

Benchmarking Qualitative Spatial Calculi for Video Activity Analysis

Muralikrishna Sridhar, Anthony G Cohn and David C Hogg

University Of Leeds, UK,
{krishna,agc,dch}@comp.leeds.ac.uk

Abstract. This paper presents a general way of addressing problems in video activity understanding using graph based relational learning. Video activities are described using relational spatio-temporal graphs, that represent qualitative spatio-temporal relations between interacting objects. A wide range of spatio-temporal relations are introduced, as being well suited for describing video activities. Then, a formulation is proposed, in which standard problems in video activity understanding such as event detection, are naturally mapped to problems in graph based relational learning. Experiments on video understanding tasks, for a video dataset consisting of common outdoor verbs, validate the significance of the proposed approach.

1 Introduction

One of the goals of AI is to enable machines to observe human activities and understand them. Many activities can be understood by an analysis of the interactions between objects in space and time. The authors in [13][14] introduce a representation of interactions between objects, using perceptually salient discretizations of space-time, in the form of qualitative spatio-temporal relationships. Then, they apply relational learning to learn event classes from this representation. This approach to understanding video activities using a qualitative spatio-temporal representation and relational learning is an alternative to much research on video activity analysis, which has largely focussed on a low-level pixel based representations e.g. [17].

This paper expands the scope of this research in the following two ways. Firstly, building on previous work [14], that has restricted itself to just simple topological relations, this work draws from a body of research in qualitative spatial relations [11][2], and proposes that these relations provide a natural way of representing video activities. This aspect is described in section 2. Secondly, this paper presents a general way of translating standard problems in video activity analysis [9] to problems in relational graph learning [4]¹, by extending the application of a novel formulation proposed in [14]. This aspect is described in section 3. Sections 4 describes experimental analysis on real data. Section 5 concludes this chapter with pointers to future research.

¹ While this paper concentrates on graph based relational learning for reasons given below, we believe that this analysis can be carried over to logic based relational learning [7] [10].

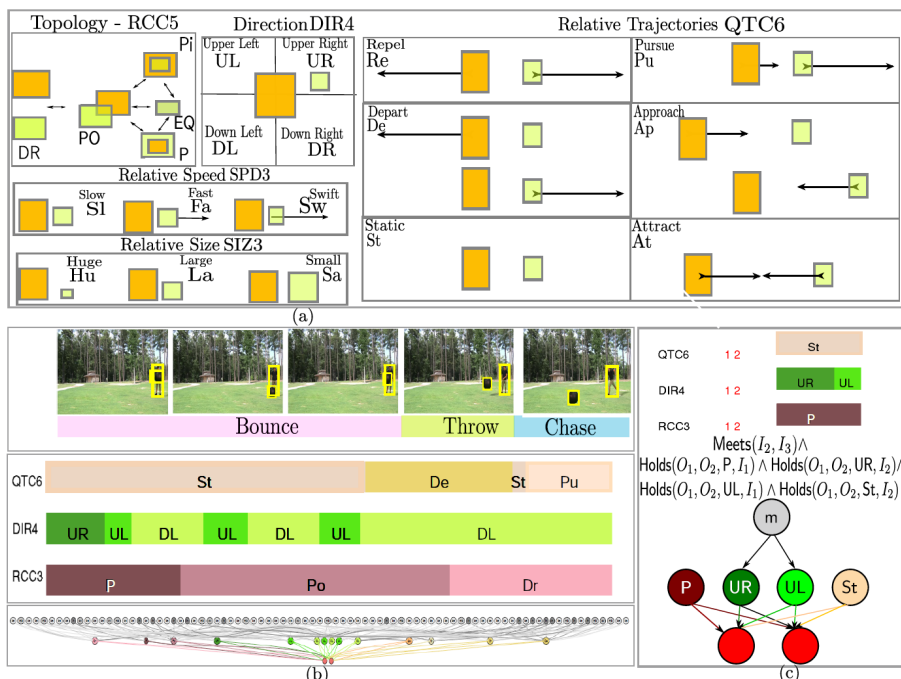


Fig. 1: (a) Five qualitative spatial relationships: (i) topology; (ii) direction; (iii) relative speed; (iv) relative size; (v) qualitative trajectories. (b) Three simple events: (i) bounce - characterized by a periodic change between the directional relationships UL and DL; (ii) throw - by the change from PO to DR and St to De; (iii) chase by a change from De to Pu. At the bottom is the corresponding spatio-temporal graph. (c) The same spatial relations in (b) for a short segment, its logical representation and equivalent relational interaction graph.

2 Graph Based Representation of Activities

We propose that qualitative spatial relations provide a natural way of representing interactions between objects participating in video activities. Qualitative relations form interesting features as they are the result of a particular way of discretizing quantitative measurements into qualitatively interesting concepts, such that these concepts signify *perceptually salient relationships* [3]. The problem of abstracting qualitative relations from noisy video data, is facilitated by the use of a Hidden Markov Model based framework described in [15].

Five types of relations are illustrated in Fig. 1 (a). Their suitability for describing interactions is illustrated in Fig. 1 (b). At the top of Fig. 1 (b) is a sequence of images representing the interaction between a person and a ball, namely bounce, throw and chase. Below that is shown, three “parallel sequences of episodes”. An episode [13] corresponds to an interval, during which a spatial relationship holds maximally, and can be described by logic (e.g. Holds(O_1, O_2, UR, I_2) as shown in Fig. 1(c) for a shorter sub-interval of the interval shown in Fig. 1(b)). Each sequence of episodes in Fig. 1(b) and (c) correspond to one of the three different types of qualitative relations, namely topology (RCC5), relative directions (DIR4) and relative trajectories (QTC6).

An alternative to the above “sequence of episodes” based representation is to relate the intervals corresponding to each pair of episodes, using Allen’s temporal relationships [1], e.g. $\text{Meets}(I_2, I_3)$, as shown in Fig. 1(c). This leads to a fully relational representation capturing many, if not all, qualitatively interesting temporal dependencies.

An alternative relational representation to logical predicates is to use interaction graphs [14], as shown in Fig. 1(c). They are three layered graphs, in which the layer 1 nodes are mapped to the interacting objects. Layer 2 nodes of the interaction graph represent the 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). The layer 3 nodes of the activity graph are labelled with Allens temporal relations (e.g. m : meets, in Fig. 1(c)) between intervals corresponding to certain pairs [12] of layer 2 nodes.

Interaction graphs are a computationally efficient alternative to logical predicates, as they avoid repetition of object and episode variables and also provide a well defined and computationally efficient comparison of interactions, by means of suitable similarity measure. This measure is defined using a kernel on a feature space obtained by expressing a interaction graph in terms of a bag of sub-interaction subgraphs [14].

An *activity graph* is an interaction graph that captures the spatio-temporal relationships between all pairs of co-temporally observed objects that are involved in activities for an extended duration. Note that the activity graph may also represent the spatio-temporal graph for activities in several unrelated videos for the same domain, and not necessarily one single video.

3 Graph Based Relational Learning of Activities

The authors in [14] proposed a novel relational graph based learning formulation for video activity understanding, in the context of a specific unsupervised learning task. In the following, we use this formulation to describe a general way of translating standard problems in video activity analysis to standard problems in relational graph learning. We show how it can be more generally applied, in order to address many of the standard video activity understanding tasks.

One of the key underlying hypotheses in research on video activity understanding [9] is that activities are composed of events of different types. Based on this hypothesis, tasks such as learning event class models, event classification, clustering and detection are defined. In this work, we characterize events by a set of co-temporal tracklets (a tracklet is a one-piece segment of a track). Events having similar spatio-temporal relationships between their constituent tracklets tend to belong to the same event class. The set of all event classes is called \mathcal{C} . A set of events E is a “cover” of a set of tracks \mathcal{T} iff the union of all tracklets in E is isomorphic to \mathcal{T} . In general there may be coincidental interactions between objects that that would not naturally be regarded as part of any event in an event class². This notion of an event cover can be regarded as an *global explanation* of the activities in a video in terms of instances of event classes.

A set of tracks \mathcal{T} can be abstractly represented using an activity graph \mathcal{A} , as described above. An event corresponds to a subgraph of \mathcal{A} , such that this subgraph is also

² We ignore this complexity here but see [12], [14]. The final paper would contain details of how coincidences can be incorporated into the learning algorithms. Here, there is no space to give further details here.

an interaction graph. An event cover in this formulation, thus becomes a set of interaction graphs, whose union is \mathcal{A} . These interaction graphs are called event graphs³. Similar event graphs tend to belong to the same event class. An *event model* is defined for the set of event classes \mathcal{C} , according to which, each class is a probability distribution over a finite set of interaction graphs. Finally, *observation noise* is modelled by allowing multiple possible activity graphs \mathcal{A} , for the same set of observed tracks \mathcal{T} . This formalism has been used to model the joint probability distribution of the above variables $\{\mathcal{C}, \mathcal{G}, \mathcal{A}, \mathcal{T}\}$ as:

$$P(\mathcal{C}, \mathcal{G}, \mathcal{A}, \mathcal{T}) \approx P(\mathcal{C})P(\mathcal{G}|\mathcal{C})P(\mathcal{A}|\mathcal{G}, \mathcal{C})P(\mathcal{T}|\mathcal{A})$$

We now apply this formulation to address the above video understanding tasks in terms of relational graph learning. The task of *learning an event model* translates to learning an event model for event classes \mathcal{C} given \mathcal{A} and a corresponding \mathcal{G} . A MAP formulation of this problem is

$$\hat{\mathcal{C}} = \arg \max_{\mathcal{C}} P(\mathcal{C})P(\mathcal{G}|\mathcal{C})$$

In this work, we learn a generative event model in the form of a simple mixture of Gaussians, in both supervised and unsupervised settings. In the unsupervised setting, we used a Bayesian Information Criterion to automatically determine the number of classes. More generally, techniques related to graph classification [5] [6] [8] and clustering [16], may be applied.

The *video event detection task* corresponds to the case, when given an event model \mathcal{C} , the goal is to detect the events, or more generally, learn a labelled cover \mathcal{G} , where the labels correspond to one of the event classes in \mathcal{C} , that is:

$$\hat{\mathcal{G}} = \arg \max_{\mathcal{G}} P(\mathcal{G}|\mathcal{C})P(\mathcal{A}|\mathcal{G}, \mathcal{C})$$

In this work, we form the cover \mathcal{G} , by simply searching for subgraphs in the activity graph that are most likely given the event model \mathcal{C} . That is, we find those graphs $g \in \mathcal{G}$ for which the likelihood $P(g|\mathcal{C})$, is above a threshold. We also simply assume a uniform distribution $P(\mathcal{A}|\mathcal{G}, \mathcal{C})$ for all possible event graph covers \mathcal{G} .

In a more general unsupervised video understanding setting, the goal is to learn the unknowns: \mathcal{G} , \mathcal{C} and \mathcal{A} , given only the observed tracks \mathcal{T} , that is:

$$(\hat{\mathcal{C}}, \hat{\mathcal{G}}, \hat{\mathcal{A}}) = \arg \max_{\mathcal{C}, \mathcal{G}, \mathcal{A}} P(\mathcal{C})P(\mathcal{G}|\mathcal{C})P(\mathcal{A}|\mathcal{G}, \mathcal{C})P(\mathcal{T}|\mathcal{A})$$

A Markov Chain Monte Carlo (MCMC) procedure is used in [14] to find the MAP solution. MCMC is used to efficiently search the space of possible activity graphs, possible covers of the activity graph and possible event models, in order to find the MAP solution.

4 Experiments

A real video dataset consisting of activities representing simple verbs such as throw (a ball), catch etc is used to evaluate the proposed approach. The dataset consists of 36 videos. Each video lasts for approximately 150–200 frames and contains one or more

³ In practical situations, with co-temporal events, there will be co-incidental interaction graphs, which are a part of \mathcal{A} , but not a part of any event graph. We leave further details of this to the full paper.

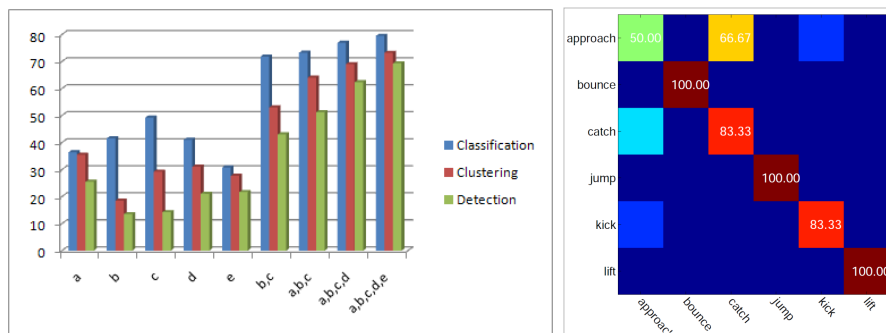


Fig. 2: Left: Accuracies for three tasks - classification, clustering and detection - for possible combinations of spatial relationships are shown. In order to make the results visually legible, only the top ranking combination for a fixed number of combinations are shown. The letters are given by letters: a - RCC5, b - QTC6, c - DIR4, d - SPD3, e - SIZ3 (See Fig. 1 (a) for further explanation of these acronyms). Right: Confusion matrix for the classification task.

of the following 6 verbs: approach, bounce, catch, jump, kick and lift. A ground truth, in terms of labelled intervals corresponding to each of the constituent verbs, in each of these videos is available. We process the dataset by detecting objects of interest using a multi-class object detector and then track the detected blobs.

This dataset is used to evaluate how possible combinations of these features perform for three of the learning tasks - event classification, event clustering and event detection - that arise out of the proposed formulation described above. In order to evaluate the performance of event recognition, a leave-one out cross validation scheme is adopted. For the classification task, an event model in terms of the interaction graphs, is learned from the training videos, in a supervised way using the available class labels. The interaction graph for the video corresponding to the test segment is classified using the learned event model. The classified label for the test segment evaluated against the ground truth label for this segment, in order to compute the average accuracy across different folds. In order to evaluate clustering, the segments for all the available videos are clustered and the accuracy of clustering is evaluated using Rand Index. Finally, the detection task is evaluated by a leave one out procedure, which uses 35 videos for training the event model. The event model is used to detect the events in the remaining video. An event is regarded as being detected if the detected interval overlaps the ground-truth interval by more than 50%.

The results for the classification, clustering and detection tasks are shown in Fig. 2 (left), for different combinations of spatial relationships. These results show that for all three learning tasks, the combination of all five types of qualitative spatial relations results in maximum accuracies. The results for the classification task for each of the six verbs is shown with the help of a confusion matrix in Fig. 2 (right). It can be seen that apart from the verb “approach”, which gets confused with “catch”, the rest of the verbs are classified with reasonably high accuracies.

5 Summary and Future Work

This paper firstly demonstrates the role of different types of qualitative spatio-temporal relations in bridging the gap between low level video input and high level activity under-

standing has been demonstrated. One direction for future research is to investigate the role of other qualitative relations and their role in representing activities. Another interesting direction is to model human actions by considering relationships between body parts. These body parts could be obtained using part-based models.

Another contribution is that this paper presents a general way of addressing problems in video activity understanding using graph based relational learning. In the future, it would be interesting to extend this formalism to other tasks in activity understanding such as anomaly detection, scene description and gap filling.

References

1. Allen, J.: Maintaining knowledge about temporal intervals. *Commun. ACM* 26(11), 832–843 (1983)
2. Cohn, A.G., Hazarika, S.M.: Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae* (2001)
3. Cohn, A.G., Magee, D., Galata, A., Hogg, D.C., Hazarika, S.: Towards an architecture for cognitive vision using qualitative spatio-temporal representations and abduction. pp. 232–248 (2003)
4. Cook, D.J., Holder, L.B.: *Mining Graph Data*. Wiley-Interscience (2007)
5. Deshpande, M., Kuramochi, M., Hale, N., Karypis, G.: Frequent substructure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 17(8), 1036–1050 (2005)
6. Gärtner, T., Flach, P.A., Wrobel, S.: On graph kernels: Hardness results and efficient alternatives. In: *Proceedings of the Conference On Learning Theory (COLT)*. pp. 129–143 (2003)
7. Getoor, L., Taskar, B.: *Introduction to Statistical Relational Learning* (2007)
8. Kudo, T., Maeda, E., Matsumoto, Y.: An application of boosting to graph classification. In: *Proceedings of Neural Information Processing Systems (NIPS)* (2004)
9. Lavee, G., Rivlin, E., Rudzsky, M.: Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Transactions on Systems, Man, and Cybernetics* pp. 489–504 (2009)
10. Raedt, L.D., Kersting, K.: *Probabilistic inductive logic programming* (2008)
11. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: *Proceedings of the Conference on Knowledge Representation and Reasoning (KR)* (1992)
12. Sridhar, M.: *Unsupervised Learning of Event and Object Classes from Video*. University Of Leeds, <http://www.comp.leeds.ac.uk/krishna/thesis.pdf>
13. Sridhar, M., Cohn, A.G., Hogg, D.C.: Learning functional object-categories from a relational spatio-temporal representation. In: *Proceedings of the European Conference on Artificial Intelligence (ECAI)* (2008)
14. Sridhar, M., Cohn, A.G., Hogg, D.C.: Unsupervised learning of event classes from video. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (2010)
15. Sridhar, M., Cohn, A.G., Hogg, D.C.: From video to RCC8: exploiting a distance based semantics to stabilise the interpretation of mereotopological relations. *Proc. COSIT*, In Press (2011)
16. Tsuda, K., Kurihara, K.: Graph mining with variational dirichlet process mixture models. In: *Proceedings of SIAM International Conference on Data Mining* (2008)
17. Wang, X., Ma, X., Grimson, E.: Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2009)