Complex Activity Recognition using Granger Constrained DBN (GCDBN) in Sports and Surveillance Video

Eran Swears[1], Anthony Hoogs[1], Qiang Ji[2], and Kim Boyer[2]

[1]Kitware Inc., {eran.swears|anthony.hoogs}@kitware.com
[2]ECSE Department, Rensselaer Polytechnic Institute, {qji|kim}@ecse.rpi.edu

Abstract

Modeling interactions of multiple co-occurring objects in a complex activity is becoming increasingly popular in the video domain. The Dynamic Bayesian Network (DBN) has been applied to this problem in the past due to its natural ability to statistically capture complex temporal dependencies. However, standard DBN structure learning algorithms are generatively learned, require manual structure definitions, and/or are computationally complex or restrictive. We propose a novel structure learning solution that fuses the Granger Causality statistic, a direct measure of temporal dependence, with the Adaboost feature selection algorithm to automatically constrain the temporal links of a DBN in a discriminative manner. This approach enables us to completely define the DBN structure prior to parameter learning, which reduces computational complexity in addition to providing a more descriptive structure. We refer to this modeling approach as the Granger Constraints DBN (GCDBN). Our experiments show how the GCDBN outperforms two of the most relevant state-of-the-art graphical models in complex activity classification on handball video data, surveillance data, and synthetic data.

1. Introduction

Many scenes in sports, surveillance, and other video domains involve complex multi-agent activities where the agents co-exist and interact in a complex manner. A complex activity is defined as a collection of sequential and co-occurring events. The core challenge here is to define the temporal interactions (temporal dependencies) between event occurrences in a manner that will improve classification. The temporal interactions between events are key to discriminating among activities that have similar sets of events. As an example, consider the Person-Unload-Vehicle (P-Unload-V) and Delivery activities from the VIRAT Ground [13] and Ocean City (OC) webcam [19] datasets, Figure 1 (L) and (R), respectively. For the P-Unload-V activity a V-Stops and a P-Exits the vehicle while another person approaches to retrieve the object. They then unload the object from the vehicle and one of them enters the vehicle and drives away while the other walks away with the object. The main difference between this activity and a Delivery activity is that one person does all of the work in the Delivery activity and also enters/exits-buildings. Unfortunately, the unique building based events and even one of the pedestrians in the P-Unload-V activity may not be detected or associated with the same activity, thus increasing the reliance on the temporal interactions for discriminating between them.

Most complex activity modeling methods [4,5,6,7,14, 15] can theoretically model any number of co-occurring agents or events. However, the methods used for learning the temporal interactions among agents have one or more limitations. That is, they require manual definition, are generatively learned using only data from the class of interest, and/or are computationally complex or impose restriction on the structure of the interactions.

Our solution is the introduction of a Dynamic Bayesian Network (DBN) structure learning approach that addresses all of these issues. Our method automatically learns the temporal interactions in a discriminative and efficient data driven manner without imposing restrictions on the structure. Note, prior knowledge of DBN theory and usage is assumed when reading this paper.

Our main contribution is the use of a Granger Causality (GC) statistic [1] to explicitly define the temporal dependencies of the DBN without needing to be incorporated into the model's parameter learning process. Granger Causality explicitly measures the temporal dependence between two time sequences, making it ideal for selecting the temporal links in a DBN, which by definition represent temporal dependence. This enables us

Figure 1: (L) VIRAT Ground surveillance video showing an example of the person-unload-vehicle activity with overlaid annotations and background clutter. (R) Ocean City webcam video with Delivery activity example annotated in yellow.

to learn the links prior to the parameter learning process, which significantly reduces computational complexity and in turn reduces the need for a large number of training examples. Using Granger Causality also has the benefit of not imposing a tree or acyclical structure, which makes it more expressive and better able to characterize activities.

Our second contribution is the novel fusion of Adaboost feature selection [22] with Granger Causality to define the most discriminative temporal links for the DBN, which is also performed prior to parameter learning. The resulting DBN is referred to as the Granger Constrained DBN (GCDBN). Note, temporal dependence does not imply "true" causality, but this is not required since our objective here is to improve classification by using discriminative temporal links, not to capture the true causal network.

At a high level, our structure learning process starts with temporal sequences derived from the events that occur in the activity of interest, see step 1 in Figure 2 for an example. This temporal sequence stores the number of times that the event occurs on a given frame over all frames. Causal analysis of pairs of temporal sequences is performed in step 2, which results in GC statistics that are then used in an Adaboost feature selection algorithm for discriminatively defining the temporal links. The resulting Granger Cause graphical model is shown in 3, which completely defines the temporal links in the GCDBN 4. Further details are in Section 3.

Experiments show how the GCDBN outperforms two of the most relevant state-of-the-art graph based activity recognition models, the Dynamic Multi-Linked Hidden Markov Model (DML-HMM) [7] and the Time Delayed DBN (TDDBN) [5] (an extension of the Time Delayed Probabilistic Graphical Model (TDPGM) [15] to DBNs).

2. Relevant Work

The first well-known graphical model for capturing the interactions among events in activity recognition is the Coupled Hidden Markov Model (CHMM) [4], which models co-occurring agents as different layers of HMMs that are fully connected. The DML-HMM [7] was then introduced, which uses a data driven DBN Structural Expectation Maximization (SEM) learning algorithm [9]. Both methods automatically define the temporal structure, but the CHMM does not use the data to do this and the DML-HMM incorporates it into an iterative learning algorithm, which is computationally expensive and requires extensive training data. Additionally, both techniques are generative, so the links are learned without taking the classification performance into account.

There have been many recent developments in activity recognition. In particular, a variety of bag-of-words (BoW) techniques that use Histogram Intersection for matching have shown great success; see [14] for details. These methods work particularly well when there is one person and/or one main activity existing at a time, such as in the TRECVID-MED 2011 dataset [3]. However, they do not generalize well to dense scenes that have several multi-agent activities occurring near each other throughout the scene, such as in the VIRAT Ground [13] or Ocean City [19] datasets. This is because they lack the discriminative capabilities of well-structured probabilistic models, which can result in high false alarm rates. Most success in this dense activity domain is attributed to methods that explicitly model these types of well-defined activities in dense scenes, such as [6]. However, most of this work, including [6], manually define their model parameters and/or the interactions between agents to ensure a proper representation. Because of the BoW limitations, we focus on probabilistic graphical models

Other structure learning techniques, [10] and [15], also apply constraints for automatically learning links. The Campos algorithm [10] guarantees a globally optimal fit of the model to the data. However, it relies on domain expert knowledge and, as with the DML-HMM, may be a good fit to the current activity's data, but does not necessarily improve classification performance.

The TDPGM [15] automatically determines the spatial links in a single time slice graphical model. It uses the Time Delayed Mutual Information (TDMI) measure to search for delayed copies of a time series, which does not capture causal/temporal dependencies among co-occurring agents. The number and type of links are initialized using Prim's Minimum Spanning Tree (MST) algorithm [16], which is restricted to acyclic tree structures, and refined using a modified K2 algorithm. This technique was extended to DBNs [5] to produce the TDDBN, which was used for recognizing complex American football plays.

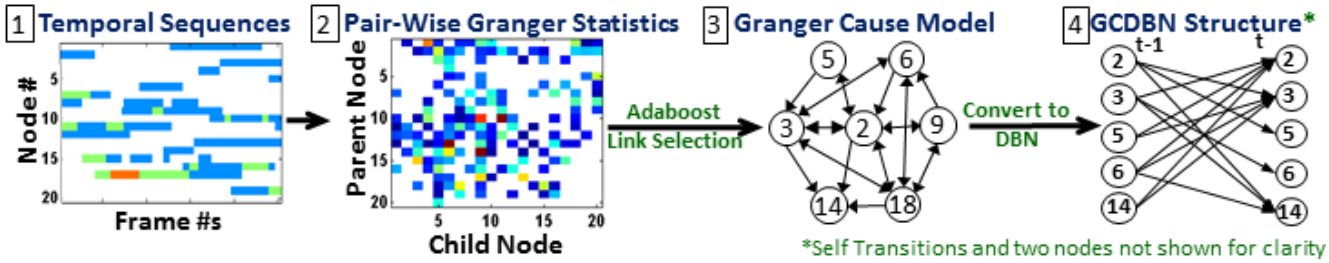


Figure 2: Overall approach: ingest temporal sequences for each event (node) 1, calculate pair-wise Granger causality statistics 2 and use their normalized versions as features in an Adaboost feature selection algorithm to discriminatively select the links. These links define the Granger Cause model 3, which are converted to temporal links in the GCDBN 4. The lighter blue colors in 1 represent smaller number of occurrences while darker orange is larger. Similarly, the darker blue colors in step 2 are weaker causal relationships and the darker red are stronger.

There are two main differences between the TDDBN and our GCDBN structure learning algorithms. The GCDBN captures more temporal relationships since it is not restricted to the tree structure of the MST and is allowed to have cycles. Second, the TDDBN is designed to capture links between sequential events caused by the same agent. On the other hand, our GCDBN captures these and links between any number of agents leading to more descriptive links and improved classification.

To our knowledge, the GC statistic has not been used to learn the structure of a DBN for multi-agent complex activity recognition in computer vision. It was only recently that the GC statistics have found applications in the computer vision domain [2,11,23]. The pairwise trajectory analysis performed in [11] uses Granger statistics as features in a Support Vector Machine to classify activities having two interacting agents. This method is restricted to two interacting agents, whereas the GCDBN can model a large number of interacting agents.

Alternatively, a spectral version of the GC statistic [12] is used in [2] and [23]. Prabhakar et al. [2] cluster visual words into independent causal sets that can be used for social game retrieval and classification of video segments. Jiang and Loui [23] created Audio-Visual Grouplets (AVGs), which are sets of audio and visual codewords that are grouped together according to their temporal dependence. These are interesting pieces of work; however, the spectral version of the Granger statistics is computationally expensive, taking about 10 times as long to produce results as the continuous time version. The spectral version also does not scale to multiple co-occurring events of the same type, since the signals are binary. This makes it impossible to determine the root cause of a change in a time series, which is common in closely spaced or high activity environments.

3. Overall Approach

The datasets are analyzed using 5-fold cross-validation, where the training data is first used to learn the GCDBN's structure and then its parameters. The overall training approach is shown in Figure 2. The first step, 1, is to extract the temporal sequences for each event in the training data. These can be extracted by counting the detections from event detectors or by counting the amount of activity assigned to clusters as a function of time [5,7]. The dataset characteristics determine which technique is used; see Section 8 for details.

The sequences from step 1 are used to calculate the pair-wise Granger statistics shown in step 2. Binary versions of the temporal sequences are used to learn the GCDBN's parameters, where each event's temporal sequence corresponds to one observation and one hidden node (one layer). The event-types (nodes) are down-selected to the top M most active nodes among the activities. This reduces computation and emphasizes the classifier's reliance on the temporal links for model comparison purposes. The most active nodes are determined by counting the number of times that each event type occurs over all frames and activities, ranking them in descending order, and selecting the top M nodes to partially define the model's structure.

The Granger statistics for the M nodes are normalized to have a score between zero and one and are then passed into an Adaboost feature selection algorithm. The top 10% most discriminative links are then chosen, which defines the Granger Cause model 3 and in turn the GCDBN structure 4. Note, the GCDBN's structure is identical for all activities, but the parameters are learned using only the data from the activity of interest. The parameters are learned using maximum likelihood Expectation Maximization (EM) with a junction tree inference engine. The EM algorithm is initialized in a data driven manner for the observation distribution parameters and randomly for the transition parameters and runs for five iterations.

The testing process is a simple maximum likelihood classification, where a pre-segmented unknown activity example is tested against each model and is assigned the ID of the most likely model.

4. Granger Causality Theory

Clive Granger [1] stated that if the variance of the autoregressive prediction error of time series X at the present time is reduced by including the joint history of X and another time series, Y, then Y has a causal influence on X. This theory is directly applied here to determine the Granger Cause of one event's temporal sequence on another. The time domain formulation of the Granger Cause test, as discussed in [1] and [12], is summarized below and implemented using Geweke's method [12].

When there are two jointly stationary stochastic processes, X_t and Y_t, they can be represented by an autoregressive model: $X_t = \sum_{j=1}^{\infty} a_{1,j} X_{t-j} + \epsilon_{1t}$ and $Y_t = \sum_{j=1}^{\infty} d_{1,j} Y_{t-j} + \eta_{1t}$, where $var(\epsilon_{1,t}) = \Sigma_1$ and, $var(\eta_{1,t}) = \Gamma_1$, respectively. The model consists of parameters, $a_{l,j}$ and $d_{l,j}$, along with noise terms, $\epsilon_{1,t}$ and $\eta_{1,t}$, with variances Σ_1 and Γ_1, respectively. The joint autoregressive models of X_t and Y_t are:

$$\boldsymbol{X_t} = \sum_{j=1}^{\infty} \boldsymbol{a_{2,j} X_{t-j}} + \sum_{j=1}^{\infty} \boldsymbol{b_{2,j} Y_{t-j}} + \boldsymbol{\epsilon_{2,t}};\ \boldsymbol{var(\epsilon_{2,t}) = \Sigma_2} \quad (1)$$

$$\boldsymbol{Y_t} = \sum_{j=1}^{\infty} \boldsymbol{c_{2,j} Y_{t-j}} + \sum_{j=1}^{\infty} \boldsymbol{d_{2,j} X_{t-j}} + \boldsymbol{\eta_{2,t}};\ \boldsymbol{var(\eta_{2,t}) = \Gamma_2} \quad (2)$$

where there are new parameters with respect to X and Y, as well as noise terms. The causal influence of Y_t on X_t is determined by analyzing the amount of reduction in Σ_2 relative to Σ_1 and is quantified by:

$$\boldsymbol{F_{Y \to X}} = \boldsymbol{\ln\left(\frac{\Sigma_1}{\Sigma_2}\right)} \quad (3)$$

When there is no causal influence from Y to X then theoretically $F_{Y\to X} = 0$, and when the strength of the causal influence increases $F_{Y\to X}$ also increases. A similar statistic can be calculated for the causal influence from X to Y by replacing the Σs with Γs and interchanging X and Y in (3). Note, mutual causality can exist here, where X influences Y and vice-versa. Equation (3) and its dual are referred to as F-statistics in the following sections. The variance of the noise terms in X_t and Y_t are considered stationary, but in practice will vary over time. Therefore, the F-statistic is calculated at every time instance, t, and the maxima are used to represent the degree of causality from X to Y and Y to X.

5. Granger Constrained DBN (GCDBN)

The GCDBN is a DBN with any node structure and spatial links, but with its temporal links constrained based on the GC statistic. The GCDBN is generally defined as having one hidden node as the parent of one observation node for each of the selected nodes. The temporal links are defined based on the Granger constraints between hidden nodes, see Section 6. No spatial links are defined here for simplicity. Without loss of generality, both the hidden and observed nodes are binomial-distributed for model simplicity and to reduce computational complexity. The observation nodes ingest binary versions of the temporal sequences, referred to as the observation profile, while the hidden nodes provide a noise buffer to compensate for errors in the observation sequences.

Using all N nodes to define the GCDBN can be computationally expensive during testing, $O(T(2N)^2)$, where T is the length of the observation profiles being classified. This is of particular concern when dealing with on-line systems and the time needed to produce results. One solution to this is to reduce the number of nodes to the M most active nodes, where $M < N$, as defined in Section 3. After defining the number and type of nodes, the temporal links are chosen to finalize the GCDBN's structure. A cross-validation approach for determining the number of links can be used to improve performance, but arbitrarily using the top 10% most discriminative Granger Cause links was sufficient for these experiments.

6. From GC to Temporal Link Selection

Given that each of the down-selected event-types i, where $i=1...M$, has a corresponding temporal sequence (observation profile) $X^i_{1:T}$, there are M^2 F-statistics from all pairwise combinations of the sequences. The F-statistics, $F_{Y\to X}$ and $F_{X\to Y}$, are stored in an $M \times M$ causality matrix, $z(i,j)$. Each cell in z represents the causal influence from event i to j, thus preserving directionality and defining a weighted adjacency graph.

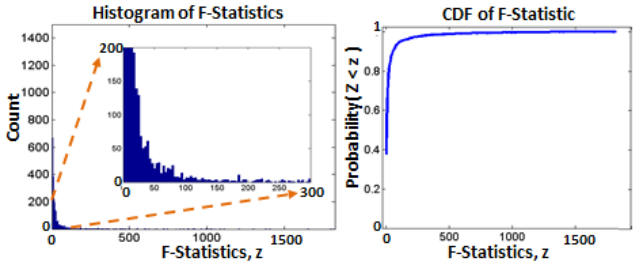


Figure 3: (L) Distribution of all F-statistics for the Handball experiments. (R) CDF of F-statistics.

Since each event type represents a node in the GCDBN, z can be used to explicitly define its temporal dependencies.

The raw F-statistic values are not between zero and one, which makes it difficult to intuitively determine a causality threshold if one wishes to define temporal links strictly based on causality. Therefore, we perform a normalization process to convert the F-statistics, z, to probabilities, $P = \mathrm{Pro}b(Z < z)$, using its Cumulative Distribution Function (CDF), which is similar to histogram equalization. This probability preserves ranking and can be interpreted as the strength or weight of Granger Causality with values between zero and one, where one is completely causal. Figure 3 shows the raw F-Statistic distribution and the normalization transform, CDF.

These normalized F-statistics, $P \in \mathbb{R}^{M\times M}$, are converted to a $1 \times M^2$ dimensional feature vector for all classes. This results in a $E \times M^2$ feature vector for input into the Adaboost feature selection algorithm [22], where E is the total number of training examples from all classes. In short, an Adaboost classifier runs for 100 iterations using a Decision Stumps weak classifier that chooses the single most discriminative feature for classification on each iteration. This vote for the most discriminative feature is accumulated over all iterations and the most frequently chosen features define the Granger Cause modal links.

Figure 2 step 3 shows the most discriminative Granger Cause model that is derived from the most active nodes for one of our experiments (Handball). This is transformed into the GCDBN model structure, Figure 2 step 4, by treating the causal links as temporal links in a two time-slice DBN. Note, the cyclical links are retained and the GCDBN structure is the same for all activity models to be consistent with its discriminative form.

Further refinement of the structure was initially performed using the SEM algorithm [9]. However, classification performance degraded since it is not a discriminative learning method and is a data intensive learning algorithm that is sensitive to dataset sizes.

7. Datasets

Three video datasets and one simulated dataset are used to analyze the performance of the various models and to demonstrate the robustness and capabilities of the

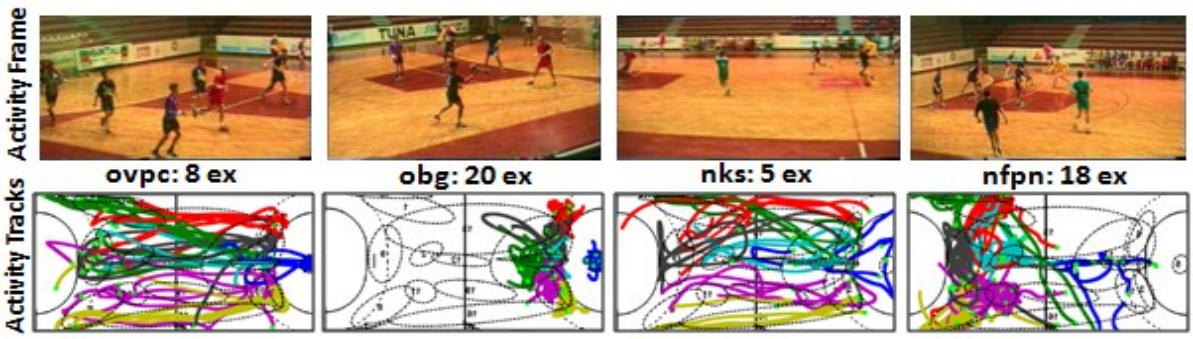


Figure 4: (Top) Selected frames of the four activities from marked frames. (Bottom) All position estimates (color coded according to player Id) in ground plane overlaid on the court image with Gaussian clusters. The activities are, ovpc: defense-returning, obg: basic defense, nks: offense-on-fast-break, and nfpn: offense-against-setup-defense, where "<#> ex" refers to the number of examples.

GCDBN. The CVBASE06 handball video dataset [17] is used to represent the sports domain, while the VIRAT Ground dataset [6,13] and Ocean City webcam video [19] represent the surveillance domain. A fourth synthetic dataset is created to emulate the surveillance domain and includes more activity types and examples. Other comparable datasets [4,5,7,15] would be ideal for these experiments, but are not publicly available. The three video datasets analyzed here are comparable in the number of activities and examples as those in [4,5,7,15] while exceeding the dataset sizes in [6].

The handball video data, Figure 4, is part of the public CVBASE 06 dataset [17] which captures ten minutes of a full court team handball match at 25fps. The handball activities can be separated into two offensive and two defensive complex activities: nfpn, nks, ovpd, and obg, with 18, 5, 8, and 20 examples, respectively. Each activity has seven players from one team with each example lasting between 100 and 200 frames.

The nfpn activity has the players passing the ball back and forth while attempting to score on the defense. The nks activity has all of the offensive players running down the court and passing the ball to each other who then attempt to score. The ovpc activity is where the team is returning to their defensive positions after the other team gets the ball. The obg activity is where the team is strictly defending their goal. Further details on the dataset can be found on the CVBASE website [17].

The Ocean City (OC) dataset [19], Figure 1(R), is a surveillance webcam video focused on a main street in Ocean City NJ. This dataset contains various types of complex pedestrian-vehicle activities, which have 17 examples each with annotated events. We focus on complex activities that have similar sets of events, but different temporal dependencies: vehicle-drop-off-person, vehicle-pickup-person, and vehicle-delivers-package.

In an effort to add another activity type to our analysis, clips from the VIRAT Ground surveillance dataset [13], Figure 1(L), were also included. This complex activity, person-unload/load-vehicle, has a similar set of events as those in the OC data, see Figure 5. Figure 5 shows the annotated event sequences and the temporal relationships for one example of the four activities as an image matrix, where the rows are the event-types and columns are the frame numbers. The events in Figure 5, from top to bottom, are: person-walk, vehicle-stopping, vehicle-starting, person-exit-building, person-enter-building, person-exit-vehicle, person-enter-vehicle, and person-near-vehicle (unload/load).

The synthetic dataset includes five randomly generated activities where each has 10 event sequences that represent the detected events and temporal dependencies between pedestrian and vehicle activities. Figure 6 shows their true temporal sequences and event durations.

Notice in Figure 6 that the event durations are constrained to only 5, 15, or 35 frames in order to reduce their separability simply based on event duration. Similarly, all temporal sequences are 45 frames long in order to eliminate separability based solely on activity duration and all the activities have the same number and types of events. The resulting activities have an increased reliance on temporal dependencies for discrimination,

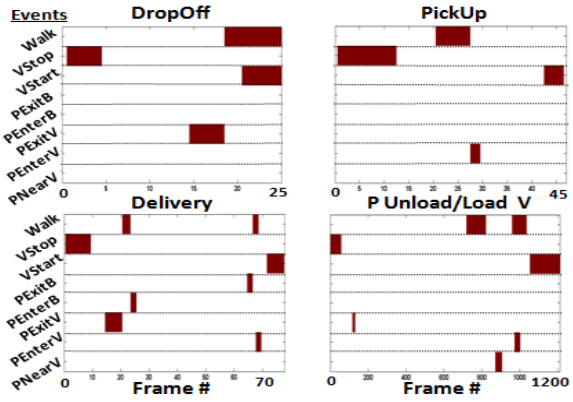


Figure 5: Event profiles for one example of each of the four surveillance complex activity types, where the rows are the events and the columns are the frame numbers. Event profiles are binary, where zero is white and one is red.

which makes this dataset much more challenging than the other two. Twenty examples of each activity are created by perturbing the truth event sequences' start and end times based on a Gaussian distribution, resulting in 100 examples, almost twice the size of the other datasets.

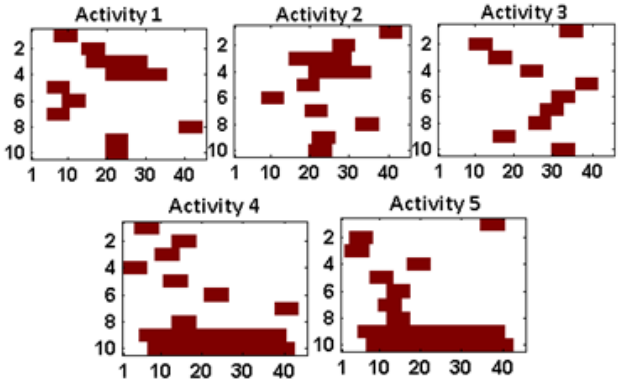


Figure 6: True temporal sequences for the ten events involved in the synthetic data's five randomly generated activities.

8. Experiments and Results

The experiments show that using the Granger Constrained DBN improves classification while being robust to a smaller number of nodes as compared to the TDDBN and DML-HMM for the three types of datasets. To demonstrate robustness to a smaller number of nodes, we varied the number of active nodes within a predefined range and recorded the resulting classification performance for a 5-fold cross-validation framework. A final experiment (All) selects the number and type of nodes from this entire range based on which experiments achieve the greatest performance on the training data. The confusion matrix and Probability of Correct Classification (PCC) are used to characterize the performance, where the PCC is the number of correctly classified examples divided by the number of total examples.

The number of states for each node and the observation profiles are identical between the three model types (GCDBN, TDDBN, and DML-HMM) resulting in models that only differ because of their learned temporal links. In order to focus on the impact of the key differences in the algorithms, the temporal links, neither the GCDBN nor the TDDBN refine their structures after being set.

Unlike the surveillance activities, the Handball activities do not have well defined events, but they do have persistent spatial patterns and dynamics. So, instead of running event detectors on the Handball data we create temporal sequences based on the amount of activity in Gaussian clusters, as is done in [5,7]. The cluster based temporal sequences make the models independent of tracks and their errors while providing both a spatial and temporal encoding of the activities. The temporal sequences in this case represent the number of moving object detections assigned to a particular cluster at time *t*, as opposed to the number of occurrences of an event type.

The clusters are formed by performing hierarchical divisive clustering [20] on all the training data using the features derived from the track's detections, *i.e.* 2D position estimates. The clustering algorithm starts by assigning all moving object detections from all tracks to a single cluster, which is then bifurcated, independent of track ID, into two more clusters. This splitting process continues by bifurcating the cluster with the largest area first, where the area is defined as the determinant of the feature's covariance matrix. The bifurcation process continues until we reach the desired number of clusters, or until the model fit to all the data *vs.* complexity no longer improves, as measured with the Bayesian Information Criterion (BIC) [8]. The leaf nodes are used to represent the set of spatial behaviors for all Handball activities.

Figure 7 shows 20 clusters with temporal sequences for eight of the most active clusters in the nfpn activity: 1,2,3,8,9,10,11, and 18. The y-axis for these sequences is the number of movers, ranging from zero to six, and the x-axis is the frame number. The track overlay in Figure 4(bottom) shows how these clusters are the most active.

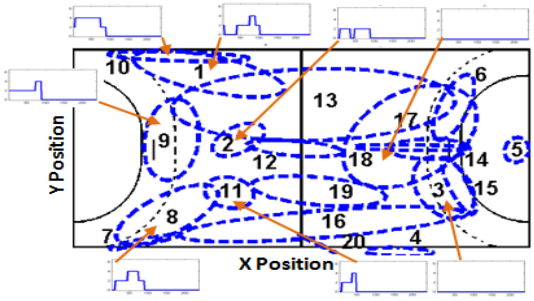
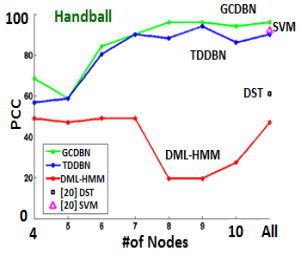


Figure 7: 20 Gaussian clusters overlaid on the Handball court with several temporal sequences from an nfpn example.

CVBASE06 Results

The experiments on the CVBASE06 Handball dataset vary the number of active nodes from four to ten. Figure 8(L) shows the PCCs for the GCDBN, TDDBN, and DML-HMM as a function of the number of active nodes, as determined across all 5-folds. Notice, the GCDBN has a consistently higher classification performance than the other two models for all experiments. In particular, when the optimal set of nodes is used (All) the GCDBN has a PCC of 96.08%, the TDDBN is 90.20%, and the DML-HMM is 47.06%. Given the relatively high PCCs the improvement are better represented by the amount of reduction in error. That is, the GCDBN has a 60% reduction in error compared to the TDDBN and a 92.6% reduction compared to the DML-HMM.

Figure 8 also shows the current state-of-the-art results on the CVBASE06 handball data [18], where five activities are modeled using a Support Vector Machine (SVM), PCC= ~92%, and a Dynamical System Tree (DST), PCC= ~61%. These methods are simpler than



GCDBN

Handball		Predicted				
		nfpn	nks	obg	ovpc	#Ex
True Activity	nfpn	**1.0**	0	0	0	18
	nks	0	**1.0**	0	0	5
	obg	0	0	**1.0**	0	20
	ovpc	0.13	0.12	0	**0.75**	8

Figure 8: (L) Handball dataset PCCs as a function of the number of active nodes across activities for the three DBN models and approximate PCCs from [18]. (R), GCDBN confusion matrices along with number of examples, *#Ex*.

DBN based approaches, but they require absolute position features as inputs, making them scene dependent, and they also do not learn temporal interactions. Therefore, these methods cannot translate to spatially independent activities, such as those in a surveillance environment.

The DML-HMM in Figure 8 chooses its temporal links based on a randomly initialized SEM algorithm that iterates over the structure 25 times and the parameters five times. Due to the random initialization, the SEM algorithm is repeated six times and the model structure that produces the highest PCC on the training data is used to report the results in Figure 8. We believe the GCDBN and TDDBN have higher PCCs because they explicitly define temporal links and because the DML-HMM is sensitive to smaller number of training examples.

The GCDBN's confusion matrix is shown in Figure 8(R) and is based on the final experiment (All). We observe difficulty classifying the ovpc activity, where it was confused with the nfpn and nks activities due to both spatial and temporal overlap; see Figure 4.

The top layer of Figure 9 shows the links for the TDDBN, while the lower layer shows the GCDBN links. Notice the noncyclical tree structure of the TDDBN compared to the more representative GCDBN links.

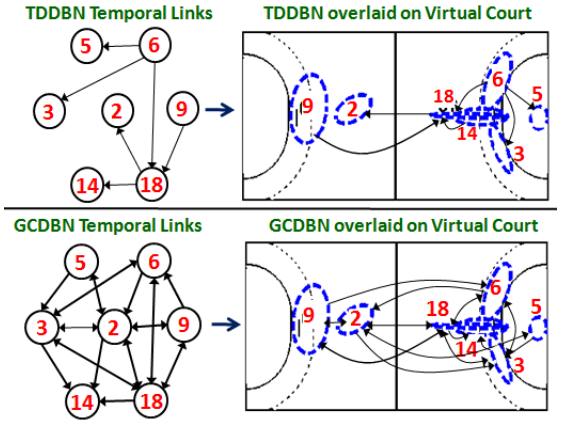
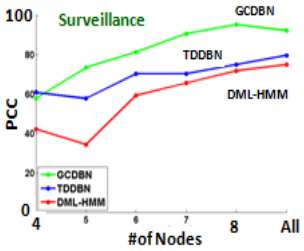


Figure 9: (L) TDDBN and GCDBN temporal links when the seven most active nodes are chosen from the handball data. (R) Graphs overlaid on handball court with temporal links.

OC and VIRAT Ground Results

The experiments on the surveillance datasets vary the number of active nodes from four to eight, where eight corresponds to all of the events across the activities. Figure 10(L) shows the PCCs for the GCDBN, TDDBN, and DML-HMM as a function of the number of active nodes for the surveillance datasets. Notice, the GCDBN consistently has a much higher performance than the other two models for most of the experiments. In particular, the PCCs for the optimal set of nodes (All) experiment are 92.2%, 79.7%, and 75% for the GCDBN, TDDBN, and DML-HMM, respectively. Notice, we do not compare against the models from [18] for the surveillance or synthetic datasets because they do not translate to activities that are location independent (occur at any location in the scene).



GCDBN

Surveil.		Predicted				
		DO	PU	DY	PULV	#Ex
True Activity	DO	**0.94**	0	0.06	0	17
	PU	0	**0.82**	0	0.18	17
	DY	0	0	**0.94**	0.06	17
	PULV	0	0	0	**1.0**	13

Figure 10: Surveillance dataset results, (L) PCCs as a function of the number of active nodes for the three models. (R) Confusion matrix for GCDBN along with number of examples, #Ex.

The confusion matrix for the GCDBN is shown in Figure 10(R) as determined from the five cross-validation iterations. The activities are abbreviated as: Drop-Off (DO), Pick-Up (PU), Delivery (DY), and Person-Unload/Load-Vehicle (PULV). The confusion matrix shows good performance overall but with a slightly lower performance on the PU activity. Comparatively, the TDDBN performs considerably lower on the PU activity (0.59), where it is confused with the DY activity 35% of the time. We believe this confusion comes about because the TDDBN structure does not characterize the PU activity well compared to the GCDBN, due to its restrictions on the structure. This can be seen in the GCDBN and TDDBN graph models, Figure 11. Notice, the TDDBN

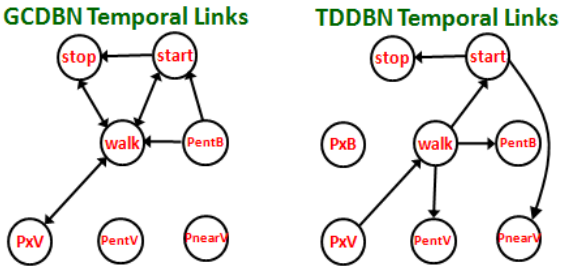
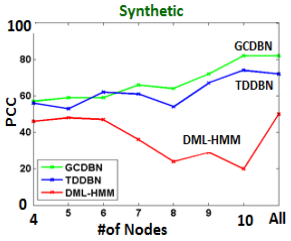


Figure 11: (L) Most-discriminative GCDBN links between hidden nodes/events for the surveillance activities, (R) TDDBN links for the surveillance activities

does not include links between the vehicle-stopping (stop) and person-walking (walk) events, and is missing links necessary to discriminate the DY activity, i.e. person-enter-building (PentB) to vehicle-start (start).

Synthetic Dataset Results

The experiments on the Synthetic dataset further confirm the higher performance of the GCDBN. Figure 12(L) shows the PCCs for the three models vs. the number of active nodes. The PCCs when using the optimal set of nodes (All) are 82%, 72%, and 50% for the GCDBN, TDDBN, and DML-HMM, respectively. The overall GCDBN confusion matrix is shown in Figure 12(R).



GCDBN

Synthetic		Predicted					
		1	2	3	4	5	#Ex
True Activity	1	**0.90**	0.1	0	0	0	20
	2	0.15	**0.80**	0.05	0	0	20
	3	0	0.05	**0.95**	0	0	20
	4	0.05	0	0	**0.70**	0.25	20
	5	0	0.05	0	0.20	**0.75**	20

Figure 12: Synthetic dataset results, (L) PCCs vs. number of active nodes for three models. (R) Confusion matrices for GCDBN along with number of examples, #Ex.

Our 82% PCC on this dataset is very significant considering it is a much more challenging dataset than the other three. In particular, the GCDBN offers a 35.7% reduction in error when compared to the TDDBN.

9. Conclusion

We introduced a Granger Constrained DBN (GCDBN) model for recognizing complex activities that consist of multiple interacting objects. The novelty of our approach comes from the fusion of the Granger Causality statistic and the Adaboost feature selection algorithm to explicitly define the GCDBN links in an automatic, efficient, descriptive, and discriminative manner. We showed how the GCDBN consistently achieves higher classification performance with the sports, surveillance, and synthetic datasets. The experiments demonstrate how our method improves classification while being robust to activities with smaller event sets, particularly when compared to other state-of-the-art graphical modelling techniques.

10. Acknowledgments/Disclaimer

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under contract nos. HR0011-10-C-0112 and W91CRB-10-C-0098. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of DARPA or the U.S. Government. Approved for public release; distribution unlimited.

11. References

[1] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3):424–438, 1969

[2] K. Prabhakar, S. Oh, P. Wang, G. Abowd, and J. Rehg, "Temporal Causality for the Analysis of Visual Events," CVPR, 2010

[3] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, and G. Quenot, "TRECVID 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms, and Metrics," Proceedings of TRECVID 2011

[4] N. Oliver, B. Rosario, and A. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions," PAMI 2000

[5] E. Swears and A. Hoogs, "Learning and Recognizing Complex Multi-Agent Activities with Applications to American Football Plays," WACV , 2012

[6] S. Kwak, B. Han, and J.H. Han, "Multi-Agent Event Detection: Localization and Role Assignment," CVPR 2013

[7] T. Xiang and S. Gong, "Beyond Tracking: Modeling Activity and Understanding Behavior," IJCV 2006

[8] G. Schwarz, "Estimating the Dimension of a Model," Annals of Statistics, 6(2), 461-464, 1978

[9] N. Friedman, K. Murphy, S. Russell, "Learning the Structure of Dynamic Probabilistic Networks," UAI, 1998

[10] C. Campos, Z. Zheng, and Q. Ji, "Structure Learning of Bayesian Networks using Constraints," ICML 2009

[11] Y. Zhou, S. Yan, and T. Huang, "Pair-Activity Classification by Bi-Trajectories Analysis," CVPR 2008

[12] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," Journal of American Statistical Association, 77(378):304-313, 1982

[13] Oh S., et al., "A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video," CVPR, 2011

[14] W. Li, Q. Yu, H. Sawhney, and N. Vasconcelos, "Recognizing Activities via Bag of Words for Attribute Dynamics," CVPR 2013

[15] C.C. Loy, T. Xiang, and S. Gong, "Modeling Activity Global Temporal Dependencies using Time Delayed Probabilistic Graphical Model," ICCV, 2009

[16] R.C. Prim, "Shortest Connection Networks and Some Generalizations," Bell Sys. Tech. J., 36:1389-1401, 1957

[17] J. Pers, M. Bon, and G. Vuckovic, CVBASE 06 Dataset, available online: http://vision.fe.uni-lj.si/cvbase06 /dataset.html

[18] S. Blunsden, R. Fisher, and E. Andrade, "Recognition of coordinated multi-agent activities, the individual vs. the group," CVBASE workshop, ECCV, 2006

[19] E. Swears and A. Hoogs, "Functional Scene Element Recognition for Video Scene Analysis," WMVC, 2009

[20] A. Guenoche, P. Hansen, and B. Jaumard, "Efficient algorithms for divisive hierarchical clustering with diameter criterion," Journal of Classification, 8(1):05-30, 1991

[21] J. Geweke, "Measurement of Linear Dependence and Feedback Between Multiple Time Series," Journal of the American Statistical Association, 77(378):304-313, 1982

[22] P. Silapachote, D. R. Karuppiah, and A. R. Hanson, "Feature Selection using Adaboost for Face Expression Recognition," VIIP, 2004

[23] W. Jiang and L. Loui, "Audio-Visual Grouplet: Temporal Audio-Visual Interactions for General Video Concept Classification," ACM Multimedia, 2011