# Discovering an Event Taxonomy from Video using Qualitative Spatio-temporal Graphs

**Muralikrishna Sridhar** and **Anthony G Cohn** and **David C Hogg** [1]

**Abstract.** This work proposes a graph mining based approach to mine a taxonomy of events from activities for complex videos which are represented in terms of qualitative spatio-temporal relationships. A Hidden Markov Model to obtain stable qualitative spatial relations from noisy measurements is introduced. The effectiveness of the approach is demonstrated through experimental results for a complex aircraft turnaround apron scenario.

## 1 Introduction

Imagine a camera observing a scene such as an aircraft apron or a kitchen. In such complex scenes, activities are usually composed of multiple events that may occur in parallel, and where overlapping events may share participating objects. Another source of complexity is the presence of spurious and missing objects and relationships, arising either due to instability in image processing or due to coincidental occurrences. We address the important problem of *unsupervised discovery* of a part-of event taxonomy from such complex videos.

Our earlier work [2] addressed the challenge of making a bridge between low level input from a video stream and high level representation of events using a relational qualitative spatio-temporal representation called an *activity graph* to represent interactions between all objects in a scene. We have very recently developed [3] a more robust variable free representation of interactions and a generic focus mechanism called *interactivity*, both of which we adopt in this work. In [3], we focussed on learning the most probable interpretation of a video using a generative model. In this work, we adopt a complementary approach of learning an event taxonomy in a bottom up graph mining fashion by characterizing *events* as sufficiently *frequent* and *interactive* subgraphs of the activity graph. The underlying hypothesis is that non-events which may be observation noise or coincidences do not tend to possess these systematic properties. We also present a novel HMM based technique that is used to obtain a more robust qualitative spatial representation from noisy video input.

## 2 Qualitative Spatio-Temporal Graphs

Given a video, object tracks (e.g. $\tau_1, \tau_2$ in Fig. 1.a) are obtained using techniques in [4]. The following summarizes the construction of a 3 layer activity graph that represents interactions between all the tracks in a scene. For the purpose of brevity, we use an over simplified scene with just two tracks $\tau_1, \tau_2$ shown in Fig. 1.a and represent their interaction in an activity graph shown in Fig. 1.c. The layer 1 nodes of the activity graph (Fig. 1.c) map injectively to the tracks $(\tau_1, \tau_2)$, but

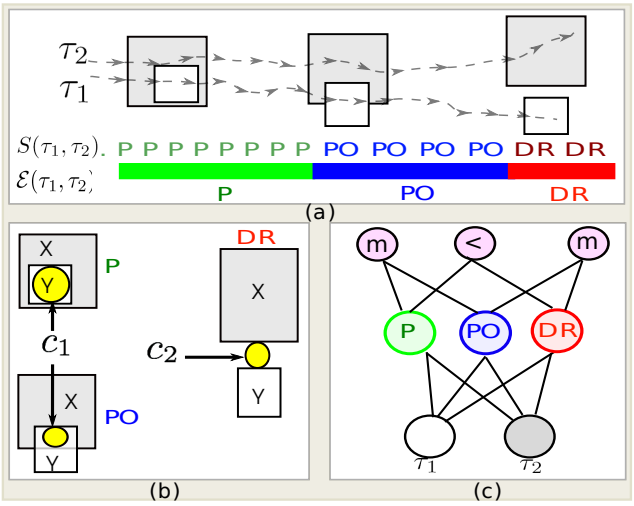


**Figure 1.** (a) Two tracks, $\tau_1, \tau_2$, sequence of spatial relations $S(\tau_1, \tau_2)$ and episodes $\mathcal{E}(\tau_1, \tau_2)$. (b) Spatial relations $\{\mathsf{P}, \mathsf{PO}, \mathsf{DR}\}$. (c) Activity graph for the interaction between $\tau_1, \tau_2$.

are not explicitly labelled with these tracks, in order to abstract away details specific to these tracks. For each pair of tracks (e.g. $\tau_1, \tau_2$), a sequence (e.g. $S(\tau_1, \tau_2)$ in Fig. 1.a) of qualitative spatial relations, which are either $\{\mathsf{P}, \mathsf{PO}, \mathsf{DR}\}$[3] (Fig. 1.b) are computed using a HMM which is detailed in section 4.

For each pair of tracks, the sequence of spatial relations is aggregated to a sequence of *episodes* (e.g. $\mathcal{E}(\tau_1, \tau_2)$ in Fig. 1.a), such that within each episode the same spatial relation holds, but a different spatial relation holds immediately before and after the episode. Layer 2 nodes of the activity graph represent these episodes between the respective pairs of tracks 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. The layer 3 nodes of the activity graph are labelled with Allen's temporal relations (e.g. $\mathsf{m}$ : meets, $<$: before in Fig. 1.c) between intervals corresponding to pairs of layer 2 nodes.

While we have adopted a graph based representation, the interaction in Fig. 1.a can also be expressed as a logical formulae with $X, Y$ as object variables and $I_1, I_2, I_3$ as temporal variables:

$$holds(X, Y, \mathsf{P}, I_1) \wedge holds(X, Y, \mathsf{PO}, I_2) \wedge holds(X, Y, \mathsf{DR}, I_3) \\ \wedge \mathsf{meets}(I_1, I_2) \wedge \mathsf{meets}(I_2, I_3) \wedge \mathsf{before}(I_1, I_3)$$

## 3 Mining an Event Taxonomy

*Event graphs* are subgraphs of the activity graph that represent significant interactions, i.e. *events* which serve to structure the activities

[1] University of Leeds, UK, {krishna,agc,dch}@comp.leeds.ac.uk. This work is supported by the EPSRC (EP/D061334/1) and the EU FP7 (Project 214975, Co-Friend). We also thank colleagues in the Co-friend project.

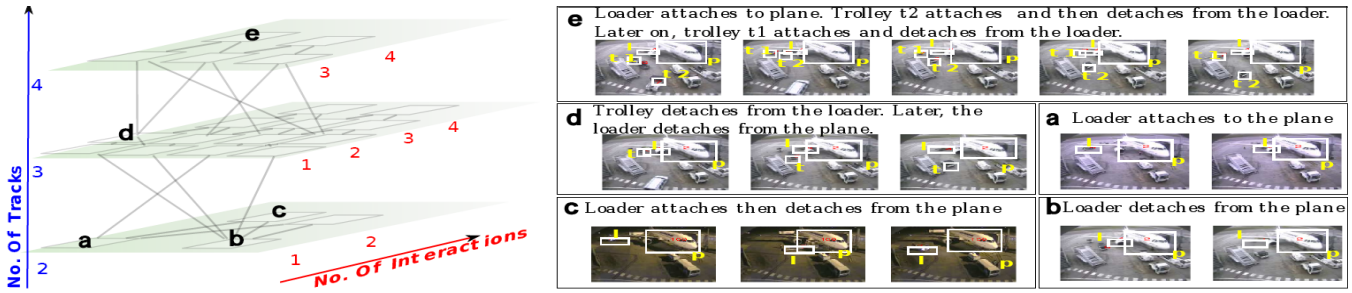


**Figure 2.** Sample events *a–e* from the taxonomy on the left are shown on the right. In the image sequences on the right, next to the respective bounding boxes are symbols that stand for the object types : b - bridge, p - plane, pp - plane puller, t - trolley, l - loader.

in a domain. It is reasonable to assume that interactions arising from observation noise and coincidences are random whereas events tend to be *frequent* (in a self similar [3]/identical manner, though we assume identity here) and are composed of actively interacting objects with higher values of *interactivity* [3].

We apply a level-wise graph mining procedure [1] to obtain event classes with the top k% of frequent and interactive subgraphs at each level of the taxonomy. Each level is given by a pair of numbers (i, j), where i represents the number of tracks and j is the number of interactions between these tracks (e.g. *i*=2 tracks and *j*=2 interactions in Fig. 1.a). Once the level-wise search terminates, all event classes which are *non-maximal* are eliminated, i.e. those classes all of whose instances appear as instances of a subgraph of a superclass. Finally, we eliminate those graphs from the remaining event classes, which are not a subgraph of any graph of the respective superclasses.

Thus, we obtain a taxonomy of event classes all of whose graphs naturally compose their respective superclasses and thus contribute to the entire structure of the taxonomy. By preferring both frequent and maximal subgraphs, the event taxonomy can be regarded as a *compressed description* of the activities in the video.

## 4 Qualitative Spatial Relations via a HMM

An HMM offers a way of computing a sequence of qualitative spatial relations (e.g. $S(\tau_1, \tau_2)$ in Fig. 1.a) between a pair of tracks (e.g. $\tau_1, \tau_2$) with the benefit of smoothing the rapidly flipping transitions between spatial relations that arise from visual noise. For each frame in the video sequence, a quantitative *measure of separation* $\delta(x, y) = \frac{d(c_2)-d(c_1)}{min(d(x),d(y))}$ is computed between the bounding boxes $x, y$ of objects corresponding to pair of tracks (e.g. $\tau_1, \tau_2$). In this expression, as illustrated in Fig. 1.b, $d(c_1)$ is the diameter of the largest circle in the intersection of two regions $x$ and $y$, $d(c_2)$ is the diameter of the smallest circle that connects two disconnected regions $x$ and $y$, $d(x), d(y)$ are the diagonals of $x, y$ respectively. $\delta(x, y)$ is positive when $x, y$ are disconnected, zero when they touch and decreases further as the relations change from DR to PO towards P. The denominator makes $\delta(x, y)$ independent of the sizes of $x, y$.

The observation models of the HMM for the three spatial states $\{\mathsf{DR}, \mathsf{PO}, \mathsf{P}\}$ are modelled by three corresponding logistic functions. The parameters of the observation models are learned for each state, using maximum likelihood estimation on training data, which are composed of manually labelled sequences of spatial relations. The transition probabilities are learned from the statistics of bi-grams of spatial relations in the annotation. With a trained HMM, a Viterbi decoder is used to predict the most likely sequence of spatial relationships (e.g. $S(\tau_1, \tau_2)$ in Fig. 1.a), given a sequence of observed measures of separation between a pair of tracks (e.g. $\tau_1, \tau_2$ in 1.a).

## 5 Experiments

The proposed graph mining approach was applied with top $k = 70\%$ for frequency and interactivity, to an activity graph with 749700 nodes, corresponding to activities for approximately 12 hours of video showing servicing of aircraft between flights.

A qualitative evaluation of the resulting taxonomy in Fig. 2 demonstrates that the technique has been able to discover event classes and organize them in a hierarchy. To start with, very simple but significant interactions, such as in *a* (attach) and *b* (detach), and their combinations (attach-detach) in *c* form the bottom of the taxonomy. At the middle level (3,.), interactions between a trolley,plane and loader such as in *d* represent events that are typical for aircraft handling scenarios, as these interactions are central to the loading operation. The event given by letter *e* usually takes place in the middle of a turnover when multiple trolleys arrive and depart with baggage.

A quantitative evaluation is performed with respect to a predetermined set of three event classes prescribed by domain experts. The proposed technique was able to discover (i) $51\%$ of the occurrences for the *unloading* event class without the HMM and $73\%$ with the HMM; (ii) $66\%$ of the occurrences for the *bridge attach-detach* event class without the HMM and $83.3\%$ with the HMM; (iii) $83.3\%$ of the occurrences for the *plane puller attach* event class without the HMM and $100\%$ with the HMM. From the quantitative evaluation, we conclude that despite the imperfect inputs from tracking and absence of any supervision for each of these event classes, the proposed technique has discovered these pre-defined events with a reasonably high accuracy, which is boosted further by the use of HMM.

## 6 Future Work

In the future, we plan to use the learned classes for detecting events in an unseen video or classifying unseen events as normal or abnormal. We also plan to compare the proposed technique to [3] and extend the definition of event classes to include similarity as in [3]. Finally, we plan to investigate the idea of learning event classes and functional object classes [2] simultaneously.

## REFERENCES

[1] D J Cook and L B Holder, *Mining Graph Data*, Wiley-Interscience, 2007.
[2] Muralikrishna Sridhar, Anthony G. Cohn, and David C. Hogg, 'Learning functional object-categories from a relational spatio-temporal representation', in *In Proc. ECAI*, (2008).
[3] Muralikrishna Sridhar, Anthony G. Cohn, and David C. Hogg, 'Unsupervised learning of event classes from video', *In Proc. AAAI*, (2010).
[4] Q Yu and G Medioni, 'Integrated detection and tracking for multiple moving objects using data-driven mcmc data association', *In Proc. WMVC*, (2008).