Chapter 6

Conclusion and Future Work

The thesis presents qualitative and geometric reasoning within a graph representation of spatio-temporal activities to recognize human activities from video. The contribution of this thesis can be seen to have several components. The first component is a geometric reasoning framework for extracting qualitative relations between extended objects called Extended CORE9. A graph based representation schema, TAG, that keeps track of the temporal evolution of relations between extended objects is then discussed. For classification of activities within such a representation, a TAG kernel that can be used with an SVM is then presented. Finally, a TAG grammar is presented for modelling activity classes. The rules of the grammar are learned using grammar induction algorithm and activities can be recognized as a TAG. The novelty, drawbacks, and possible future works and extensions for each component of this thesis are discussed in the ensuing sections.

6.1 Extended CORE9

In literature, there exists several qualitative formalisms that discuss the theoretical aspects of binary qualitative relations between spatial regions/objects [70, 71, 72]. However, such theoretical discussions usually concentrate on a single aspect of space, such as topology [70] or direction [71] or distance [72]. CORE9 was presented as a comprehensive framework that allows computing qualitative relations pertaining to several important aspects of space between a pair of objects [13]. CORE9 discusses computation of qualitative relations for topology, direction, distance, size, and motion. However, CORE9 is designed to deal with only "axis-aligned single-piece rectangle" objects. It has been noted that abstracting objects using a single axis-aligned rectangle often leads to computation of inaccurate relations [41]. In literature, extensions of CORE9 exist such as, Angled CORE9 [41]

and a volume based extension of Angled CORE9 [44]. In Angled CORE9, the emphasis was on extending CORE9 so as to deal with *oriented* rectangles. In a similar vein, an extension of Angled CORE9 was presented to deal with oriented rectangles over spatio-temporal volumes [44]. Both of these works have been shown to improve classification results when used to classify human activities in video data. However, neither of these extensions are equipped to handle extended objects.

In this work, Extended CORE9 is presented as a geometric reasoning framework for computing qualitative relations pertaining to topology, direction, and distance between a pair of *extended objects*. It has been argued as well as experimentally shown in Chapter 3, that an extended object abstraction leads to fewer inaccuracies, particularly for human bodies. Further, it has been shown in Chapter 4 that when combined with an appropriate learning mechanisms, Extended CORE9 outperforms Angled CORE9 for HAR.

Other Applications or Extensions

In this thesis, Extended CORE9 has been discussed from the perspective of HAR in video. Extended CORE9 can also serve as a standalone framework for computing qualitative relations between any pair of extended objects. Thus, it can potentially be applied to any scenario wherein objects have multiple components. Most existing work on computing qualitative relations between extended objects involve a single aspect of space such as topology [16] or direction [76]. Extended CORE9 serves to fill that gap by discussing multiple aspects of space for extended objects. One typical application that require an extended object abstraction is Geographical Information System. Discrete geographical locations are easily viewed as extended objects and reasoning within such systems can be based on relations computed using Extended CORE9.

Another unconventional use of qualitative spatial relations is in anomaly detection for firewalls [130, 131]. Firewalls inspect incoming and outgoing traffic and block or allow packets based on the security requirements. The filtering rules in a firewall consist of a condition and an action. The condition part has fields corresponding to *source IP range* and *destination IP range* alongwith discrete fields corresponding to *protocol*, *source-port* and *destination-port*. The action is usually either *accept* or *deny* based on the security requirements of the network. A sample set of firewall rules is shown in Table 6.1.

In the condition part of the rules, the source IP range and destination IP range would allow the rule to be seen as a two-dimensional spatial region. The

Order	Protocol	Source IP Addr.	Source Port	Destn. IP Addr.	Destn. Port	Decision
1	tcp	192.170.21.[10,30]	any	192.170.26.[50,90]	any	deny
2	tcp	192.170.21.[0,50]	any	192.170.26.[70,120]	any	deny
3	tcp	192.170.21.[15,25]	any	192.170.26.[50,110]	any	accept
4	tcp	192.170.21.[10,30]	any	192.170.26.[20,45]	any	accept
5	tcp	192.170.21.[20,60]	any	192.170.26.[25,35]	any	deny
6	tcp	192.170.21.[30,70]	any	192.170.26.[20,45]	any	deny
7	tcp	192.170.21.[15,45]	any	192.170.26.[25,30]	any	accept

Table 6.1: A sample set of firewall rules

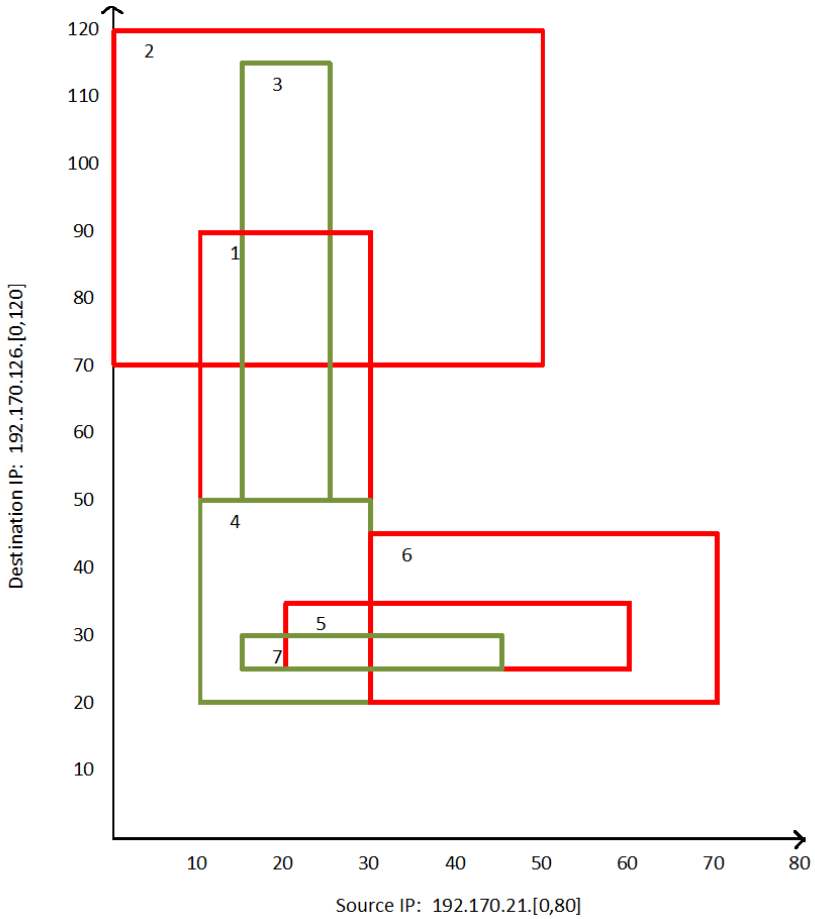


Figure 6-1: Extended object based abstraction of the set of firewall rules in Table 6.1

set of spatial regions corresponding to an action (*accept* or *deny*) can be seen as an extended object. As shown in Figure 6-1, the rules cover rectangular regions in a two dimensional space. Further, the rules of order 1, 2, 5, and 6 from the sample set in Table 6.1 have the same action, *deny*. The rectangles corresponding to these rules constitute the components of one extended object. Similarly, the rectangles corresponding to rules 3, 4, and 7 constitute components of a second extended objects. In this example, all rules have the same values for *protocol*, *source-port*, and *destination-port*. If the firewall contains multiple values for these fields, separate extended objects can be used to abstract the rules corresponding

to a single combination of protocol, source-port and destination port values. In this manner, the entire set of firewall rules can be modelled as a set of extended objects. Anomaly detection under such a design would be a matter of inspection of the qualitative relations between the extended objects obtained using Extended CORE9 [130, 131] or on similar line as done for qualitative spatio-temporal logics.

Drawbacks and Possible Future Work

One major challenge faced in the implementation of Extended CORE9 is the non-availability of video datasets with part-based tracking data. There are several part-based tracking mechanisms available in literature [29, 31]. However, most of these tracking systems are not publicly available. Therefore, at present the objects are manually tracked and annotated. It would be interesting to investigate the integration of Extended CORE9 with a full-fledged part based tracking system.

Currently Extended CORE9 deals with extended objects wherein components of the objects are assumed to be axis aligned rectangles. An interesting direction of future work would be a modification in Extended CORE9 to include objects whose components are oriented rectangles.

Another drawback in the current framework is that qualitative size and motion relations can not be computed. Working on this drawback to allow computation of size and motion relations of extended objects would be another interesting direction of future work.

6.2 Temporal Activity Graph

Temporal Activity Graph or TAG has been presented in Chapter 4 as a means for representing human activities. Human activities have been represented using graphs by various researchers [45, 46, 49, 86, 87]. Some of these representations model only the temporal nature of activities [49, 86] or only the spatial features of the video activities [51]. This is resolved in TAG by modeling spatial and temporal features of activities within the same graph structure. Spatial features are modeled using *spatial edges* and the temporal structure is modeled by the *temporal edges* of TAGs.

In literature, both spatial and temporal features of an activity have been modeled simultaneously [45, 46, 87]. Some have used temporal sequences of *structured feature graphs* [46] for representing activities. However a sequence of disjoint graphs is unable to keep track of how individual spatial features evolve over time.

This is resolved in TAG by connecting the subgraphs representing spatial features at a single instant of time in the activity by using temporal edges. A volumetric view of the video has also enabled spatio-temporal graph representation [45]. Such approaches rely on accurate spatio-temporal segmentation of the video. Whereas only tracking data for the video is used for constructing the TAG. Activities have also been represented as hierarchical qualitative spatio-temporal graphs [87]. However, when such a representation is used for an extended object based abstraction, the number of nodes in the graph increases substantially. The TAG was especially designed for extended object abstraction of the interacting entities. The novelty of TAG lie partly in its ability to encode qualitative spatial relations between extended objects. Furthermore, the temporal edges in the TAG allow encoding the temporal evolution of spatial relations between a pair of components over time.

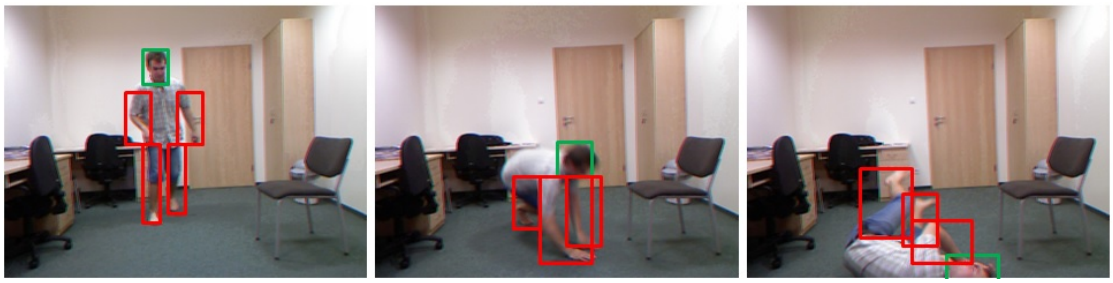
Other Applications or Extensions

In this thesis, TAGs have been used for representing human *interactions* (see Chapter 2, Section 2.1.1). However, we believe TAGs can potentially be used to model more complex *behavior* of humans from video or even simpler *actions* of humans from video.

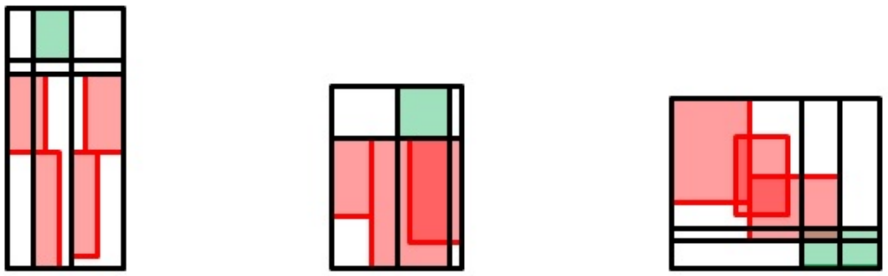
TAGs can be used to represent single-person activities or *actions*. To represent an *action* using a TAG for action classification, the set of components of the single human body are divided into two sets - each set constituting a single extended object. This can be particularly useful for representing and detecting certain categories of human activities, such as in *human fall detection*. When a human body falls, the relation of the *head* of the body changes with respect to all of the rest of body parts. In this case the *head* constitutes a single component extended object, and the *hands* and *legs* constitute the second extended object. This idea is depicted for an instance from the single person action dataset, UR Fall Detection Dataset [132], in Figure 6-2. Figure 6-2a is a sequence of frames for an instance of the *fall* action in the dataset. Figure 6-2b shows how two extended objects are formed for the components of the same person. Figure 6-2b shows the TAG representation of the single person action.

Drawbacks and Possible Future Work

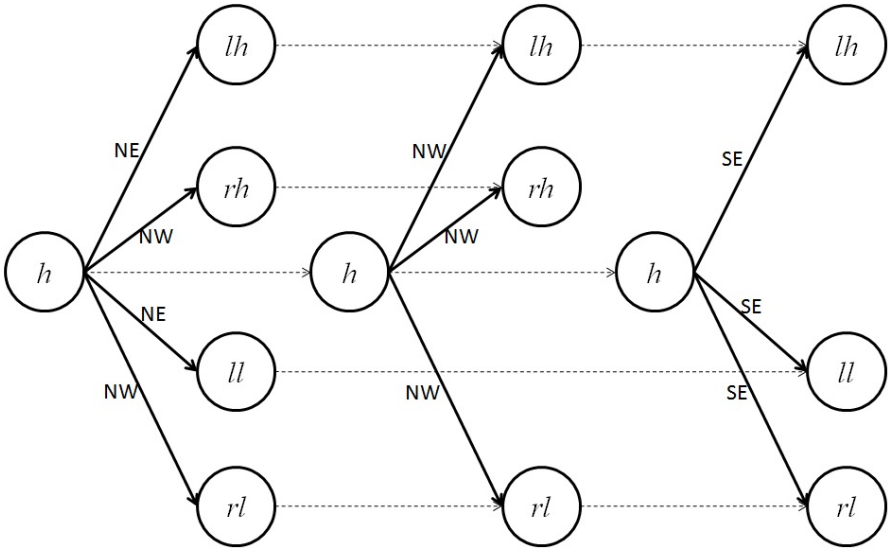
The nodes in a TAG correspond to components of extended objects involved in the activity. In turn, a *component* is a part of an entity that can have an individual identity, such as a hand or a leg for a human body. The rectangles constitut-



(a)



(b)



(c)

Figure 6-2: (a) Fall activity from the UR Fall Detection dataset [132] (b) Extended object abstraction (c) TAG representation

ing an extended object are obtained from the part-based tracking data of the video [29, 30, 31]. As is the case with any high-level system, the accuracy of a TAG representation depends on the accuracy of the underlying low-level tracking system.

An interesting direction of future work could be to address this limitation. Instead of statically generating a TAG after the tracking is done, one can attempt to dynamically generate the TAG during the tracking process. Such a dynamic approach would allow giving feedbacks to the tracking system. The feedbacks could be used to inform any mismatch in the expected spatial relation for a pair of components and actual computed relation based on the tracking data[1]. Such a feedback mechanism could potentially improve tracking results as well as the accuracy of the TAG representation.

As noted in Section 4.4.3 and Section 5.4.3, the work reported in [118], outperforms the work reported in the thesis. In [118], videos are described as a bag of ST-phrases that are combinations of spatio-temporal description for related local features. It has been noted that one of the reasons [118] outperforms the work reported in the thesis could be because the TAG representation fails to combine the spatio-temporal descriptions of related local features. To address this, one may use a combination of spatial relations for related local features - such as the *right hand* and *left hand* of the human body. Addressing this issue within the TAG representation could be another interesting direction of future work.

6.3 Temporal Activity Graph Kernel

The TAG kernel is presented in this thesis for classification of activities represented as TAGs using an SVM. While there exists several approaches for learning from a graph representation, graph kernels are a generic solution that can be used with any kernel based method [90]. For learning from graph representations of human activities, graph based relational learning [87] and graph kernels [46] have been reported in literature among others.

In [87], graph based relational learning is used to search for similar sub-graphs within a larger activity graph, to discover event classes in an unsupervised manner. Although in this work, the TAG kernel has been used for learning in a supervised setting, it is possible to use kernel function for unsupervised learning as well.

A graph kernel based classification has been used for classification of activ-

[1]The expected spatial relation is usually a neighbouring relation from the conceptual neighbourhood graphs (see Chapter 2 Section 2.4.1)

ities represented as a temporal sequence of *structured feature graphs* [46]. The researchers define a random-walk based kernel that is a combination of subgraph matching and time sub-sequence matching. A random walk kernel computes similarity of two graphs according to the frequency of their common walks [91]. It is to be noted that activities are represented as a temporal sequence of disjoint graphs. This is in contrast to the TAG kernel defined in this thesis. TAG kernel computes similarity of TAG graphs based on a comparison of the set of label-sequences. A label sequence, as defined herein, encode the temporal evolution of spatial relations between a pair of components. In this context, a label sequence can also be seen as a walk along the temporal edges of the TAG. Thus, the TAG kernel can be interpreted as a comparison of the *temporal walks* of two graphs. The TAG kernel directly takes advantage of the temporal structure of activities in computing the similarity of graphs.

Other Applications or Extensions

It is widely accepted that classifiers such as SVMs provide a good discriminative classification of the data. As has been discussed in Chapter 4, the results of using TAG kernel within an SVM classifier for HAR is comparable to results reported in literature. In this thesis, TAG kernel has been used for classification of activities involving more than one person or object. Experiments have also been conducted for classification of single-person activities represented as TAGs as shown in Figure 6-2c. The UR Fall Detection dataset [132] consists of instances of *fall* activities (30) and instances of various *activities of daily living* (40). The activities of daily living videos include instances that involve a single person *walking, sitting, lying down* and *picking up an object from the floor*. The single person actions are represented using TAGs. Thereafter, using the same experimental set-up as discussed in Chapter 4, Section 4.4, a classification accuracy of 91.47% is achieved for classification of human *fall* action. The confusion matrix for the experiments is given in Table 6.2.

	ADL	Fall
ADL	0.925	0.075
Fall	0.10	0.90

Table 6.2: Confusion matrix for classification of Fall activities in the UR Fall Detection Dataset [132]

The classification accuracy obtained by [132] is 90.00% but is reported for 30 instances of fall activity and only 30 instances of activities of daily living (for

Kinect depth data). This is in contrast to the results obtained using TAG kernel based classification where only RGB data is used and the entire dataset is taken for the experiments.

Drawbacks and Possible Future Work

One drawback of the TAG kernel is that it is completely deterministic and does not consider the probability distribution of the label sequences. Incorporating such probabilistic values could be one possible future direction for this work. Label sequences could be associated with probability values indicating likelihood of such a sequence occurring in general for any class. Say, P_{ls} is the probability of the label sequence ls occurring for any class. Such a probability value can be used to determine whether a label sequence is informative enough. A higher value of P_{ls} would indicate that the label sequence ls commonly occurs in all activity classes. Therefore, such a label sequence would have lower utility in discriminating activity classes. Similarly, the label sequences could be associated with probability values indicating likelihood of such a sequence occurring within a particular activity class. Say, P_{ls}^{C} is the probability of the label sequence ls occurring for the activity class C. A higher value of P_{ls}^{C} would indicate that the label sequence ls commonly occurs within the activity class C. Such a label sequence would have high utility in discriminating activity classes. The probability values of the label sequences when coupled with the interestingness factor could potentially lead to better classification results.

Along similar lines, other possible research directions include identification of a minimum cardinality label sequence set for better interpretation of activities. A minimum cardinality set of label sequence could further improve efficiency of the approach.

It has been noted in Section 4.4.3, that deep learning based work such as [58] outperforms the TAG kernel based work reported in the thesis. Deep learning based techniques are known to achieve high classification accuracy for activity recognition. However, the thesis does not explore further into deep learning based approaches because "they often require large training sets and high computation power devices". This has been referred to in Chapter 4. It has also been emphasized that our work focuses on the "high level knowledge obtained from the tracking data, that allows one to learn richer models from smaller datasets".

6.4 Temporal Activity Graph Grammar

The Temporal Activity Graph Grammar or TAG grammar is presented in this thesis for a grammar based recognition of activities represented as TAGs. While discriminative classifiers produce good classification of data they offer little in terms of modelling the underlying structure of the activity classes [125]. To model such underlying structures (with activities represented as TAGs) a TAG Grammar is presented as a generative learning tool. In literature, there exist reports of researchers who prefer grammars as a tool for encoding the recursive and hierarchical nature of human activities [6, 52, 53, 54]. Some of these works represent the hierarchical nature of human activities using a Context Free Grammar(CFG) but rely on hand coded grammar rules [6]. For TAG grammar, an induction algorithm is presented that learns the grammar rules from a set of positive examples. Others have used Stochastic Context Free Grammars (SCFG) extended with temporal relations to encode the parallel *sub-events* within activities [52]. Recognition within such grammars is based on specialized multi-thread parsing mechanisms. In the TAG Grammar proposed herein, such concurrently occurring sub-events are easily captured by the TAG. Consequently, treating the TAGs as a string of insubTAGs has a similar effect without having to fall back on complex parsing mechanisms.

The TAG Grammar, presented in this thesis, is a probabilistic context free graph grammar. To the best of our knowledge, such a probabilistic graph grammar is the first of its kind to be used for HAR. One of the advantages of TAG grammar is the amalgamation of elements from graph grammars and string grammars. The elements of graph grammar allow complex structure of activities to be maintained, including concurrently occurring sub-events. The elements of string grammar allow uncluttered grammar induction and parsing algorithms.

Other Applications or Extensions

In literature, there exist reports of works wherein context free grammars have been used for higher-level human behavior recognition [55]. While HAR involves interpreting *what* is going on in a video, Human Behavior Recognition is concerned with interpreting *why* and *how*. Since TAG Grammar is essentially a context free grammar, it can potentially be used for human behavior analysis in a manner similar to [55].

Drawbacks and Possible Future Work

One of the biggest challenges faced in the implementation of the TAG grammar is during parsing. Currently the parsing is done using a modified LR(0) parser. However, the LR(0) parsing requires the grammar rules to be transformed into a tabular form called a *parse table*. Traditionally, tools called *parser generators* are used for this purpose. Existing parser generators can be modified to automate the construction of parse tables for a TAG grammar as well. Overcoming this challenge is one major direction of our future work.

One of the drawbacks of TAG Grammar is that it learns only from positive examples in a supervised setting. A supervised learning algorithm that takes into account both positive and negative examples is known to learn better models. Modifications to the grammar induction algorithm that considers negative examples as well can be part of our future work. Learning grammar rules in an unsupervised manner is another desirable property; investigations towards such a grammar induction algorithm could be another interesting direction of research.

Furthermore, the proposed TAG Grammar uses a non-incremental learning algorithm. An incremental induction algorithm would allow the grammar to be dynamically modified when new positive or negative examples are encountered. Development of such a grammar induction algorithm can be part of our future work.