Discovering Bayesian Causality among Visual Events in a Complex Outdoor Scene

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

Abstract

Modelling events is one of the key problems in dynamic scene understanding when salient and autonomous visual changes occurring in a scene need to be characterised as a set of different object temporal events. we propose an approach to understand complex outdoor scenarios which is based on modelling temporally correlated events using Dynamic Bayesian Networks (DBNs). A Partially Coupled Hidden Markov Model (PCHMM) is exploited whose topology is determined automatically using Bayesian Information Criterion (BIC). Causality discovery and events modelling are also tackled using a Multi-Observation Hidden Markov Model (MOHMM).

1. Introduction

Modelling visual behaviour for dynamic scene understanding has received increasing attention from computer vision researchers for decades. Different sub-areas of computer vision research such as visual surveillance and monitoring, facial behaviour modelling, gesture modelling and multiple object tracking are essentially solving the same underlying problem, although they may adopt seemingly very different representations. Consequently, same or very similar modelling approaches are utilised in different sub-areas. For example, Hidden Markov Models (HMMs) and their extensions have been widely used in human interaction modelling [8], gesture modelling [11] and traffic monitoring [3].

Instead of modelling the behaviour of only a single object/person in isolation, it has become increasingly necessary that visual behaviour involving multiple objects/people either in interaction or as a group must be modelled simultaneously. We consider that a complex dynamic scene consists of activities which are often composed of spatially and temporally structured autonomous visual events and activity units. By autonomous events, we imply that both the number of meaningful events and their whereabouts in the scene are automatically learned and detected rather than manually labelled or hypothesised as usually reported in the literature. At a higher level, spatio-temporally correlated events form activities. In some cases, the temporal structure of events follows a certain temporally repeated pattern, which is referred as an activity unit. On top of activities, we define a scene level, which is composed of temporally correlated activities. For example, if human behaviour is the dynamic scene we want to model, gestures such as 'clapping hands' can be modelled as activities. The 'clapping hands' activity is composed of temporally ordered events 'hands move towards each other' and 'hands leave each other'. An activity unit in this case corresponds to a 'hands move towards each other' event followed by a 'hands leave each other' event.

The work presented in this paper focuses on modelling events towards activity understanding in a complex outdoor scene. In a typical outdoor scenario, there are multiple moving objects. The movements of these objects can be simultaneous with the number of objects changing constantly. To avoid the difficulties associated with tracking multiple objects, we detect automated visual events and classify them into classes which correspond to different movement patterns. We believe that the semantics to be extracted from dynamic scenes are encoded in the evolution of events and the temporal correlations among them. A realistic outdoor scenario in general offers more challenges than a well controlled indoor scenario due to factors such as the unstable lighting conditions. As a consequence, the detected visual events are often contaminated by noise. It becomes critical to take into account of these errors when modelling the temporal relationships among events. Dynamic Bayesian Networks (DBNs) [6, 7] are ideally suited to associate correlated temporal events in a complex outdoor scene which deal with the errors in the observed data explicitly under a probabilistic framework by introducing hid-

Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03)
0-7695-1971 3 $17.00 © 2003 IEEE

den states. We exploit the use of a Partially Coupled Hidden Markov Model (PCHMM) to model temporal events whose topology is determined automatically by the result of event causality discovery using Bayesian Information Criterion (BIC) [9]. Multi-Observation Hidden Markov Model (MOHMM) is also employed to perform causality discovery and events modelling using a single network. Experiments are conducted on an airport cargo unloading scenario to demonstrate that meaningful event modelling can be performed for the application of abnormality detection using our approach.

2. Event Detection and Recognition

We adopt the approach proposed in [12] to detect and recognise events. Pixel Change History (PCH) [12] and adaptive Gaussian mixture background model [10] are combined to detect pixel level visual changes. Pixel level events are then grouped into blobs represented by bounding boxes to form autonomous events.

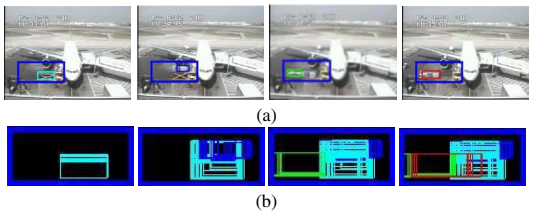


Figure 1. Event detection and classification during an aircraft cargo unloading activity. (a) Detected and classified events with the cargo service area highlighted. (b) Highly overlapped events were detected over time, including `movingTruck`, `movingCargo`, `movingTruckCargo` and `movingCargoLift`,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

Each event needs to be represented using a feature vector before the classification in the feature space is performed. Usually four types of features can be extracted from images based on location, shape, colour and motion respectively. The selection of features is largely dependent on the context of the particular scenario to be modelled which determines the stability of available features. In this paper, we consider the outdoor scenarios with large field of view such as the cargo unloading scenario depicted in Figure 1. Although colour is widely believed to be a stable feature, we found that for typical outdoor scenarios, colour information can be highly unstable. The lose of colour information can be caused by far away camera and the conversion from composite analogue videos into RGB digital image sequences. Consequently, the final features we have chosen are: (1) centroid of the location of pixel level events $(\bar{x}, \bar{y})$, (2) width and height of the bounding box (w, h), (3) filling ratio R_f, representing the percentage of the bounding box occupied by pixel-level events and (4) first order moments of the PCH image within each bounding box $(M_p x, M_p y)$. Among these features, (1) are location features; (2) and (3) are shape features; and (4) are motion features which aim to capture the direction of the movement. Considering that all these features are computed based on the detected pixel level visual changes, even the location and shape features contain motion information. A seven dimensional feature vector is then used to represent each event as follows:

$$\mathbf{v} = [\bar{x}, \bar{y}, w, h, R_f, M_p x, M_p y] \quad (1)$$

In order to detect the presence of meaningful events and their whereabout in the scene, clustering is performed in a 7-D feature space. For clustering, we adopt a Gaussian Mixture Model (GMM) [1] classification with automatic model order selection using modified Minimum Description Length (MDL) principle [4]. The obtained parameters of the mixture model are used to classify events into different classes.

3. Modelling Temporal Relationships among Events

Event detection in a busy outdoor scene can be subject to large errors due to object occlusion and trajectory discontinuities, as well as a great degree of sensory noise and poor resolution in outdoor scenes. To address this problem, we wish to model groups of events as observational input to a Dynamic Bayesian Networks (DBN).

3.1. Dynamic Bayesian Networks

Dynamic Bayesian Networks (DBNs) are BBNs that have been extended to model time series data [6, 7]. More specifically, hidden nodes have been introduced in the topology of DBNs to represent hidden temporal states. This is similar to that of a sequential graph model like HMMs. A 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 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 $j \in \{1, ..., N_o\}$ and N_h and N_o are the number of hidden state and observation variables respectively. The second set of parameters $\mathbf{\Theta}$ quantifies the state transition models $P(S_t^{(i)}|Pa(S_t^{(i)})$, the observation

models $P(O_t^{(j)}|Pa(O_t^{(j)})$ and the initial state distributions $P(S_1^{(i)})$ where $Pa(S_t^{(i)})$ are the parents of $S_t^{(i)}$ and similarly, $Pa(O_t^{(i)})$ for observations. In this paper, unless otherwise stated, $S_t^{(i)}$ are discrete and $O_t^{(j)}$ are continuous random variables. Each observation variable has only hidden state variables as parents and the conditional probability distributions (CPDs) of them are Gaussian for each state of their parent nodes.

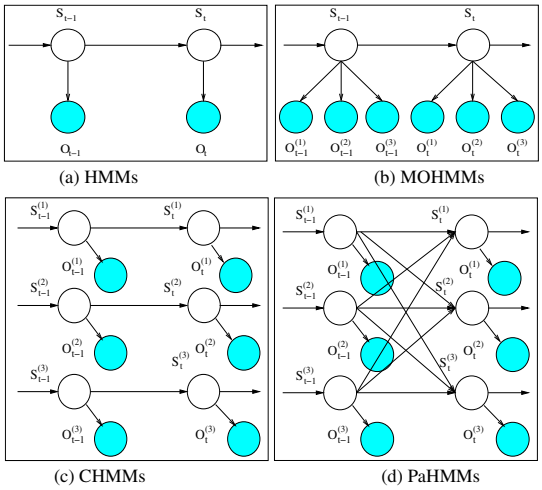


Figure 2. Graphical representations of Hidden Markov Models and their extensions. Observation nodes are represented as shaded circles and hidden nodes clear circles.

As shown in Figure 2(a), a standard HMM has only one hidden state node and one observation node at each time instance modelling a single temporal process, which often results in the high dimensionality of both the state space and observation space and requires large number of parameters if it is to model multiple temporal processes simultaneously. Unless the training data set is very large and relatively 'clean', poor model learning is expected. To address this problem, various topological extensions to the standard HMMs can be considered to factorise the state and observation space by introducing multiple hidden state variables and multiple observation variables. Vogler and Metaxas [11] proposed Parallel Hidden Markov Models (PaHMMs). The hidden state space is factorised into 'state channels' corresponding to multiple independent temporal processes. Figure 2(c) shows a PaHMM of three independent processes. Clearly this assumption is invalid in most cases, especially when dealing with group or interactive activities. Brand *et al.* [2, 3] proposed Coupled Hidden Markov Models (CHMMs) to take into account the temporal causal relationships among hidden state variables (Figure 2(d)). It is assumed that each hidden state variable is conditionally dependent on all hidden state variables in the previous time instance. Both PaHMMs and CHMMs require the observation space to be factorised by its corresponding processes.

3.2. Partially Coupled Hidden Markov Models

Instead of being fully connected as in the case of a CHMM, a Partially Coupled Hidden Markov Models (PCHMM) aims to *only* connect a subset of relevant hidden state variables across multiple temporal processes. This reduces the number of unnecessary parameters and caters for better network structure discovery.

We wish to learn the causal and temporal relationships among events simultaneously by finding a DBN $B = (\mathbf{m}, \mathbf{\Theta})$ that can best explain the observed events $\mathbf{D}$. Such a best explanation is quantified by the minimisation of a cost function. For a Maximum Likelihood Estimation (MLE), the cost function is $-\ln P(\mathbf{D}|\mathbf{m}, \mathbf{\Theta_m})$, the negative logarithm of the probability of observing $\mathbf{D}$ by model $\mathbf{m}$ where $\mathbf{\Theta_m}$ are the parameter settings for the candidate structure $\mathbf{m}$ that maximise the likelihood of the data. $\mathbf{\Theta_m}$ are estimated through Expectation-Maximisation in order to determine the distribution of the hidden states. A MLE of the structure of B in the most general case results in a fully connected DBN, which implies that any class of events would possibly cause all classes of events in the future. Therefore adding a penalty factor in the cost function to count for the complexity of a network is essential for extracting meaningful and computationally tractable causal relationships. To this end, we adopt Schwarz's Bayesian Information Criterion (BIC) [9] to measure the goodness of one hypothesised network model against that of another in describing a given dataset. For a model $\mathbf{m}_i$ parameterised by a K_i-dimensional vector $\mathbf{\Theta}$, the BIC is defined as:

$$BIC = \{L(\mathbf{\Theta}_{\mathbf{m}_i}) - L(\mathbf{\Theta}_{\mathbf{m}_0})\} - \frac{\ln(N)}{2}(K_i - K_0) \quad (2)$$

where $L(\mathbf{\Theta}_{\mathbf{m}_i})$ and $L(\mathbf{\Theta}_{\mathbf{m}_0})$ are the maximal likelihoods under $\mathbf{m}_i$ and a reference model $\mathbf{m}_0$ respectively, K_0 and K_i are the dimension of the parameters of $\mathbf{m}_0$ and $\mathbf{m}_i$ and N is the size of the dataset. For our case of an activity consisting of a group of events, the BIC can be rewritten as

$$\begin{aligned} BIC = -\ln \Bigg\{ & \prod_{i=1}^{N_h} P\left(S_1^{(i)}\right) \prod_{j=1}^{N_o} P\left(O_1^{(j)}|Pa(O_1^{(j)})\right) \\ & \prod_{t=2}^{T} \prod_{i=2}^{N_h} P\left(S_t^{(i)}|Pa(S_t^{(i)})\right) \\ & \prod_{t=2}^{T} \prod_{j=2}^{N_o} P\left(O_t^{(j)}|Pa(O_t^{(j)})\right) \Bigg\} \\ & -\frac{\ln T}{2} K_i + C \end{aligned} \quad (3)$$

where $S^{(i)}$ are hidden states, $O^{(j)}$ are events as observations, $Pa(S^{(i)})$ and $Pa(O^{(j)})$ are the parents of $S^{(i)}$ and $O^{(j)}$ respectively, T is the length of a training sequence and C is a constant. We consider that the number of hidden processes is the number of event classes extracted through automatic model order selection in the event detection process (see Section 2 for details on event detection and classification). We also consider two states for each hidden state variable, i.e. a binary variable switching between true and false. A model B estimated using a structured EM [5] that produces the minimal BIC value gives the desired PCHMM topology.

Compare PCHMM with CHMM, it is clear that PCHMM 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. In addition, a more subtle but perhaps also more critical advantage of PCHMM over CHMM is its ability to cope with noise. Given sufficiently noise-free data, it is possible for CHMM to learn the correct relationships between coupled hidden temporal processes. However, with noisy data, probability propagation travels freely among all the hidden state variables during the EM parameter estimation, the CHMM can be led to capture structures heavily biased by noise, especially when there are large number of hidden processes.

3.3. Multi-Observation Hidden Markov Models

A Multi-Observation Hidden Markov Model (MOHMM) (Figure 2(b)) can also be adopted to model temporal events and learn the causal relationships among events simultaneously. If there are K event classes, we have 2^K states which correspond to the occurrences of events of different classes. We wish that the state transition matrix of MOHMM can provide us with information on the causal relationships among events of different classes. It has been shown in [3] that the transition matrix of a standard HMM trained using Expectation-Maximisation (EM) is heavily affected by initialisation and is unable to capture accurately the true structure of the data. However, our experiments (presented in Section 4) suggest that when the observation space is factorised, the state transition matrix learned by EM is insensitive to initialisation and reveals useful information on the data structure. However, compared with PCHMM, MOHMM needs more parameters and is thus more likely to become computationally intractable when modelling complex scenarios.

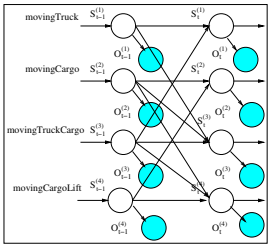
4. A Case Study

Experiments were conducted on an outdoor cargo unloading scenario occurred at one particular ramp area of an airport. A fixed CCTV analogue camera took continuous image sequences over two weeks period. The video was sub-sampled by a factor of 8. After digitisation, the final video sequences have the frame rate of 2Hz. Each image frame has a size of 320×240 pixels. Various activities were happening in the busy ramp area (Figure 1). Our experiments were concentrated on one particular activity, the frontal cargo unloading activity, because this is perhaps one of the most significant activities in the entire aircraft turn around circle and it usually occurs over relatively long time span in the ramp sequence and thus can provide rich and plenty of modelling data for us. We have manually segmented 13 complete frontal cargo unloading sequences, whose lengths range from 1000 to 3000 frames (12–25 minutes). Fast moving clouds were common in the daytime, which resulted in very unstable lighting condition. The low frame rate made the unstable lighting condition even worse and could also cause discontinuous object motion. The camera was very far away from the observed scene which means low resolution of the observed objects, especially in the region where frontal cargo unloading service took place (Figure 1). All the experiments were conducted on an Athelon 1.5G dual processor platform.

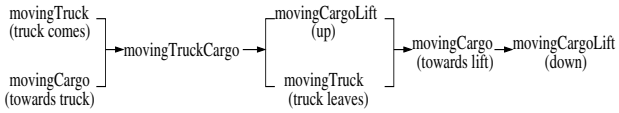
Six of the 13 frontal cargo unloading sequences were chosen as training set and the rest of the sequences as test set. There were 5197 and 5315 blob-level events detected from the training and test sets respectively. GMM classification was performed on the training set to obtain the parameters of the mixture model. It was combined with a modified MDL method to determine the number of the classes of significant events in the scene. Figure 1(a) shows that 4 classes of events were automatically detected unsupervised. By observation, the four different event classes correspond to the important stages of the frontal cargo unloading service, and hereafter labelled to `movingTruck`, `movingCargo`, `movingCargoLift` and `movingTruckCargo`. As reflected by the labels for these events, the first three classes of events correspond to the movements of the three objects, truck, cargo and cargo lift. The last class of events corresponds to the overlapped movements of truck and cargo. It is observed that our classification made mistakes. It is also noted that different classes of events did occur simultaneously.

A PCHMM was constructed to model the temporal correlations among these four event classes and the optimal topology of the PCHMM that can best *explain* the observed temporal event data was searched using the Bayesian Information Criterion. The discovered causal relationships are illustrated in Figure 3(a) with the directed arcs representing the directed causal relationships. To verify these results, by observation we summarise the temporal orders of different classes of events in the most general cases, as shown in Figure 3(b). It can be observed that most of the causal

relationships among events were discovered correctly.



(a) A PCHMM



(b) The true event temporal order by observation

Figure 3. Modelling the airport cargo unloading activity using a PCHMM.

The frontal cargo unloading activity is obviously composed of repeated activity units—the process of 'the truck comes, picks the cargo and goes' (see Figure 3(b)) repeats several times in a completed service sequence. Each process corresponds to an activity unit. Between activity units there are non-activity intervals whose durations can vary significantly. Non-activity intervals also exist within each activity unit which make it difficult to segment the activity unit. We ultimately wish to develop an automated system which can detect the starting and ending stage of activity units and thus segment them. In the experiments reported here, we manually segmented the frontal cargo unloading activity into activity units. We obtained 47 and 46 activity units from the training and test sets respectively. Each unit was manually labelled as normal or abnormal unit based on the temporal order of events. There are in general different criteria for labelling normal and abnormal units. The temporal evolution of events can follow different patterns. From the point of view of a safety monitoring officer, all patterns observed in our data are normal because nobody was injured and no cargo was stolen. However, an operational officer may think some of the patterns are abnormal because events following certain orders can cause delay and affect arrangement for other activities occurred in the scene or at other unseen ramp areas. Here we simply choose the most commonly occurred pattern as normal and treat all other patterns as abnormal. Normal activity units account for about half of the activity units in both the training and test set. Each unit contains about 100–300 frames.

We compare four DBNs models for the abnormal activity unit detection task. They are PaHMM, CHMM, PCHMM and MOHMM. For the MOHMM, we have one hidden state node with 16 states at each time instance. The four observation variables for each models are 7-D continuous observation vectors expressed by Equation (1). Their distributions are mixtures of Gaussian with respect to the states of their discrete parent nodes. The parameters of the GMM classifier were used to initialise the distribution of the observation vectors. The priors and transition matrices of states were initialised randomly. The normal activity units from the training set were employed to train the model. It took roughly 40 seconds to train the PaHMM, CHMM and PCHMM and 90 seconds to train the MOHMM on about 6000 frames using MATLAB. The experiments show that the learned parameters of all the models were insensitive to initialisation. The learned models were then applied to the test set to detect abnormal activity units. We use Receive Operating Characteristic (ROC) curve to measure the performance of our abnormal activity unit detectors. Figure 4 shows the ROC curves for the four DBNs models we tested. Figure 5 shows the initial and the EM learned transition matrices of the MOHMM.

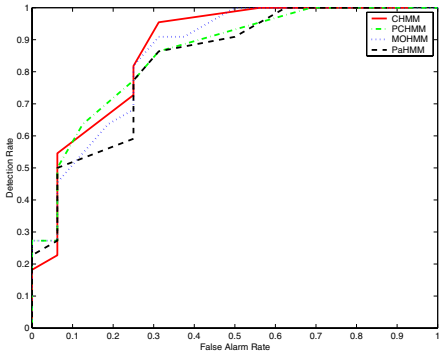


Figure 4. ROC curves for different DBNs.

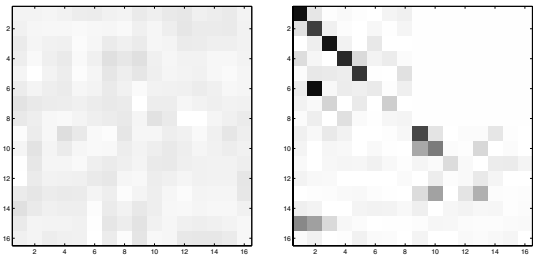
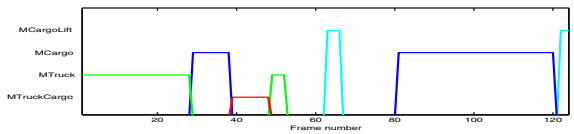


Figure 5. The transition matrix at initialisation (left), and sparsified by EM learning (right). Each row represents the transition probabilities from a single state (black represents one and white represents zero).

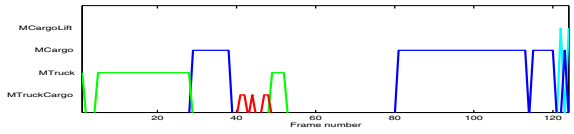
Comparing the ROC curves obtained, it appears that CHMM, PCHMM and MOHMM had similar performances which were slightly better than that of PaHMM as expected. This verifies our argument that since the arcs cut in the topology of PCHMM represent those weak correlations,

they have little influence on the selectivity of the model. The frontal cargo unloading scenario is relatively simple in terms of the number of event classes. The scalability problem is thus not severe which can probably partially explain why there was no big difference among the performances of CHMM, PCHMM and MOHMM. We expect that the strength of PCHMMs would be more clear when a more complex scene is being modelled. Overall, all the models gave modest performances. The main reason for the low detection rate and high false alarm rate is due to the nature of the data— some of the abnormal activity units in the test set has only subtle difference from the normal activity units used for training. Figure 5 shows that even being initialised randomly which means no prior knowledge on the possible state transition was employed, a very sparse states transition matrix was obtained. Each state corresponded to the simultaneous occurrences of the four event classes and the state transitions embodied the temporal correlations among different event classes. It is argued in [3] that states transitions learned by standard EM can hardly capture the meaningful structure of data. Our experiment shows that when the observation space is factorised meaningfully, the states and their transitions of a MOHMM can capture the hidden regularities and structure of data.

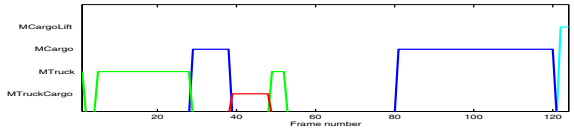
DBNs can also be used to perform event prediction and explanation. Here we show a simple example of how the explanation characteristic of DBNs can be used to improve the event detection and recognition results. Figure 6(a) shows the ground truth of event occurrences for a normal activity unit from the test set. The event recognition result contains errors, as can be seen in Figure 6(b). We inferred the hidden states of PCHMM which corresponded to the occurrences of the four classes of events. Figure 6(c) shows that the improved event detection and recognition results.



(a) Ground truth



(b) GMM classifier result



(c) PCHMM result

Figure 6. Improving event detection and recognition accuracy using a PCHMM.

5. Conclusions

In this paper, an approach is proposed to understand complex outdoor scenes which is based on modelling temporally correlated events using Dynamic Bayesian Networks (DBNs). A Partially Coupled Hidden Markov Model (PCHMM) is proposed to remove the unnecessary arcs in the topology of Coupled Hidden Markov Models (CHMMs) in order to reduce the number of parameters. The topology of PCHMM is determined automatically by event causality discovery using Bayesian Information Criterion (BIC). Causality discovery and temporal events modelling are also tackled using a Multi-Observation Hidden Markov Model (MOHMM). Experiments are conducted on an airport cargo unloading scenario to demonstrate that meaningful event modelling can be performed for the application of abnormality detection using our approach.

References

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

[2] M. Brand. Coupled hidden Markov models for modelling interacting processes. *Neural Computation*, 1996.

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

[4] M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. PAMI*, 24(3):381–396, 2002.

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

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

[7] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995.

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

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

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

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

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