to recognize the hand gestures matching event generations. Two types of recognition are studied in hand gesture recognition, static posture recognition and dynamic gesture recognition. This paper focuses on the dynamic gesture recognition with time-temporal variation. Therefore, both Hierarchical Hidden Markov Models (HHMMs) and Dynamic Bayesian Networks (DBNs) are used for modelling dynamic recognition system in order to investigate the effects of dynamic gesture recognition.

With the successful experience and enlightenment in speech recognition [16], hidden Markov models [15] has been extensively applied to activity or gesture recognition. A hidden Markov model (HMM), a statistical model for ordered data, is assumed to be a Markov process with unknown parameters and which has become the one of the most significant methods for modeling stochastic processes and sequences in applications, e.g. speech recognition, handwriting recognition, and natural language modelling [8, 9, 14, 15]. In activity recognition, the HMM has been employed in many approaches such as Parametric Hidden Markov Models which recognizes the hand gestures by the parameters during the spatial execution period [18]. In most of the recognition applications, the topology of a HMM model is predefined and the model parameters are estimated by an EM procedure [4] such as Baum-Welch algorithm [1] which is used to find the unknown parameters of a HMM by using the forward-backward algorithm.

A hierarchical hidden Markov model (HHMM) [6], a generalization of the HMM, is implemented for modelling structured multi-level stochastic processes, e.g. activity recognition [10] and DNA sequences [7]. The difference between a standard HMM and a HHMM is that individual states in the hierarchical model can traverse to a sequence of production states, whereas each state in the standard model corresponds is a production state that contains a single observation [2]. Dynamic Bayesian networks (DBNs) [5], extended of Bayesian

networks and directed graphical models of stochastic process, models probability distributions among random variables over time. Inspired by the earlier work, dynamic stochastic processes of a hierarchical model can be modeled in a dynamic Bayesian network [11]. With the idea of Murphy and Paskin [12], a HHMM can be converted to a DBN and apply both inference and learning techniques of a general DBN in a model.

However, accuracy and consistency of gesture recognition are still insufficient. From this standpoint, the present HMM can not handle large number of states efficiently. One solution to this problem is offered by HHMM which provides hierarchically structured to improve processing time and extraction accuracy. To reduce the training time in the dynamic gesture recognition system, HHMM is represented to DBN. Moreover, standard inference techniques of Bayesian networks, e.g. junction-tree algorithm, can be used to replace Baum-Welch algorithm for more efficiency in the inference process of DBN. The proposed recognition approach of the HHMM-based DBN was performed for recognizing three different types of hand gestures, and its techniques and results are reported here.

This study demonstrates that the DBN representation of the HHMM can provide an accuracy, consistency and robust dynamic framework for the dynamic gesture recognition system. The results indicate the HHMM-based DBN model would be suitable to cope with so-called high level recognition in a dynamic system.

2 Hierarchical Hidden Markov Models

The DBN representation of the HHMM for recognizing hand gestures is the aim of the study. A HHMM is a generalization of a HMM that is implemented for modeling structured multi-level stochastic processes. In a HHMM, sequence generations are recursively from a root state to one of the substates of a state, and each substate is also an HHMM. During the sequence generation process of a HHMM, the process of recursive activations ends while it reaches a special state termed a production state.

HHMM can be represented by a tree-like structure. Therefore, production states are same as the leaf nodes in a tree structure and are the only states which actually emit output symbols like a HMM. There are two types of states in a HHMM. One is production state, and the other one is internal state. The internal state does not emit observable symbols, and which only emits substates. The internal state activates, moreover, a substate termed a vertical transition. Within the same level of a state transition, we term the relationship between states a horizontal transition. When a vertical transition is completed, a horizontal transition is then executed. There is a state called final state in each level of HHMM, except the top level, which is represented the end index during a recursive activation.

In this present model description, the notations are similar to Fine et al. [6] for achieving consistency. The discrete model $\lambda(\xi,\vartheta)$ is defined with 2-tuple: the model structure ξ and a set of parameters ϑ. The model structure ξ, a hierarchical structure, is consisted of several elements: the model depth $\xi(d)$, the internal state q_i^d, and the production state q_i^D .The model depth d, where $d < D$ is specified with the model. The internal state in a level is denoted by q_i^d, where $d \in \{1,2\}$, and the production state is denoted by q_i^D, where $D = 3$. The number of substates of an internal state is denoted by $|q_i^d|$. The parameters ϑ contain the state transition probability functional $A^d(\cdot)$, initial probability functional $\pi^d(\cdot)$, and observation probability functional $B^D(\cdot)$. Following by Fine et al. [6], the state transition probability matrix $A^d = (a_{i,j}^d) = P\{q_j^{d+1} | q_i^{d+1}\}$ represents the probability of the substates of q^d. The initial probability $\pi_i^d = P(q_i^{d+1} | q^d)$ represents the initial probability of the substates of q^d. The observation probability represents the observation probability in the production state q^D.

Therefore, a recognition model based on the HHMM consists of three levels to recognize different gestures. In the HHMM recognition model, top one level called the perception level, second level called the motion level, and third level called the feature level as shown in Fig.1. The state in perception level is predefined goals which are extracted from the sequence of hand gestures in motion level.

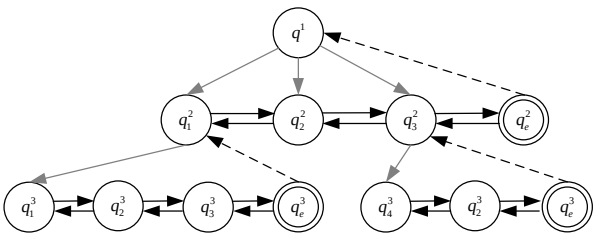


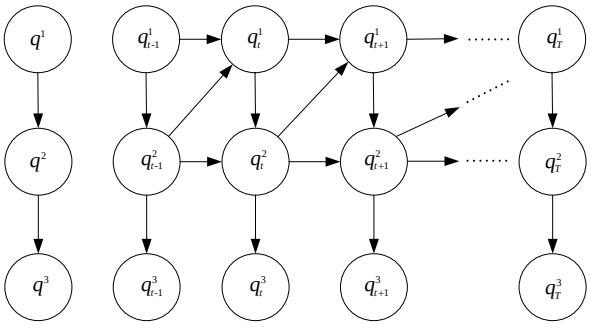
Fig.1 The representation of the HHMM for recognizing hand gesture activities

3 Modeling and Recognizing Hand Gestures

DBNs, dynamic Bayesian networks, are directed graphical models of stochastic process, which represent the hidden and observed state with more complex interdependencies inside a stochastic model. DBNs can be seen an extension of Bayesian networks to achieve temporal process over time for modeling a dynamic model. A DBN $(B_0, B_{\rightarrow})$ [13] is consisted of two components: a Bayesian network $B_0 = P\{q_t\}$ and a two-slice temporal Bayes net $B_{\rightarrow} = P\{q_t^i \,|\, q_{t-1}^i\}$, where q_t^i is a state variable at time slice t and q_{t-1}^i is the parent state of q_t^i as shown in Fig.2. Therefore, a DBN is denoted as Eq.1:

$$(B_0, B_{\rightarrow}) = P\{q_t \,|\, q_{t-1}\} = \prod_{i=1}^{N} P\{q_t^i \,|\, \mathrm{Pa}(q_t^i)\} \quad (1)$$

where $\mathrm{Pa}(q_t^i)$ are the parents of q_t^i in the DAG. The parent state of a state can be in either one of same time slice and previous time slice.



(B_0) $(B_{\rightarrow})$

Fig.2 A DBN is consisted of a BN (B_0) and a temporal Bayesian network with T-slice $(B_{\rightarrow})$

3.1 Model framework

The HHMM can be represented as the DBN [11, 13] as shown in Fig.3. In our DBN model, the internal state is denoted by $q_t^d\,(d \in \{1,2\})$, where t is the time index and d is the level index. The index of top level is 1 and of the production state (bottom level) is D. The end state is denoted by $e_t^d\,(d \in \{1,\ldots,D\})$ for terminating the state activation process. The conditional probability distributions between the states of two time slices $P\{f_t^D \,|\, f_{t-1}^D,\ m_t^{D-1},\ e_{t-1}^D\}$ and the termination probability of the production state $P\{e_t^D \,|\, m_t^{D-1},\ f_t^D\}$ in feature level are expressed partially as Eq.2 and Eq.3:

$$P\{f_t^D = j \,|\, f_{t-1}^D = i, m_t^{D-1} = k, e_{t-1}^D\} = \begin{cases} A_{i,j,k}^D, & if\ e_{t-1}^D = 0 \\ \pi_{j,k}^D, & if\ e_{t-1}^D = 1 \end{cases} \quad (2)$$

$$P\{e_t^D = 1 \,|\, m_t^{D-1} = k,\ f_t^D = i\} = A_{i,k,end}^D \quad (3)$$

where $f_t^D = \{1,2,\cdots,8\} = i, j$ and $m_t^{D-1} = \{1,2,3\} = k$. The matrix $A_{i,j,k}^D$ represents the transition probability from f_{t-1}^D to f_t^D given the parent state of f_t^D. Similarly, $\pi_{j,k}^D$ represents the initial probability of f_t^D given the parent. $A_{i,k,end}^D$ represents the termination probability of f_t^D given the parent. The conditional probability distributions between the states of two time slices $P\{m_t^{D-1} \,|\, m_{t-1}^{D-1},\ r_t^{D-2},\ e_{t-1}^{D-1,D}\}$ and the termination probability of the production state $P\{e_t^{D-1} \,|\, m_t^{D-1},\ r_t^{D-2},\ e_t^D\}$ in motion level are expressed partially as Eq.4 and Eq.5:

$$P\{m_t^{D-1} = j \,|\, m_{t-1}^{D-1} = i, r_t^{D-2} = k, e_{t-1}^{D-1,D}\} = \begin{cases} \delta(i,j,k), & if\ e_{t-1}^{D-1} = 0\ and\ e_{t-1}^D = 0 \\ A_{i,j,k}^{D-1}, & if\ e_{t-1}^{D-1} = 0\ and\ e_{t-1}^D = 1 \\ \pi_{j,k}^D, & if\ e_{t-1}^{D-1} = 1\ and\ e_{t-1}^D = 1 \end{cases} \quad (4)$$

$$P\{e_t^{D-1} = 1 \,|\, m_t^{D-1} = i, r_t^{D-2} = k, e_t^D\} = \begin{cases} 0, & if\ e_t^D = 0 \\ A_{i,k,end}^{D-1}, & if\ e_t^D = 1 \end{cases} \quad (5)$$

where $\delta(\cdot)$ is the Dirac delta function and $A_{i,j,k}^{D-1}$ is the transition matrix for the motion level. Based on the parent states, the probability of a particular state can be calculated with Eq. 2, Eq. 3, Eq. 4, and Eq. 5.

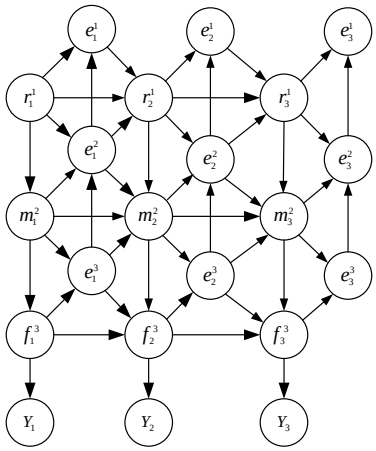
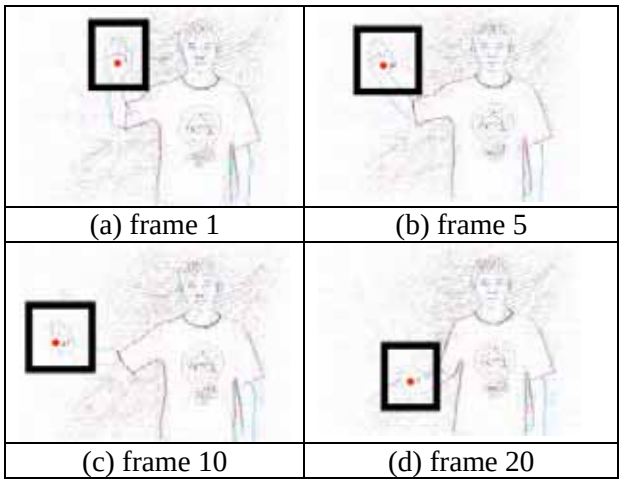


Fig.3 The proposed DBN model with 3-level

3.2 Data acquisition

In our proposed system, the motion trajectories of hand gestures are detected in the image sequence by a special tag as shown in Fig.4. There are three types of information are stored in the observation vector with 3-tuple $O_T = < \gamma_t, \Delta\ell_t, \Delta\upsilon_t >$, the hand position $\gamma_t = (S_x^t, S_y^t)$, the hand direction $\Delta\ell_t$, and the velocity $\Delta\upsilon_t$. The central point of the hand position γ in a video frame is recorded with S_x and S_y, where $1 \le x \le 320$ and $1 \le y \le 240$. In the study, a scale schematic direction representation of a single motion tag is used to measure the change direction of a motion trajectory. The moved direction of the hand (represented with tag) is in one of eight directions, including the cardinal directions and oblique directions $\ell_t^N (N \in \{1,2\cdots 8\})$ as shown in Fig.5 and Fig.6. Therefore, the hand moved direction in a frame is denoted by $\Delta\ell_t = \ell_t^{1:8}$, where $\ell_t^{1:8}$ is one of the eight directions.



(a) frame 1 (b) frame 5

(c) frame 10 (d) frame 20

Fig.4 The trajectory of the hand moved is detected with a tag in each frame

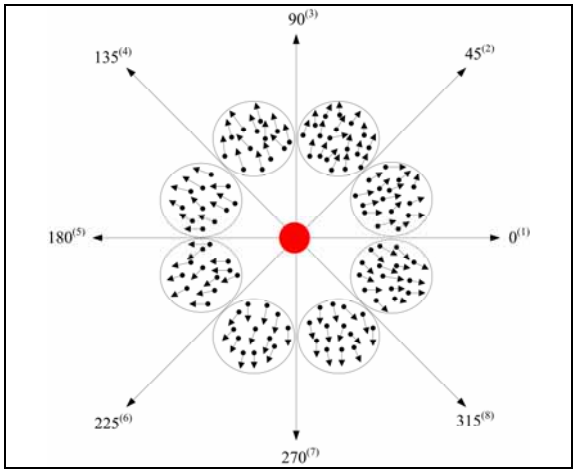


Fig.5 A scale schematic direction representation for a single motion tag

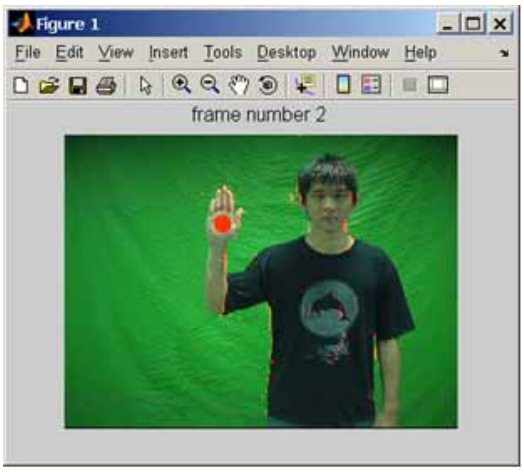


Fig.6 a single motion tag

3.3 Inference

The goal of inference in a DBN model is to compute the posterior probability $P\{q_t^d | o_{1:t}\}$, where q_t^d is a random variable in time t and $o_{1:t}$ present an observation sequence. To achieve the goal of inference, four main computations have to be calculated (Murphy, 2002): Filtering, Smoothing, Prediction, and Decoding. Filtering, $P\{q_t^d | o_{1:t}\}$, is used for monitoring the state over time. Smoothing, $P\{q_{t-k}^d | o_{1:t}\}$ $0 < k < t$, estimates the state of the past. Prediction, $P\{q_{t+k}^d | o_{1:t}\}$ $k > 0$, predicts the future of states with the estimation of either the current states and the past states. Decoding, $q_{1:t}^* = \arg \max_{q_{1:t}^d} P(q_{1:t}^d | o_{1:t})$, finds out the most likely explanation.

In the proposed HHMM-based DBN model, forward algorithm is used to estimate the belief state recursively during the filtering process (Rett and Dias, 2006). The estimation of the posterior probability $P\{q_t^d | o_{1:t}\}$, given observation sequences, is calculated recursively by Bayes rule as Eq.6:

$$P\{q_t^d, e_t^d | o_{1:t}\} = P\{o_t | q_t^d, e_t^d\} \sum_{q_{t-1}^d} P\{q_t^d, e_t^d | q_{t-1}^d\} P\{q_{t-1}^d | o_{1:t-1}\} \quad (6)$$

where, $d \in \{1,2,3\}$ is an indicator of the model level, and e_t^d is a terminating factor of the state activation process.

To find out the most likely sequence of hidden states, the best solution $q_{1:t}^*$ is found by max Eq.6. Therefore, Eq.6 can be rewritten for the purpose as shown in Eq.7.

$$q^*_{1:t} = \arg \max_{q^d_{1:t}} P(q^d_{1:t}, e^d_{1:t} | o_{1:t})$$
$$= P\{o_t | q^d_t, e^d_t\} \max_{q^d_{1:t}} (P\{q^d_t, e^d_t | q^d_{t-1}\} P\{q^d_{1:t-1} | o_{1:t-1}\}) \quad (7)$$

In the learning phase, given observation sequence $o_{1:t}$, the posterior probability of the parameters Θ of the HHMM-based DBN were estimated iteratively by Expectation- Maximization (EM) algorithm [2, 4] as shown in Eq.8, Eq.9, and Eq.10.

$$\Theta^* = \arg \max_{\Theta} P\{\Theta, q^d_t, e^d_t | o_{1:t}\} \quad (8)$$

where, Θ^* are the maximum likelihood parameters.

$$E\text{-}step: f^t(q^d_t, e^d_t) = P\{q^d_t, e^d_t | o_{1:t}, \Theta^t\} \quad (9)$$

$$M\text{-}step: \Theta^{t+1} = \arg \max_{\Theta} [Q^t(\Theta) + \log P(\Theta)] \quad (10)$$

where, $Q^t(\Theta)$ is calculated in the *E-step* (Eq.9) by evaluating $f^t(q^d_t, e^d_t)$ using Θ^t, whereas $Q^t(\Theta)$ is optimized in the *M-step* (Eq.10) to obtain the new Θ^{t+1}.

4 Experimental Results

The experiment was conducted to recognize the human hand gesture with motion trajectories in an indoor scene. To test the approach, a HHMM-based DBN with 3-levels was introduced to recognize the different motion types of hand gestures as shown in Table 1. The frame size of video data was 320×240 captured by Fujifilm digital camera.

In this paper, three types of motion were considered, namely, circle-motion with anticlockwise, square-motion with anticlockwise, and triangle-motion with anticlockwise. For each of the different motion, 20 samples were used for the experiment, and 30 frames of each sample were captured in one minute. In both the model learning and testing phase, half of the samples were used for parameters estimating and the other samples for testing. The preliminary recognition results are summarized in Table 2.

Comparing the recognition rates, 10 samples of triangle-motion are recognized exactly. No significant differences in recognition rates between circle-motion and square-motion recognition are found, and both recognition rates are lower than triangle-motion. This may be due to the similarity of trajectory path between circle-motion and square-motion. To overcome the problem, a similarity measure of trajectory, i.e. Euclidean distance (Zhang et al., 2006), can be considered in the classification phase. From these results, thus, HHMM-based DBN may be a useful approach for recognizing dynamic hand gestures.

Table 1 The proposed DBN model and different motion types.

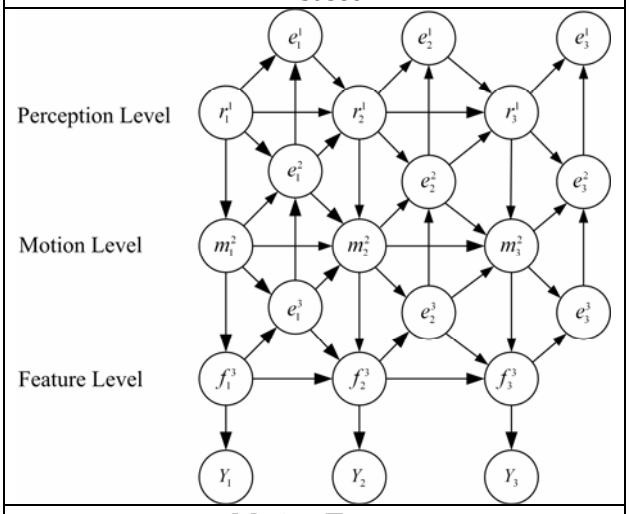
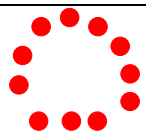
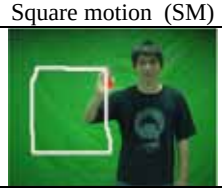
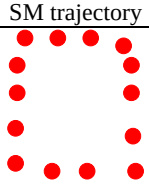
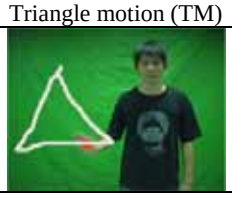
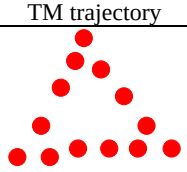
HHMM-based DBN	
	
Motion Types	
Circle motion (CM)	CM trajectory
	
Square motion (SM)	SM trajectory
	
Triangle motion (TM)	TM trajectory
	

Table 2 Gestures recognition rates

Recognition Rates			
	CM	SM	TM
CM	60%	40%	
SM	30%	70%	
TM			100%

5 Conclusions

In conclusion, we propose the dynamic model of the HHMM topology of the DBN with three levels shown in Figure 3 to recognize the three dynamic hand gestures in an offline dynamic recognition system. The results indicate the dynamic Bayesian networks base on the hierarchical hidden Markov model would be suitable to cope with so-called high level recognition in a dynamic system.

The proposed approach may be embedded to a security system for attack early warning. For this purpose, particle filtering would be used to replace the forward filtering in the proposed approach. To improve the recognition rates, future research would aim to study structure learning which comes up with a proper structure to DBN for dynamic gesture recognition.

References:

[1] Baum, L.E., Petrie, T., Soules, G. and Weiss, N, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, Vol.41, No.1, 1970, pp.164-171.

[2] Chou, L.Y., Techniques to incorporate the benefits of a Hierarchy in a modified hidden Markov model. In *Proc. of the COLING/ACL Main Conference Poster Sessions*, 2006, pp. 120-127.

[3] Dellaert, F., *The Expectation Maximization Algorithm*. Technical Report, College of Computing, Georgia Institute of Technology, 2002

[4] Dempster, A.P., Laird, N.M. and Rubin, D.B., Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, Vol.39, 1977, pp.1-38.

[5] Dean, T., Kanazawa, K., A model for reasoning about persistence and causation. *Artificial Intelligence*, Vol.93, 1989, pp.1-27.

[6] Fine, S., Singer, Y. and Tishby, N., The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, Vol.32, 1998, pp.41-62.

[7] Hu, M., Ingram, C., Sirski, M., Pal, C., Swamy, S. and Patten, C., *A Hierarchical HMM Implementation for Vertebrate Gene Splice Site Prediction.* Technical Report, Dept. Computer Science, Univ. Waterloo, 2000.

[8] Jelinek, F., *Self-organized language modeling for speech recognition*. Technical Report, IBM T.J. Watson Research Center, 1985.

[9] Kinjo, T. and Funaki, K., HMM Speech Recognition Based on Complex Speech Analysis. In *Proc. IEEE Conf. Industrial Electronics*, 2006, pp. 3477-3480.

[10] Kawanaka, D., Okatani, T. and Deguchi, K., HHMM Based Recognition of Human Activity. *IEICE Trans Inf. & Syst.*, Vol.89, No.7, 2006, pp.2180-2185.

[11] Li, Z., Hofemann, N., Fritsch, J. and Sagerer, G., Hierarchical Modeling and Recognition of Manipulative Gesture, In *First Workshop On Modeling People and Human Interaction*, 2005.

[12] Murphy, K. and Paskin, M., Linear time inference in hierarchical HMM. In *Proc. Neural Information Processing Systems*, Vol.14, 2001, pp.833-840.

[13] Murphy, K.P., *Dynamic Bayesian networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, 2002.

[14] Nag, R., Wong, K.H. and Fallside, F., Script recognition using hidden Markov models. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing*, 1985, pp.2071-2074.

[15] Rabiner, L.R. and Juang, B.H., *An introduction to hidden Markov models*. IEEE ASSP Magzine, Vol.3, No.1, 1986, pp.4-16.

[16] Rabiner, L.R., A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. IEEE*, Vol.77, No.2, 1989, pp.257-285.

[17] Rett, J. and Dias J., Gesture Recognition Using a Marionette Model and Dynamic Bayesian Networks (DBNs). In *Proc. Int. Conf. on Image Analysis and Recognition*, 2006, pp.69-80.

[18] Wilson, A.D. and Bobick, A.F., Parametric Hidden Markov Models for Gesture Recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.21, No.9, 1999, pp. 884-900.

[19] Zhang, Z., Huang, K. and Tan, T., Comparison of Similarity Measures for Trajectory Clustering in Outdoor Surveillance Scenes. In *Proc. IEEE Int. Conf. on Pattern Recognition*, Vol.3, 2006, pp.1135-1138.