On the Structure of Dynamic Bayesian Networks for Complex Scene Modelling

Tao Xiang, Shaogang Gong and Dennis Parkinson
Department of Computer Science
Queen Mary, University of London, London E1 4NS, UK
{txiang,sgg,dennisp}@dcs.qmul.ac.uk

Abstract

We introduce the idea of constructing Dynamic Bayesian Networks (DBNs) with hierarchical structures for modelling complex scenes at both the event level and the activity level simultaneously. Practical issues regarding the structure design of a DBN with multiple hidden processes and hierarchical structure are identified and discussed. Experiments are presented to compare a Multi-Observation Hidden Markov Model (MOHMM), a Hierarchical MOHMM, a Hierarchical Dynamically Multi-Linked Hidden Markov Model (DML-HMM), and a Hierarchical 2-layer DML-HMM (2L-DML-HMM) for complex scene modelling. It is demonstrated that only the Multi-Observation Hidden Markov Model is able to perform meaningful factorisation in the activity state space and to extract the deterministic temporal structure of activities occurred in a complex dynamic scene.

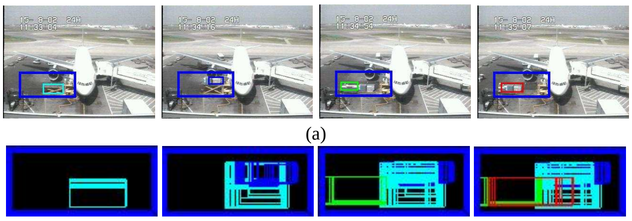
1. Introduction

Modelling visual behaviour of activities captured by video has received enormous attention in recent years due to its great potential in applications such as threat assessment, abnormal behaviour detection and public facility management [5, 2, 11, 13, 8, 3, 16, 14, 17, 9]. A complex dynamic scene often involves multiple objects moving interactively. Instead of modelling the activities of only a single object/person in isolation, it has become increasingly necessary that grouped activities should be modelled simultaneously. Dynamic Bayesian Networks (DBNs), which are capable of decomposing a complex system into simpler parts and learning the hidden dependencies among these simpler parts from data, appear to be suitable for complex scene modelling [5, 11, 13, 8, 3, 14, 17, 9].

Typically, object behaviours are modelled based on the tracked trajectories in a state space in which behaviour interpretation is critically based on the discovery and subsequent modelling of the underlying temporal structures of the trajectories. The recognition of activities then becomes the problem of trajectory recognition, either continuously using particle filters such as the CONDENSATION algorithm [12] or discretely using Hidden Markov Models (HMMs) [13]. However, a number of implicit assumptions are often made with this approach. Firstly that the videos are of high enough quality that allows for elaborative object models to be built using local image features and colour. Secondly that objects can be tracked consistently in space and over time. The first assumption is often not true for the video data captured in surveillance and for visual communication, which are characterised by low resolution and being highly noisy. The second assumption is normally invalid in busy scenes (e.g. outdoor and public places) involving activities of multiple objects with frequent overlapping motion patterns resulting in discontinuous object trajectories and inconsistent labelling. Multiple object tracking is thus ill-conditioned and remains one of the biggest challenges for computer vision research.

Rather than decomposing a dynamic complex scene into tracked trajectories, we can decompose the scene into activities consisting of correlated visual events [18, 9, 19]. Events are defined as significant scene changes and the criterion for event detection can be different for different applications. Detected events are represented and classified into different event classes. An event is labelled by its class, instead of the identity of the object causing the event. Tracking is thus avoided. Scene modelling is achieved based on interpreting the relevances and correlations of events of different classes. In [9, 19], Dynamic Bayesian Networks were constructed to model group activities involving correlated multiple temporal processes. Each temporal process is characterised by the temporal and spatial occurrences of one class of visual events. It has been shown that activity recognition can be performed successfully based on event modelling.

In the real world, a complex dynamic scene often involves multiple activities that may occur simultaneously. Taking into account the fact that each activity is composed of multiple events, it is natural to think of developing a hierarchical model which is capable of modelling a complex scene at both the event level and the activity level. To this end, we introduce the idea of constructing DBNs with hierarchical structures for complex scene modelling. Such DBNs enable activities to be represented in a state space where each discrete state corresponds to one important stage of activity. In Section 2, we illustrate how visual events can be modelled for the recognition of activities involving multiple objects. Examples are given on modelling the cargo loading and unloading activities occurred in an outdoor airport ramp scene. In Section 3, we focus on the issue of structure design for DBNs with multiple hidden temporal processes and hierarchical structure. Practical issues regarding the structure design of a DBN such as the topology design and the determination of the number of states for each hidden state variable are considered. Experiments are presented to compare hierarchical DBNs with different topologies and hidden states. Conclusions are presented in Section 4.


(a)

(b)

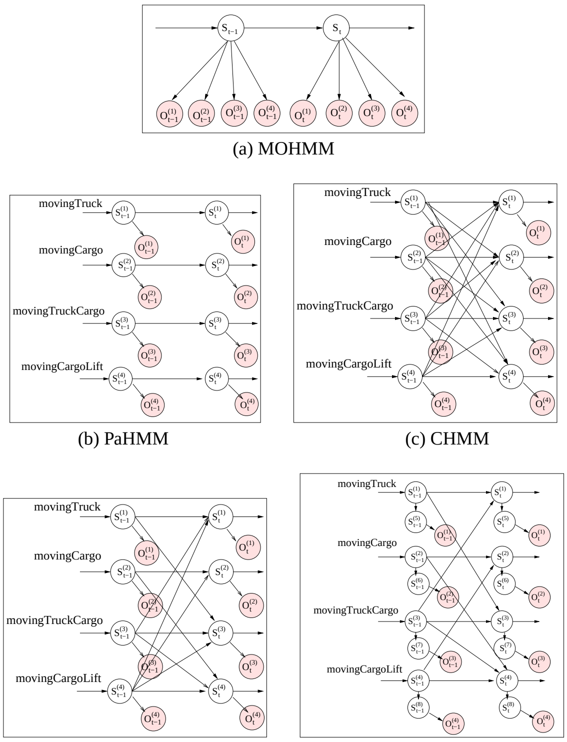
Figure 1: Event detection and classification during an aircraft cargo unloading activity. (a) Detected and classified events with the cargo service area highlighted. (b) Highly overlapped events were detected over time, including `movingTruck`, `movingCargo`, `movingTruckCargo` and `movingCargoLift`, illustrated using green, blue, red and cyan bounding boxes respectively.

2. Activity Modelling

A Dynamic Bayesian Network (DBN) B is described by two sets of parameters $(\mathbf{m}, \mathbf{\Theta})$. The first set $\mathbf{m}$ represents the structure of a DBN which includes the number of hidden state variables and observation variables per time instance, the number of states for each hidden state variable and the topology of the network (set of directed arcs connecting nodes). The structure of a DBN can be either manually set based on *a priori* knowledge or learned from data. In this paper, we are interested in how the structure of a DBN can affect its ability to model complex scenes. To this end, let us first consider a specific problem of modelling group activities in a complex outdoor airport ramp scene based on discrete object event recognition.

In an airport scene with ground based cargo loading and unloading operations, events that reflect significant changes in the scene are to be detected automatically over time *without* manual labelling or top-down hypothesising. To this end, we focus on the cargo service area of the scene and adopt an approach proposed by [18] for event detection and recognition. As shown in Figure 1, four different classes of events are automatically detected. They are `movingTruck`, `movingCargo`, `movingCargoLift` and `movingTruckCargo`. It is observed that they correctly correspond to four key elements that contribute towards frontal cargo service activities. It is also noticed that such an event detection and recognition mechanism makes mistakes. Mis-detection and wrong labelling can be caused by discontinuous movement due to low frame rate and overlap of different objects. The problem of erroneous event detection and recognition can only be effectively addressed by interpreting groups of autonomous events in correlation and as a result, explaining away the errors in the detection and labelling of individual events.

Although the cargo loading and cargo unloading activities consist of the same classes of events, these events follow different occurrence patterns. In other words, the same classes of events are correlated in different ways. We hope that DBNs can learn these differences from data. Various DBNs can be considered for factorising the state and observation space by introducing multiple observation variables and/or multiple hidden state variables. Figure 2 shows five different types of DBNs for aircraft cargo activity modelling. Observation nodes are shown as shaded circles and hidden nodes as clear circles. The ith hidden state variable and the jth observation variable at time instance t are denoted as $S_t^{(i)}$ and $O_t^{(j)}$ respectively where $i \in \{1, ..., N_h\}$ and $j \in \{1, ..., N_o\}$, N_h and N_o are the number of hidden state and observation variables respectively. In this paper, unless otherwise stated, $S_t^{(i)}$ are discrete and $O_t^{(j)}$ are continuous random variables. One of advantages of DBNs is that the *a priori* knowledge of the problem domain can be easily incorporated into the model via topology design. Different topologies of the DBNs shown in Figure 2 imply different understandings of the scene. More specifically, a Multi-Observation Hidden Markov Model (MOHMM) only factorises the observation space with each observation variable corresponding to one event class (Figure 2(a)). In contrast, all the other four DBNs aim to factorise both the state and observation space. The hidden state space is factorised into 'state channels' corresponding to multiple temporal processes. Figure 2(b) shows a Parallel Hidden Markov Models (PaHMMs) [17]. The temporal processes are assumed to be independent with each other which in this case implies that the four different classes of events occurred in the cargo service area are independent from each other. This assumption is clearly invalid. A Coupled Hidden Markov Model (CHMM) [4] assumes that each hidden state variable is conditionally dependent on all hidden state variables in the previous time instance (Figure 2(c)). Instead of being fully connected as in the case of a CHMM, a Dynamically Multi-Linked Hidden Markov Model (DML-HMM) aims to *only* connect a subset of relevant hidden state variables across multiple temporal processes [9] . This is achieved by factorising the state transition matrices using Schwarz's Bayesian Information Criterion [15]. The factorisation reduces the number of unnecessary parameters and caters for better network structure discovery. Comparing DML-HMM with CHMM, it is clear that DML-HMM will always consist of more optimised factorisation of the state transition matrices and most likely have less connections. This allows for more tractable computation when reasoning about complex group activities. It has been noted that the factorisation in the observation space, which is achieved by the event classifier, would have a significant effect on the states of hidden variables when the observation functions are continuous [7]. To alleviate the effect of an inaccurate factorisation in the observation space on the factorisation in the state space, a second layer of hidden variables are introduced in the topology of DML-HMM, resulting in a 2-layer DML-HMM (2L-DML-HMM) [19]. A 2L-DML-HMM for modelling the airport cargo activities is shown in Figure 2(e).



(a) MOHMM

(b) PaHMM

(c) CHMM

(d) DML-HMM

(e) 2L-DML-HMM

Figure 2: Five different types of Dynamic Bayesian Networks for activity modelling.

The topology of a DBN has a direct influence on the physical meaning of its states. States of each hidden state variable in PaHMM, CHMM, DML-HMM and 2L-DML-HMM represent the status of the occurrences of one class of events, while states in MOHMM correspond to the status of the occurrences of all the four classes of events because there is only one hidden state variable at each time instance. The physical meaning of states can also be affected by the number of hidden states, which is another important aspect of structure design for DBNs. For example, all the state variables in PaHMM, CHMM, DML-HMM and 2L-DML-HMM are binary due to the consideration of computational efficiency. After parameter learning, these states correspond to whether events of certain class are occurring in the scene. In contrast, the state variables in MOHMM have 2^C states where C is the number of event classes. After parameter learning, these states correspond to the collective status of the occurrences of events of different classes.

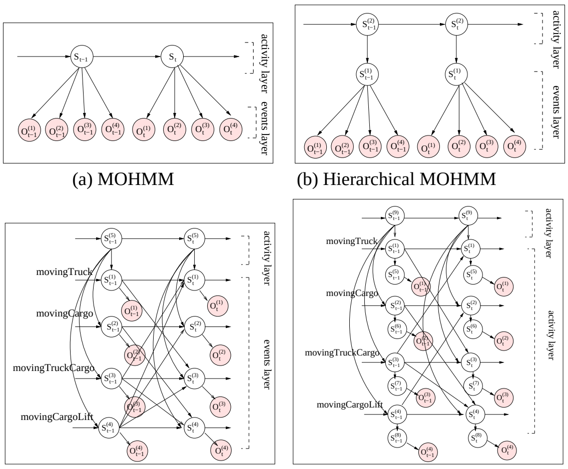
In summary, topology design and the determination of the number of states of each hidden state variable are the two important aspects of structure design for a DBN. These two aspects determine the meaning of states of each hidden state variable. The structures of the DBNs shown in Figure 2 have determined that all the hidden states would correspond to the occurrences of events. As we mentioned before, a complex dynamic scene often involves multiple activities that may occur simultaneously. To model the correlation at the activity level, it is necessary to represent activities in a state space where each discrete state corresponds to one important stage of activity, meaning that activities are modelled explicitly and separately from events. In terms of structure design, it means that hierarchy needs to be introduced in the topology of DBNs.

3. Scene Modelling

A hierarchical Dynamic Bayesian Network for complex scene modelling has two layers: an activity layer and an event layer, with the activity layer built on top of the event layer. As far as topology design is concerned, a hidden state node representing the activity state space is introduced at each time instance.

Such a topology can be implemented based on any of the DBNs shown in Figure 2. However, experiments in [9, 19] suggested that in terms of learning the correlations among events for activity recognition the performances of MOHMM, DML-HMM and 2L-DML-HMM are superior compared to those of PaHMM and CHMM, with 2L-DML-HMM being the best. We thus conduct the topology extension only to MOHMM, DML-HMM and 2L-DML-HMM. The resultant hierarchical DBNs are shown in Figure 3.

As shown in Figures 3 (b), (c) and (d), we can simply add an activity state layer to the topologies of MOHMM, DML-HMM and 2L-DML-HMM. The resultant DBNs are called Hierarchical MOHMM, Hierarchical DML-HMM and Hierarchical 2L-DML-HMM respectively. The activity layer consists of a single hidden state variable at each time instance which acts as the parent node of the event state nodes.



(a) MOHMM (b) Hierarchical MOHMM

(c) Hierarchical DML-HMM (d) Hierarchical 2L-DML-HMM

Figure 3: Four different Dynamic Bayesian Networks for scene modelling.

This type of topology determines that each state of the activity state variable would correspond to one important activity stage. As we discussed in Section 2, the number of states also plays a part in determining the meaning of states. If the number of states for the hidden state variables in MOHMM is automatically determined instead of being manually set as 2^C, the event state nodes (S_t in Figure 2(a)) turn into activity state nodes and we have a two-layer structure using the same topology as the MOHMM for activity modelling. The MOHMM for scene modelling is shown in Figure 3(a). Comparing Figures 3(a) and (b), the MOHMM for scene modelling can be seen as a simplified version of Hierarchical MOHMM, with events being modelled only in the observation space.

The remaining structure design problem is to determine the exact number of states of the activity state variable, which is essentially a model selection problem. We propose to use Schwarz's Bayesian Information Criterion (BIC) [15] to automatically determine the number of states from data. For a model $\mathbf{m}_i$ parameterised by a K_i-dimensional vector $\mathbf{\Theta_{m_i}}$, the BIC is defined as:

$$BIC = -2 \log L(\mathbf{\Theta}_{\mathbf{m}_i}) + K_i \log N \qquad (1)$$

where $L(\mathbf{\Theta}_{\mathbf{m}_i})$ is the maximal likelihood under $\mathbf{m}_i$, K_i is the dimension of the parameters of $\mathbf{m_i}$ and N is the size of the dataset. For the DBNs shown in Figure 3 , the BIC can be rewritten as:

$$BIC = -2 \log \left\{ \sum_{S_t^{(i)}} \left\{ \prod_{i=1}^{N_h} P\left(S_1^{(i)}\right) \prod_{t=2}^{T} \prod_{i=1}^{N_h} P\left(S_t^{(i)} | Pa(S_t^{(i)})\right) \prod_{t=1}^{T} \prod_{j=1}^{N_o} P\left(O_t^{(j)} | Pa(O_t^{(j)})\right) \right\} \right\} + K_i \log N \qquad (2)$$

where $S^{(i)}$ are hidden state variables, $O^{(j)}$ are events as observations, and $Pa(S^{(i)})$ and $Pa(O^{(j)})$ are the parents of $S^{(i)}$ and $O^{(j)}$ respectively. We consider two states for each event state variable except for variable $S_t^{(1)}$ in Figure 3(b) which has 16 states. The search for the optimal number of activity states that produces the minimal BIC value involves parameter learning. More specifically, for each candidate state number, the corresponding parameters are learned iteratively using Expectation-Maximisation (EM) algorithm. The E step, which involves the inference of hidden states given the parameters estimated in the last M step, can be implemented using an exact inference algorithm such as the junction tree algorithm [10]. After parameter learning the BIC value can be computed using Equation (1) where $L(\mathbf{\Theta}_{\mathbf{m}_i})$ has been obtained from the final M step of EM for parameter learning. Alternatively, parameter and structure learning can be performed within a single EM process using a structured EM algorithm [6].

The hierarchical structure of the DBNs shown in Figure 3 are designed such that the temporal structure of an activity is modelled explicitly in the activity state space. Can meaningful temporal structure of activities be discovered from real data by these DBNs? We shall find out through experiments. Our database for the experiments consists of 24 (10 loading and 14 unloading) continuous aircraft loading/unloading activity sequences selected from the 2 weeks recording giving in total 44490 frames of video data that cover different time of different days under changing lighting conditions, from early morning, midday to late afternoon. The length of each sequence was between 828 to 3449 frames, accounting for 7-29 minutes video footage. Typically sequences taken in the early morning contained indistinct objects, reflecting poor lighting, whilst those taken during the midday had strong sunshine causing strong shadows in the scene. Fast moving clouds, exacerbated by the low frame rate of 2Hz, were common during the daytime, which resulted in very unstable lighting condition and discontinuous object motion. The camera was more than 50 meters away from the activities, giving low resolution images of the objects concerned. Among the 24 sequences, there are 8 clean loading and 8 clean unloading, 2 noisy loading and 6 noisy unloading sequences. By 'clean' we imply that the lighting change in the duration of a sequence is tolerable with limited error in event detection. We used different combinations of different subsets from the 24 sequences data set to train the models in order to avoid any bias in the results. Four training sets were constructed using randomly selected 4 clean loading and 4 clean unloading sequences added with the 2 noisy loading and 6 noisy unloading sequences.

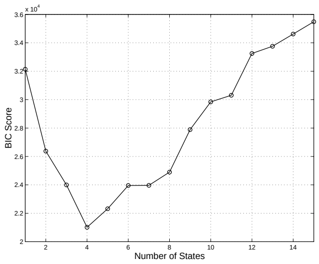
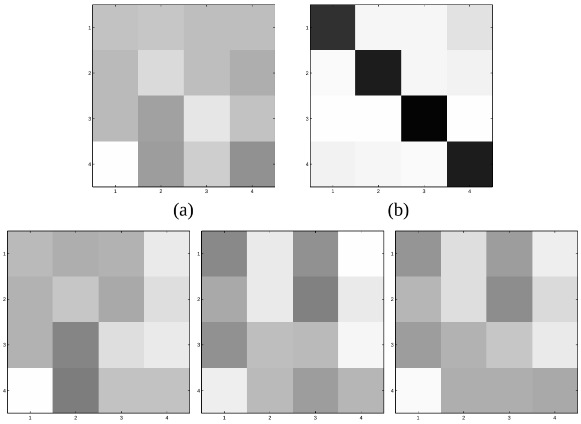


Figure 4: Search for the optimal number of activity states for MOHMM using BIC.

The optimal numbers of activity states for MOHMM, Hierarchical MOHMM, Hierarchical DML-HMM and Hierarchical 2L-DML-HMM were determined using BIC. As shown in Figure 4, the optimal number of activity states for MOHMM was determined as 4 [1]. However, it was found that larger numbers of activity states always produced higher BIC score for the other three models which means that BIC failed to determine the meaningful optimal number of activity states for them.

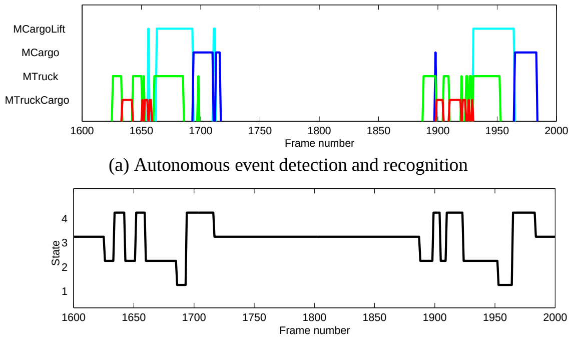


(a) (b)

(c) (d) (e)

Figure 5: (a) is a randomly initialised transition matrix and (b), (c), (d), (e) are the activity state transition matrices for MOHMM, Hierarchical MOHMM, Hierarchical DML-HMM and Hierarchical 2L-DML-HMM learned from cargo unloading data respectively. Darker entries represent higher state transition probabilities.

It is desirable for a DBN to be insensitive to parameter initialisation [7]. Since the parameters of a DBN are estimated using EM, the estimated parameters correspond to a local minimum on the error surface in the optimisation space. Being sensitive to initialisation indicates that there are many local minimal on the error surface. Figure 5 shows an example of how the initialisation of parameters affected the learning of the activity state transition matrices ($P(S_t|S_{t-1})$ for MOHMM, $P(S_t^{(2)}|S_{t-1}^{(2)})$ for Hierarchical MOHMM, $P(S_t^{(5)}|S_{t-1}^{(5)})$ for Hierarchical DML-HMM and $P(S_t^{(9)}|S_{t-1}^{(9)})$ for Hierarchical 2L-DML-HMM.). Figures 5(c), (d), and (e) show that when the numbers of activity states were set to 4 and the activity state transition matrices were randomly initialised, parameter estimated for Hierarchical MOHMM, Hierarchical DML-HMM and Hierarchical 2L-DML-HMM did not move far from their initialised values. This indicates that these three models are sensitive to parameter initialisation and thus unable to discover the temporal structure of activity. In contrast, a sparse activity state transition matrix was obtained for MOHMM (Figure 4(b)) which was found to be independent of the initialised parameters. Figure 6(b) shows the inferred activity states when the learned model was applied on one cargo unloading sequence. Comparing these states with the detected events (see Figure 6(a)), it is clear that state 3 in Figure 5(b) corresponded to 'no activity', while states 1, 2, and 4 corresponded three important activity stages in the cargo unloading activity respectively.



(a) Autonomous event detection and recognition

(b) Inferred activity states

Figure 6: Inferred activity states of MOHMM on one cargo unloading sequence. The physical meaning of each state can be discovered by comparing (b) with (a).

Our experiments demonstrate that only MOHMM is able to perform meaningful factorisation in the activity state space and to extract the deterministic temporal structure of activities occurred in a complex dynamic scene. As shown in [9, 19], MOHMM, DML-HMM and 2L-DML-HMM are capable of extracting the correlations among events towards activity recognition. Why cannot factorisation be performed correctly in the activity state space when a separate activity layer is added to the topology of these three models? We consider that there are two possible explanations. Firstly that adding more hidden state variables

[1] Unless otherwise stated, experiment results illustrated in this paper were obtained using unloading sequences in training set 1. Similar results were obtained using other training sets.

in a DBN means that more parameters are needed to describe the model. The DBN is thus more likely to suffer the "curse of dimensionality" problem [1]. Consequently, the model tends to both under-fit, being unable to capture the structure of the system, and over-fit, being sensitive to the initialisation of parameters and the noise in the data. Secondly that for MOHMM the activity state space is built directly on top of the continuous observation space, rather than the discrete event state space as in the case of Hierarchical MOHMM, Hierarchical DML-HMM and Hierarchical 2L-DML-HMM. Although the errors in event detection and recognition could have a direct influence on the factorisation of the activity state space, continuous child nodes (event observations) put more constraints on the transition structure of their parent nodes (activity state variables) compared to discrete child nodes (event state variables), preventing the factorised activity state space from being misled by the parameter initialisation.

4. Conclusions

In this paper, we introduce the idea of constructing Dynamic Bayesian Networks (DBNs) with hierarchical structures for complex scene modelling. Object temporal events are detected and labelled with automatic model order selection. Hierarchical DBNs are then constructed to model complex scenes at both the event level and the activity level simultaneously. Practical issues regarding the structure design of a DBN with multiple hidden processes and hierarchical structure are identified and discussed. Experiments are presented to compare a Multi-Observation Hidden Markov Model (MOHMM), a Hierarchical MOHMM, a Hierarchical Dynamically Multi-Linked Hidden Markov Model (DML-HMM), and a Hierarchical 2-layer DML-HMM (2L-DML-HMM) for complex scene modelling. It is demonstrated that only the Multi-Observation Hidden Markov Model is able to perform meaningful factorisation in the activity state space and to extract the deterministic temporal structure of activities occurred in a complex dynamic scene. Our analysis shows that scalability should be a major concern for designing the structure of a hierarchical DBN for complex scene modelling.

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*. Cambridge University Press, 1995.

[2] A. Bobick and A. Wilson. A state-based approach to the representation and recognition of gesture. *PAMI*, 19(12):1325–1337, December 1997.

[3] M. Brand and V. Kettnaker. Discovery and segmentation of activities in video. *PAMI*, 22(8):844–851, August 2000.

[4] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *CVPR*, pages 994–999, Puerto Rico, 1996.

[5] H. Buxton and S. Gong. Visual surveillance in a dynamic and uncertain world. *Artificial Intelligence*, 78:431–459, 1995.

[6] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Uncertainty in AI*, pages 139–147, 1998.

[7] Z. Ghahramani. Learning dynamic bayesian networks. In *Adaptive Processing of Sequences and Data Structures. Lecture Notes in AI*, pages 168–197, 1998.

[8] S. Gong, M. Walter, and A. Psarrou. Recognition of temporal structures: Learning prior and propagating observation augmented densities via hidden markov states. In *ICCV*, pages 157–162, Corfu, 1999.

[9] S. Gong and T. Xiang. Recognition of group activities using dynamic probabilistic networks. In *ICCV*, 2003.

[10] C. Huang and A. Darwiche. Inference in belief networks: a procedural guide. *International Journal of Approximate Reasoning*, 15(3):225–263, 1996.

[11] S. Intille and A. Bobick. Representation and visual recognition of complex multi-agent actions using Belief networks. In *ECCV Workshop on Perception of Human Action*, Freiburg, Germany, June 1998.

[12] M. Isard and A. Blake. CONDENSATION – conditional density propagation for visual tracking. *IJCV*, 29(1):5–28, 1998.

[13] N. Johnson, A. Galata, and D. Hogg. The acquisition and use of interaction behaviour models. In *CVPR*, pages 866–871, Santa Barbara, USA, 1998.

[14] N. Oliver, B. Rosario, and A. Pentland. A bayesian computer vision system for modelling human interactions. *PAMI*, 22(8):831–843, August 2000.

[15] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[16] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22(8):747–758, August 2000.

[17] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. *CVIU*, 81:358–384, 2001.

[18] T. Xiang, S. Gong, and D. Parkinson. Autonomous visual events detection and classification without explicit object-centred segmentation and tracking. In *BMVC*, pages 233–242, 2002.

[19] T. Xiang, S. Gong, and D. Parkinson. Outdoor activity recognition using multi-linked temporal processes. In *BMVC*, 2003.