Recognition of Group Activities using Dynamic Probabilistic Networks

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

Abstract

Dynamic Probabilistic Networks (DPNs) are exploited for modelling the temporal relationships among a set of different object temporal events in the scene for a coherent and robust scene-level behaviour interpretation. In particular, we develop a Dynamically Multi-Linked Hidden Markov Model (DML-HMM) to interpret group activities involving multiple objects captured in an outdoor scene. The model is based on the discovery of salient dynamic interlinks among multiple temporal events using DPNs. Object temporal events are detected and labelled using Gaussian Mixture Models with automatic model order selection. A DML-HMM is built using Schwarz's Bayesian Information Criterion based factorisation resulting in its topology being intrinsically determined by the underlying causality and temporal order among different object events. Our experiments demonstrate that its performance on modelling group activities in a noisy outdoor scene is superior compared to that of a Multi-Observation Hidden Markov Model (MOHMM), a Parallel Hidden Markov Model (PaHMM) and a Coupled Hidden Markov Model (CHMM).

1. Introduction

Probabilistic graph models including both temporal sequential models such as Hidden Markov Models (HMMs) and static causal models such as Bayesian Belief Networks (BBNs) have received enormous attention in recent years for modelling and recognising visual behaviours of activities captured in video, ranging from visual surveillance, gesture recognition, visually mediated human-machine interaction, sport analysis to virtual character synthesis [7, 4, 1, 11, 12, 8, 2, 16, 13, 15, 17].

Given a representation, either as object trajectories or labelled discrete events, activities can be modelled using a probabilistic graph model as a set of structured states in a state space. These states are linked by a set of causal or temporal connections referred to as the structure of the model. The model requires both the determination of the states, through unsupervised clustering of a training dataset, and the discovery of the underlying structure performed by the factorisation of the state space.

Instead of modelling the activities of only a single object/person in isolation, it has become increasingly necessary that activities involving multiple objects/people either in interaction or as a group must be modelled simultaneously. Both conventional BBNs and HMMs are unsuitable for modelling activities underpinned by not only causal but also clear temporal correlations among multiple hidden processes. Despite that BBNs have been shown to be capable of reasoning about behaviours of object activities, they are limited to modelling static causal relationships without taking into consideration the temporal ordering [4, 11]. This is only applicable for well structured activities with clear causal semantics. For modelling less structured group or interactive activities involving multiple temporal processes, Dynamic Probabilistic Networks (DPNs) are required [6, 9].

One way to construct a DPN is to extend a standard HMM to a set of interconnected multiple HMMs. A Multi-Observation-Mixture+Counter Hidden Markov Model (MOMC-HMM) was introduced by Brand and Kettnaker [2] to represent multiple observations of different objects at each state. Vogler and Metaxas [17] proposed Parallel Hidden Markov Models (PaHMMs) that factorise state space into multiple independent temporal processes without causal connections in-between. Any interconnection among temporal processes is implicitly assumed to be by strict zero-order synchronisation, i.e. simultaneousness. This is generally untrue. Brand and Oliver *et al.* [3, 13] exploited Coupled Hidden Markov Models (CHMMs) to take into account the causal connections among multiple temporal processes. They are essentially fully coupled pairs of HMMs such that each state is conditionally dependent on all past states of all processes at the previous time instance. However, it can be shown that such a fully connected state space cannot be factorised effectively therefore leading to poor network topology [6].

In this work, we develop a Dynamically Multi-Linked

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

Hidden Markov Model (DML-HMM) for the recognition of group activities involving multiple different object events in a noisy outdoor scene, with its topology being intrinsically determined by the underlying causality and temporal order discovered automatically using Schwarz's Bayesian Information Criterion based factorisation. In Section 2, we introduce the general framework of Dynamic Probabilistic Networks and in particular a Dynamically Multi-Linked Hidden Markov Model. In Section 3, we develop a specific model suitable for the recognition of behaviours of multiple objects involved in cargo loading and unloading activities in an outdoor airport ramp scene. We present in Section 4 experiments to evaluate the performance of a DML-HMM against alternative schemes including MOHMMs, PaHMMs and CHMMs, and conclude in Section 5.

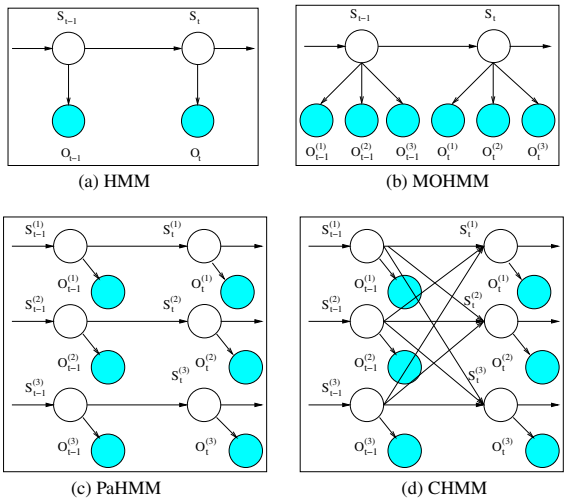
2. Dynamic Probabilistic Networks

Visual events of group and interactive activities are necessarily noisy largely due to object occlusion and trajectory discontinuities. There is also a greater degree of sensory noise and poor resolution in outdoor scenes. Figure 1 shows an example of aircraft cargo loading/unloading activities in which trucks, cargo lift and cargo container boxes are moving or being moved purposively to transfer cargoes to and from an docked aircraft on the ground. In this scenario, objects appeared in the scene often move simultaneously and the number of moving objects at any time can vary significantly.

overcast dark strong shadow flooding light

Figure 1. Typical scenes of aircraft cargo activities under different lighting conditions during a day. Normally there are about 10 objects on the ground and 0-4 events caused by object movements at any given time.

Static causal relationships represented by a conventional BBN are limited for modelling correlations among temporal states of multiple processes. Dynamic probabilistic networks and in particular Dynamic Bayesian Networks (DBNs) are BBNs that have been extended to model time series data [6, 9]. 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 the 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 $j \in \{1, ..., N_o\}$, N_h and N_o are the number of hidden state variables 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 $S_t^{(i)}$ at the previous time instance (assuming first-order Markov models) 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. The conditional probability distribution (CPD) of each observation variable is Gaussian for each state of their parent nodes.



(a) HMM (b) MOHMM (c) PaHMM (d) CHMM

Figure 2. Four different types of Dynamic Bayesian Networks (DBNs).

As shown in Figure 2(a), observation nodes shown as shaded circles and hidden nodes as clear circles, 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. For example, a straightforward Multi-Observation Hidden Markov Model (MOHMM) (Figure 2(b)) can be

used to factorise the observation space. Other extensions have been proposed to factorise both the state and observation space. Vogler and Metaxas [17] 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.* [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 according to their corresponding processes.

2.1. Discovering the Structures of Activities

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. This is achieved by factorising the state transition matrices using Schwarz's Bayesian Information Criterion [14]. The factorisation reduces the number of unnecessary parameters and caters for better network structure discovery.

For modelling group or interactive activities at a scene involving multiple objects, we consider that the scene consists of groups of dynamically linked object-centred events representing significant changes in the image over time caused by different objects in the scene. We call each such a group of events an *activity unit*. An event is represented by a multi-dimensional feature vector (see details on event detection in Section 3.1). Event detection in a busy scene such as shown in Figure 1 can be subject to large errors. To address this problem, we wish to model groups of events as observational input to a DBN so that learning causal and temporal relationships among events by finding a DBN $B = (\mathbf{m}, \mathbf{\Theta})$ can best explain the observed events $\mathbf{D}$ simultaneously. Such a best explanation is quantified by the minimisation of a cost function. For a Maximum Likelihood Estimation (MLE), the cost function is $-\ln P(\mathbf{D}|\mathbf{m}, \mathbf{\Theta}_\mathbf{m})$, the negative logarithm of the probability of observing $\mathbf{D}$ by model $\mathbf{m}$ where $\mathbf{\Theta}_\mathbf{m}$ are the parameter settings for the candidate structure $\mathbf{m}$ that maximise the likelihood of the data. $\mathbf{\Theta}_\mathbf{m}$ are estimated through Expectation-Maximisation in order to determine the distribution of the hidden states and observations. 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 account 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) [14] 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 $\mathbf{\Theta}_{\mathbf{m}_i}$, the BIC is defined as:

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

where $L(\mathbf{\Theta}_{\mathbf{m}_i})$ is the maximal likelihoods 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 our model of an activity unit consisting of a group of events, the BIC can be rewritten as:

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

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 $O^{(j)}$ at the previous time instance respectively. We consider that the number of hidden processes is the number of event classes extracted through automatic model order selection in the event classification process (see Section 3 for details on event detection and classification). We also consider two states for each hidden state variable, true and false. The search of the optimal model B that produces the minimal BIC value also involves parameter learning. More specifically, for each candidate structure, the corresponding parameters are learned iteratively using EM. The E step, which involves the inference of hidden states given parameters, 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 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 [5].

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 state 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 DML-HMM over CHMM is its ability to cope

with noise. Given sufficiently noise-free data, it is possible for CHMM to learn the correct relationships among 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 capture structures heavily biased by noise, especially when there are large number of hidden processes. Similar problems should surface for MOHMM. As for PaHMM, although it may not be easily influenced by noise, it will pay the price for discarding any correlations between multiple temporal processes. This will be shown in our experiments in Section 4.

2.2. Activity Graph of High-Level Semantics

The temporal relationships among events are quantified by the structure and parameters of DBNs learned using the training data. Once trained, DBNs aim to encode the understanding of the dynamics of the scene. The parameters of the trained DBNs can thus be utilised to extract high level semantics from the scene. One of the important semantics we wish to extract is the important stages of the activity at the correlated events level. To this end we automatically generate an activity graph from the transition matrices of the trained DBNs (see Figure 5). Each node in the graph corresponds to an important activity stage and the arcs among nodes represent the temporal order of these activity stages. In particular, for MOHMM, the transition matrix can be used directly to extract the activity graph representing important activity stages. For other models which have multiple hidden processes and hence multiple transition matrices, it is easy to convert their transition matrices into a single transition matrix with each state corresponding to the occurrences of all event classes. When the activity is composed of repeated activity units, the extracted high level semantics, compared with *a priori* knowledge can be utilised to segment the activity units from activity robustly.

3. Modelling Airport Ramp Activities

Let us now consider the specific problem of modelling group activities in a complex outdoor airport ramp 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.

It is not often the case that sufficient details about the objects of interest are available in videos of high fidelity that allows for elaborative object models to be built using local image features. Video data captured for surveillance are usually characterised by low resolution and being highly noisy. 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.

3.1. Event Detection

What constitutes an event that reflects a significant change in a scene is to be detected automatically at the pixel-level across the image over time *without* manual labelling or top-down hypothesising. To this end, an approach is adopted [18] in which Pixel Change History (PCH) was computed together with an adaptive Gaussian mixture background model in order to detect pixel level changes of significance (more than motion). Detected pixel level changes are grouped into local image blobs for event detection.

More specifically, a 7-dimensional (7D) feature vector $\mathbf{v} = [\bar{x}, \bar{y}, w, h, R_f, M_p x, M_p y]$ is constructed to capture any significant image changes in a local neighbourhood according to the PCH measurement and the adaptive Gaussian mixture background detection at each pixel. Among them, (1) $(\bar{x}, \bar{y})$ is the centroid of a local neighbourhood where changes occurred (a bounding box), (2) (w, h) are the width and height of the bounding box, (3) R_f is the filling ratio in the bounding box representing the percentage of the bounding box occupied by significant pixel changes and (4) $(M_p x, M_p y)$ are a pair of first order moments of the PCH image within each bounding box. Among them, (1) is location feature, (2) and (3) are principally shape features but also contain some indirect motion information and (4) are motion features capturing the direction of object motion.

It is worth pointing out that despite colour is widely applied elsewhere as an effective and low cost image feature, in the context of modelling the type of outdoor scenes addressed here (see Figure 1), colour information degrades dramatically. The loss of colour information is caused by a number of factors including the large distance between the camera and objects in the scene (over 50 meters), poor signal and recording media, conversion from composite to separate RGB channels. As a result, colour information is ignored in this computation.

Clustering is performed in the 7D feature space of all the feature vectors detected at each time frame using a Gaussian Mixture Model with automatic model order selection based on modified Minimum Description Length (MDL) [18]. Each cluster is labelled as a different class of event. The number of event classes in the scene is determined by the automatic model order selection. An example of event detection and classification is shown in Figure 3. At the airport scene with cargo loading and unloading operations carried out on the ground, four different classes of events were automatically detected (`movingTruck`, `movingCargo`, `movingCargoLift` and `movingTruckCargo`). They are illustrated using green, blue, red and cyan bounding box

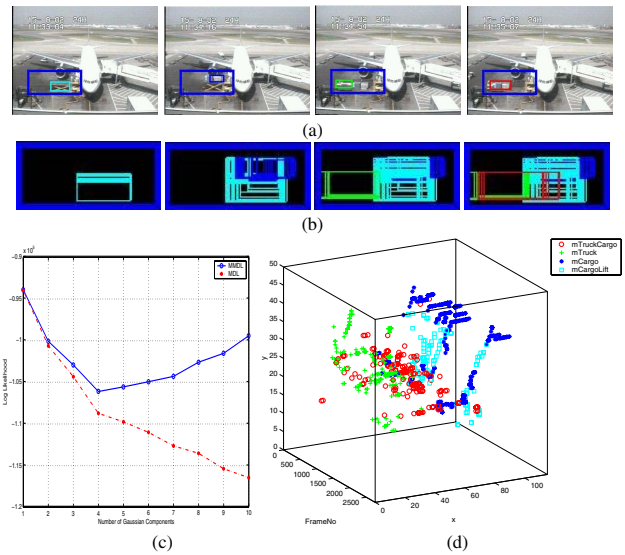


Figure 3. Event detection and classification during an aircraft cargo unloading activity. (a) Detected and classified events with the cargo service area highlighted. (b) Highly overlapped events were detected over time. (c) Automatic model order selection using MDL and modified MDL. (d) The whereabout and temporal order of the four classes of events being detected. Centroids of different classes of events are depicted using different symbols.

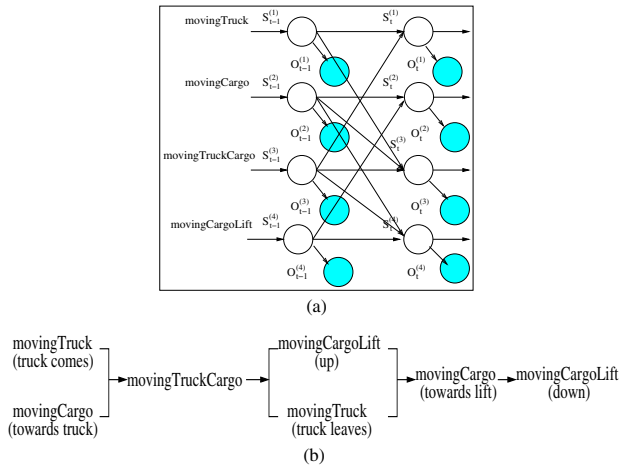


Figure 4. (a) A learned DML-HMM for modelling four temporal processes corresponding to four classes of events involved in cargo loading/unloading activities, which corresponds well to (b) the expected causal and temporal structure of the activities.

respectively in Figure 3(b). As can be seen in Figure 3(b) and (d), they correctly correspond to four key elements that contribute towards a frontal cargo service activity.

The first three events correspond respectively to a truck, a cargo container and a cargo lift moving into a specific locations with particular directions of motion and occupancies in the image space. The last event corresponds to any occurrence of simultaneous movement of the truck and the cargo container when they are overlapped. It is noted that different classes of events do occur simultaneously. It is also true that such an event detection mechanism does make mistakes. Mis-detection and wrong labelling can be caused by discontinuous movement and closeness of different objects. This 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.

3.2. A Model for Cargo Activities

For modelling the airport cargo loading/unloading activities with four different classes of events, we exploit a DML-HMM network topology as illustrated in Figure 4(a). The topology of the DML-HMM are learned from training data using the method described in Section 2.1. The causal relationships discovered among different classes of events are embodied in the topology of the DML-HMM. Figure 4(b) shows the expected structure for the airport cargo unloading activities. It can be seen that causal relationships among different classes of events have been discovered correctly.

Each of the four hidden state variables of the DML-HMM shown in Figure 4(a) has two states and takes on value 2 when corresponding class of event occurs and 1 otherwise. Each observation variable is continuous and given by a 7D feature vector representing events. Its distributions was mixture of Gaussian with respect to the states of its discrete parent nodes. For model training, the distributions of the detected autonomous events were used to initialise the distributions of the observation vectors. The priors and transition matrices of states were initialised randomly. With a trained model, the most important state transitions that minimised Criterion (2) were discovered as

$$P(S_t^{(1)}|S_{t-1}^{(1)}, S_{t-1}^{(3)}),$$

$$P(S_t^{(2)}|S_{t-1}^{(2)}, S_{t-1}^{(4)}),$$

$$P(S_t^{(3)}|S_{t-1}^{(1)}, S_{t-1}^{(2)}, S_{t-1}^{(3)}),$$

and

$$P(S_t^{(4)}|S_{t-1}^{(2)}, S_{t-1}^{(3)}, S_{t-1}^{(4)})$$

4. Experiments

Experiments were conducted on the modelling of airport cargo loading and unloading activities using MOHMM, PaHMM, CHMM, and DML-HMM networks and testing their comparative performances. A fixed CCTV analogue camera took continuous recordings over a 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.

Our database for the experiments consists of 23 (9 loading and 14 unloading) continuous activity sequences selected from the 2 weeks recording giving in total 43275 frames of video data that covers different time of different days under changing lighting conditions, from early morning, midday to late afternoon. The length of each sequence was between 1000 to 3000 frames at 2Hz, covering 12–25 minutes video footage. For the purpose of testing, we also extracted labelled ground-truth by manually identifying that each of the 9 loading and 14 unloading sequences typically has 4–9 repeated loading or unloading activity units and the entire dataset of 23 sequences captured in total 137 activity units including 55 loading and 82 unloading respectively, ranging 73–382 frames per activity unit. 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 (Figure 1). In the following we present results on (1) model training, (2) activity graphes, (3) comparative performance evaluation on activity recognition, and (4) explaining away errors in autonomous event detection.

Model training — Among the 23 sequences, there are 8 clean loading and 8 clean unloading, 1 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 23 sequences dataset to train the models in order to avoid any bias in the results. We used the remaining subsets for testing. Three different types of model training were conducted as follows. *Case I: Training by small clean sets*. We randomly splitted the 16 clean sequences into 8 small sets for which each set, consisting of one loading and one unloading sequence with average of 10 activity units, is used for training. The other 7 sets were used for testing. Each set has on average 3733 frames with the shortest being 3117 and longest 4929. This was repeated 8 times with a different set. Automatic event detection was performed on each set using both sequences and there were on average 695 events of four different classes automatically detected per training set (Figure 3(b) and (d)). These detected events (represented by 7D feature vectors) were then used as the observational input for training a DBN. The loading and unloading sequences in each set were used to train separately two sets of model parameters based on the same topology without manual activity unit segmentation in the training process. *Case II: Training by large clean sets*. Each training set now consisted of

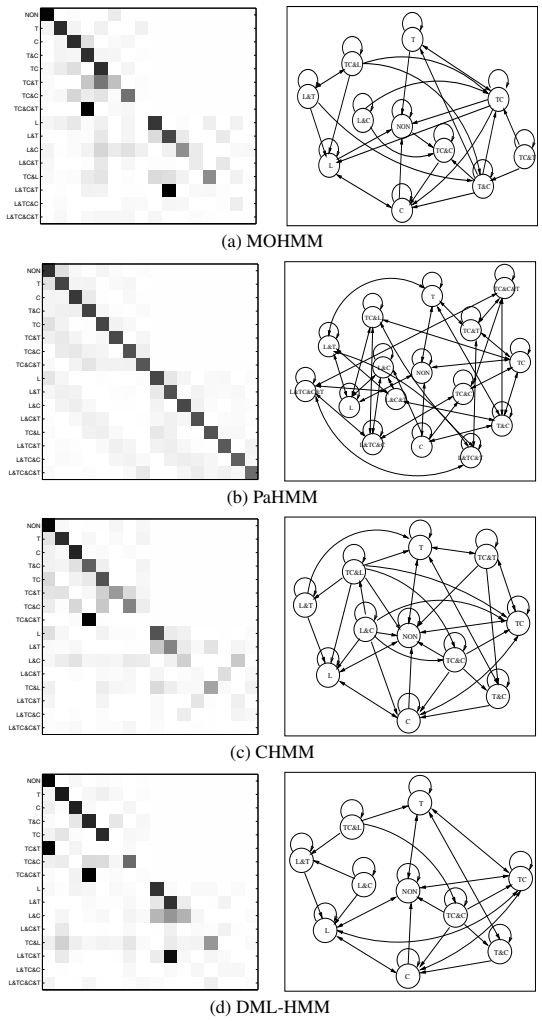
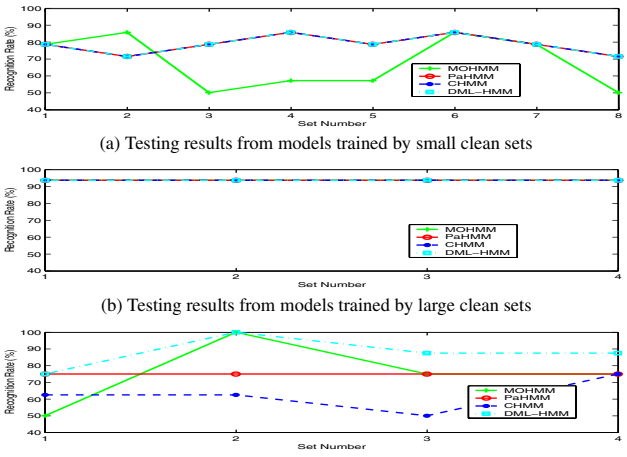


Figure 5. Left: State transition matrices learned from a noisy training set. Each entry corresponds to the transition probabilities of a hidden variable's two states (black for one and white for zero) and each state corresponds to the occurrence of one or more different classes of events. Right: Activity graphes automatically generated from the state transition matrices.

randomly selected 4 clean loading and 4 clean unloading sequences from the 16 sequences. Each set has on average 14929 frames with shortest being 13637 and longest 16221. The training was repeated as above 4 times. These training and testing were repeated four times. *Case III: Training by large noisy sets*. Four training sets were constructed using randomly selected 4 clean loading and 4 clean unloading sequences as above, but this time also included 1 noisy loading and 6 noisy unloading sequences in each set. Each set has on average 28346 frames with shortest being 27054 and longest 29638. The training was repeated 4 times again.



(a) Testing results from models trained by small clean sets

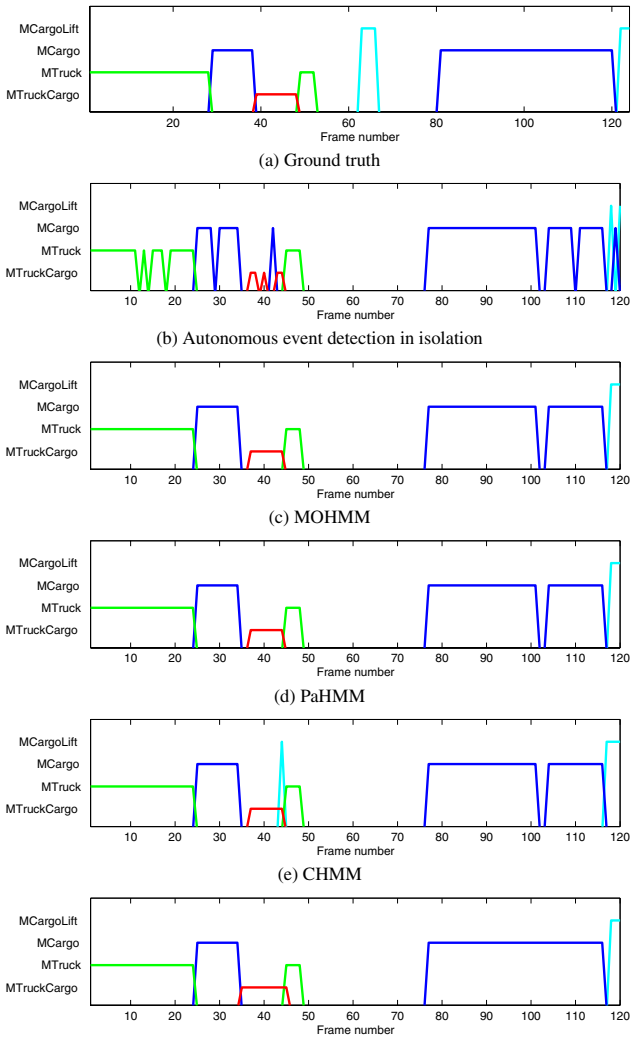
(b) Testing results from models trained by large clean sets

(c) Testing results from models trained by large noisy sets

Figure 6. Activity recognition rates.

Activity graphes — Figure 5 shows four different activity graphes automatically generated from the trained model state transition matrices of MOHMM, PaHMM, CHMM and DML-HMM. They were trained using a large noisy dataset from one of the *Case III* training sets above. States 'T', 'C', 'TC', 'L' and 'NON' correspond to `movingTruck`, `movingCargo`, `movingTruckCargo`, `movingCargoLift` and `no-activity` respectively. Simultaneous occurrence of events is indicated by '&', e.g. 'T&C' refers to `movingTruck` and `movingCargo` occurring simultaneously. From these activity graphes, important stages of activities are shown to be discovered by the models. Although these transition matrices were initialised randomly with no constraint on their transitions, the learned transition matrices have structures with sparse connections. It is also clear that among the four, the activity graph generated by the DML-HMM was least affected by noise with the cleanest connections showing the best factorised state space.

Activity recognition — The above trained four different types of models were tested for activity recognition. By activity recognition, we imply both automatic detection of the starting and ending point of individual activities in a continuous sequence of unknown number of repeated activity units and their classification into loading or unloading. The models trained using each small clean set were tested for activity recognition on the remaining 7 sets. The models trained using each of the large clean sets and each of the noisy sets were tested on the remaining sets.



(a) Ground truth

(b) Autonomous event detection in isolation

(c) MOHMM

(d) PaHMM

(e) CHMM

(f) DML-HMM

Figure 7. Improving autonomous event detection accuracy using different DBNs.

Figure 6 shows comparative testing results. As expected when small data sets were used for training, PaHMM, CHMM and DML-HMM achieved higher average recognition rate over the 7 testing sets (79%) than that of MOHMM (68%) (Figure 6 (a)). This is due to that the latter's large number of parameters were poorly estimated without enough data. Given sufficiently large sets of clean data for training, all the models were able to give a fairly high and similar average recognition rate over the 4 testing

sets at about 94% (Figure 6 (b)). However, if noisy data were used, the average recognition rate over the 4 testing sets of MOHMM (75%), PaHMM (75%) and in particular CHMM (62%) dropped significantly compared to that of DML-HMM (88%) (Figure 6 (c)).

Explaining away errors in autonomous event detection — DBNs can also be used to perform event prediction and explanation. Here we show a simple example of how DBNs can be utilised to explain away errors in event detection. Figure 7(a) shows the ground truth of event occurrences for an activity unit from the test set which lasted 124 frames. The detected autonomous events contained fair amount of errors as shown in Figure 7(b). The hidden states of four different DBNs were used to infer (generate) occurrences of events and their classes. Figure 7(c)-(f) show that the event detection results were improved using the inferred hidden states of the DBNs. The result from the DML-HMM was the nearest to the ground truth shown in (a).

5. Conclusion

In this paper, we presented an approach using Dynamic Probabilistic Networks and in particular Dynamically Multi-Linked Hidden Markov Model (DML-HMM) to interpret group activities involving multiple objects captured in an outdoor scene. The model is based on the discovery of salient dynamic interlinks among multiple temporal events using DPNs. Object temporal events are detected and labelled using Gaussian Mixture Models with automatic model order selection. A DML-HMM is built using Schwarz's Bayesian Information Criterion based factorisation resulting in its topology being intrinsically determined by the underlying causality and temporal order among different object events. Experiments are presented to demonstrate that its performance on modelling group activities in a noisy outdoor scene is superior compared to that of other DBNs including a Multi-Observation Hidden Markov Model (MOHMM), a Parallel Hidden Markov Model (PaHMM) and a Coupled Hidden Markov Model (CHMM).

Currently, the states of the DBNs used simply correspond to the occurrences of event classes which in turn constrains the number of states. It would be worthwhile to further investigate whether we can learn additional higher-level semantics through state transitions. Our future work will be focused on developing a hierarchical DBN topology in order to model the underlying temporal processes of groups of different activity units at the scene level.

Acknowledgements

We shall thank Huw Farmer and Mark Ealing at BAA for providing us with the data under the DTI/EPSRC MI LINK project ICONS.

References

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

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

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

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

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

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

[7] S. Gong and H. Buxton. On the visual expectations of moving objects: A probabilistic approach with augmented hidden markov models. In *ECAI*, pages 781–786, Vienna, August 1992.

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

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

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

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

[15] J. Sherrah and S. Gong. VIGOUR: A system for tracking and recognition of multiple people and their activities. In *ICPR*, pages 179–182, Barcelona, 2000.

[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.