# Unsupervised Learning of Event Classes from Video

**Muralikrishna Sridhar** and **Anthony G. Cohn** and **David C. Hogg**
School of Computing, University of Leeds, UK
{krishna,agc,dch}@comp.leeds.ac.uk*

**Abstract**

We present a method for unsupervised learning of event classes from videos in which multiple actions might occur simultaneously. It is assumed that all such activities are produced from an underlying set of event class generators. The learning task is then to recover this generative process from visual data. A set of event classes is derived from the most likely decomposition of the tracks into a set of labelled events involving subsets of interacting tracks. Interactions between subsets of tracks are modelled as a relational graph structure that captures qualitative spatio-temporal relationships between these tracks. The posterior probability of candidate solutions favours decompositions in which events of the same class have a similar relational structure, together with other measures of well-formedness. A Markov Chain Monte Carlo (MCMC) procedure is used to efficiently search for the MAP solution. This search moves between possible decompositions of the tracks into sets of unlabelled events and at each move adds a close to optimal labelling (for this decomposition) using spectral clustering. Experiments on real data show that the discovered event classes are often semantically meaningful and correspond well with ground-truth event classes assigned by hand.

## 1 Introduction

Many human *activities* are planned and structured in terms of units called *events*. Events play an integral part in serving the overall purpose of activities. For example, events such as chopping, drilling and unloading play a significant role in activities for kitchen, workshops and aircraft *domains* respectively.

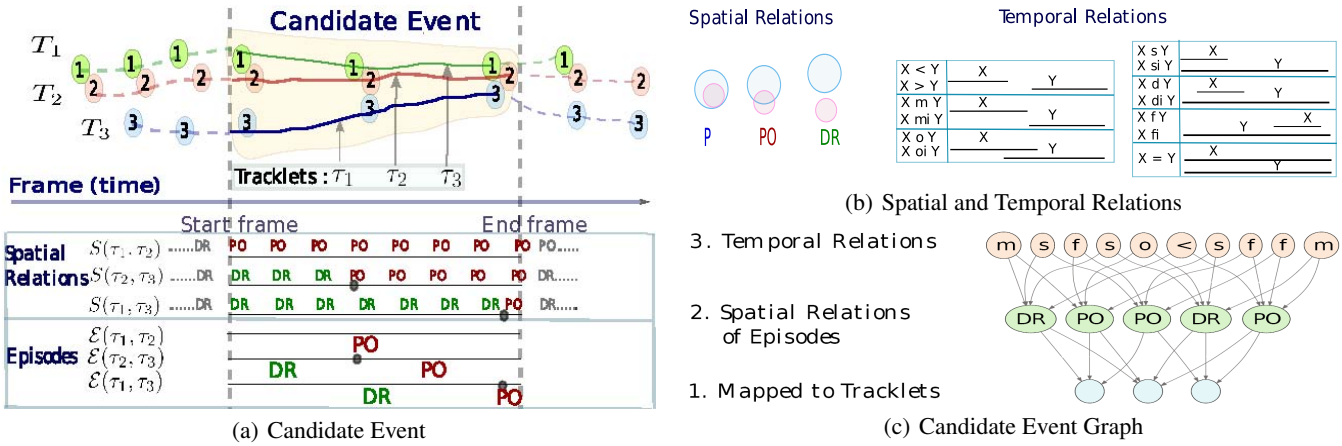
This paper addresses the following important problem in Artificial Intelligence : If we were to point a camera at activities for a certain domain over an extended period of time, is it possible for a computer program to discover (i) the events that compose the activities; (ii) the process that generated these events ?

*This work is supported by the EPSRC (EP/D061334/1) and the EU FP7 (Project 214975, Co-Friend). We also acknowledge the support of colleagues in the Co-friend project.


The following analysis starts with the assumption that many activities are a result of *interactions* between a set of objects and that these interactions are usually accompanied by changes in *qualitative spatial relations* (e.g. touches, disconnected) between them. Interactions are unique because they occur at concrete metric positions, specific time intervals and involve a particular set of objects. However, they become comparable at the level of a *qualitative relational description*, where qualitative spatio-temporal relations between the interacting objects are abstracted away from these concrete details of their occurrence. We therefore represent interactions using a graph based relational structure called a *qualitative spatio-temporal graph*, according to which two interactions can be similar (resp. identical), if their respective graphs are similar (resp. isomorphic). Intuitively, similar qualitative spatio-temporal graphs represent spatio-temporally similar ways of doing some thing, such as *preparing a hot drink* or *unloading an aircraft*.

For many activities, not all interactions are equally significant. While some interactions may be coincidental or the result of noise in observation, others are composed of actively interacting objects and they play an integral part in serving the overall purpose of activities. We regard these significant interactions as the building blocks of activities and refer to them as *events*.

We assume that events are generated according to the following three step generative process. It is supposed that for activities in a domain of interest, there is an underlying prior probability distribution over *sets of event classes*. For a *particular set of event classes*, each event class in this set is itself a probability distribution over a finite set of qualitative spatio-temporal graphs referred to as *event graphs*. In the first step, event graphs are sampled from the event classes according to this probability distribution. In the second step, a single structure called the *activity graph* is constructed by combining all the event graphs, and also specifying the spatio-temporal relationships between objects *across* different event graphs. The activity graph captures the spatio-temporal relations between all the objects that constitute the activities. The generation of activity graphs is influenced by a conditional distribution that favours certain activity graphs over others, given a set of event graphs. In the third step, the activity graph is *embedded* as tracks in space and time with concrete objects, spatial positions and temporal intervals.




(a) Candidate Event

(b) Spatial and Temporal Relations

(c) Candidate Event Graph

Figure 1: Illustrations are best viewed in colour. (a) Spatial relations and episodes for a candidate event with three tracklets $\{\tau_1, \tau_2, \tau_3\}$. (b) Spatial relations adopted from RCC-5 and Allen's temporal relations (c) Candidate event graph whose layer 1 nodes are mapped to $(\tau_2, \tau_3, \tau_1)$ respectively and whose node labels capture the spatio-temporal relations for the interaction in Fig.1(a).

The generative model provides a probabilistic framework that is used in the unsupervised event learning setting, rather than for the task of actually generating a set of tracks. In this setting, a set of tracks is observed for a video from a certain domain. The goal is to find the most likely interpretation, that is the most likely event classes, event graphs and the activity graph that could have generated the observed tracks. The posterior probability for any candidate interpretation is a measure of how likely it is that the candidate interpretation could have generated the observed set of tracks. The Maximum a Posteriori (MAP) solution is found by efficiently sampling the space of posterior distributions of candidate interpretations using MCMC.

## 2 Related Work

Much work in the area of video understanding has focussed on the supervised learning setting, where a model is trained with an annotated set of events to be applied for the task of event recognition. Graphical models such as Dynamic Bayesian Networks (Lavee, Rivlin, and Rudzsky 2009) and grammar based models (Ivanov and Bobick 2000) have been commonly used in the supervised setting.

There has been relatively less work in the unsupervised learning setting. The work described in (Hamid et al. 2009) is based on an unsupervised learning of larger semantic units of activity from a sequence of atomic actions. However, these atomic actions are manually specified for a video. Moreover, sequence based approaches do not naturally generalize to the case of multi-threaded (Lavee, Rivlin, and Rudzsky 2009) events with possibly shared and interacting objects, which are challenging to mine even with possible extensions to relational sequence based mining techniques (Kersting et al. 2008). Recent work in (Wang, Ma, and Grimson 2009) directly learn atomic actions from video images, where the atomic actions are certain statistically significant patterns of change in image pixel values. Topic models are used to learn larger semantic units which are co-occurrences of these atomic patterns. This approach is used to learn such co-occurring patterns in crowded scenes such as the arrival of a train followed by people getting out and moving away from the train.

In contrast to the pixel based abstraction used in the work described above, this work focusses on an object based abstraction for multi threaded events with shared and interacting objects. A graph based relational representation that relies on variable instantiation and generalization was introduced in (Sridhar, Cohn, and Hogg 2008) to represent spatio-temporal relations between objects for an entire activity and events were mined from this activity graph. This work generalizes the technique in (Sridhar, Cohn, and Hogg 2008) in the following ways : (i) a variable free graph based representation of events with which an event class is represented as a distribution over a set of event graphs; (ii) a probabilistic model for the generation of activities; (iii) an efficient procedure for recovering the most likely model that generated the activities for a given video.

## 3 Candidate Event Graphs

This section describes *candidate events* as interactions between certain sets of tracklets and their corresponding graph based descriptions which are called *candidate event graphs*.

**Detection and Tracking.** Given a video, a set of tracks $T$ are obtained using the techniques in (Ott and Everingham 2009), (Yu and Medioni 2008). Each track is a sequence of detected regions for a set of moving objects. Three tracks, $T_1, T_2, T_3$ for objects $1, 2, 3$ are illustrated in Fig. 1(a).

**Qualitative Spatial Relations.** A sequence of distances $\delta(T_i, T_j)$ is computed for every pair of tracks $(T_i, T_j)$, by measuring the distance at each image frame, between the detected regions of the respective objects, which constitute these tracks. A sequence $S(T_i, T_j)$ (Fig. 1(a)) of qualitative

spatial relations which are either $\{\mathsf{P}, \mathsf{PO}, \mathsf{DR}\}$(Fig. 1(b))[1] are computed from the sequence of distances $\delta(T_i, T_j)$ using a Hidden Markov Model (HMM)[2] $\theta$. For example, spatial relations $S(T_2, T_3) = (\mathsf{DR}, \mathsf{DR}, \mathsf{DR}, \mathsf{DR}, \mathsf{PO}, \mathsf{PO}, \mathsf{PO}, \mathsf{PO})$ are shown for the interval marked by dashed lines in Fig. 1(a).

**Episodes.** For each pair of tracks $(T_i, T_j)$, the sequence of spatial relations $S(T_i, T_j)$ can be aggregated to a sequence of *episodes* $\mathcal{E}(T_i, T_j)$ such that within each episode the same spatial relation holds, but a different spatial relation holds immediately before and after the episode, as shown in Fig. 1(a). For example, episodes $\mathcal{E}(T_2, T_3) = (\mathsf{DR}, \mathsf{PO})$ are shown for the interval marked by dashed lines in Fig. 1(a).

**Tracklets.** Using these episodes, *tracklets* can be formed from tracks. A tracklet $\tau_i$ is a contiguous subsequence of a track $T_i$ which starts and ends on an episode boundary associated with the tracked object. Three tracklets $\tau_1, \tau_2, \tau_3$, which are segments of $T_1, T_2$ and $T3$ respectively, are illustrated in Fig. 1(a).

**Candidate Event.** The set of all tracklets is given by $\mathcal{T}$. A *candidate event* is defined as a subset of tracklets $\mathcal{T}$ such that (i) no two tracklets are from the same track (ii) the latest start frame of any tracklet in this subset is before the first end frame of any other tracklet in this subset. These two conditions define a notion of "non-gappedness" which has been found useful in ensuring that events which are deemed to be *similar* (see below) do indeed have intuitive similarity. Any subset of tracklets that satisfy these two conditions qualifies as a candidate event as illustrated by the shaded region in Fig. 1(a).

**Abstracted Candidate Event Graphs.** Temporal relations between episodes gives rise to a candidate event graph. An abstracted *candidate event graph* is defined as a connected directed node-labelled graph in which the vertices are partitioned into 3 layers and edges exists only between adjacent layers. In this graph, the layer 1 nodes map injectively to tracklets in $\mathcal{T}$ but are *not explicitly labelled* with these tracklets. Layer 2 nodes represent episodes $\mathcal{E}(\tau_i, \tau_j)$ between the respective pairs of tracklets $\tau_i, \tau_j$ pointed to at layer 1 and *are labelled* with their respective maximal spatial relation as shown in Fig. 1(c) for the episodes in Fig. 1(a). Layer 3 nodes *are labelled* with temporal relations between the pairs of layer 2 episode nodes as illustrated. The *key feature* of the candidate event graph is that it is a *level of description* in which qualitative spatio-temporal relations hold between *abstract* entities which denote the tracklets for *concrete* object but without representing the *locations* or *intervals* metrically. This facilitates the formation of event classes since similarities become more pronounced at this level of description. A candidate event is regarded as an *embedding* of the corresponding candidate event graph.

**Similarity Measure between Candidate Event graphs.** In order to compare any two candidate events, an appropriate similarity measure between their respective candidate event graphs $g_j$ and $g_k$ needs to be defined. Accordingly, any two graphs $g_j$ and $g_k$ are re-represented as a bag of graphemes (BoG), analogously to the way in which a paragraph might be represented in terms of a bag of words. Graphemes capture several possible interactions between subsets of objects. The grapheme dictionary $(\upsilon_1, ..., \upsilon_l, ...\upsilon_n)$ is constructed off-line by synthetically generating candidate event graphs for all possible distinct interactions between subsets of objects, by varying the number of objects and number of interactions up to some fixed bounds. Using this dictionary, an event graph $g_j$ is re-represented as a histogram of length $n$:

$$\Phi(g_j) = [f_{j1}..., f_{jl}, ..., f_{jn}]$$

The term $f_{jl}$ is the frequency with which a grapheme $\upsilon_l$ occurs in event graph $g_j$. The BoG kernel $\mathcal{K}_{jk}$ measures the similarity between two graphs $(g_j, g_k)$, in terms of the extent to which they share common graphemes.

$$\mathcal{K}_{jk} = \langle \Phi(g_j), \Phi(g_k) \rangle = \sum_{l=1}^{n} f_{jl} f_{kl}$$

## 4 A Generative Model for Activities

*Events* are a subset of interactions that are significant, while other interactions may be the result of coincidence or noise in observations, as noted in §1. In order to distinguish events from all other interactions, the following model of event generation is assumed to underlie activities for domains of interest. The conditional probabilities given by the generative model are then used to formulate the unsupervised task of finding the optimal model, that is presumed to have generated a given set of tracks, as described in §5.

**Prior Distribution Over Event Classes.** It is supposed that for activities in a domain of interest, there is an underlying prior probability distribution $P(\mathcal{C})$ over *sets of event classes* $\mathcal{C} = \{c_1, ..., c_p\}$. This just means that certain sets of event classes are more likely to have generated the activities than others.

For a *particular set* of event classes $\mathcal{C}$, each event class $c_i$ itself is a probability distribution over a finite set of event graphs $\Gamma$. More precisely, $P(g_j|c_i, \mathcal{C})$ is the conditional probability of $g_j$ given $c_i$ and $\mathcal{C}$. $P(c_i|\mathcal{C})$ is the probability of an event class $c_i$ given $\mathcal{C}$.

The *prior distribution* $P(\mathcal{C})$ is a normalized exponential function of four independent components[3]:

$$P(\mathcal{C}) = \frac{1}{z_1} \exp\left(-\lambda_1 \mathcal{N}(\mathcal{C}) + \lambda_2 \omega(\mathcal{C}) + \lambda_3 \chi(\mathcal{C}) + \lambda_4 \mathcal{L}(\mathcal{C})\right) \quad (1)$$

First, it is assumed that activities are generated by a relatively small number of event classes that are compact (wrt.

[1] These relations are adapted from RCC-5 (Cohn and Hazarika 2001) by collapsing the $\mathsf{PP}, \mathsf{EQ}$ distinction.

[2] The HMM is trained to learn a mapping from the distance metric between object regions, to qualitative spatial relations between objects; this has the benefit of smoothing the transitions between spatial relations (i.e. inhibits rapid flipping between relations owing to visual noise).

[3] These probabilities are used to search for a maximum in §7. Therefore the normalizing factor $z_1$ for $P(\mathcal{C})$ and $z_2$ that appears further below need not be evaluated. The parameters $\lambda$s are discussed further in the experiment §8

their similarity metric). Therefore the first component is an exponentially decreasing function of $\mathcal{N}(\mathcal{C})$ and therefore favours *fewer event classes*. The second component favours event classes $c_i$ that are a set of structurally very similar event graphs i.e. those with higher values of *within class similarity* $\omega(c_i)$, as motivated further and made more precise in §6. Finally, the events that compose activities are assumed to be planned in terms of larger units than smaller ones. Therefore the fourth component favours event classes that have a high probability of generating interactions between a larger set of tracklets. This is measured by $\mathcal{L}(\mathcal{C})$ for a set of event classes $\mathcal{C}$, which is the sum of the number of layer 1 node in each event graph $g_j \in \Gamma$ weighted by $P(g_j|c_i,\mathcal{C})P(c_i|\mathcal{C})$.

**Generation of Event Graphs.** For a set of event classes $\mathcal{C}$, a bag of event graphs $G = g_1, ..., g_m$ are sampled i.i.d according to the joint probability distribution $P(g_j, c_i|\mathcal{C})$. First an event class $c_i$ is sampled according to the distribution $P(c_i|\mathcal{C})$. Then an event graph $g_j$ is sampled from $c_i$ according to $P(g_j|c_i,\mathcal{C})$. Given a set of event classes $C$, the conditional probability $P(G|C)$ for a set of event graphs $G$ is given by

$$P(G|\mathcal{C}) = \prod_{g_j} \sum_{c_i} P(g_j|c_i,\mathcal{C})P(c_i|\mathcal{C}) \quad (2)$$

**Generation of an Activity Graph.** The spatio-temporal relations between objects across different event graphs i.e. *inter event graph* interactions have not been specified so far in the generative process. In the third step, objects across different events are spatio-temporally related in a structure called the activity graph $AG$, by allowing for the possibility of both sharing of objects and interaction between objects across different event graphs. The layer 1 nodes in the $AG$ correspond to tracks such that each event graph $g \in G$ is a subgraph of the $AG$. Additionally, the *tracklets* mapped to by the layer 1 nodes of $g$ are segments of the *tracks* mapped to by the corresponding nodes of the $AG$.

Given a set of event graphs $G$, a conditional distribution $P(AG|G)$ is assumed to hold over the space of activity graphs $AG$, according to which certain $AG$s are more probable than others for activities across many domains. This distribution has two components $\chi_{inter}(AG)$ and $\mathcal{O}(AG)$.

$$P(AG|G) = \frac{1}{z_2} \exp\left(-\lambda_5 \chi_{inter}(AG) - \lambda_6 \mathcal{O}(AG)\right) \quad (3)$$

We assume that objects *across* the event graphs interact *less actively* than objects represented *within* each event graph. Therefore, the first component favours activity graphs with less interactivity between objects across the event graphs i.e. *lesser inter event graph interactivity* $\chi_{inter}(AG)$. Interactivity $\chi_{inter}(AG)$ is the mean of the interactivity $\chi(g)$[4] of all the graphs $g$ that represent interactions between objects across the event graphs. The second component favours activity graphs with a smaller proportion of overlapping objects between different events $\mathcal{O}(AG)$.

[4]Interactivity for any set of interactions is defined in §6.

**Generation of Tracks.** In the final step, the activity graph is *embedded* as tracks in space and time with concrete objects. Certain sets of tracks are regarded as being better embeddings of the activity graph in space-time than others if they are more prototypical of the spatial relations. The degree of typicality is measured using the distance measure underlying the HMM model described in §3. Embeddings which are highly typical should be generated with a high probability. The motivation for this is that such embeddings will be a perceptually clearer instantiation of the AG.

The probability $P(T|AG)$ of generating a set of tracks $T$ given an activity graph $AG$ can be expressed by decomposing the $AG$ into independent HMMs for each pair of tracks $\{T_i, T_j\}$. The activity graph $AG$ captures a sequence of states $S(T_i, T_j)$ between the tracks $T_i$ and $T_j$ as predicted by a HMM model $\theta$. The probability $P(\delta(T_i, T_j)|S(T_i, T_j), \theta)$ is the probability of a sequence of distances $\delta(T_i, T_j)$ between the tracks given the sequence of states $S(T_i, T_j)$ in the $AG$ and $\theta$ respectively. Thus $P(T|AG)$ can be expressed as the following product.

$$P(T|AG) = \prod_{(T_i,T_j)} P(\delta(T_i, T_j)|S(T_i, T_j), \theta) \quad (4)$$

**Events.** When the activity graph is embedded as tracks, the embedding corresponding to each of the event graphs are the *events*. More formally, an event $\varepsilon_j$ is an embedding of an (abstract) event graph $g_j \in c_i$ with a particular subset of tracklets. This process is formalized by a mapping $\mu(\varepsilon_j) = g_j$ as illustrated with arrows in Fig.2(a). An *event set* is a grouping of tracklets into events such that each event corresponds to some event graph in some event class. More formally, $E \equiv \{\varepsilon : g \in \mathcal{C} \wedge \mu(g) = \varepsilon\}$.

## 5 Unsupervised Activity Understanding.

In this unsupervised framework for activity understanding, a set of tracks $T$ is given for the video(s) of a domain. A candidate *interpretation* $I$ is defined as a tuple $\mathcal{I} = \langle \mathcal{C}, G, AG \rangle$ of candidate event classes $\mathcal{C}$, candidate event graphs $G$ and a candidate activity graph $AG$, that could have possibly generated the tracks $T$.

The goal of unsupervised event learning is regarded as the task of finding the most likely interpretation $\hat{\mathcal{I}}$ given a set of tracks. Accordingly, the existence of a *target distribution* of possible interpretations is assumed, according to which each interpretation $\mathcal{I}$ has some posterior probability (given tracks) and the optimal interpretation $\hat{\mathcal{I}}$ has the highest posterior probability.

$$\hat{\mathcal{I}} = \arg\max_{\mathcal{I}} P(\mathcal{I}|T) \quad (5)$$

$$= \arg\max_{\mathcal{C},G,AG} P(\mathcal{C})P(G|\mathcal{C})P(AG|G)P(T|AG) \quad (6)$$

This factorization is derived by making the following assumptions : (i) $T$ is conditionally independent of $G$ and $\mathcal{C}$ given $AG$; (ii) $AG$ is conditionally independent of $\mathcal{C}$ given $G$. The probability $P(T)$ is not represented in the factorization as the set of tracks $T$ are given and so does not influence the comparison between the interpretations.

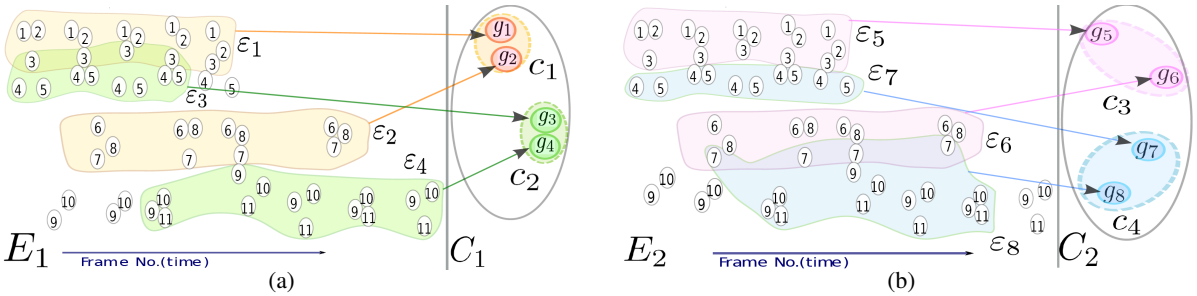



Figure 2: *Overview :* Illustrating two possible interpretations for the same set of tracklets. The interpretation $\mathcal{I}_1$ is better because it has more compactly clustered event classes in comparison to those of $\mathcal{I}_2$ as described further in §6. Note the following which make the learning problem complex: (i) the events can have overlapping object tracklets e.g $\varepsilon_1, \varepsilon_3$ share tracklet 3.; (ii) There are interactions between tracklets across events e.g. between objects 7,9 of events $\varepsilon_2, \varepsilon_4$ respectively.

**Candidate Interpretations.** In the learning phase, candidate interpretations are generated bottom up from an observed set of tracks $T$ as follows. First, a candidate event set $E = \{\varepsilon_1, ..., \varepsilon_m\}$ is chosen as a possible decomposition of a set of tracks into events. Two candidate event sets $E_1 = \{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4\}$ and $E_2 = \{\varepsilon_5, \varepsilon_6, \varepsilon_7, \varepsilon_8\}$ for the same set of tracks are shown in Fig. 2(a) and 2(b) respectively.

A bag of event graphs $G = g_1, ..., g_m$ are induced from $E$ using the procedure described in §3 with a corresponding mapping $\mu(G) = E$. A candidate activity graph $AG$ is also induced from $T$ using the same procedure. The conditional probabilities $P(AG|G)$ and $P(T|AG)$ are computed using equations 3 and 4 respectively.

The bag $G$ may contain several identical (isomorphic) candidate event graphs. The unique graphs in the bag $G$ are grouped into a candidate set of event classes $\mathcal{C} = \{c_1, c_2, ..., c_p\}$, such that each class $c_i$ contains a set of similar candidate event graphs. The probabilities $P(g_j|\mathcal{C}, c_i)$ and $P(c_i|\mathcal{C})$ are estimated for each $g_j \in G$ and $c_i \in \mathcal{C}$ as follows. Let $\|g_j\|$ be the frequency (or the number of isomorphic instances) of graph $g_j$ in the bag $G$. The cardinality $\|c_i\|$ of a candidate event class $c_i$ is measured as $\sum_{g_k \in c_i} \|g_k\|$, which is just the sum of the frequencies $\|g_k\|$ of all the graphs $g_k$ in $G$, whenever $g_k \in c_i$. Then $P(g_j|\mathcal{C}, c_i)$ can be estimated simply by normalizing $\|g_j\|$ with $\|c_i\|$. Similarly, $P(c_i|\mathcal{C})$ can be estimated simply by dividing $\|c_i\|$ by the number of graphs $m$ in $G$. Finally, the conditional distribution $P(G|C)$ is computed using the probabilities $P(g_j|\mathcal{C}, c_i)$ and $P(c_i|\mathcal{C})$ as given in equation 2.

Two sets of candidate event classes (i) $\mathcal{C}_1 = \{c_1, c_2\}$ where $c_1 = \{g_1, g_2\}$ and $c_2 = \{g_3, g_4\}$ corresponding to $G_1 = g_1, g_2, g_3, g_4$ (ii) $\mathcal{C}_2 = \{c_3, c_4\}$ where $c_3 = \{g_5, g_6\}$ and $c_4 = \{g_7, g_8\}$ corresponding to $G_2 = g_5, g_6, g_7, g_8$ are shown in Fig. 2(a) and Fig. 2(b) respectively. Note that the frequencies for all these event classes in the simplified illustration is equal to one.

Once the candidate event classes are constructed, the probability $P(\mathcal{C})$ is computed using equation 1. In this way a candidate interpretation $\mathcal{I} = \langle \mathcal{C}, G, AG \rangle$ is formed from a set of tracks $T$. The illustration in Fig. 2(a) and 2(b) correspond to two interpretations $\mathcal{I}_1 = \langle C_1, G_1, AG_1 \rangle$ and $\mathcal{I}_2 = \langle C_2, G_2, AG_2 \rangle$ respectively, for the same set of tracks. For any candidate interpretation $I$, the posterior probability $P(I|T)$ is computed using the factorization given in equation 6. An efficient strategy to search for the optimal interpretation is described in §7.

## 6 Within Class Similarity and Interactivity

The precise notions of the two key components – within class similarity and interactivity - that influence the posterior probability for candidate interpretations are detailed below.

**Within Class Similarity $\omega$.** This term favours interpretations which are characterized by compact event classes. Compactness can be characterized using *within class similarity*. The event classes were shown to be derived from the bag representation. In the bag based representation $B$, the within class similarity for a bag is just the average of all pairwise similarities (as expressed by the BoG kernel) between the event graphs that constitute the bag. It can be shown through algebraic manipulations that the within class similarity $\omega(\mathcal{C})$ for the corresponding set of event classes $\mathcal{C}$ can be expressed by weighting the kernels with the corresponding probabilities. Using the following notations $P_i = P(c_i|C), P_{ij} = P(g_j|c_i, \mathcal{C}), P_{ik} = P(g_k|c_i, \mathcal{C})$, within class similarity is expressed as follows.

$$\omega(C) = \sum_{c_i \in \mathcal{C}} \frac{P_i}{1 - P_i} \left[ 2 \sum_{g_j, g_k \in c_i} P_{ij} P_{ik} \mathcal{K}_{jk} + \sum_{g_j \in c_i} P_{ij}(1 - P_{ij}) \mathcal{K}_{jj} \right]$$

The reason for using within class similarity as a property of the prior, can be explained by comparing two possible interpretations $\mathcal{I}_1$ (Fig. 2(a)) and $\mathcal{I}_2$ (Fig. 2(b)), for the same set of tracks.

For the first interpretation $\mathcal{I}_1$ in Fig. 2(a), it can be observed that the object interactions in event $\varepsilon_1$ are similar to those in $\varepsilon_2$. Since they share similar interactions i.e. similar graphemes, the induced event graphs $g_1, g_2$ are similar with respect to the BoG kernel. Analogously, a similar set of interactions in $\varepsilon_3$ and $\varepsilon_4$ gives rise to a similar set of event graphs $g_3, g_4$. Therefore the interpretation $\mathcal{I}_1$ results in compactly clustered set of event classes $\mathcal{C}_1 = \{c_1, c_2\}$, where

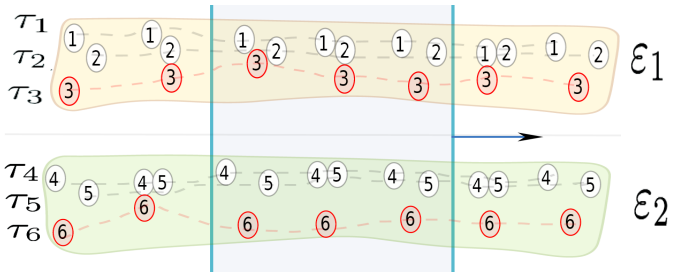



Figure 3: Interactions between tracklets $\tau_1, \tau_2, \tau_3$ and $\tau_4, \tau_5, \tau_6$ for two events $\varepsilon_1$,$\varepsilon_2$ are shown. Changes in spatial relations are more evenly distributed across all subsets tracklets for the event $\varepsilon_1$ than for the event $\varepsilon_2$.

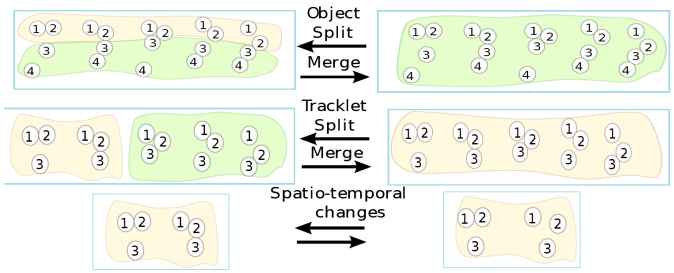



Figure 4: Illustrates the three types of moves for the MCMC search. The first type of move splits or merges events based on objects. The second type of move splits or merges tracklets at episode boundaries. The third type of move results in changes in spatio-temporal relations.

$c_1 = \{g_1, g_2\}$ and $c_2 = \{g_3, g_4\}$ with high *within class similarity*, as illustrated in Fig. 2(a).

In contrast, for the interpretation $\mathcal{I}_2$ in Fig. 2(b), it can be observed that the interactions for the events $E_2 = \{\varepsilon_5, \varepsilon_6, \varepsilon_7, \varepsilon_8\}$ are less similar to one another, producing no obvious grouping of event graphs $\{g_5, g_6, g_7, g_8\}$ into compact event classes. Therefore, if a near optimal set of event classes $\mathcal{C}_2 = \{c_3, c_4\}$ where $c_3 = \{g_5, g_6\}$ and $c_4 = \{g_7, g_8\}$ respectively is assumed, the *total within class similarity* of the interpretation $\mathcal{I}_2$ in Fig. 2(b) is expected to be *lower than* the corresponding measure for the interpretation $\mathcal{I}_1$ in Fig. 2(a).

A greater prior probability is assigned to the interpretation with a higher within class similarity, since this would favour interpretations that characterize more patterns in the data, thereby enabling us to describe the video with a more compressed representation, in terms of compact event classes.

**Interactivity $\chi$ .** This term prefers candidate event graphs $g$ for which all objects engage actively to those in which some objects interact far more than others. For simplicity, interactivity $\chi(g)$ for an event graph $g$ is formulated below in terms of the corresponding event $\varepsilon = \mu(g)$. Interactivity $\chi(\varepsilon)$ for an event $\varepsilon$ can equivalently be thought of as a measure that has higher values whenever a greater number of subset of objects $e \subseteq \varepsilon$ have a high amount of interaction.

To compute this quantity consider a candidate event $\varepsilon \subseteq T$ for a set of tracks $T$ amongst which there is a total of $r$ interactions. *Interaction probability* $P_\chi(e)$ is defined for the subsets $e \subseteq \varepsilon$, as the probability that the objects in $e$ are actively interacting with each other. This is computed by considering $r - \sigma$ windows, where $\sigma$ is equal to the number of interactions in each of these windows. One such window is illustrated in Fig. 3, where $\sigma = 2$ for the two events $\varepsilon_1$ and $\varepsilon_2$. The interaction probability $P_\chi(e)$ is defined as the proportion of such $r - \sigma$ windows in which each object in $e$ undergoes at least one change in spatial relationships with other objects in $e$. It can be observed from Fig. 3, that all subsets of 2 objects $(\{(1,2),(2,3),(1,3)\})$ in $\varepsilon_1$ have a larger proportion of such windows than the subsets of 2 objects $(\{(4,5),(5,6),(4,6)\})$ in $\varepsilon_2$ .

*Interactivity* is defined as a measure of the extent to which the interaction probability $P_\chi(e)$ is distributed uniformly across interactions between all subsets of objects $e \subseteq \varepsilon$ for an event $\varepsilon$. Interactivity is measured by extending total correlation as described in (Watanabe 1960) to *pointwise total correlation* $\chi(\varepsilon)$ which aggregates the interaction probabilities $P_\chi(e)$ of all $e \subseteq \varepsilon$ :

$$\chi(\varepsilon) \quad := \quad \sum_{e:e\subseteq\varepsilon} P_\sigma(e) \log \left( \prod_{d:\epsilon\subseteq e} P_\sigma(d)^{q_d} \right), q_d = (-1)^{\|d\|}$$

Interactivity $\chi_{intra}(\mathcal{C})$ for a set of event classes $\mathcal{C}$ is the mean of the interactivity $\chi(g_j)$ of all the event graphs $g_j$ that constitute these classes.

## 7 Search for the Most Likely Interpretation

The space of interpretations is searched to find the most likely interpretation using the following procedure. In practice, enumerating all possible interpretations is infeasible. Therefore MCMC with simulated annealing (Kirkpatrick et al. 1983) is used to generate a Markov chain of event sets $(E_1, ..., E_t, ...)$, using three types of moves (Fig. 4) which transforms an event cover $E_t$ to another event cover $E_{t+1}$.

For any event set $E_t$ that is generated, the corresponding activity graph $AG_t$ and the bag of event graphs $G_t = ..g_i..$ are induced. Enumerating all possible class labellings for a bag of event graphs $G_t$ is also infeasible. Therefore, a *class labelling* $\mathcal{C}_t$ that is likely to be near optimal is obtained by clustering the event graphs using self tuning spectral clustering (Zelnik-Manor and Perona 2005) with the BoG kernel. Thus a candidate interpretation $I_t = \langle C_t, G_t, AG_t \rangle$ corresponding to the event cover $E_t$ is obtained. The candidate interpretation $I_t$ of the Markov chain is accepted or rejected as given by the acceptance probability. Simulated annealing is used to speed up the convergence of the MCMC sampling. In this manner the posterior distribution of the candidate interpretations is sampled and the optimal interpretation is chosen from this sample after convergence.

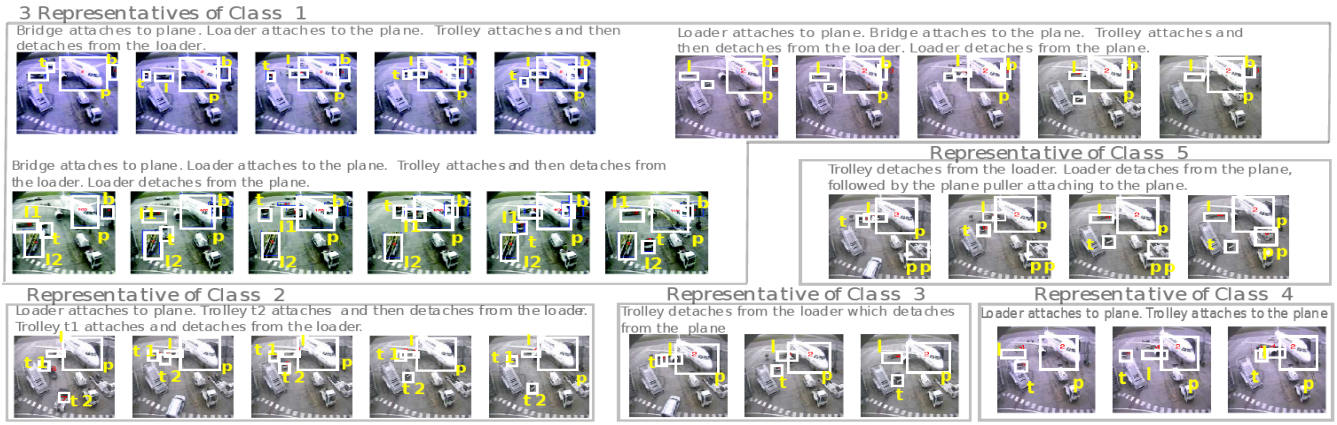



Figure 5: Representative samples from 5 out of 14 learned event classes are shown within the regions given by the gray boxes. Events are shown as a sequence of images with bounding boxes on the relevant objects and corresponding descriptions above these images. The following short forms are used for the corresponding object types : p - plane, l - loader, t - trolley, b - bridge, pp - plane puller.

## 8 Experiments

The proposed framework was evaluated on real data of a video showing servicing of aircraft between flights. A suitable set of parameters ($\lambda_1$ to $\lambda_6$) given in §4 for application on to the real data is determined from synthetic data for the following reason. These parameters influence the relative importance of the various factors in the generative process such as (i) within class similarity; (ii) intra event graph interactivity; (iii) number of classes; (iv) size of the event graphs;(v) inter event graph interactivity; (vi) overlap. Even though the relative importance may vary from domain to domain, it is assumed that many interesting activities are generated by a preference for greater values of factors (i),(ii),(iv) and smaller values for (iii),(v),(vi). These assumptions offer a promising solution for distinguishing events from other less significant interactions for many domains.

**Parameter Determination from Synthetic Data.**

Synthetic data is constructed according to the generative process outlined in §4. The number of classes varies between 3 and 5. The number of event graphs in each class varies between 2 and 3, each with certain probability. Forty event graphs are sampled from this distribution of event classes and embedded in an activity graph such that 10% of the event graphs have shared objects and 10% have interactions between objects across different events. The activity graph thus generated is embedded as tracks. Since the data is generated synthetically, the optimal interpretation is known and the parameters ($\lambda_1$ to $\lambda_6$) can be determined as follows. The optimal interpretation is first degraded using the set of moves described in §7 and a set of 12 interpretations is obtained. For these 12 interpretations, the parameters that produce a posterior distribution that most favours the the optimal interpretation are regarded as a set of suitable parameters for real data. The 12 interpretations form the horizontal axes of the two plots in Fig. 6 (interpretation 3 is the one predetermined as the optimal). The vertical axes are their posterior probabilities in the first plot and the within class similarity and interactivity in the second plot respectively.

**Real Data Set.** The proposed method is evaluated on approximately 12 hours of video showing servicing of aircraft between flights. The camera positioning for all the eight turnarounds is the same, so we obtain the same view. This dataset was chosen since it clearly contains structured events. However the problem of learning is complex as these may occur in parallel with objects shared between events (e.g. the plane) and interactions between objects across different events. Moreover, the tracking output introduces more complexities that arise due to unstable bounding boxes, missing detections, mistakes in tracking and the presence of noisy blobs.

**Detection and Tracking.** First, six visual appearance models are learned for object classes (1.Plane 2.Trolley 3.Carriage 4.Loader 5.Bridge 6.Plane Puller) from two hours of video. Instances of these object classes are detected using the technique in (Ott and Everingham 2009) and tracked using techniques in (Yu and Medioni 2008) for the rest of the 10 hours of video. Note that although the tracked objects have types as a result of the detection based tracking technique that is used, the event learning procedure deliberately ignores these in order not to be dependent on them. Thus in principle, it could work equally with untyped tracks.

**Evaluation by Qualitative Inspection** A total of 14 event classes with varying number of events were obtained with the proposed framework. A qualitative inspection informs that the framework has been able to discover several interesting events which compose the set of activities in the aircraft domain. A representative set of events are shown in Fig. 5 from which the following observations can be made. First, it can be seen that class 1 has been able to capture very similar interactions between trolleys, loaders, planes and bridges and these interactions usually span the entire servicing a plane over 70000 image frames. A representative

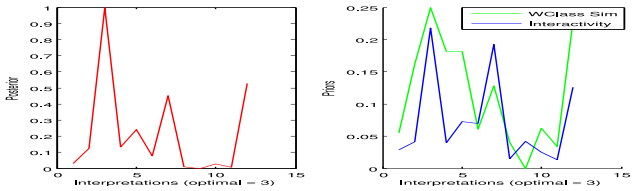



Figure 6: Left : Plot of 12 Interpretations (optimal = 3) and their posteriors. Right : Plot of Within class Similarity and Interaction for the 12 interpretations (optimal = 3)]

event from class 2 is very typical for aircraft scenarios and usually takes place in the middle of a turnover when multiple trolleys arrive and depart with baggage. From a similar inspection of representatives from the other event classes in Fig. 5, it can be concluded that the proposed technique has discovered event classes whose events represent significant interactions that take place in aircraft handling scenarios.

**Evaluation with Respect to Predefined Set of Events**

The performance of the proposed framework was evaluated with respect to a pre-determined set of event classes: 1.Unloading(Un) 2.Bridge/loader(Br) attaches and detaches from the plane 3.Plane Puller(PP) attaches to the plane. These classes were predefined as *interesting* with respect to monitoring tasks that were prescribed by independent domain experts. A ground truth of the 6 turnarounds for these three classes was defined by other domain experts. The results obtained are summarized in table 1 whose rows are the predefined event classes.

The columns of table 1 are the five mined event classes[5] in which these three predefined classes are present. The *first* of the pair of entries of table 1 are the fraction of these classes that are discovered with respect to the total number of events present in the test data (as obtained from ground truth). It can be observed that the proposed framework has been able to discover and group $58\%$ of the unloading events into cluster $C_1$ and a smaller percentage $16\%$ into cluster $C_3$. In total it has been able to discover $75\%$ of the unloading events from the video. The *second* of the pair of entries of table 1 are the fraction of these classes that are discovered with respect to the total number of events present in cluster (as obtained from ground truth). For the unloading operation, it can be seen that $77\%$ of the cluster $C_1$ and $66\%$ of cluster $C_3$ are unloading operations. A similar analysis can be carried out for the other two classes. We can conclude from these results that the proposed framework gives a promising performance on these predefined event classes.

## 9 Summary and Future Work

In this work we have proposed a framework for unsupervised learning of event classes from videos based on a probabilistic model for the generation of activities. Experimental results have shown that the proposed framework offers a novel and promising direction for discovering semantically meaningful events on challenging videos with complex events, despite visual noise in the tracked input.

In the future, we plan to use the learned event classes for detecting events in an unseen video or classifying unseen events as normal or abnormal. We also plan to extend the proposed framework with techniques in (Sridhar, Cohn, and Hogg 2008) to have a unified model for learning functional object classes and event classes.

[5]The other mined event classes do not correspond to the "official IATA" events determined by the domain experts, but do correspond to semantically meaningful and interesting events.

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $Total$ |
|---|---|---|---|---|---|---|
| $Un$ | .58,.77 | 0,0 | .16,.66 | 0,0 | 0,0 | .75,75 |
| $Br$ | 0,0 | .66,.80 | 0,0 | .16,1.0 | 0,0 | .86,.38 |
| $PP$ | 0,0 | 0,0 | 0,0 | 0,0 | .83,1.0 | .83,1.0 |

Table 1: Evaluation with respect to a predefined set of events

## References


Cohn, A. G., and Hazarika, S. M. 2001. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae* 46(1-2):1–29.

Hamid, R.; Maddi, S.; Johnson, A.; Bobick, A.; Essa, I.; and Isbell, C. 2009. A novel sequence representation for unsupervised analysis of human activities. *Art Intell.* 173(14):1221–1244.

Ivanov, Y., and Bobick, A. 2000. Recognition of visual activities and interactions by stochastic parsing. *IEEE T. PAMI* 22(8):852–872.

Kersting, K.; Raedt, L. D.; Gutmann, B.; Karwath, A.; and Landwehr, N. 2008. *Relational Sequence Learning*.

Kirkpatrick, S.; Gelatt, C. D.; Jr.; and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Lavee, G.; Rivlin, E.; and Rudzsky, M. 2009. Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. *Trans. Sys. Man Cyber Part C* 39(5):489–504.

Ott, P., and Everingham, M. 2009. Implicit color segmentation features for pedestrian and object detection. *In Proc. ICCV*.

Sridhar, M.; Cohn, A. G.; and Hogg, D. C. 2008. Learning functional object-categories from a relational spatio-temporal representation. In *Proc. ECAI 2008*, 606–610.

Wang, X.; Ma, X.; and Grimson, E. 2009. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE T. PAMI* 31(3):539–555.

Watanabe, S. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development* 4:66+.

Yu, Q., and Medioni, G. 2008. *Integrated Detection and Tracking for Multiple Moving Objects using Data-Driven MCMC Data Association*. IEEE Computer Society.

Zelnik-Manor, L., and Perona, P. 2005. Self-tuning spectral clustering. In Saul, L. K.; Weiss, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press. 1601–1608.